

European Council for Modelling and Simulation

www.scs-europe.net

ECMS Digital Library

<http://www.scs-europe.net/dlib/dl-index.htm>

Copyright

© ECMS

ISBN 978-3-937436-77-7 (Print)

ISSN 2522-2414 (Print)

ISBN 978-3-937436-76-0 (CD-ROM)

ISSN 2522-2422 (Online)

ISSN 2522-2430 (CD-ROM)

**Cover pictures front and
and back side**

**© pictures: Norwegian Maritime
Competence Center**

Printed by

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

Communications of the ECMS

Volume 36, Issue 1, June 2022

Proceedings of the 36th ECMS International Conference on Modelling and Simulation ECMS 2022

May 30th – June 3rd, 2022
Ålesund, Norway

Edited by:

Ibrahim A. Hameed
Agus Hasan
Saleh Abdel-Afou Alaliyat

Organized by:

ECMS - European Council for Modelling and Simulation

Hosted by:

NTNU – Norwegian University of Science and Technology

Collocated with:

OSC – Offshore Simulation Centre - Conference 2022

Sponsored by:

NTNU – Norwegian University of Science and Technology
IDUN from PhD to Professor
Sparebanken Møre
Research Council of Norway
Norwegian Maritime Competence Center (NMK)
Offshore Simulator Centre (OSC)
Augment City
DigiCat

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 2022 ORGANIZATION

Conference Chair

Ibrahim A. Hameed

Norwegian University
of Science and Technology
Ålesund, Norway

Programme Chair

Agus Hasan

Norwegian University
of Science and Technology
Ålesund, Norway

Programme Co-Chair

Saleh Abdel-Afou Alaliyat

Norwegian University
of Science and Technology
Ålesund, Norway

Editors in Chief

Ibrahim A. Hameed

Norwegian University of Science and Technology Ålesund, Norway

Mauro Iacono

Università degli Studi della Campania *Luigi Vanvitelli*, Italy

Managing Editor

Martina-Maria Seidel

St. Ingbert, Germany

Editorial Advisory Board

Andrzej Bargiela

Nottingham, United Kingdom

Khalid Al-Begain

Kuwait College of Science & Technology, Kuwait

Zuzana Kominkova Oplatkova

Tomas Bata University in Zlin, Czech Republic

Frank Herrmann

OTH Regensburg, Germany

Evtim Peytchev

Nottingham Trent University, United Kingdom

Lars Nolle

Jade University of Applied Sciences, Germany

Editorial Board

Khalid Al-Begain	Kuwait	Zuzana Kominkova Oplatkova	Czech Republic
Romeo Bandinelli	Italy	Michael Manitz	Germany
Andrzej Bargiela	United Kingdom	Lars Nolle	Germany
Umit Cali	Norway	Evtim Peytchev	United Kingdom
Lelio Campanile	Italy	Janos P. Radics	Hungary
Erlend Coates	Norway	Rostislav Razumchick	Russia
Ricardo da SilvaTorres	Norway	Christian Reiter	Germany
Benoit Eynard	France	Frederic T. Stahl	Germany
Mohamed Gaber	United Kingdom	Kornel Tamas	Hungary
Henrique M. Gaspar	Norway	Christoph Tholen	Germany
Marco Gribaudo	Italy	Marco Trost	Germany
Marwan Hassani	Netherlands	Kata Varadi	Hungary
Frank Herrmann	Germany	Agnes Vidovics-Dancs	Hungary
Mauro Iacono	Italy	Jens Werner	Germany
Agnieszka Jakobik	Poland	Edward J. Williams	USA
Anniken Th.Karlsen	Norway	Di Wu	Norway
Eugene Kerckhoffs	Netherlands	Peter T. Zwierczyk	Hungary
Joanna Kolodziej	Poland		

INTERNATIONAL PROGRAMME COMMITTEE

Business Process Modelling and Simulation for Industrial Operations

Track Chair: **Romeo Bandinelli**
University of Florence, Italy

Co-Chairs:

Benoit Eynard
Technical University of Compiègne, France

Edward J. Williams
University of Michigan-Dearborn, USA

Finance, Economics and Social Science

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

Co-Chair: **Agnes Vidovics-Dancs**
Corvinus University of Budapest, Hungary

Simulation and Optimization

Track Chair: **Frank Herrmann**
OTH Regensburg, Germany

Co-Chairs:

Michael Manitz
University of Duisburg-Essen, Germany

Marco Trost
Technical University Dresden, Germany

Christian Reiter
Roche Diagnostics, Germany

Finite – Discrete – Element Simulation

Track Chair: **Peter T. Zwierczyk**
Budapest University of Technology and Economics, Hungary

Co-Chair: **Kornel Tamas**
Budapest University of Technology and Economics, Hungary

Modelling and Simulation of Cyber-Physical-Systems

Track Chair: **Umit Cali**

Norwegian University of Science and Technology, Norway

Co-Chair: **Erlend Coates**

Norwegian University of Science and Technology, Norway

Special Student Track on AI, Machine Learning, Simulation and Visualization

Track Chair: **Di Wu**

Norwegian University of Science and Technology, Norway

Co-Chair: **Anniken Th. Karlsen**

Norwegian University of Science and Technology, Norway

Modeling and Simulation for Performance Evaluation of Computer-based Systems

Track Chair: **Mauro Iacono**

Universita degli Studi della Campania

Luigi Vanvitelli, Italy

Co-Chairs:

Agnieszka Jakobik

Cracow University of Technology, Poland

Lelio Campanile

Universita degli Studi della Campania

Luigi Vanvitelli, Italy

Rostislav Razumchik

Federal Research Center "Computer Science and Control" of the Russian Academy of Science, Russia

IPC Members

Kolos Csaba Agoston, Corvinus University of Budapest, Hungary

Saleh Alaliyat, Norwegian University of Science and Technology, Norway

Sara Antomarioni, Universita Politecnica delle Marche, Italy

Walailak Atthirawong, King Mongkut's Institute of Technology Ladkrabang, Thailand

Anna Bagirova, Ural Federal University, Russia

Romeo Bandinelli, University of Florence, Italy

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

Enrico Barbierato, Politecnico di Milano, DEIB, Italy

Robin T. Bye, Norwegian University of Science and Technology, Norway

Lelio Campanile, Universita degli Studi della Campania *Luigi Vanvitelli*, Italy

Damian Fernandez Cerero, University of Seville, Italy

Tamas Doka, Budapest University of Technology and Economics, Hungary

Thomas Farrenkopf, Technische Hochschule Mittelhessen, Germany

Nora Felföldi-Szucs, Corvinus University of Budapest, Hungary

Massimo Ficco, Universita degli Studi di Salerno, Italy

Ingo Frank, Ostbayerische Technische Hochschule Regensburg, Germany

Mouhsene Fri, Universite Euro-Mediterraneenne de Fes, Marokk

Melinda Friesz, Corvinus University of Budapest & Keler Ltd., Hungary

Maria Ganzha, Warsaw University of Technology, Poland

Marton Groza, Karman Mechanics, Hungary

Alexander Grusho, Federal Research Center "Computer Science and Control" of the Russian Academy of Science, Russia

Daniel Grzonka, Cracow University of Technology, Poland

Michael Guckert, Technische Hochschule Mittelhessen, Germany

Stefan Haag, HS Worms University of Applied Sciences, Germany

Ibrahim A. Hameed, Norwegian University of Science and Technology, Norway

Marc Hanfeld, University of Applied Sciences Emden-Leer, Germany

Agus Hasan, Norwegian University of Science and Technology, Norway

Muhammad Umair Hassan, Norwegian University of Science and Technology, Norway

Lars Ivar Hatledal, Norwegian University of Science and Technology, Norway
Frank Herrmann, Ostbayerische Technische Hochschule Regensburg, Germany
Karl Hribernik, BIBA - Bremer Institut für Produktion und Logistik GmbH, Germany
Mauro Iacono, Università degli Studi della Campania *Luigi Vanvitelli*, Italy
Agnieszka Jakobik, Cracow University of Technology, Poland
Bogumil Kaminski, SGH Warsaw School of Economics, Poland
Joanna Kolodziej, Cracow University of Technology and NASK, Poland
Petia Koprinkova-Hristova, IICT, Bulgarian Academy of Sciences, Bulgaria
Andreasz Kosztopulosz, University of Szeged, Hungary
Mateusz Krzyszton, Research and Academic Computer Network, Poland
Frederick Lange, Maschinenfabrik Reinhausen GmbH, Germany
Alexander H. Levis, George Mason University, USA
Lidija Lovreta, EADA Business School, Spain
Andrea Marin, Università Ca' Foscari Venezia, Italy
Stefano Marrone, Università degli Studi della Campania *Luigi Vanvitelli*, Italy
Michele Mastroianni, Università degli Studi di Salerno, Italy
Nicolas Meseth, University of Applied Sciences Osnabrueck, Germany
Lusine Meykhanadzhyan, Financial University under the Government of the Russian Federation, Russia
Frank Morelli, University of Applied Sciences in Pforzheim, Germany
Marco Nobile, Università Ca' Foscari Venezia, Italy
Laszlo Oroszvary, Knorr-Bremse Research, Hungary
Ottar Osen, Norwegian University of Science and Technology, Norway
Marcin Paprzyck, Polish Academy of Science, Poland
Oucheikh Rachid, Lund University, Sweden
Rostislav Razumchick, Federal Research Center "Computer Science and Control" of the Russian Academy of Science, Russia
Christian Reiter, Roche Diagnostics, Germany
Simone Righi, University College London, United Kingdom
Maximilian Selmair, Tesla Manufacturing Brandenburg SE, Germany
Oleg Shestakov, Moscow State University, Russia
Oksana Shubat, Ural Federal University, Russia

Markus Siegle, University of the German Armed Forces, Germany
Carlo Simon, HS Worms University of Applied Sciences, Germany
Grazyna Suchacka, University of Opole, Poland
Andras Suplicz, Budapest University of Technology and Economics, Hungary
Ferenc Szabo, Budapest University of Technology and Economics, Hungary
Armando Tacchella, University of Genoa, DIBRIS, Italy
Jacek Tchorzewski, Cracow University of Technology, Poland
Enrico Teich, DUALIS GmbH, Germany
Hajo Terbrack, Technical University Dresden, Germany
Marco Trost, Technical University Dresden, Germany
Anete Vagale, Norwegian University of Science and Technology, Norway
Agnes Vaskoevi, Corvinus University of Budapest, Hungary
Oystein Volden, Norwegian University of Science and Technology, Norway
Andrzej Wilczynski, Cracow University of Technology, Poland
Edward J. Williams, University of Michigan-Dearborn, USA
Alexander Zeifman, Vologda State University, Russia
Peter T. Zwierczyk, Budapest University of Technology and Economics, Hungary

PREFACE

The 36th edition of the ECMS conference is organized this year by the Norwegian University of Science and Technology (NTNU), the largest university in Norway, in cooperation with Offshore Simulator Center (OSC) AS, the most advanced provider of simulators for demanding offshore operations. This year the chosen conference venue is in the Norwegian Maritime Competence Center (NMK), Ålesund, Norway.

NTNU is a university with an international focus, with headquarters in Trondheim and campuses in Ålesund and Gjøvik. NTNU has a main profile in science and technology, a variety of programmes of professional study, and great academic breadth that also includes the humanities, social sciences, economics, medicine, health sciences, educational science, architecture, entrepreneurship, art disciplines and artistic activities. NTNU has eight faculties, 55 departments, and NTNU University Museum with a total budget of NOK 10 billion, of which NOK 2,6 billion from external sources (2021). NTNU offers 356 study programs, as well as further education to more than 42000 from Norway and abroad enrolled in various bachelor, master, and PhD programs.

NTNU in its current form was established by the King-in-Council in 1996 by the merger of the former University of Trondheim and other university-level institutions, with roots dating back to 1760, and has later also incorporated some former university colleges. NTNU is consistently ranked in the top one percentage among the world's universities, We are honored to offer, besides the scientific program of the conference, consisting in more than 45 high quality accepted papers in 8 scientific tracks, a rich and various series of social events, as usual for the spirit of this conference.

Thanks to the precious effort of our track chairs and to the submissions coming from 15 countries and three special contributions by our excellent invited speakers, Hussein Abbas, full professor at University of New South Wales Canberra, Australia and the Founding Editor-in-Chief of the IEEE Transactions on Artificial Intelligence (IEEE-TAI), Manish Gupta, the Director of Google Research India and Infosys Foundation Chair Professor at IIT Bangalore, and Knut-Andreas Lie, visiting professor at the Norwegian University of Science and Technology (NTNU), and chief scientist at SINTEF, the largest independent research organization in Scandinavia.

We hope that this first volume of "Communications of ECMS" will be a good reference for the readers and the practitioners in simulation and modelling, and that our efforts to organize a pleasant stay, a cultural social program and a fruitful environment may

be enough to thank all authors and reward them for their contributions. We also wish to thank for their always precious work: Martina-Maria Seidel, who runs the ECMS office, Mauro Iacono, President of ECMS, and all board members, all track chairs and referees, Ingrid Schjøberg, the Dean of the Faculty of Information Technology and Electrical Engineering, Rune Voden, the head of the department of ICT and Natural Science, Annik Fet, the vice rector of NTNU in Ålesund, and all institutions and sponsors who supported this conference.

Special thanks go to Webjørn Rekdalsbakken, Amela Paro, Gunn Helen Hellevik, and students of the international master program in simulation and visualization, who voluntarily donated their time for local organization matters.

Wishing all the best for ECMS,

Ibrahim A. Hameed, Agus Hassan, Saleh Abdel-Afou Alaliyat
Ålesund, Norway, May 2022

TABLE OF CONTENTS

Plenary Talks - Abstracts

On The Role Of Modelling And Simulation For Artificial Intelligence
Hussein Abbass05

Modelling Using Deep Learning
Manish Gupta07

Building An Open-Source Community For Subsurface Flow Simulation
Knut-Andreas Lie09

Business Process Modelling and Simulation for Industrial Operations

Optimization Of Internal Logistics Using A Combined BPMN And Simulation Approach
Maximilian Wuennenberg, Benjamin Wegerich, Johannes Fottner.....13

Simulating The Automation Of Sorting Crates – A Stepwise Approach
Armand Misund, Henrique M. Gaspar.....20

Towards A Meta-Modeling Approach For An IoRT-Aware Business Process
Najla Fattouch, Imen Ben Lahmar, Khouloud Boukadi29

A Data-Driven Approach For Process Simulation Optimization: A Case Study
Romeo Bandinelli, Andrea Nunziatini, Virginia Fani, Bianca Bindi.....36

Finance and Economics and Social Science

Reflections On Assumptions For A Simulation Model Of Dental Caries Prevention Planning In A Primary School

Maria Hajlasz, Bozena Mielczarek.....45

European Quality Of Life In Retirement Analyzing Personal Differences Based On SHARE Data

Sara Szanyi-Nagy, Erzsebet Kovacs, Agnes Vaskoevi.....51

Regional Models Of Corporate Sector Development In Russia: Where Does Family-Friendly Policy Matter Most? (A Study Based On Cluster Analysis)

Anna Bagirova, Oksana Shubat.....58

Rising Regional Importance Of The Renminbi In The Asia-Pacific Area: A Panel Analysis

Eszter Boros, Gabor Sztano64

Implied Volatility Based Margin Calculation On Cryptocurrency Markets

Balazs Kralik, Nora Felfoeldi-Szuecs, Kata Varadi70

Identifying Regional Models Of Active Grandparenting In Russia Based On Cluster Analysis

Oksana Shubat, Irina Shmarova.....78

Comparison Of Separated Families' Standard Of Living In Germany Analyzing The Equalised Incomes In Simulated Families After Child Support And Child Benefit Paid

Erzsebet Terez Varga.....84

Forecasting Models For First Year Premium Of Life Insurance

Somsri Banditvilai90

Simulation and Optimization

A Simple Algorithm Selector For Continuous Optimisation Problems

Tarek A. El-Mihoub, Christoph Tholen, Lars Nolle.....99

Optimised Bumblebee Paths As Search Strategy For Autonomous Underwater Vehicles

Christoph Tholen, Lars Nolle, Tarek A. El-Mihoub, Oliver Zielinski..... 107

Emission Reduction Through Production Scheduling By Priority Rules And Energy Onsite Generation

Hajo Terbrack, Thorsten Claus, Frank Herrmann 114

Supply Chain Resilience Management Using Process Mining

Frank Schaeffer, Frank Morelli, Florian Haas 121

Stratification Of Timed Petri Nets At The Example Of A Production Process

Carlo Simon, Stefan Haag, Lara Zakfeld 128

Microbial Growth Of *Lactobacillus Delbrueckii Ssp. Bulgaricus* B1 In A Complex Nutrient Medium (MRS-Broth)

Georgi Kostov, Rositsa Denkova-Kostova, Vesela Shopska, Bogdan Goranov, Zaprjana Denkova..... 135

Synergy Between Shuttles And Stacker Cranes In Dynamic Hybrid Pallet Warehouses: Control Strategies And Performance Evaluation

Giulia Siciliano, Yue Yu, Johannes Fottner..... 143

Calibration Model For Perceptual Compensation Of Defective Pixels Of Self-Emitting Display

Olga A. Basova, Anton Grigoryev, Dmitry P. Nikolaev..... 150

How To Run A World Record? A Reinforcement Learning Approach

Sajad Shahsavari, Eero Immonen, Masoomah Karami, Hashem Haghbayan, Juha Plosila..... 159

Elongated Boundaries Detector Parameters Optimisation Based On Generation Of Synthetic Data From Aerial Imagery

Ekaterina Panfilova, Anton Grigoryev, Vladimir Burmistrov 167

A Model For Predicting The Amount Of Photosynthetically Available Radiation From BGC-ARGO Float Observations In The Water Column

Frederic Stahl, Lars Nolle, Ahlem Jemai, Oliver Zielinski..... 174

Taking Randomness For Granted: The Complexities Of Applying Random Number Streams In Simulation Modelling

Maximilian Selmair..... 181

Finite – Discrete - Element Simulation

FEM Study On The Strength Increasing Effect Of Nitrided Spur Gears

Jakab Molnar, Peter T. Zwierczyk, Attila Csoban 189

Application Of The Finite Element Method To Determine The Velocity Profile In An Open Channel

Daria Wotzka 196

Development Of A 2D Discrete Element Software With LabVIEW For Contact Model Improvement And Educational Purposes

Laszlo Pasty, Jozsef Graeff, Kornel Tamas 203

Application Of The Extended Finite Element Method In The Aim Of Examination Of Crack Propagation In Railway Rails

Daniel Bobis, Peter T. Zwierczyk, Tamas Mate 210

Modelling and Simulation of Cyber-Physical-Systems

Modelling AGV Operation Simulation With Lithium Batteries In Manufacturing

Ozan Yesilyurt, Marius Kurrle, Andreas Schlereth, Miriam Jaeger, Alexander Sauer 219

Digital Twins For Lighting Analysis: Literature Review, Challenges, And Research Opportunities

Muhammad Umair Hassan, Stavroula Angelaki, Claudia Viviana Lopez Alfaro, Pierre Major, Arne Styve, Saleh Abdel-Afou Alaliyat, Ibrahim A. Hameed, Ute Besenecker, Ricardo da Silva Torres 226

On The Use Of Graphical Digital Twins For Urban Planning Of Mobility Projects: A Case Study From A New District In Ålesund, Norway

Pierre Major, Ricardo da Silva Torres, Andreas Amundsen, Pernille Stadsnes, Egil Tennfjord Mikalsen 236

Special Student Track on AI, Machine Learning, Simulation and Visualization

GENOR: A Generic Platform For Indicator Assessment In City Planning

Leo Leplat, Ricardo da Silva Torres, Dina Aspen, Andreas Amundsen.....245

Heuristic Techniques For Reducing Energy Consumption Of Household

Sarah M. Daragmeh, Anniken Th. Karlsen, Ibrahim A. Hameed.....254

Learned Parameterized Convolutional Approximation Of Image Filters

Olga Chaganova, Anton Grigoryev.....262

Modeling and Simulation for Performance Evaluation of Computer-based Systems

Causal Analysis Graph Modeling For Strategic Decisions

Alexander H. Levis, Amy Sliva.....271

Predicting Performance Of Heterogeneous AI Systems With Discrete-Event Simulations

Vyacheslav Zhdanovskiy, Lev Teplyakov, Anton Grigoryev.....278

Epistemic Games With Conditional Beliefs For Modelling Security Threats Defence In Cloud Computing Systems

Lukasz Gaza, Agnieszka Jakobik285

Simulating The Programmable Networks For HLA Compatible High-Performance Simulators

Kayhan M. Imre291

Agent And Evolutionary-Based Modelling And Simulation Of A Simplified Living System

Adrian Sosnicki, Daniel Grzonka, Lukasz Gaza.....296

Time Series Clustering With Different Distance Measures To Tell Web Bots And Humans Apart

Grazyna Suchacka.....303

Formal Verification Of Neural Networks: A Case Study About Adaptive Cruise Control

Stefano Demarchi, Dario Guidotti, Andrea Pitto, Armando Tacchella.....310

A DSL-Based Modeling Approach For Energy Harvesting IoT / WSN

Lelio Campanile, Mauro Iacono, Fiammetta Marulli, Marco Gribaudo, Michele Mastroianni317

Some Ergodicity And Truncation Bounds For A Small Scale Markovian Supercomputer Model

Rostislav Razumchik, Alexander Rumyantsev.....324

Effect Of Impurities On Stability Of The Skyrmion Phase In A Frustrated Heisenberg Antiferromagnet

Mariia Mohylina, Milan Zukovic.....331

Massive Degeneracy And Anomalous Thermodynamics In A Highly Frustrated Ising Model On Honeycomb Lattice

Milan Zukovic.....336

Author Index342

ECMS 2022

SCIENTIFIC PROGRAM

Plenary Talks

On the Role of Modelling and Simulation for Artificial Intelligence

Hussein Abbass

University of New South Wales Canberra, Australia

Hussein Abbass is a Full Professor and Acting Deputy Head of School – People in the School of Engineering and Information Technology, University of New South Wales, Canberra (UNSW Canberra) Campus at the Australian Defence Force Academy. He has been with UNSW Canberra since March 2000 and a full professor since 2007. Before joining UNSW Canberra, he was an academic with Cairo University since 1995. Before that, he worked in the IT industry. He was a visiting scholar or professor at Imperial College London, the University of Illinois at Urbana-Champaign, the National Defence Academy of Japan, and the National University of Singapore.

Prof. Abbass is a Fellow of IEEE, a Fellow of the Australian Computer Society, a Fellow of the UK Operational Research Society, a Fellow of the Australian Institute of Managers and Leaders, and a Graduate Member of the Australian Institute of Company Directors (GAICD). He is a Distinguished Lecturer for the IEEE Computational Intelligence Society. Prof. Abbass is the Founding Editor-in-Chief of the IEEE Transactions on Artificial Intelligence, an Associate Editor of several IEEE journals, and a Senior AE for ACM Computing Surveys. He was the National President (2016-2019) for the Australian Society for Operations Research (ASOR), the Vice-President for Technical Activities (2016-2019) for the IEEE Computational Intelligence Society, a member of the Australian Research Council (ARC) College of Experts (2013-2015), and a Chair (2013-2014) of the Emerging Technologies Technical Committee, IEEE Computational Intelligence Society. His current research focuses on AI-enabled swarm systems, shepherding-based swarm guidance, human-AI teaming, and machine education.

Abstract

Artificial Intelligence (AI) is the ubiquitous revolutionary technology of this century. AI has revolutionised humanity, including industry and government organisations, and transformed our world into smart digital spaces. As a discipline, one of the definitions of AI is the automation of cognition; or, put simply, the set of technologies required to support the design and implementation of artificial cognition in artificial systems. To design an intelligent system/machine, technologists need to transform the algorithms of AI into a system-of-systems design of cognition, whereby the artificial agent can sense, make sense, make decisions, take decisions, and learn about the contexts it is situated within. Modelling and Simulation (M&S) sit at the core of an artificial agent's design and implementation components.

In this presentation, I will cover the use of M&S within AI from different angles . In doing so, I will bring elements from my research to showcase how M&S not only contributes to AI but also shows that without M&S, AI can't operate. I will paint futures that range from very concrete and narrowly defined uses of AI to a world of human and artificial cognitive agents. Humans educate AI agents, and AI agents educate humans. I will then conclude with some challenges for the M&S community to support the effort in advancing AI.

This presentation will be drawn from many of my published works; below are some critical references for interested readers.

1. Abbass, H. (2021). What is Artificial Intelligence?. *IEEE Transactions on Artificial Intelligence*, 2(2), 94-95.
2. Tang, J., Leu, G., & Abbass, H. A. (2019). *Simulation and Computational Red Teaming for Problem Solving*. John Wiley & Sons.
3. Abbass, H. A. (2015). *Computational red teaming*. Springer.
4. Abbass, H., Petraki, E., Hussein, A., McCall, F., & Elsawah, S. (2021). A model of symbiomemesis: machine education and communication as pillars for human-autonomy symbiosis. *Philosophical Transactions of the Royal Society A*, 379(2207), 20200364.

Modelling using Deep Learning

Manish Gupta

International Institute of Information Technologie (IIIT -) Bangalore, India

Prof. Manish Gupta is the Director of Google Research India. He holds an additional appointment as Infosys Foundation Chair Professor at IIIT Bangalore. Previously, Manish has led VideoKen, a video technology startup, and the research centers for Xerox and IBM in India. As a Senior Manager at the IBM T.J. Watson Research Center in Yorktown Heights, New York, Manish led the team developing system software for the Blue Gene/L supercomputer. IBM was awarded a National Medal of Technology and Innovation for Blue Gene by US President Barack Obama in 2009. Manish holds a Ph.D. in Computer Science from the University of Illinois at Urbana Champaign. He has co-authored about 75 papers, with more than 7,000 citations in Google Scholar, and has been granted 19 US patents. While at IBM, Manish received two Outstanding Technical Achievement Awards, an Outstanding Innovation Award and the Lou Gerstner Team Award for Client Excellence. Manish is a Fellow of ACM and the Indian National Academy of Engineering, and a recipient of a Distinguished Alumnus Award from IIT Delhi.

Abstract

Machine learning, and in particular, deep learning has emerged as an important tool for advancing science, in addition to its broad based impact on the world. This talk describes three research efforts that illustrate how deep learning can complement modeling and simulation to pursue scientific discoveries and to tackle societal problems. We begin by describing a flood forecasting initiative that has already led to hundreds of thousands of alerts being sent to people in India. It utilizes a new hydrologic model that has been built using an LSTM (long short-term memory) architecture and a physics based inundation model whose effectiveness has been enhanced using machine learning methods. We also describe how self-supervised learning is being applied to study several interesting aspects of the organization of the human brain. The generated embeddings can be used to rapidly annotate new structures and develop new ways of clustering and categorizing brain structures based on purely data-driven criteria. Finally, we present a deep learning based modeling of human behavior in a specific game-based setting, which has very interesting implications if we are able to generalize that approach to broader settings.

Building an open-source community for subsurface flow simulation

Knut-Andreas Lie

Norwegian University of Science and Technology, Norway

Knut-Andreas Lie is a chief scientist at SINTEF, the largest independent research organization in Scandinavia, where he leads the Computational Geosciences group at the Mathematics and Cybernetics department in Oslo. He also holds a part-time Professor II position at the Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU). He was elected as a life-time member of the Norwegian Academy of Technological Sciences in 2014 and SIAM Fellow in 2020.

Abstract

The MATLAB Reservoir Simulation Toolbox (MRST) is a unique research tool for rapid prototyping and demonstration of new computational methods for flow in porous media. The software has a large user base all over the world in both academia and industry. My group is also one of the key contributors to OPM Flow, the world's first open-source reservoir simulator aimed at full industry use. In addition, we are currently developing Jutul, an upcoming Julia code for high-performance demonstrators of subsurface flow. In the talk, I will briefly describe the MRST and OPM Flow platforms, how they came to be, compare and contrast their development and ownership models, and outline some of the factors that have contributed to their current success.

MRST was originally an internal research tool that has, over the last ten years, morphed into the de-facto standard tool for researchers who want to learn about subsurface flow, and obtain a head start for their own research prototypes. MRST is organized as a minimal core module offering basic data structures and functionality for representing grids and physical parameters relevant to porous media flow, and a large set of add-on modules offering discretizations, solvers, physical models, and a wide variety of simulators and workflow tools. In the modules, you will find many tutorial examples that explain and showcase how MRST can be used to make general or fit-for-purpose simulators and workflow tools. The modular structure makes it easy to add new or modify existing functionality, and many of the 60 currently released

modules are authored entirely or in part at other institutions. The software is licensed under the viral GPL license, with copyright of the core parts vested at SINTEF, where we heavily rely on the code for both our contract work and academic research. The software exists as an add-on to the commercial MATLAB product, which greatly simplifies the process of getting started, but also poses certain restrictions on usage and performance potential. (Except for graphical interfaces, large parts of MRST also work well with GNU Octave.)

OPM Flow aims to simulate many of the same physical processes as MRST, but its disruptive power comes from being a drop-in replacement for commercial reservoir simulators that have been developed and validated over decades. Such simulators, which are used for development planning and day to day operations of oil and gas fields and CO₂ storage operations, are highly complex, expensive and difficult to modify as the source codes are not available. OPM Flow is a collaborative effort, developed in collaboration with Equinor, NORCE, and others, to provide an open-source alternative that enables more open innovation. The simulator is written in high-performance C++ and is released under the GPL license. A key requirement for the development has been that the new simulator should reproduce virtually identical simulation results as contemporary commercial tools both in hindcasting decades of production histories and in forecasting future production. Achieving this has been a vast undertaking filled with stumbling blocks and a lot of reverse engineering.

The two platforms are approaching the same goal of open-source simulation of subsurface flow from opposing ends: MRST educates users in subsurface simulation and embeds them directly at the research front by giving access to the codes behind scientific papers, while OPM's greatest achievement is that users may not notice that they have switched from an expensive closed code to a modern open-source simulator with high potential for computational speedup.

Business Process Modelling and Simulation for Industrial Operations

OPTIMIZATION OF INTERNAL LOGISTICS USING A COMBINED BPMN AND SIMULATION APPROACH

Maximilian Wuennenberg
Benjamin Wegerich
Johannes Fottner

Chair of Materials Handling, Material Flow, Logistics
TUM School of Engineering and Design
Technical University of Munich
Boltzmannstraße 15, 85748 Garching bei Muenchen, Germany
E-mail: max.wuennenberg@tum.de
E-mail: benjamin.wegerich@tum.de
E-mail: j.fottner@tum.de

KEYWORDS

BPMN, Internal Logistics, Material Flow Systems, Model-based Systems Engineering, Planning, Simulation.

ABSTRACT

The optimization of material flow systems requires a profound understanding of the underlying processes. Business Process Model and Notation (BPMN) is an established way of creating a process model that allows an interdisciplinary analysis and optimization. Quantitative exploration of systems using discrete-event simulation can help to enrich these insights. For that reason, this paper introduces a combined BPMN-simulation approach that connects the advantages of both modeling frameworks. By synthesizing systems from generic modules, a comprehensive yet structured optimization process chain is developed. A case study evaluation based on key metrics for material flow operations proves the applicability of the methodology.

INTRODUCTION

Creating a virtual model of a material flow (MF) system promises higher availabilities and shorter throughput times. To achieve this, the model needs to contain data from various sources in the MF domain (e. g. stacker cranes or conveyor belts). However, the application of this approach in real-world systems is often impaired by complex and distributed processes in heterogenous organizations (Pires et. al. 2019). The necessary transparency can be generated by defining and implementing a proper process visualization. Business Process Model and Notation (BPMN) is a widespread standard for this task since the created models can be understood by experts from various domains. However, the focus of this modeling language primarily lies on administrative processes. (Muehlen and Recker 2013) The most common process type modeled with it is the information flow. In theory, BPMN can also be used to represent MFs as well as the movement of workers, forklift trucks, and other mobile resources. Common notations in this domain, such as flow charts or value stream mapping, often cannot meet the requirements for

these particular use cases. Flow charts, for instance, suffer from a low level of standardization and development, and are unsuitable for depicting more complex process properties. Value stream mapping, on the other hand, does not depict sequence flows as accurately and in detail as BPMN, for example because events are not mapped (Garcia et. al. 2012). That makes it difficult to subsequently create a discrete-event simulation (DES). Additionally, this notation focuses heavily on manufacturing settings and is therefore difficult to understand for users outside of the domain of production management (Forno et. al. 2014).

But although BPMN possesses several properties that make it attractive to be applied to MF processes, there are only a few examples where it is actually used in this field (Zor and Leymann 2011). One potential reason for this is the lack of a scientific foundation for modeling strategies in the internal logistics domain (Robinson 2006). A generalized and well-structured approach that considers the specific characteristics of both modeling frameworks can offer guidance for practitioners and help them to model MF systems in a predictable amount of time.

Central Concepts & Related Research

When modeling administrative processes with BPMN, the sequence flow (SF) is used to show the chronological order in which events and activities take place. However, MF systems are characterized by a greater variety of different process types, which raises the question of how these can be mapped in BPMN. An intuitive option is to use the SF as a representation of the MF, resulting in a material-oriented model. Alternatively, the model can be resource-oriented, meaning that the SF represents the movements of mobile resources, e. g. workers. A third option is the addition of MF-specific elements to the BPMN syntax (Zor and Leymann 2011). While this somewhat reduces the ambiguity of the SF, it also makes the models more complex and limits the choice of modeling software. That is why the BPMN models shown in this article use the standard BPMN syntax and are either material- or resource-oriented.

Another challenge arises from the fact that the activities in MF systems are usually object-constrained, meaning that their execution requires the availability of certain objects (Wagner 2021). Depicting these relationships between objects and activities is essential for the modeling of MF systems, but it is also beyond the possibilities of the BPMN syntax. The different approaches discussed above, being exclusively visual, do not address this problem. That is why this article proposes the use of an additional tool in the form of DES. Being widely used in the MF domain, DES includes various possibilities to represent object-constrained activities. Since BPMN does not cover these activities, DES has the potential to work as a complement for BPMN. By synthesizing the two modeling approaches, the high variety of MF processes can be represented. The question of how those two tools can be combined has already been partly explored outside the manufacturing domain, e. g. by extending the DES framework to include tools for business process modeling (Wagner et al. 2009). However, as of today, this is not yet supported by existing DES software. The reverse approach, adding DES elements to the BPMN syntax, is developed as a BPMN variation called “DPMN” (Guizzardi and Wagner 2011, Wagner 2018 & 2021). While this new modeling language can depict object-constrained activities in a qualitative way, existing DES software is still required to perform simulation experiments and generate quantitative results. Moreover, it is uncertain to what extent DPMN is applicable to MF systems.

Summing up, the ever-recurring goal of an MF operator to optimize the system could be met more effectively by creating support in the shape of a proper system model. It seems to be a promising approach to combine BPMN and DES using the existing modeling syntax and established software. However, there is no current research which sufficiently covers this topic.

Research Questions

Therefore, the first objective of this article is to assess if and to what degree BPMN is a suitable tool to model and illustrate MF systems, especially as a complement for DES. Based on that, a standardized methodology is proposed that combines both techniques to increase their usability and reduce the effort spent for modeling. This approach is expected to provide better insights into the process. Hence, the following two questions are investigated in this article:

1. How must a modeling approach for material flow systems be designed to ensure a sensible combination of BPMN process modeling and DES?
2. How can improvements for a material flow system be found based on its BPMN process model?

MATERIALS

Characteristics of Material Flow Systems

MF systems contain all operations which are necessary for the processing and distribution of goods within a defined area. Thus, they execute the physical component of an enterprise logistics process. MF processes lead to a transformation of transported goods regarding time, location, quantity, composition, and quality. For different types of transformations, MF systems contain different subsystems like conveying, storage, or handling. Although very different in terms of their technical design, these subsystems follow similar requirements from a process-overarching perspective. Key performance indicators (KPIs) for logistics contain, for instance, the throughput, which is the number of TUs processed in a certain timeframe. (Hompel et al. 2018) An important concept from lean management are the seven forms of waste. They allow for a distinction between those activities which directly contribute to the value of goods and those which do not. Particularly in the field of MF, unnecessary transportation processes are one form of waste. (Guenther and Boppert 2013) Those KPIs allow the controlling of how well an MF system can be optimized with a certain improvement. They therefore are an important aspect of an optimization methodology.

A typical challenge that needs to be dealt with in the MF domain is queueing systems. If a certain process can only handle one object at a time, but several objects require the availability of that process, all of these objects but one form a queue. Depending on whether – over a certain period of time – the number of arriving objects is smaller or larger than the number of processed objects, the queue either shrinks or grows. (Hompel et al. 2018) Since neither arrival nor processing time are constant but rather can follow a complex distribution, analytical modelling of queueing system is a challenging task. (Arnold and Furmans 2019)

Business Process Model and Notation

The process models in this article follow the standard BPMN syntax and have been created with the software “Modelio” (SOFTEAM 2020, GitHub 2021).

In BPMN, a process is represented as a sequence of events and activities, connected by the SF. Various gateways allow for branching and merging of the SF to depict process variations. Additional elements like messages and data objects make BPMN useful for the modeling of information flow. (OMG 2011) Applied to MF systems, the SF can represent the movement of different objects, making the model material- or resource-oriented. However, BPMN does not offer any possibility of representing the scarcity of these objects resulting from object-constrained activities. While so-called pools and lanes can be used in BPMN to represent certain resources (e. g. the worker who is responsible for a task), the usefulness of this option is limited by the fact that each element can be part of only one lane (and each

lane can be part of only one pool) (OMG 2011). This does not properly represent the complexity that is typical for more detailed models of MF systems.

Discrete-Event Simulation

DES on the other hand focuses more on the objects that a system consists of. How they are represented can vary depending on the simulation software, but common elements include TUs, containers, conveyors, workstations, assembly stations, resources, and submodels (JaamSim 2021a). A DES model also contains information about processes, but usually without an explicit visualization. However, in contrast to BPMN, visual representation is not the main purpose of a DES model anyway. Instead, the model is used to receive quantitative information about the system by performing simulation experiments. (VDI 2018) An example for this would be the simulation of a queuing system: Instead of trying to calculate dynamic queuing lengths and waiting times analytically, the system behavior is modeled and simulated using simple elements like entity generators, statistical distributions, queues, and servers. Based on these, various simulation experiments can be conducted to represent different operating scenarios and predict the system behavior quantitatively and in detail. In many cases, DES can deliver very accurate solutions for this type of problem while requiring less modeling effort than other tools (Arnold and Furmans 2019).

The simulation models for this article have been created with the DES software “JaamSim” (JaamSim 2021a & 2021b).

METHODOLOGY

Combining BPMN & DES

As discussed above, each of the two modeling tools focuses on its own aspects of an MF system. Combining them makes it possible to create a more comprehensive model by using their respective advantages, and to reduce the modeling effort by using similarities and synergies. Figure 1 shows a methodology for the modeling and simulation of MF systems in which the BPMN model is used both as a result in itself and as a starting point for the creation of the DES model. A key element of this methodology is the use of generic modules, which is illustrated with an example later in this article.

The purposes of the system analysis as the first step of the methodology are to document external requirements, to gather information about the system and to decide on a sensible substructure. Qualitative information covers the different process paths which simulation entities can follow, the order of conveying and processing operations and the connected paths of simulation entities in assembly processes. This kind of information is necessary for the process modeling. That is, for creating a BPMN model, the modeler needs to understand all different process variants in the system but without the necessity of specific values, e.g. process times. To

generate the DES model; however, quantitative data is necessary as well. The parametrization of the simulation requires inputs for conveying times, process times, and inter-arrival times of entity generators. In addition to that, if downtimes of processes are supposed to be represented, statistical distributions of breakdowns and maintenance must be provided.

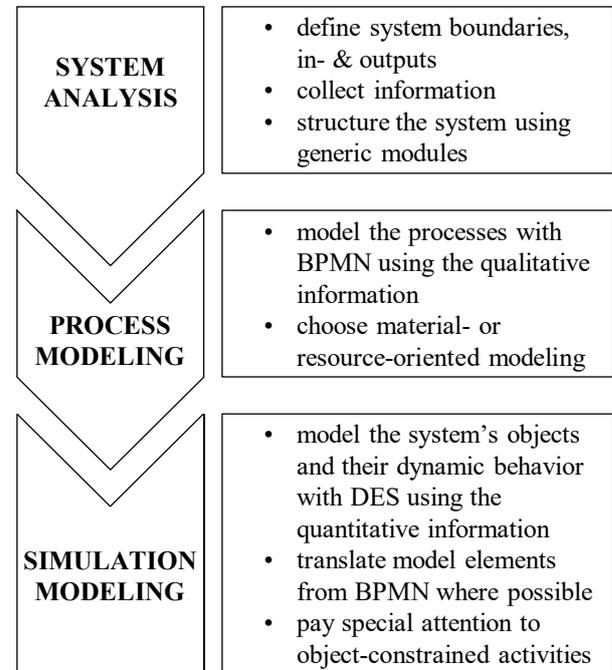
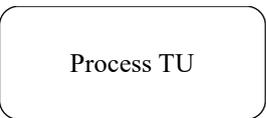
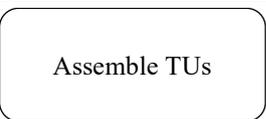
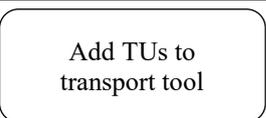
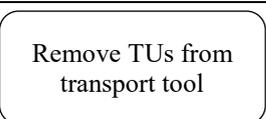
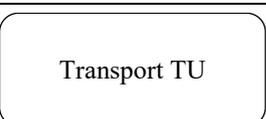
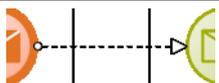


Figure 1: Systematic Modeling and Simulation of an MF System using both BPMN and DES

Regarding the second step of the methodology and the decision for material- or resource-oriented modeling in BPMN, a sensible approach is to use the SF to represent the more complicated (or more important) process type. Some examples for this are shown in this article.

Lastly, when creating the simulation model, the necessary effort can be reduced by translating between BPMN and DES. This is especially useful in material-oriented BPMN models, where the visualized sequence of activities shows strong similarities to the material flow in a DES model. Although the specific use always depends on the system, the modeling perspective, and other factors, it is generally possible to map certain elements between the two modeling tools. Table 1 shows selected examples. They are intended not only to show the underlying idea of translating between BPMN and DES, but also to clarify the content of figures 2 – 6. Although this table contains only a small subset of BPMN elements, it can be assumed that the vast majority of processes in MF systems can be mapped with it. This is firstly because only 20 % of the BPMN syntax is regularly used in practice (Muehlen and Recker 2013), and secondly because many qualitative and data-based relationships are mapped in BPMN via naming and comments, without requiring additional modeling elements.

Table 1: Comparison of Model Elements between BPMN & DES

BPMN	DES
 Start Event	 SubModelStart
 End Event	 SubModelEnd
 Intermediate Event	 Queue
 Process TU	 Server
 Assemble TUs	 Assemble
 Add TUs to transport tool	 AddTo
 Remove TUs from transport tool	 RemoveFrom
 Transport TU	 EntityConveyor
 Subprocess	 SubModel
 Branching Gateway	 Branch
 Sequence Flow	 Entity Flow
 Message Flow between End and Start Event	 Entity Flow

Generic Modules

When modeling complex systems in BPMN, it is recommended to identify subprocesses that are similar to each other and model them by creating and reusing a

generic module, much as the source code of a computer program may define a function once and then call it multiple times (White 2004). Similarly, in DES, submodels are used to create a hierarchical structure, using generic modules here as well. It is therefore possible to translate not only individual elements, but even entire compound modules between BPMN and DES. In the logistics domain, most subsystems can be attributed to one of the basic functions mentioned above. This opens the possibility of creating a selection of generic modules that can be used for many different MF systems using different parameters, much like a software library. This, too, promises a reduction in the amount of work that must be invested in modeling an MF system with the described methodology. Generic modules provide support when structuring systems or modeling processes, objects, and their dynamic behavior.

An example for this is shown in the following. This module called “Deliver TUs” describes the movement of a vehicle transporting TUs between different workplaces in a recurring sequence. Figure 2 shows the depiction of this process in BPMN, where a resource-oriented model is used. Those activities that take place at the same location can be grouped into a subprocess (see Figure 3), which can then be translated into a DES submodel (see Figure 4).

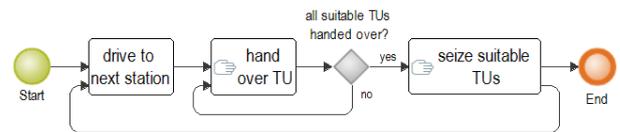


Figure 2: Module “Deliver TUs”: BPMN Process Model

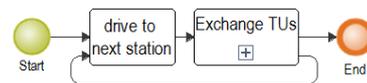


Figure 3: Module “Deliver TUs”: BPMN Process Model with subprocess “Exchange TUs”

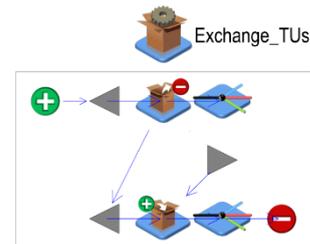


Figure 4: Module “Deliver TUs”, subprocess “Exchange TUs”: DES Model

When a simulation entity (representing a vehicle) enters the DES submodel, it will first pass through the upper part, releasing all suitable TUs to the Branch object, and then move on to the lower part, seizing all suitable TUs from the queue in the middle.

CASE STUDY

To evaluate the methodology described above, it has been applied to several real MF systems. One of these case studies, a production plant for motorcycles in Berlin, is shown in this section (BMW 2021, Welt 2017). Both qualitative and quantitative information for this system was gathered in workshops with process owners and cross-checked with specifications provided by developers of MF technology (e. g. conveying speed of belt conveyors). This MF system includes two assembly lines, “engine assembly” and “assembly”, that each consist of several stations. The supply of components from the manufacturing department to these assembly

stations is realized using a milk run delivery. This milk run is not used for the transport between the two assembly departments. Figure 5 and Figure 6 show parts of the BPMN and DES model for this system, respectively. The milk run delivery is a typical element of MF systems and can therefore be modeled generically, using the module “Deliver TUs” shown above, and additional EntityConveyors between the departments in the DES model, which represent the movement of the transport vehicles. In the BPMN model, activities labeled “[...]” are placeholders for additional processes, the illustration of which would go beyond the limits of this article.

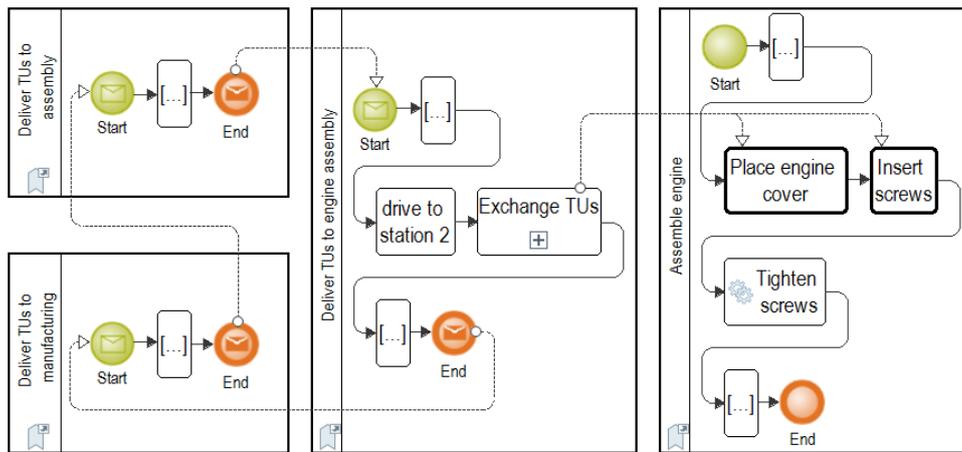


Figure 5: Motorcycle Production: Extract from the BPMN Process Model

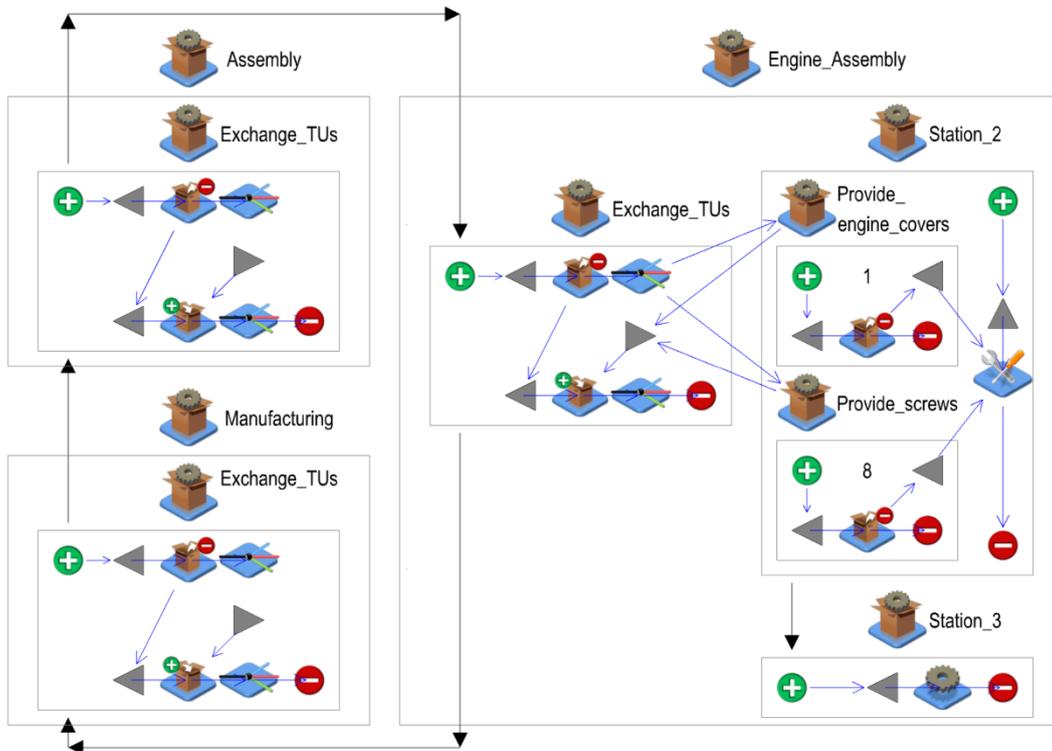


Figure 6: Motorcycle Production: Extract from the DES Model

As previously mentioned, modeling administrative processes with BPMN can be used to visualize weaknesses and help with improving the system. Using this case study, two examples show that the same is true for BPMN models of MF processes.

Firstly, Figure 5 shows that all the vehicles that deliver components pass through the departments of manufacturing, assembly, and engine assembly in the same order. There is no movement of individual components between assembly and engine assembly, so this procedure is inefficient. It can be improved by dividing the schedule so that the two assembly departments are supplied separately. The quantitative evaluation of this process change in the DES model shows that it could reduce the required number of vehicles by 25 %, thereby avoiding a waste of resources in the form of unnecessary WIP and transportation of material.

Secondly, a further consideration of the BPMN model reveals a potential improvement regarding equipment downtimes: In the engine assembly, a breakdown of the machine responsible for tightening the screws will result in a heavy accumulation of material right after station 2 or a shutdown of the assembly line. Depending on the reliability of this process, it could be sensible to implement a preventive measure, e. g. adding the activity “tighten screws” to the tasks of station 2 in case of a machine failure. This suggestion can be implemented into the BPMN model using additional gateways and intermediate events. Applying this change in the DES model shows that the “emergency plan” keeps the average throughput during a representative machine downtime at 60 % of its normal value. Without it, the machine failure would result in an increase in queue lengths and throughput times until station 3 is running again and the system can level off again.

DISCUSSION

Based on the results of the previous section, the two initial questions can be answered. For the combined modeling approach, it was first examined to what extent the BPMN syntax enables the mapping of different logistical processes. In this context, the principles of material- and resource-orientation were introduced. While the former is more suitable for linear connections of manufacturing and assembly steps, the latter should be used to depict transport processes in which material is picked up and delivered at several points. Especially when modeling in a material-oriented manner, most elements can be translated into DES without major problems. However, as the evaluation of important approaches from the literature showed, BPMN has proven to be largely unsuitable for mapping object-constrained activities and their quantitative effects, given that pools and lanes usually cannot represent the relationships in complex MF processes.

Lastly, it could be shown that BPMN models of MF systems are suitable to visualize typical improvement potentials of these systems as well as the qualitative advantages of the optimized processes. Changes in the MF can be modeled in BPMN by re-arranging the sequence flow between activities and events. Due to the lack of quantitative information in the BPMN models, DES is then used to test and evaluate the identified process improvements based on MF KPIs. It can be concluded that existing methodologies for systematic optimization are also applicable within this framework.

Although the presented methodology has a significant potential to increase the system performance while reducing the modeling effort, there are some limitations for its application. Modeling object-based activities with BPMN is hardly feasible and the scarcity of resources and other objects cannot be illustrated. Nonetheless, the approach can be applied to real-world industrial scenarios and process insights can be enhanced. The methodology clearly separates requirements and solutions and improves the usability, also by incorporating generic modules.

Compared to the related publications mentioned above, the methodology presented in this article applies BPMN specifically to MF systems without requiring any modifications of the existing syntax. Quite the reverse, combining BPMN and DES makes it possible to concentrate only on the common, useful, and well-understood modeling elements in each tool. This also results in a greater variety of eligible software, which is beneficial for applications both in industrial and academic settings.

SUMMARY

In this paper, a methodology for the modeling of MF systems using both BPMN and DES was presented. As a first part of it, the comparison between elements from both approaches allows for a translation from one framework into the other. Secondly, by synthesizing MF systems from generic modules, a plannable and time-saving modeling process could be created. The combination of standardized and well-known elements from both modeling languages allows for a methodology which is easy to understand and suitable for multidisciplinary teams. A case study at an MF process within a manufacturing system showed that the approach enables the identification and assessment of optimization potentials without the need for a costly real-world test run. Further research in this area could extend the “translations” to include rarer BPMN and DES elements, as well as expand the selection of generic modules to include other typical use cases in the MF domain. Regarding the detection and elimination of process weaknesses, the developed methodology could benefit from a more systematic approach specifically focused on the combination of BPMN and DES in the field of MF systems.

REFERENCES

- Arnold, D.; and K. Furmans. 2019. *Materialfluss in Logistiksystemen*. Springer, Berlin / Heidelberg.
- BMW Group. 2021. *BMW Group Werk Berlin*. bmwgroup-werke.com/berlin/de/unser-werk.html (last accessed on 2021-10-15).
- Forno, A. J. dal; F. A. Pereira; F. A. Forcellini; and L. M. Kipper. 2014. "Value Stream Mapping: a study about the problems and challenges found in the literature from the past 15 years about application of Lean tools". *The International Journal of Advanced Manufacturing Technology* 72, 779-790.
- García-Domínguez, A.; M. Marcos; and I. Medina. 2012. "A comparison of BPMN 2.0 with other notations for manufacturing processes". In *AIP Conference Proceedings* 1431 (Cadiz, Spain, 2011-09-21 – 23), 593-600.
- GitHub (unknown authors). 2021. *Modelio User Documentation*. github.com/ModelioOpenSource/Modelio/wiki/Modelio-User-Documentation (last accessed on 2020-10-07).
- Guizzardi, G.; and G. Wagner. 2011. "Can BPMN Be Used for Making Simulation Models?" In *Enterprise and Organizational Modeling and Simulation*, J. Barjis; T. Eldabi; and A. Gupta (Eds.). Springer, Berlin / Heidelberg, 100-115.
- Guenther, W. A.; and J. Boppert. 2013. *Lean Logistics – Methodisches Vorgehen und praktische Anwendung in der Automobilindustrie*. Springer, Berlin / Heidelberg.
- Hompel, M. ten; T. Schmidt; and J. Dregger. 2018. *Materialflusssysteme*. Springer, Berlin / Heidelberg.
- JaamSim Development Team. 2021a. *JaamSim – Discrete-Event Simulation Software*. Version 2021-05. jaamsim.com (last accessed on 2022-01-20).
- JaamSim Development Team. 2021b. *JaamSim User Manual*. jaamsim.com (last accessed on 2022-01-10).
- Muehlen, M. zur and J. Recker. 2013. "How Much Language Is Enough? Theoretical and Practical Use of the Business Process Modeling Notation". In *Seminal Contributions to Information Systems Engineering*, J. Bubenko et al. (Eds.). Springer, Berlin / Heidelberg, 429-443.
- OMG (Object Management Group). 2011. *BPMN (Business Process Model and Notation)*. Version 2.0.
- Pires, F.; A. Cachada; J. Barbosa; A. P. Moreira; and P. Leitao. 2019. "Digital Twin in Industry 4.0: Technologies, Applications and Challenges". In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. IEEE, 721 – 726.
- Robinson, S. 2006. "Conceptual Modeling for Simulation". In *Proceedings of the 2006 Winter Simulation Conference*, L. F. Perrone et al. (Eds.), 792-800.
- SOFTEAM Group. 2020. *Modelio*.
- VDI (Verein Deutscher Ingenieure). 2018. *Simulation von Logistik-, Materialfluss- und Produktionssystemen*. VDI Technical Rule No. 3633.
- Wagner, G.; O. Nicolae; and J. Werner. 2009. "Extending Discrete Event Simulation by Adding an Activity Concept for Business Process Modeling and Simulation". In *Proceedings of the 2009 Winter Simulation Conference*, M. D. Rossetti (Ed.). IEEE et al., Piscataway, NJ 2951-2962.
- Wagner, G. 2018. "Information and Process Modeling for Simulation – Part I: Objects and Events". *Journal of Simulation Engineering*, 1 (2018/2019).
- Wagner, G. 2021. *Information and Process Modeling for Simulation – Part II: Activities and Processing Networks*.
- WELT Nachrichtensender. 2017. *Die Motorradfabrik – Ein Superbike entsteht*. welt.de/mediathek/reportage/automobile/sendung171325367/Die-Motorradfabrik-Ein-Superbike-entsteht (last accessed on 2021-10-15).
- White, S. 2004. *Introduction to BPMN*. IBM Corporation.
- Zor, S.; D. Schumm; and F. Leymann. 2011. "A Proposal of BPMN Extensions for the Manufacturing Domain". In *New Worlds of Manufacturing*, CIRP (International Academy for Production Engineering) (Ed.).

AUTHOR BIOGRAPHIES

MAXIMILIAN WUENNENBERG is currently a research associate pursuing his Ph.D. at Technical University of Munich (TUM), Chair of Materials Handling, Material Flow, Logistics (fml), where he received his M.Sc. in Mechanical Engineering in 2020. He is responsible for the research project "Consistent Development of Material Flow Systems using a Model-based approach". His main research interests are Model-based Systems Engineering, Material Flow Systems and Data Analytics. His e-mail address is max.wuennenberg@tum.de and his web page can be found at <https://www.mec.ed.tum.de/fml/ueber-den-lehrstuhl/mitarbeitende/maximilian-wuennenberg/>

BENJAMIN WEGERICH received his B.Sc. in Mechanical Engineering in 2020. He is currently pursuing his M.Sc. in Mechanical Engineering at TUM, with his main research interests in manufacturing, logistics, digitalization, and Industry 4.0. His e-mail address is benjamin.wegerich@tum.de

JOHANNES FOTTNER is a professor for technical logistics at TUM, chair fml. His research areas are innovative identification technologies, digital planning of logistics systems and human factors in logistics. After obtaining his Ph.D. at TUM, chair fml in 2002, he worked in several management positions at Swisslog before becoming managing director of MIAS Group. Since 2015, he also has worked at the Association of German Engineers (Verein Deutscher Ingenieure, VDI) as chairman for Bavaria and vice-chairman for manufacturing and logistics. His e-mail address is j.fottner@tum.de and his web page can be found at <https://www.mec.ed.tum.de/fml/ueber-den-lehrstuhl/mitarbeitende/prof-dr-ing-johannes-fottner/>

SIMULATING THE AUTOMATION OF SORTING CRATES – A STEPWISE APPROACH

Armand Misund¹ and Henrique M. Gaspar¹

¹ Dept. of Ocean Operations & Civil Engineering, Norwegian University of Science & Technology, Norway
Email: henrique.gaspar@ntnu.no

KEYWORDS

automation, sorting crates, stepwise simulation

ABSTRACT

This paper presents an initial study on the task of automating the sorting of crates. It was paramount in the study that the technology used for the automation process is already existing and available to the client in Norway, therefore an existing pallet sorting system was used as state of the art. The simulation was performed via a stepwise approach, that is, first we simulated the process as it is now. Based on this manual case we simulated the same system under a new automated task at time, until the final case (close to) full automated. A set of KPIs was defined and used in the assessment of all cases, including the manual one. These are Opex [NOK], Capex [NOK], Reliability [Stops over time], Robustness [%], Probability of consequence, and Efficiency [crates per. hour]. Our finding overall is that the time it took to sort a given number of crates decreased with increasing implementation of automation through the various change cases. With reduced time consumption the efficiency and number of crates per hour increased, from 300 in the manual case, to 1200 in the fully automated case. The increase in automation also resulted in increased costs, both capital- and operating costs. If we look at the cost increase in relation to the capacity increase, the cost per sorted crate decreases by about 50%. The manual process sorts 300 crates per hour, and if we look at the cost per sorted crate, the fully automated case must sort 600 crates per hour to breakeven and justify the automation.

SORTING CRATES PROBLEM

The starting point of our work was based on the needs from a real grocery distributor in Norway, supplying supermarkets with food. The food is transported to the grocery stores on pallets and in various reusable plastic crates (Figure 1). Upon delivery, they also take used pallets and crates back, and deliver these to a sorting department. All pallets and crates are visually inspected for damage and defects, before being sorted by type. These are then shipped to the distributor for reuse. The crates studied comes in two sizes, the IFCO6420 and IFCO4314 (half size), Figure 2.

The core of our work was to be able to simulate the benefits of automating a currently manual process. The simulation will be done by identifying the main process and sub- processes of the actual manual case. The available technology for automation is then studied, with a gradual implementation of the automation in the current processes, keep as benchmark the same assumption and constraints.



Figure 1 – Reusable crates returned from grocery stores (above) and crates after the manual sorting process (below)

AUTMATION TECHNOLOGY FOR SORTING

Automatic sorting takes place in several different industries, typically industries with high volumes like food processing (sorting), parcel sorting, and waste

management. Although much robotic technology is already developed, the true integration in sorting is still in its infant stage. While most of the system is made of standardized parts, the integration of all the parts, the software that controls the system and in many cases the grippers, are custom made for each case. (Automated parcel sorting - An introductory guide, 2021)



Figure 2 - Unfolded crate IFCO6420 (above) and two crates folded and stacked (below).

In food processing, automation is found in high volume sorting jobs, like finding and picking small pits or stalks in fruits and nuts. Increasingly, they are also used in fruit and vegetable harvesting, where robots can select perfectly ripe fruit accurately and fast. (Industrial automation : TOMRA, 2021). In waste management sorting is often done for recycling purposes, where the different waste types can be cardboard, electronics, organic, metal, different types of plastic, and glass, and among these there is a large variation in sizes and shapes. (Bonello, Saliba and Camilleri, 2017). We comment in the following subsections the solutions currently available, divided into the functions that are relevant for the crate problem, that is, movement of objects, sensors and gripping

Movement of objects

The most used technology for transporting objects in a production and sorting environment are belt conveyor or roller conveyor. Belt conveyor systems are some of the most universally used and recognized machines in any industrial setting. (Conveyor Types & Configurations, 2021) The belt conveyors use a series of powered pulleys to move a continuous belt. This belt can be made of natural or synthetic fabrics such as polyester or nylon. In extreme temperatures or aggressive parts, a wire mesh or fiberglass belting can be used. In the modular belt

conveyors, the belting is made of individual, interlocked segments, usually made of hard plastic. These are easier to repair than flat belts models, easier to wash, and more resilient to sharp, abrasive, or otherwise problematic materials. Conveyor belt systems can be configured with back-lit belts for visual inspection and quality control, and vacuum belts for holding flat products to the belts surface. (Conveyor Types & Configurations, 2021)

Roller conveyors are a series of rollers supported within a frame where objects can be moved either manually, by gravity, or by power. Because of their adaptability roller conveyors are used widely in numerous industries, but mostly in logistics and manufacturing.

Gravity roller conveyors are useful as they use gravity force to move objects by putting the conveyor at a decline angle. This is a cost-effective solution, requires no power source which reducing the cost of operation, the need for maintenance, and time in maintaining the conveyor. This also provides a more environmentally friendly solution compared to a powered roller conveyor.

It is generally harder to control the conveyor speed, especially with heavy objects on the conveyor. Powered roller conveyors are more suitable when transporting object over a longer distance, there is a need to control the speed, or split the conveyor in zones with different speeds. The motion can also be controlled by sensors. Powered systems are more expensive and needs more maintenance than passive systems.

Sensors

In automatic sorting, different sensors are used, either alone or in combination. The choice of sensors is dependent on what feature to identify, material, shape, or colour. A combination of sensors can identify both material and colour/shape of the same object. In the (MRF) material recovery facility in Marsaskala, Malta, they combine a NIR sensor (Near-infrared) with visual imaging (2D) to identify PET plastic and to differentiate between white and clear versions. Near-infrared (NIR) spectroscopy is effective, and a common technology to identify various materials like paper, cardboard, metallic objects, plastic, and various foam products. (Bonello, Saliba and Camilleri, 2017)

Gripping

While the technology used for gripping is standardized, how it is used, in which combination and in what shape, can differ from project to project. Often the grippers are specially made for the object(s) to be handled, and even a combination of technologies is used like the universal gripper design proposed in (Bonello, Saliba and Camilleri, 2017) and seen in the picture below. Here both mechanical jaws and a retracting vacuum tube is used both individually and simultaneously, depending on the object to be lifted.

To extract the metallic objects, magnets or electromagnets can be used on ferrous objects, while eddy current separation technology sorts out 90% of non-ferrous metallic objects. Some objects can be sorted by air stabilization systems that pins the object to the conveyor and allows it to exit through dedicated outlets. In some cases, a cooperation between man and machine is a good solution, where some objects can be extracted manually to increase sensing precision. Figure 3 summarizes the automation technology found relevant for our study.

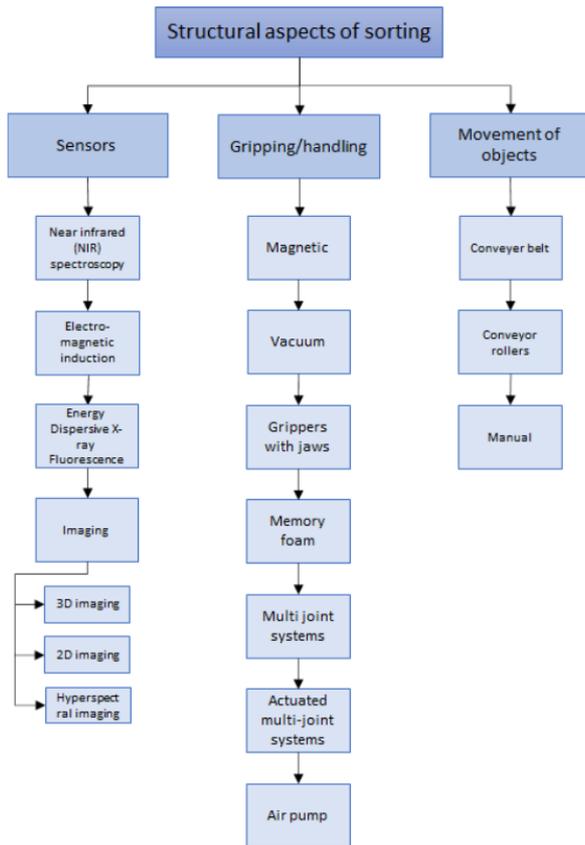


Figure 3 – Automation technology relevant for the sorting crates problem

PALLET SORTING

Sort Machine

Another relevant input was the successful automation of pallets sorting in the same type of industry, a robotic pallet sorting system called *Sort*, developed by Currence Robotics (2021). This system is now coming out of beta and is ready for production. According to the white paper from the company, *Sort* is a modular and scalable system, making it suitable for both small warehouse hubs and regional district centres, as it can handle a variety of pallets, Euro-, plastic-, eco-, half sized- and quarter sized pallets. It can process about 180 pallets per hour, working 24 hours a day. The future goal is 400-500 pallets per hour. (Currence robotics, 2021).

The machine is built up from five main parts, the infeed conveyor system, infeed tower with vision equipment,

all the output towers, robot with gripper, and outfeed conveyors. The machine can be controlled from the main cabinet, located next to the machine (Figure 4). This is a large machine that requires substantial space, as well as additional space in the front for the access of the forklifts. This system is designed as a modular system, so at any given point there can be added more towers for different types of pallets. By adding towers and types of pallets to sort, the sorting time will decrease, as there is still only one robot to do the sorting (Figure 4).



Figure 4 – Sort pallet sorting system (Currence Robotics, 2021)

One of these five main parts is the outfeed towers, one for each type of pallet to be sorted. It is these output towers that make the system modular, the ability to add more towers as needed, and in this way handle new types of pallets. What is called the robot, grabs, lifts, and moves the pallets to the correct outfeed tower. This is based on the assessment made by 3D vision sensors. This robot moves between the output towers on a horizontal traverser crane mounted on top of the towers. Figure 5 presents the automation technology from *Sort* also divided in the choose taxonomy.

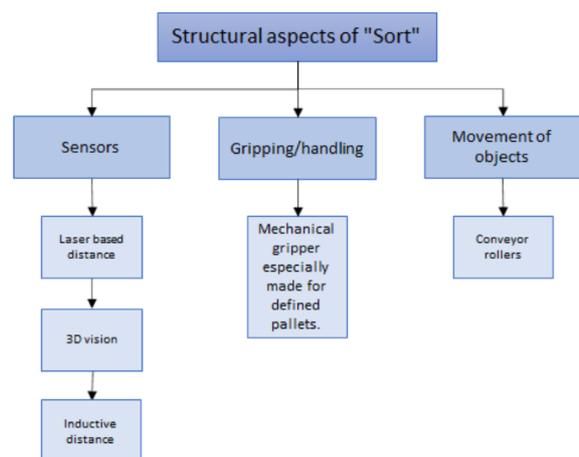


Figure 5 – Sort automation technology in terms of sensors, gripping and movement of objects.

Movement of pallets

Stacks of unsorted pallets are transported to the sorting area by trucks. These stacks are picked up by forklifts and moved in to the *Sort* buffer area, pallet infeed. The

infeed buffer is made of roller conveyors. These conveyors are slightly lifted at one end, giving it a slight slope towards the machine, making the pallets travel to the machine due to gravity. A mechanical stopper holds back the next stack of pallets, so the robot gripper can work with the front stack. In the same way, the output conveyors are sloped from the machine, making the pallets slide to the end of the roller conveyor. Here the mechanical stopper prevents the stack from sliding forward, until the stack has reached its predefined max height of 17 pallets. At the end of these output conveyors, forklift pick up the stack of pallets, moves it to the storage area where it is wrapped in plastic for stabilisation before it is placed on storage.

Sensors

The first sensor to detect new pallets are the laser based top limit sensor at the infeed tower. This is to prevent the stack of pallets from being higher than the machine can handle. The most important sensors are the six 3D vision cameras. They are all placed in the infeed tower and are used to examine and evaluate all the pallets, one by one. The pallets are then picked up by the robot and placed in the correct outfeed tower. In each of these output towers there are inductive limit sensors to detect the stack height. When the stack contains 17 pallets, it is released and will slide down the outfeed conveyor. The number of stacks on the output conveyor is monitored by a laser-based outfeed sensor. There are absolute encoders in all motors on the robot and the towers, this to know the exact position of the robot.

Gripping

When unsorted pallets are fed the system through the infeed buffer, the 3D vision sensors do the quality assurance and sort identification before the robot with gripper grips and lifts the pallet. The 3D vision sensors do a second inspection from underneath, before the robot transports the pallet to the correct tower, sorted by type or due to damage. The gripper can handle seven different types of pallets.

SIMULATING SORTING OF CRATES – FROM MANUAL TO FULLY AUTOMATED

Overall Assumptions and Methodology

This work builds on a stepwise procedure for simulation. All change cases will contain different degrees of automation, the first as it is now, (fully manual) until the last one being (close to) fully automated. Our premise is based on the fact that the limit of the manual case currently lies on 300 crates per hour with two operators. This is already not enough, especially since there are some large seasonal variations. To meet future requirements from main stakeholders, it is a requirement that an automated solution must be able to sort 1200 crates per hour.

The first technology to consider was automatic visual quality assessment as this is the sub-process that takes the most time. In addition to the use of time, the quality

of this process is also important. Later automatically sorting the crates according to the quality control is introduced. The crates are then ready to be transported to the designated area. The next case includes all the machinery for full automation for the sorting process. This may seem like a big step in relation to previous case, but to transport pallets with crates over relatively short distances of 5 -6 meters, and at the same time act as a buffer for the automatic sorting process, a roller conveyor will be a robust, simple, and good solution. As a buffer, there will be room for 5- 6 pallets of crates on a roller conveyor used today. If a larger buffer is needed, the roller conveyor can simply be expanded. The pallets with crates have enough weight that gravity roller conveyors can be used. The pallets that are waiting are held back with stop blocks integrated in the roller conveyors, which are controlled by sensors. Different sensors are used for different tasks. To detect the stacking height, both laser and inductive sensors are available. Several 3D vision cameras, all of which are located in the infeed tower, are used to verify both type and quality.

KPIs

The key performance indicators (KPIs) are decided upon stakeholders' analyses, based on conversations and interviews (Misund, 2021). The manual process is based on the time it takes to sort 51 crates, which is the average of IFCO- 4314 and 6420 on a pallet with several types of unsorted crates. This is then multiplied up to find the number of crates sorted per. hour. For the forklifts, which are included in all cases, the purchase price is divided by the estimated useful life, plus service and operating costs for the same period. The operators driving the forklifts is also to varying degrees included in all cases and here salaries, personal protective equipment, work clothes, sick leave, injuries are included in the various KPIs.

Opex [NOK]

Operating costs is the day-to-day expenses necessary for the process. This includes maintenance of equipment, and salary for the operators. The maintenance costs are divided into working time cost and material cost. In the processes described, there are two operators working, and, salaries, sick leave, personal protective equipment, and work clothes for the operators are included.

Capex [NOK]

Capital cost are major purchases designed to be used over a long term, included the installation. This is different for the different processes, especially the manual one in relation to those where automation is implemented to different degrees. For the manual case, the forklifts are responsible for the capital cost, as these are the only machines in use. For the three subsequent automated cases, the increased implementation of automated solutions (new equipment) will increase the capital cost for the processes. The installation costs will also increase with the amount of install equipment

through the various cases, these costs are divided into working time cost and material cost.

Reliability [Stops over time]

In this context, a reliable process is a process with few stops and if a stop occurs, the process restarts again quickly. This indicator gives an average number of stops in the process per unit of time. A lower number equals a more reliable system. This average number is based on interviews of the main stakeholders, experiences made by the operators (Misund, 2021).

Robustness [%, Probability of consequence]

System robustness is here understood as the ability to remain functioning under a range of disturbances. (Mens *et al.*, 2011). To measure robustness to the process, three levels of failure with increasing consequence have been defined. The probability of the different failures for each case is then established equal for all cases.

- **Failure 1** is a simple failure that can be quickly corrected without stopping the process or the process restarting in seconds. This can be a box or crate falling on the floor, and an operator uses one or two second picking it up and continues the process.
- **Failure 2** is a medium failure. This can be equipment that fails and needs to be fixed, either within a few minutes or within a few hours. Standardized equipment in stock is replaced within minutes, while many spare parts are currently in centralized remote warehouses in the county and can be delivered within a few hours.
- **Failure 3** is a large failure. This type of error stops the process for a long time, often several days. This can be caused by major errors that require service personnel and ordering and waiting days for parts. A larger or smaller fire in the plant will be able to stop the process for a long time and is considered a major fault.

Efficiency [crates per. hour]

The efficiency indicator tells us how many crates are sorted per hour. There is today a desire to sort approx. 300 crates per. hour, but all main stakeholders agree that an automated system must handle more than this in the future to be considered viable. We established that an automatic sorting process must handle 1200 crates per hour when designing our final automated case.

CASES

Case 1 – Manual Sorting (as it is today)

The manual process is sketched in Figure 6. Its elements are:

1. Truck with unsorted crates and pallets from grocery stores.
2. Pallets with crates unloaded front rucks.
3. Move pallets with forklift to sorting station.
4. Sorting station
5. Pallets with sorted crates by type and size.

6. Pallets with sorted crates wrapped in plastic.
7. Storage.
8. Pallets with sorted crates ready for shipping.
9. Trucks transporting sorted crates to grocery suppliers.

Trucks collect pallets and crates at grocery stores and bring these to the sorting facility. The different crates arrive mixed and stacked on Euro-pallets. Truck drivers collect the pallets and bring them to the sorting area. Here the different crates are separated and stacked on new euro-pallets. When the stacks have reached the desired height, they are wrapped in plastic film for safety, and put in stock. The process to be investigated is the sorting, and not the processes before and after. The product delivery from this process is both the sorted and quality-controlled stacks of crates, the storage, and the distribution of crates when needed (numbers 7, 8 and 9, Figure 6).

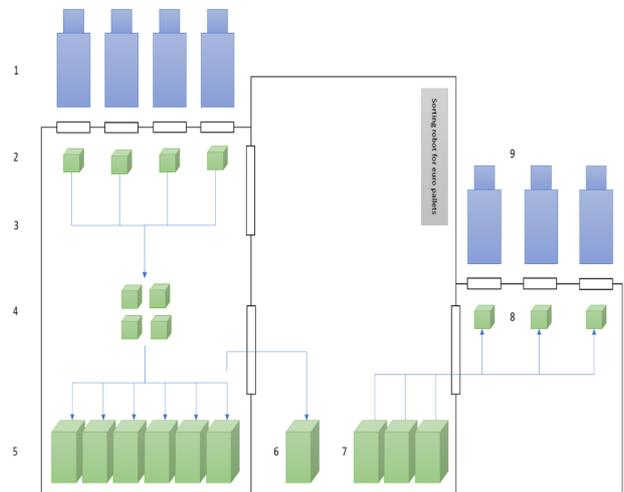


Figure 6 – Elements of the manual process of sorting crates

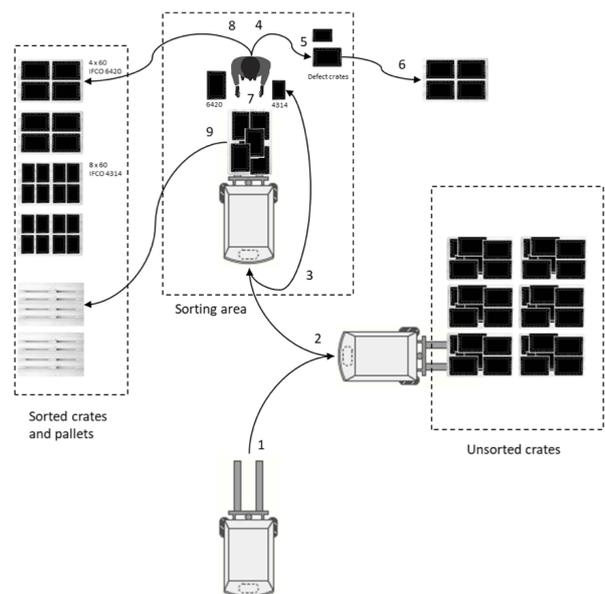


Figure 7 – Case 1 Manual Sorting

This process is nowadays mostly manual, with the help of forklifts to move pallets full of crates. The manual solution is working well if the volumes is moderate. When the volumes increase, which they do seasonally, the manual solution becomes a bottleneck. It simply takes too long to sort all the crates that arrive in a day, and overtime must be used to solve the challenge. The current elements of this process are therefore traditional, and consist of the forklift, the pallet with crates, and the crates themselves. This arrangement is observed in Figure 7.

The process is explained in the flowchart from Figure 8. With a forklift, one person picks up the pallet with unsorted crates (1) and drive as close to the pallets with sorted crates as is convenient (2). Here the person walks in front of the forklift and pallet (3) and starts to sort the crates by type and stack them on the floor in small and manageable stacks (7). All times while the person is collecting and sorting crates, the quality and functionality of the crates must be considered (4). Defect crates are sorted out (5) As the pallet with unsorted crates is empty, and there are several small stacks on the floor, the small stacks are then picked up by hand, and place on top of the already sorted crates (6 and 8). On each pallet there are 4 by 60 IFCO 6420 crates, and 8 by 60 IFCO 4314 crates (half size). The now empty pallet is then driven by forklift and placed on the designated place (9). The process then repeat itself, until all crates are sorted.

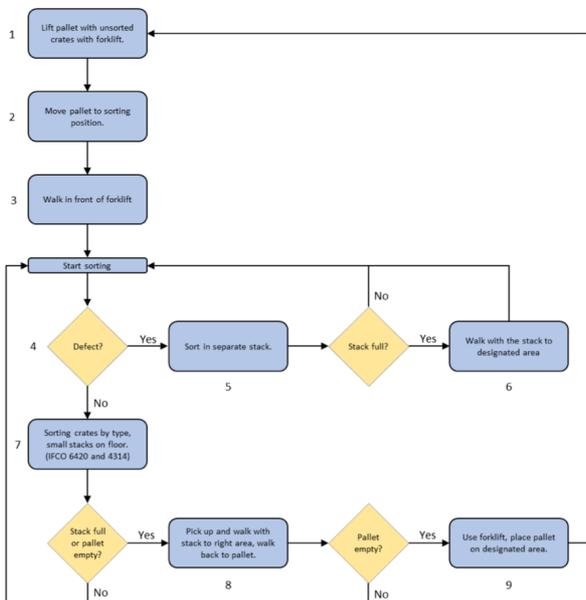


Figure 8 – Manual sorting process

Case 2 – Automatic Quality Control

The first step towards a fully automated process is to implement an automatic quality control of the crates. This is a consequence on the study that it takes an average of 255 sec to inspect 51 crates (Misund, 2021). This is the sub-process that takes the longest time to complete manually and both time use, and workload will

hopefully benefit from automation. A sub-process has been added, marked with the number 4, which symbolizes the automatic quality control that is performed on all crates before sorting (Figure 9). The elements that are new in this case are the parts that the automatic quality control consists of. This is a large implementation as it contains most of the sensors, the gripper for the crates and the electronics that control it all. Namely, a laser based top limit sensor, several 3D-vision sensors to perform the quality control and a gripper handling the crates.

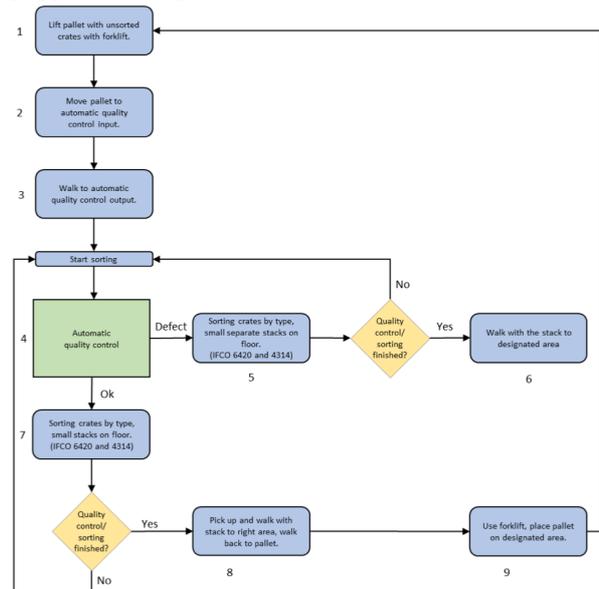


Figure 9 - Process with automatic quality control

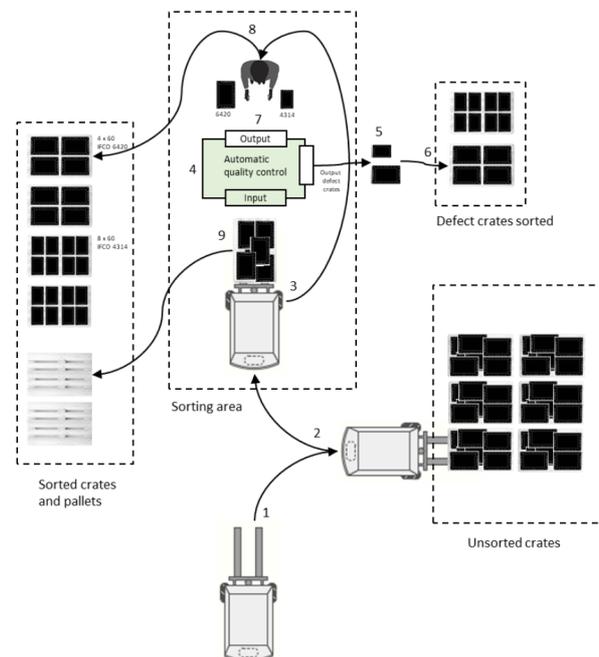


Figure 10 – Case 2 Automatic Quality Control

The first thing that meets the pallet crates are the laser based top limit sensor at the infeed. This sensor prevents the stack of crates from being higher than the machine can handle. The quality assurance is done by the 3D

vision sensors first from the top before the gripper lifts the pallet and the 3D vision sensors do a second inspection from underneath. The defect crates are then sorted out in a separate stack (Figure 10).

Case 3 – Automatic Sorting of Size and Quality

The next step towards fully automated solution is automatic sorting by size and quality. There are two sizes of the crates, the IFCO 6420 and the half size 4314. It will be the same gripper used in the automatic quality control, which is used to sort crates. There are also the same 3D vision cameras that are used as sensors to determine size. The additional feature is that The gripper must be given greater freedom of movement to be able to place the different crates in different places. This is solved by placing the gripper on an overhead crane that has a horizontal movement. To be able to pick crates at different heights from the pallets, the gripper already can move vertically. To distinguish the different sizes, the software that performs the quality control must be expanded to separate the crates by size. Otherwise, the same sensors are in use (Figure 11).

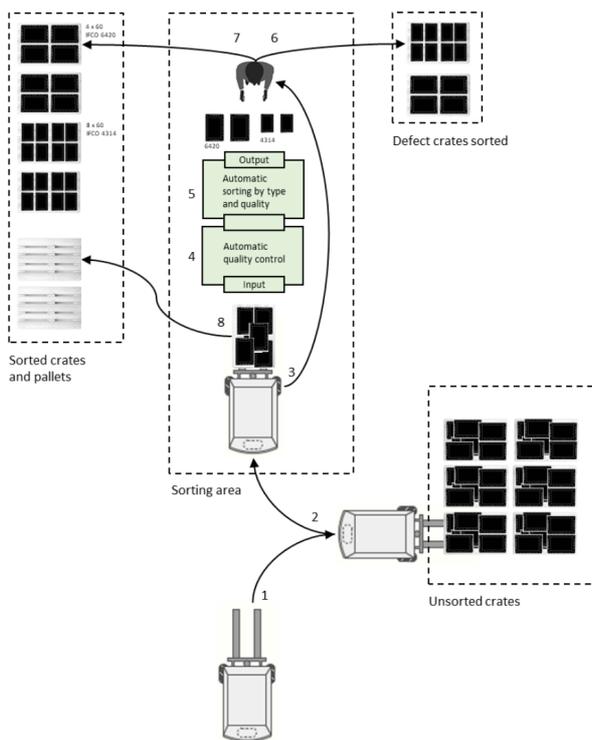


Figure 11 - Case 3 Automatic Sorting of Size and Quality

The crates are now stacked and presented to the operators according to size and quality. A predefined number of crates come stacked out of the machine ready to be carried to the specified place. The product from the automatic processes is now four different stacks, two stacks with approved crates one for each size, and two stacks with unapproved crates, one for each size.

Case 4 – Fully Automated

This is the third and final step towards fully automation of the sorting process. Although the process is now called fully automated, it yet requires manual work, as pallets with unsorted crates must be delivered to the infeed by forklift, in the same way the pallets with sorted crates must be picked up by forklift at the outfeed. The solution consists of 5 main elements. The first is the infeed conveyor which is fed with pallets with unsorted crates. These pallets are detected by a laser-based sensor that also detects the height of the unsorted crates, so they do not exceed the maximum height. The second main unit is the main sensors which consist of several 3D vision cameras which are used to identify the size and quality of the crates. The third main unit is the gripper which lifts the crate for examination from the underside. This gripper is mounted on the fourth main unit, an overhead crane that moves the gripper with crate to the right outfeed conveyor (Figure 12).

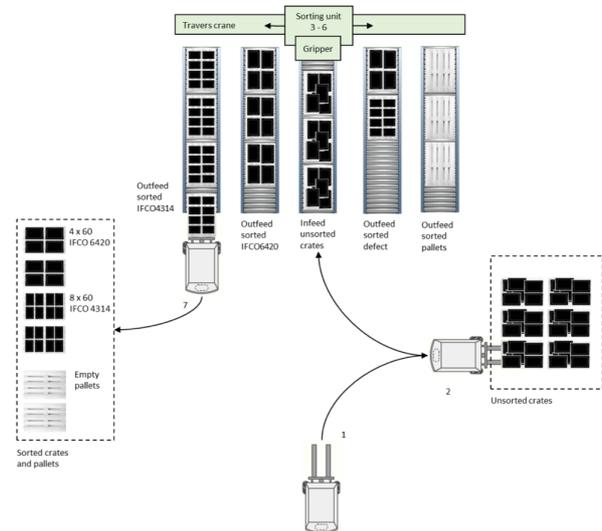


Figure 12 – Case 4 Fully Automated Sorting

The detailed process is presented in Figure 13. (1) The forklifts pick up pallets with unsorted crates, (2) which are lowered onto the infeed conveyor. Depending on the volume, the infeed conveyors can be adjusted in length to accommodate as many pallets as are necessary for the overall process. These conveyors are slightly lifted at one end, giving it a slight slope towards the machine, making the pallets travel to the machine due to gravity. A mechanical stopper holds back the next stack of pallets, so the robot gripper can work with the front stack. (3) When the pallet with crates arrives at the sorting unit, one by one the crate is identified, and quality checked from the top. Gripper then picks up the selected crate, lifts it up, for quality control from the underside. 3D vision sensors are constantly being used to assess size and quality. (4)(5) Once size and quality are identified, the entire gripper moves along the overhead crane, placing the crate on a pallet on the correct outfeed roller conveyor. (6)(7) When the pallet on an outfeed conveyor is full, it is released, and it rolls forward so that it can be picked up by a forklift. The machine will work as long as there are unsorted crates

on the infeed conveyor, and the sorted crates are picked up from the outfeed conveyor and this does not become full.

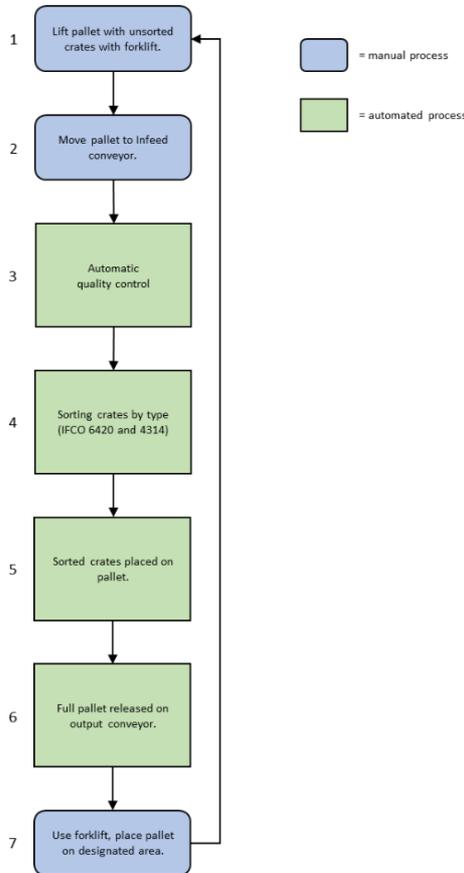


Figure 13 – Case 4 Overall process for the fully automated sorting

Evaluation

The starting point for this work has been a manual sorting process, where several different reusable plastic crates are sorted. Results are observed in Tables 1-4.

The unsorted crates arrive on pallets, and these pallets contain several different crates. During the observations, it was found that each pallet contained an average of 51 crates of 6420 and 4314. It is therefore based on the time it takes to sort 51 crates, which takes 612 seconds in the manual process, and which is gradually reduced to 153 seconds in the fully automated case 3, a reduction of 75% (Table 1). With reduced time consumption comes the ability to sort more crates per hour. From 300 in the manual case via 409 pieces and 712 pieces in change cases 1 and 2, to 1200 in change case 3. This is an increase of 400% from the manual case. Despite this increase in capacity, both operators from previous cases are retained in the cost calculations. This is because it will be impossible for one person to both feed the machine and empty it of sorted crates fast enough when it sorts 1200 crates per hour. In addition, there are the two forklifts, one for each operator. We observed an Opex raise costs of 77% due to an annual maintenance

cost of the automated solution (Table 2, Case 4). For capex costs, the purchase price for the structural components is included. We observed an 803% increase in capital costs. Capex costs are spread over 10 years, as this is a conservative estimate of the lifespan to the machine. Despite this increase in costs, the price of each sorted crate decreases. This price comes from the sum of Capex and Opex, divided by the number of sorted crates in each case.

Reliability is only described as stops per unit of time. With a slightly unclear definition, and divided opinions about what is a stop in the manual process, but at the same time with the certainty that some stops are, it is difficult to say if one stop per day, 0.13 per hour is a valid assessment. As for Robustness, probability of failure 1 is defined as a simple failure that can be quickly corrected like a box or crate falling on the floor, and an operator uses one or two second picking it up. These are small errors that are directly related to how large a part of the process the operators are part of. With an increasing degree of automation, the probability of such errors will be reduced. Therefore, the probability of failure 1 is reduced between the cases. The definition of failure 2 has been a bit inconsistent throughout the simulation. It has both been defined as components that fail and need to be replaced, and slightly larger work accidents such as crashing the forklift into a pallet that results in many crates falling on the floor. In definition one, with defective components, the probability should increase for each case due to increasing number of components used in the automation solution. In definition two, slightly larger work accidents primarily caused by forklifts, the probability will be unchanged between the cases because forklifts are used to about the same extent. Therefore, the probability of failure 3 is the same for all cases. This is defined as a major accident that stops the process for a long time. This could be a power outage in the area or a minor fire. Our assumption is that an automated solution requires electricity, but it is not a question of enormous power consumption. There are relatively light parts to be lifted and moved, with associated small electric motors in the various moving parts. A sufficiently dimensioned power supply in the room will not affect the power supply in the area.

Table 1 – KPIs for all cases

Task	Duration sec.				Unit	Change from manual case	Unit
	Manual process	Change case 1	Change case 2	Change case 3			
Time to sort 51 crates	612	449	258	153	sec.	-75	%
1 Opex (NOK per hour)	675	832	989	1198	NOK	77,5	%
2 Capex (NOK per hour)	26	92	157	209	NOK	803,8	%
3 Reliability (Stops per hour)	0,13	0,13	0,13	0,13	Stops per hour	0	%
4 Robustness (%)							
Probability of failure 1	85	70	60	10	%	-88	%
Probability of failure 2	20	20	20	20	%	0	%
Probability of failure 3	1	1	1	1	%	0	%
5 Efficiency (crates per hour)	300	409	712	1200	crates per hour	400	%
6 Cost per sorted crate (Sum Capex + sum Opex/sorted)	3,36	2,35	1,65	1,17	NOK per crate	-69,822	%

Table 2 -OPEX and CAPEX for Case 4

System performance indicators (SPIs)	Unit price	Operators	Days per year	Hours per day	Calculation	Sum	
						Unit	Unit
Opex							
Salary	472 080,00	2	239	4	472080 / (239*4)	494	NOK per hour
Sick leave	247,00	2	20	4	(247*2*20)/239	165	NOK per hour
Maintenance forklift	6 000,00	2	239	8	(6000*2)/(239*8)	8	NOK per hour
Personal protective equipment	4 000,00	2	239	8	(4000*2)/(239*8)	4	NOK per hour
Work clothes	3 000,00	2	239	8	(3000*2)/(239*8)	3	NOK per hour
Yearly maintenance Automatic quality control	300 000,00		239	8	300 000/(239*8)	157	NOK per hour
Yearly maintenance sorting unit	300 000,00		239	8	300 000/(239*8)	157	NOK per hour
Yearly maintenance outfeed	400 000,00		239	8	400 000/(239*8)	209	NOK per hour
Sum						1328	NOK per hour
Capex							
Purchase price forklift	250 000	2	239	8	(250 000*2)/(239*8*10 years)	26	NOK per hour
Automatic quality control	1 250 000		239	8	(1 250 000)/(239*8*10 years)	65	NOK per hour
Automatic sorting unit	1 250 000		239	8	(1 250 000)/(239*8*10 years)	65	NOK per hour
Automatic sorting unit	1 000 000		239	8	(1 000 000)/(239*8*10 years)	52	NOK per hour
Sum						208	NOK per hour

Table 3 – Efficiency per Case

Task	Efficiency (crates per. hour)				Unit
	Manual process	Change case 1	Change case 2	Change case 3	
Efficiency (crates per. hour)	300	409	712	1200	crates per. hour
Increase in number of crates sorted between cases.		36	74	69	%
Increase between manual case and case 3				400	%
Cost per sorted crate (Sum Capex + sum Opex/ sorted)	2,34	2,26	1,61	1,17	NOK per crate
Decrease in cost between cases.		-3,30	-29	-27	%
Decrease between manual case and case 3				-50	%

Table 4 – Cost per sorted crate

Task	NOK per sorted crate			
	Manual process (300)	Change case 1 (409)	Change case 2 (712)	Change case 3 (1200)
Cost per sorted crate at 300 crates available per hour	2,34	3,08	3,82	4,69
Cost per sorted crate at 400 crates available per hour		2,31	2,87	3,52
Cost per sorted crate at 500 crates available per hour		1,85	2,29	2,81
Cost per sorted crate at 600 crates available per hour			1,91	2,35
Cost per sorted crate at 700 crates available per hour			1,64	2,01
Cost per sorted crate at 800 crates available per hour			1,43	1,76
Cost per sorted crate at 900 crates available per hour				1,56
Cost per sorted crate at 1000 crates available per hour				1,41
Cost per sorted crate at 1100 crates available per hour				1,28
Cost per sorted crate at 1200 crates available per hour				1,17

Tables 3 and 4 show the efficiency and cost per sorted crate for a varying number of available crates for each change case. Breakeven point for each solution is presented in Table 4, and the fully automated solution presented itself profitable at 600 crates per hour.

DISCUSSION AND FUTURE WORK

The stepwise approach was found relevant for the given simulation. Large parts of the works were spent on mapping the current sorting process (Figure 6 and 7) as well as studying the available technology. Including changes one process at time were useful for understanding the gaps between manual and automation, as well as understand the consequences of changes. This was clearly mapped in the simulation, as every new product/process affected the KPIs.

There are many simplification, however, s made in this simulation also in connection with the estimation of KPIs. If it turns out that the real cost picture is significantly higher in change case three, it is still likely that the process is profitable, given the cost per sorted crate at 1200 crates per hour is half of what it is at 600 crates per hour. The KPI Reliability (stops per hour) did not provide the feedback that was intended, as it proved difficult to calculate or in other ways make the values probable. Robustness [%, Probability of consequence] also proved to be a theoretically difficult exercise, it was somewhat better defined than Reliability, but still not good enough. This shows the importance of well-thought-out KPIs, and possibly a test case to see what results they give and how easy it is to arrive at probable values. A simplified test case was also proposed by the supervisor, but not completed.

The chosen method provided a good understanding of the process, which is critical to be able to evaluate how the implementation of automation will affect the process. It is now clear in retrospect that these should have been defined differently. This is especially true for Reliability, where in interviews about the manual process, there have been divided opinions and perceptions about what a stop is, and what is just a normal work pace with small talk and natural breaks. Given the validity of the results in this thesis, it seems a valid risk to already automate this process. If they can

reuse parts and equipment from existing pallet sorting machine, the probability will increase that this will be an economically profitable project. The good results for the automated solution include those operators who are already performing this process manually. Together with the increased capacity, these are good arguments that one can keep today's employees while increasing sorting capacity.

Natural next steps is a 3D digital manufacturing simulation (REF). A detailed simulation of the automated process will further strengthen the value of the results, and if not, it can reveal possible errors or deficient assumptions. Building a test case of the process, a prototype, will also be a further option. By using existing parts from the pallet sorting robot, it is possible to see what can be reused. It will also be interesting to see if the available software can be reused, and to what extent it needs to be rewritten.

ACKNOWLEDGEMENTS

NTNU in Ålesund for the support for the research, as well as Currence Robotics for providing relevant data for the simulation.

REFERENCES

- Beumer, 2021; *Automated parcel sorting - An introductory guidet*
<https://knowledge.beumergroup.com/cep/automated-parcel-sorting-introduction> (Accessed: 7 April 2021).
- Bonello, D., Saliba, M. A. and Camilleri, K. P. 2017; *An Exploratory Study on the Automated Sorting of Commingled Recyclable Domestic Waste*, Procedia Manufacturing. The Author(s), 11(June), pp. 686–694. doi: 10.1016/j.promfg.2017.07.168.
- Currence robotics 2021. *Portfolio & Sort* Available at: <https://www.currence-robotics.com/> (Accessed: 23 March 2021).
- Mens, M. J. P. et al. 2011; *The meaning of system robustness for flood risk management*, Environmental Science and Policy. Elsevier, 14(8), pp. 1121–1131. doi: 10.1016/j.envsci.2011.08.003.
- Misund, A. 2021. *A System Engineering approach to automation of sorting crates*. MSc Thesis, NTNU. Available at: <https://hdl.handle.net/11250/2976427>
- MK, 2021; *Conveyor Types & Configurations*. Available at: <https://www.mknorthamerica.com/Blog/belt-conveyor-types/> (Accessed: 17 April 2021).
- Tomra, 2021. *Industrial Automation*. <https://www.tomra.com/en/sorting/food/why/industrial-automation> (Accessed: 6 April 2021)

TOWARDS A META-MODELING APPROACH FOR AN IORT-AWARE BUSINESS PROCESS

Najla Fattouch
FSEG Sfax, MIRACL Laboratory
University of Sfax
Sfax, Tunisia
Email: fattouchnajla@gmail.com

Imen Ben Lahmar
ISIM Sfax, ReDCAD Laboratory
University of Sfax
Sfax, Tunisia
Email: imen.benlahmar@isims.usf.tn

Khoulood Boukadi
FSEG Sfax, MIRACL Laboratory
University of Sfax
Sfax, Tunisia
Email: khoulood.boukadi@fsegs.usf.tn

KEYWORDS

IoRT; IoRT-aware Business Process; IoT; Robot; Business Process; Meta-modeling; Industry 4.0

ABSTRACT

In the context of Industry 4.0, the Internet of Robotic Things (IoRT) represents an attractive paradigm that aims to supply real-time data and automate tasks via human imitation. An IoRT is defined as the incorporation of IoT technology within robotic systems. In this setting, the business managers may improve performance and increase the productivity of their process by integrating the IoRT within their Business Processes (BPs). Nonetheless, this integration is not a trivial task due to the diversity of the IoT, robot, and BP concepts. In this paper, we address the incorporation of the IoRT within the BP through a lightweight extension of the Business Process Modeling Language 2.0 (BPMN 2.0) meta-model called *IoRT-aware Business Process meta-model (IoRT-aware BP2M)*. Our proposed IoRT-aware BP2M allows, on the one hand, to represent the main concepts of IoT, robot and BP in a unique meta-model, and on the other hand to specify some practical constraints to select the suitable device. As a proof of concept, we generated an IoRT-aware BP model in the agriculture field with some implementation details. Besides, we used the Bunge-Wand-Weber (BWW) ontology to prove the proposed meta-model's completeness and clarity. The obtained results show that the proposed meta-model has an acceptable completeness value and ontological expressiveness.

INTRODUCTION

With the proliferation of Internet of Things (IoT) technologies and smart devices, Industry 4.0 (I4.0) is gaining more attention. This new paradigm is revolutionizing not only the way manufacturing is working but also several areas like agriculture, health, education, etc. It aims to increase the efficiency of production that is strongly based on IoT and the Cyber-Physical Systems-enabled manufacturing in several areas. This industrial revolution gives birth to the IoRT

(Internet of Robotic Things) where IoT technologies and robotics systems should cooperate together to reach a higher level of automation. Thus, the essential pillars of an IoRT are IoT and robotics systems. The IoT allows the connection of anything with the internet via the use of some stipulated protocols to achieve different objectives, such as monitoring, smart recognition, automation, etc. However, robotics have been used for production since they can communicate and cooperate and even have learning ability. Moreover, the need of Industry 4.0 to an autonomous production can be performed only by incorporating robotic systems that can achieve repetitive tasks in the places of a human. A robot may refer either to a bot (Egger et al., 2020) or to an actuated device (machine) that can execute some tasks usually performed by users in manual ways (ISO, 2020).

In this setting, business managers can avail themselves of the IoRT advantages in their Business Process (BPs) by incorporating the latter within the traditional processes. Embedding IoRT within the traditional processes allows business managers, on the one hand, to speed up production and to avoid human errors and, on the other hand, to reduce their business costs by using IoT and robotic devices to accomplish traditional human tasks.

Nevertheless, this incorporation is not a trivial task regarding the diversity of the IoT, robot and traditional BP concepts. Most of the current works focus on integrating the IoT technology within the traditional process called an IoT-aware Business Process (IoT-aware BP) or integrating the robot within classic process called Robotic Process Automation (RPA). However, the IoT-aware BP fails to imitate human intervention, especially in the tasks that need the movements (e.g., pick the weeds, check the plant leaves). Moreover, the RPA refers to the automation of a BP based on a bot without considering the machine robots. Therefore, integrating IoRT within BP comes to bring the gap of the IoT-aware BP and the RPA. Recently, (Rebmann et al., 2020b) presented an approach for the real-time recognition of robot activities used for process assistance during a rescue mission based on IoT data. However, this model is applied to a specific field as it does not comply with any

meta-model. In (Rebmann et al., 2020a), the authors define the incorporation of IoRT within BP as a supplementary dimension to an already stressful and complicated situation. whereas (Masuda et al., 2021) defines it as a firms digitization through the automation of its process tasks. Nonetheless, they do not present any proposal to model this incorporation. We get inspired from these definitions to propose the IoRT-aware BP that designates the integration of the IoT and robots concepts within the traditional BPs. To the best of our knowledge, no existing work proposed a meta-model integrating the IoRT concepts within the traditional BP in a common process.

Therefore, this paper aims to present a generic metamodel called IoRT-aware BP2M that integrates the most common concepts related to IoT, robot and traditional BP. This meta-model represents a lightweight extension of the BPMN 2.0 meta-model. Moreover, it is built upon standards which makes it more realistic and useful for industry. The IoRT-aware BP2M specifies also some constraints useful to select the suitable executor device for a process task. As proof of concept, we used the BWW ontology based analysis to measure the completeness and clarity values of our meta-model (Wand and Weber, 1993). In addition, we generated an IoRT-aware BP model in the agriculture field, we give also some implementation details.

The remainder of this paper is structured as follows, the second section gives an overview of the related works that deal with BP's automation. In the third section, we present the IoRT-aware BP2M. The fourth section presents the implementation details of the generation of an IoRT-aware BP model based on the suggested meta-model. The fifth section presents the assessment of the proposed metamodel. Finally, the last section concludes the article with an overview of our future works.

RELATED WORK

Many research works deal with the integration issue of the traditional BP either with IoT or robot capabilities. Few of them have proposed solutions to integrate both paradigms with BP. This section gives an overview of the existing works.

Among the recent works that deal with the automation of a traditional process based on IoT technology and robots, we cite (Rebmann et al., 2020b) that designed and developed a global architecture for a real-time recognition of robot activities integrated into traditional process assistance during a rescue mission. Towards this objective, the authors use, on the one hand, an ad-hoc process model that is modeled via the BPMN language, and on the other hand, they use a rescue robot that is equipped with some IoT sensors. However, the proposed model does not refer to any meta-model through which it is possible to generate other models. Moreover, their proposal is limited to an ad-hoc process which is an unstructured process and it needs human intervention to decide which process activity to do and when to do it.

In the BP automation based on robot concepts, there are several proposed initiatives. We cite (Van Looy, 2020) that target the incorporation of the bot within the BP. Nevertheless, these works are limited to a theoretical study of integration without any modeling suggestion. In (Hindel et al., 2020), the authors introduce an exemplary process that is automated via the incorporation of a robot as a bot. During the traditional process transformation, the authors chose to keep the sophisticated and non-standardized tasks without automation. Moreover, these works do not give any detail about neither the integrated concepts nor the supported meta-model. In addition, they are limited to robots as bots and they do not consider the robots as machines which are also useful for tasks that require movement.

In the BP automation based on IoT technology setting, several approaches, are based on the BPMN standard to integrate the IoT concepts (Fattouch et al., 2020). Recently, (Seiger et al., 2021) propose a meta-model for mixed reality uses case where they use the BPMN as a defacto standard for the BPs. This meta-model can be improved by using a standard for the IoT concepts such as the ISO/ IEC 20924.

In summary, most of the existing approaches deal either with the IoT technology or the robotic one in order to automate the classic BP. Moreover, the existing solutions are limited to specific areas. To the best of our knowledge, no existing work proposed a meta-model integrating the IoRT concepts within the traditional BP in a common process.

IoRT-AWARE BP2M: METAMODEL AND CONSTRAINTS

This section is devoted to detail our generic IoRT-aware BP2M (meta-model). First, we detail the integrated concepts related to IoT and robotics. Then, we present a set of suggested constraints that are useful to select the type of an executor which can be an IoT actuator, a robot as machine or a bot.

IoRT-aware BP2M

This paradigm managed to achieve a boost in the business field thanks to its advantages. Nevertheless, this incorporation must be conformed to a meta-model that defines its main used concepts.

A meta-model is a set of concepts that are used to describe an interest area; it allows the generation of an infinite number of models as an instance in different domain. In this setting, we propose in this paper an IoRT-aware BP2M (IoRT-aware BP meta-model) which is a lightweight extension of the BPMN 2.0 meta-model as shown in Figure 1. This generic meta-model incorporates in addition to BP concepts, IoT concepts (colored in purple), robotic ones (colored in grey and orange) and some meta-classes (colored in green).

The BPMN is an Object Management Group (OMG) standard used to model BPs by adding new meta-classes (OMG, 2001). The BPMN has several advantages that

The integrated robot concepts (colored in grey) are the following:

- Robot: is an actuated mechanism that is characterized by its ability to move and it should be equipped by sensors.
- Control system: represents a set of hardware and software components having logical and control functions (e.g., launch, control, etc.).
- Manipulator: lays out a mechanism that allows the robot to perform its functions.
- EndEffector: represents a device (e.g., spray gun, welding dun, etc.) used by the robot to achieve some actions.
- RobotActuator: is a powerful mechanism within the robot device used to convert pneumatic, electronic, and hydraulic energy to the robot's motion.
- Operator: is a person that can launch, control and stop a robot.

Since a robot may also represent a software script (bot), we are interested to integrate the bot concepts in the IoRT-aware BP2M specification. Back to the literature, there are no predefined standards that specify the bot concepts. Consequently, we pinpoint the bot concepts based on some existing works such as (Romao et al., 2019). The following concepts (colored in orange) present the most used ones in the bot field according to the literature and they are presented in what follows.

- Bot: is a software program that runs on a physical or virtual machine.
- Controller: orchestrates the bot during its execution.
- Operator: represents a person who can launch, monitor and stop the bot execution.

The IoRT-aware BP2M contains some other metaclasses (colored on green) that we added to our metamodel to guarantee its simplicity and clarity. These metaclasses are here after:

- SensorDevice: is an added meta-class, modeled via a lane to display all used sensors for an IoRT-aware BP model.
- ActuatorDevice: is an added meta-class, modeled via a lane to present all used actuators within an IoRT-aware BP model.
- GatewayDevice: is an added meta-class, modeled via a lane, to show all used IoT gateway devices for an IoRT-aware BP model.
- RobotDevice: is an added meta-class, modeled via a pool, to present all used robots (machines) for an IoRT-aware BP model.
- Script: is a meta-class added to the meta-model to display all bots and their ControllerBot that are used by an IoRT-aware BP model. It is modeled via a Pool.
- ScriptBot: is a meta-class that modeled via a lane. It grouped a set of bots.
- ControllerBot: is a meta-class drawn as a lane to model all used controller bots on the IoRT-aware BP model.

- RobotManipulator: represents a meta-class modeled through a lane to indicate all used manipulators to execute a robot's tasks.

IoRT-aware BP2M constraints

The generic IoRT-aware BP2M includes also a set of constraints called *Selection Executor Constraints* as shown in Figure 2. These constraints specify which kind of an executor is recommended for a specific task. The IoRT-aware BP2M considers three main types of executor which are a robot machine, an IoT actuator and a bot. Some tasks should be executed by a robot machine where they require a movement, whereas the tasks that do not require the movement, should be performed by an IoT device (actuator). Nonetheless, the tasks that execute a script code are performed by a bot. In what follows, we detail the suggested constraints.

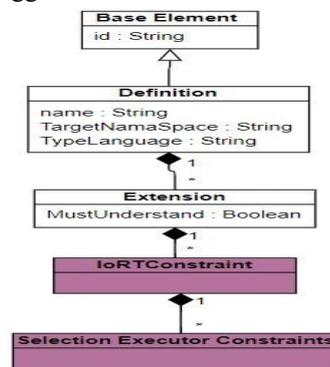


Figure 2: Selection executor constraints for the IoRT-aware BP2M.

- Constraint 1: only a robot as a machine can perform a task that requires movement. Otherwise, an IoT device (i.e., actuator) can execute task without movement.
- Constraint 2: a bot performs a task that has an executed code script. This constraint refers that a bot accomplishes only the script tasks.
- Constraint 3: a task performed by an actuator, can not be performed by a robot and vice versa.

IMPLEMENTATION

We have generated an IoRT-aware BP model in agriculture field from the suggested IoRT-aware BP2M as a proof of concept. Towards this goal, we have extended the BPMN Modeler 2.0 plug-in which is an open source Eclipse editor. This plug-in allows us to extend the traditional process concepts via the adding of new properties and/or icons. In addition, the BPMN Modeler 2.0 plug-in allows us to add new concepts. These extensions consist of adding three categories of element to the palette which are *IoT Components*, *Robot Components*, and *Bot Components* that refer respectively to the IoT concepts, robot concepts, and bot concepts as shown in Figure 3.

The generated IoRT-aware BP of a smart irrigation management system aims to enhance the water and nutrient use efficiency. The presented process starts with the sensing temperature, soil moisture and solar radiation data, which are recorded throughout the day as depicted by figure 3. These data are captured via dedicated sensors and stored in the cloud. After 12 hours, a message event launches a bot to access the data stored on the cloud and apply a machine-learning algorithm to make decision irrigation. Afterward, the bot sends a decision notification to *Receive irrigation decision* task. According to this notification, the process will end if there is no irrigation need. Otherwise, two actions are launched, which are *Start irrigation* and *Request picking weeds*. The irrigation and the picking weed tasks start at the same time. An IoT actuator performs the irrigation task while a robot device accomplishes the second task as it requires movement. Finally, both of them stop the execution as soon as they receives a notification from the *Finish irrigation* task or the *Finish picking weeds* one. The operator can interact at any time during the process execution to start, monitor and stop the robot and the bot. Without automation of some of these tasks, this process becomes more difficult due to the massive repetitiveness of tasks such as capturing data throughout the day, launching irrigation or picking weeds.

To validate the generated model, we have suggested a set of practical constraints defined in the meta-model and they are used to select the type of device to execute an IoRT-aware BP task. These constraints are described by using the Object Constraint Language (OCL). The OCL is a formal language, it intends to describe the expression on Unified Modeling Language (UML) and Meta-Object Facility (MOF) (ISO, 2020). To achieve this goal, we use both Eclipse modeling Framework (EMF) and OCLinEcore plug-ins. The EMF is a modeling framework that provides a set of modeling tools. However, the OCLinEcore plugin is used to embed the OCL constraints within an Ecore model (Meta-model) in order to enrich its provided models.

Referring back to the generated agriculture IoRT-aware BP model, Figure 4 shows the validation of the *Selection Executor Constraints*. At this level, we have validated

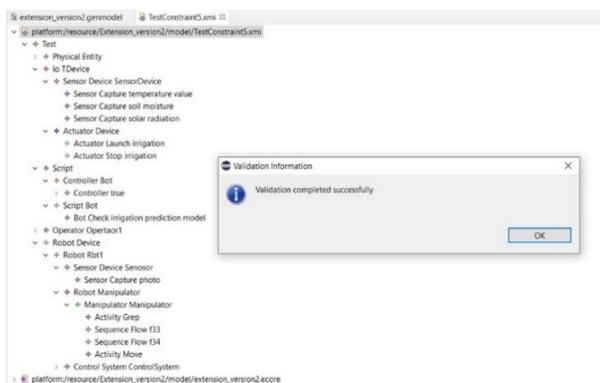


Figure 4: Validation of an agriculture IoRT-aware BP model using the *Selection Executor Constraints*

that a robot device performs only the tasks that require movement. Additionally, we have proved that each task that does not require movement is accomplished by an IoT device. Moreover, we have approved, that each used task that executes a script is achieved by a bot.

IOIRT-AWARE BP2M ASSESSMENT

The suitability of a meta-model can be measured through estimation of its completeness and clarity. The completeness measures a meta-model's capacity to cover all the IoT, robot and bot concepts. At the same time, the clarity defines the meta-model gaps by gauging some measures such as the redundancy.

To assess the completeness and clarity measures, we rely on an ontology-based analysis. Among the most important ontology-based analysis, we cite the BWW ontology proposed by Wand and Waber in 1990. BWW proposes two main mapping types: representation mapping and interpretation one. The representation mapping allows for each real-world construct, provided by the BWW ontology constructs, to be mapped to a grammatical construct provided within the design constructs (meta-model) (Wand and Weber, 1993). As a result, the first mapping will help assess the proposed IoRT-aware BP2M completeness. In contrast, the second mapping aims to evaluate the meta-model's clarity by evaluating its overload, excess, and redundancy degrees. Based on these two mappings, we estimate the value of the following ontology dependency measurements:

- **Completeness:** aims to estimate the degree to which the modeling technique can present a complete description for a real-world according to the BWW ontology's constructs. A modeling technique says complete when it covers all the BWW constructs. We measure the completeness value as one minus deficit degree (Recker et al., 2009). For our meta-model, the deficit degree is equal to the number of the BWW constructs which are not mapped to any IoRT-aware BP2M constructs, divided by the total number of the BWW constructs.
- **Clarity:** evaluates the gaps of a modeling technique. The clarity relies on the overload, redundancy and excess degrees described in what follows.
 - **Overload:** verifies whether the modeling technique provides a construct that can be mapped on one BWW construct. Following the formula proposed in (Recker et al., 2009), we measure the overload degree of the suggested IoRT-aware BP2M as the number of the constructs that are mapped to more than one BWW construct, divided by the total number of constructs. The overload occurs if each construct in the IoRT-aware BP2M is mapped to more than one BWW construct.
 - **Redundancy:** validates if each BWW ontology construct is presented through

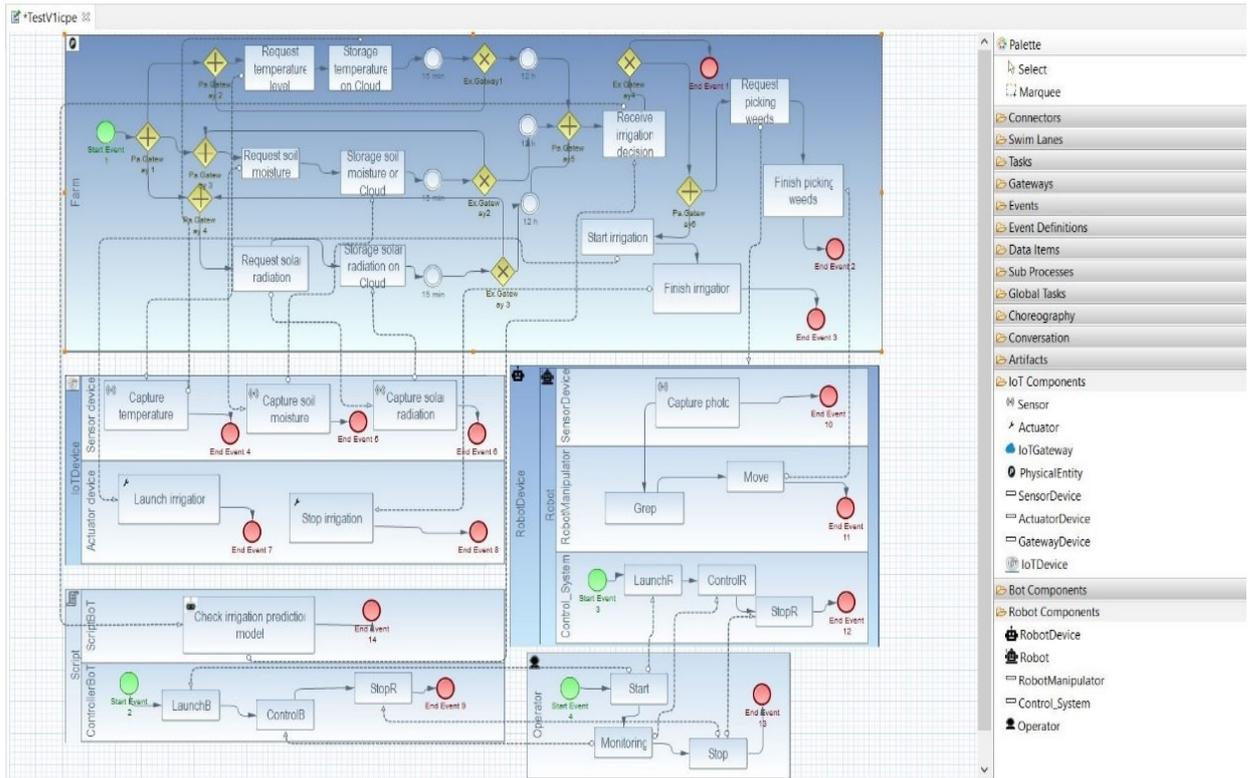


Figure 3: An extended BPMN 2.0 modeler plug-in in Eclipse tool.

exactly one modeling technique construct. The redundancy occurs when a BWB construct is mapped into several modeling technique constructs. In our case, we measure the redundancy degree as the number of the IoRT-aware BP2M constructs that have been mapped to the same BWB construct, divided by the total number of our meta-model constructs.

- **Excess:** measures whether a modeling technique construct can be mapped on at least one BWB ontology construct. The excess occurs when the modeling technique provides at least one construction that does not correspond to any BWB constructs. Following the formula presented in (Recker et al., 2009), the excess degree of the proposed meta-model can be measured by the number of the IoRT-aware BP2M constructs that have not mapped to any BWB construct divided by the total number of the meta-model constructs.

Table 1 summarizes the obtained results after the application of the ontological analysis. To evaluate the obtained results, we compare them with the threshold values found in the literature. Referring to the existing works, we notice that a modeling technique says complete when its completeness value becomes equal to 100% (Kudo et al., 2020). Nevertheless, according to table 1, we note that the completeness of the

suggested IoRT-awareBPMN is equal to 69%, which indicates that the deficit of this metamodel is equal to 31%, which is an acceptable result compared to the

Table 1: Meta-model assessment based on BWB.

	Completeness	Clarity		
		Overload	Redundancy	Excess
IoRT-aware BP2M	69 %	28.8 %	23 %	12.8 %

threshold value and the other modeling languages illustrated in (Recker et al., 2009). Besides, as highlighted in the literature, the overload can be a confusion source due to the provided constructs that may have many significations in the real world (Recker et al., 2009). Consequently, the overload value will be better when it is close to zero. Furthermore, the redundancy can be a confusion source due to several provided constructs for a real-world construct. Hence, the redundancy value also will be better when it is close to zero. Moreover, the excess can be a disarray source due to the provided constructs that do not have real meaning, according to BWB. As a result, the excess value will be better when it is close to zero. Table 1 shows that the IoRT-aware BP meta-model has a 12.8% as an overload construct value, 23% as a redundancy value, and its excess value is equal to 12.8%. These results demonstrate that the proposed IoRT-aware BP2M has an acceptable ontological

expressiveness compared to the existing threshold values.

CONCLUSION

In the context of Industry 4.0, integrating robotics systems and IoT concepts with traditional BP, will allow the business managers to improve performance and increase the productivity of their BPs. We denote by an IoRT-aware BP a business process that embodies the traditional concepts of a process within IoT and robotics ones. Nonetheless, modelling an IoRT-aware BP is not a trivial task regarding the diversity of concepts. Towards this issue, we proposed, in this paper, a generic meta-model called IoRT-aware BP2M that integrate the intended concepts referring to standards specifications. Thus, it is possible to generate an IoRT-aware BPs applied for any domain areas. In addition to the concepts specification, the proposed meta-model defines a set of useful constraints to recommend the convenient type of a process task executor (i.e. a robot as machine, an IoT actuator or a bot) for a defined IoRT-aware BP task.

As a proof of concept, we generated an IoRT-aware BP model in the agriculture field and we give some implementation details. To assess the IoRT-aware BP2M, we used the Bunge-Wand-Weber (BWW) as an ontology based analysis. The objective was to measure the completeness and the ontological expressiveness of the IoRT-aware BP2M. As a future endeavor, we are working to outsource some IoRT-aware BP parts to the Cloud and/or Fog with the intention of enhancing performance and reduce cost of the process execution.

REFERENCES

- Fattouch, N., Lahmar, I. B., and Boukadi, K. 2020. “IoT-aware business process: comprehensive survey, discussion and challenges”. In *International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. IEEE, 100–105.
- Hindel, J., Cabrera, L. M., and Stierle, M. 2020. “Robotic process automation: Hype or hope?”. In *Wirtschaftsinformatik (Zentrale Tracks)*. 1750–1762.
- ISO . 2018. In *Information technology — Internet of Things (IoT) — Vocabulary-ISO/IEC 20924*.
- ISO . 2020. In *Draft International Standard ISO/ DIS 8373*.
- Kudo, T. N., Bulcao-Neto, R. F., and Vincenzi, A. M. R. 2020. “Metamodel quality requirements and evaluation (mquare)”. In *arXiv preprint arXiv: 2008.09459*.
- Masuda, Y., Zimmermann, A., Shirasaka, S., and Nakamura, O. 2021. “Internet of robotic things with digital platforms: Digitalization of robotics enterprises”. In *Human Centred Intelligent Systems*. Springer, 381–391.
- OMG. 2001. Business process model and notation (bpmn), version 2.0.
- Rebmann, A., Rehse, J., Pinter, M., Schnaubelt, M., Daun, K., and Fettke, P. 2020a. “IoT-based activity recognition for process assistance in human-robot disaster response”. In *International Conference on Business Process Management*. Springer, 71–87.
- Rebmann, A., Rehse, J.-R., Pinter, M., Schnaubelt, M., Daun, K., and Fettke, P. 2020b. “IoT-based activity recognition for process assistance in human-robot disaster response”. In *International Conference on Business Process Management*. Springer, 71–87.
- Recker, J., Rosemann, M., Indulska, M., and Green, P. 2009. “Business process modeling—a comparative analysis”. In *Journal of the association for information systems*. 1.
- Romao, M., Costa, J., and Costa, C. J. 2019. “Robotic process automation: A case study in the banking industry”. In *Conference on information systems and technologies (CISTI)*. IEEE, 1–6.
- Seiger, R., Kuhn, R., Korzetz, M., and Aßmann, U. 2021. “Holoflows: modelling of processes for the internet of things in mixed reality”. In *Software and Systems Modeling*. Springer, 1–25.
- Van Looy, A. 2020. “Adding intelligent robots to business processes: A dilemma analysis of employees attitudes”. In *International Conference on Business Process Management*. Springer, 435–452.
- Wand, Y. and Weber, R. 1993. “On the ontological expressiveness of information systems analysis and design grammars”. In *Information systems journal*. Wiley Online Library, 217–237.

AUTHOR BIOGRAPHIES

Najla Fattouch is a PhD student at the Sfax University. She obtained her master degree in 2019 in information system and new technologies.

Imen Ben Lahmar is an associate professor in Computer Science at Higher Institute of Computer Science and Multimedia of Sfax—Tunisia. Her research fields include pervasive computing, fog computing and service oriented architecture.

Khouloud Boukadi is an associate professor in Computer Science at Faculty of Economics and Management of Sfax—Tunisia. Her research fields include business process, cloud computing, blockchain, and machine learning.

ACKNOWLEDGEMENTS

The work is carried out in the frame of the PRECIMED project that is funded under the PRIMA Programme. PRIMA is an Art.185 initiative supported and co-funded under Horizon 2020, the European Union’s Programme for Research and Innovation. (project application number: 155331/14/19.09.18).

A Data-driven approach for process Simulation Optimization: a case study

Romeo Bandinelli, Andrea Nunziatini, Virginia Fani and Bianca Bindi
Department of Industrial Engineering
University of Florence
50134, Florence, Italy
E-mail: romeo.bandinelli@unifi.it

KEYWORDS

Simulation, Data Driven, Electroplating, Process Optimization, Fashion.

ABSTRACT

The paper deals with the development of a data-driven simulation model for the process optimization of an automatic electroplating plant in the fashion industry. Starting from the process mapping of the production process using the Business Process Modelling and Notation (BPMN standard), an object-oriented simulation model has been defined using the commercial software AnyLogic®. Finally, the model has been validated and the plant has been optimized.

INTRODUCTION

The fashion industry is very popular in the literature, many contributions can be found. Despite this, most of the contributions have a brand owners' perspective, whilst labor or raw materials suppliers, mostly composed by micro and small companies, are less analyzed.

Recently, focal companies have faced with an increasing attention to delivery dates, cost reduction, and sustainability issues (May et al., 2015) (Brun et al. 2014) (Caniato et al. 2015) (Brun et al, 2017) (Brun and Castelli, 2008). As a consequence, this attention has moved toward all the Supply Chain (SC) actors, including metal accessories suppliers, that has started a process increasing their performances in terms of quality, time, and costs under the pressure of the brand owners and their increasing attitude toward a performance measurement systems implementation (Cagnazzo et al., 2010).

As widely known, production in the fashion industry is a complex process distributed between different actors operating at different levels. Production scheduling and optimization of a multi-level Supply Chain, composed by several small companies (mostly Small Medium Enterprises - SMEs) coordinated by a big company (which usually is the brand owner in the fashion industry), has been widely discussed in the literature (Fani et al., 2017). Simulation-optimization has been widely recognized as a useful tool to resolve such complex system considering finite capacity (Rahmani et al. 2013) (Ait-Alla et al. 2014).

According to this, this paper presents a data-driven simulation model for the process optimization of an electroplating automatic plant. The electroplating process

is the last job of the production process of metal accessories (e.g. buckles, chains, buttons) that have to be assembled in final products as bag, shoes, belts.

The paper is structured as follows. After a brief presentation of the metal accessories industry, the case study is introduced and deeply analyzed. Then some conclusion and future steps are reported.

METAL ACCESSORIES IN THE FASHION INDUSTRY

Metal accessories suppliers have never been deeply analyzed in the literature, despite their relevance from an economical point of view. Only few papers are related to such industry (Fani et al., 2016), even if they cover, looking at the Italian scenario, more than 3.5B€ of revenues in 2020, with more than 250,000 companies and occupying more than 14,000 employees.

One of the main reasons of the lower attention to these suppliers in the literature, compared to that one related to leathers or textile's ones, is that metal accessories do not represent at the costumers' eyes the fashion product, differently from the other components.

Despite this, the performances of these companies and the quality of the products greatly influence all fashion SC. At least a metal accessory has to be added to each bag, shoes or belt, and every delay in the delivery of this item, or a quality problem in a production batch inevitably leads to a delay or the need to reschedule production. This way, it is very important to optimize the production plan of the metal accessories suppliers, and in detail of the electroplating phase, that is the last one of the production processes (Bandinelli et al, 2021).

CASE STUDY

The analyzed company is a metal accessories producer for the fashion industry located in Florence and working for the major fashion brands in Italy. The company is composed of two independent production plants: one for manual electroplating and one for automatic electroplating.

The project carried out focuses exclusively on the production process of the automatic plant. This choice is possible thanks to the assumption of independence between the two plants. A generic representation of the layout, where the automatic plant is reported, can be seen in Figure 1.

Process Mapping

Before starting with the real modeling on the AnyLogic® software, a conceptual model of the system has been realized, in order to understand the logic of operation, the connections among the different activities carried out within the company and the principal events that characterize them.

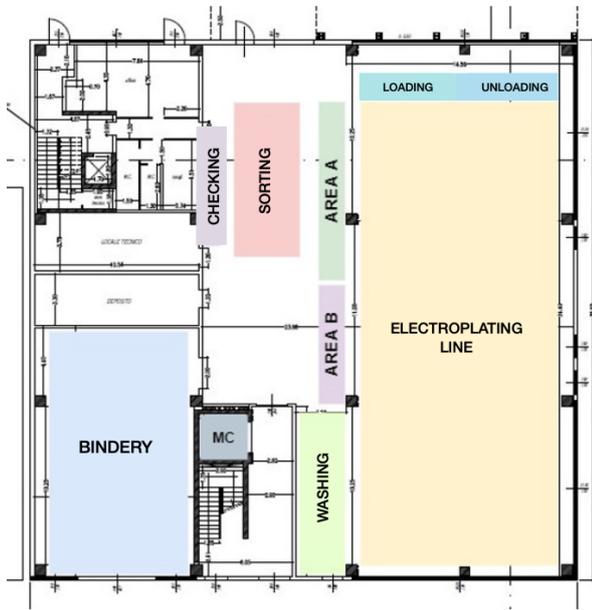


Figure 1: Plant layout

The production process begins at the moment in which the acceptance of the raw product, supplied by the customer (e.g. the brand owner), is carried out following a work order. Therefore, the arrival of raw material can be considered contextual to the order generation. This allows to neglect the operative flows related to the supplying of the material. Subsequently, the work order is transferred to the Binding department, where it is queued on shelves waiting to be taken over. The processing phase related to binding is generally divided into two main operations: the binding together of the individual pieces, using copper wires or rings, and the subsequent assembly of the latter on special frames.

Once prepared, the frames ready to be worked are transported by operators in the warehouse placed in *Area A* shown in Figure 1. From there, items are picked up one at a time by the operator of the electroplating department and transported to the *Loading* area of the plant.

To be placed inside the machinery, the frames are picked up by the trolley and hooked onto bars positioned in a special area, called *Loading*, from which they will be picked up by an overhead crane and moved into the *automatic plant*.

The plant consists of 150 galvanic treatment tanks, arranged according to a "zig-zag" folded line layout, so as to form 4 lines. The production flow is unidirectional and is forced on all four lines.

Unloading from the plant is managed in a similar way to the loading logic: once the last treatment has been completed, it is evaluated whether one of the unloading outlets is empty and therefore available to receive the bar. From this position the bars can be picked up and moved to the *Unloading* area with the same logic.

As soon as they are available in one of the loading inlets, the operator unloads the frames from the bars, places them on a trolley positioned near the unloading area and informs the machine of the unloading by means of a special button. In this way the unloading mouths can be freed from the empty bars, which are then returned to the inlet store in line 1, leaving space free for the next ones.

On exit from the lift, the trolleys are positioned in *Area B* because, before they can be used again for the processing of subsequent work orders, the frames need to be washed. This operation is carried out in a dedicated area to the side of the plant where there are washing tanks. After this operation both the trolleys and the frames can be brought back to the binding department.

Simulation model description

In this section, the simulation model is described, starting from the process mapping described in Figure 2, where the order generation is reported.

The arrival of work orders takes place physically within a designated hub from which, following an acceptance phase represented by a delay, they are transferred to the binding departments where they can be taken over and initiated.

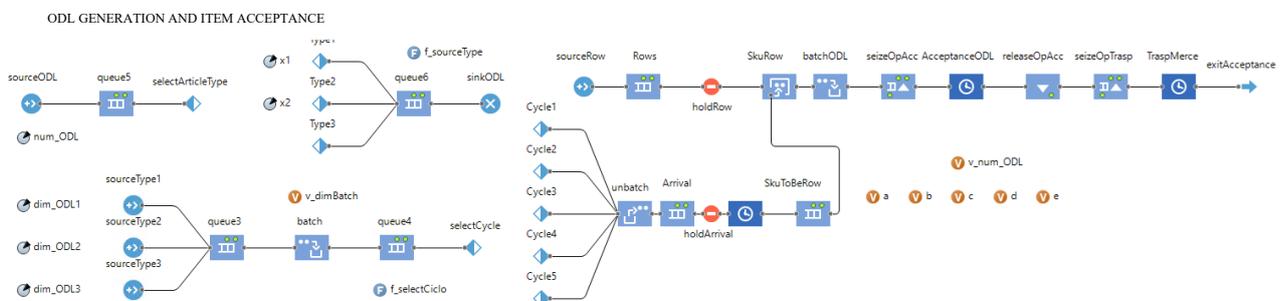


Figure 2: Process chart of the order generation section

Once the work orders have been generated and transferred into the tying department, the goods can be placed in a queue within the appropriate shelves waiting to be processed. The operating logic foresees that the work orders are taken over according to a FIFO (First In First Out) logic, as reported in Figure 2. Moreover, before the binding process can begin, for each work order the presence in the department of the resources needed to advance the goods to the next department is evaluated. In fact, according to the type of article of which the work order is composed, an appropriate type of frame and a certain number of trolleys necessary to transfer the goods towards the galvanic plant are required.

Both of them are generated when the model starts and located into an internal warehouse. They are taken and assigned to an item at the beginning of the production process and then released when the process ends, according to a circular flow.

The operations that are carried out in the bindery include the binding of the wires to the appropriate frames, the scheduling of the RFID devices placed on each frame, the subsequent loading of each of them on the trolley and its transportation to the automatic plant.

The operations of insertion of a specific item inside a container and its subsequent extraction have been managed throughout the process through the functional blocks *PickUp* and *DropOff*.

Since each type of item requires a specific frame and its capacity is variable, it was necessary to create a specific function that appropriately regulates both the release of the required resource and the number of threads to be tied on each of them.

Once the operations of this department have been completed, the looms loaded on the trolleys can be transferred to a dedicated warehouse close to the machinery, queued and waiting to be fed into the plant. In this phase the production flow is divided into two different ways: on one side the frames are picked up to be introduced into the automatic line, on the other side, once the trolley is emptied, it is led inside its dedicated area.

After this phase, the loading phase of the frames into the machinery takes place. The process involves the execution of three consecutive operations. The first involves the transportation of empty bars, initially located inside the inlet store at the beginning of the automatic line, towards the two loading mouths. Subsequently, once the bars are positioned correctly, they are loaded by an operator with the frames ready to be fed into the system. The operator at this time informs the machinery that the operation has been completed by pressing a special button. At this point the bars are once again picked up by the gantries and taken back to the input warehouse, following the movements previously carried out.

The logic of these operations envisages a preference for use of *loading mouth 1* since, through it, the bars can be led more quickly towards their destination, requiring fewer movements. Handling of bars in both directions is

carried out by the bridge cranes and shifters located into the line.

A schema of this part of the model is reported in Figure 3, where it is possible to note the use of the Material Handling Library of AnyLogic®.

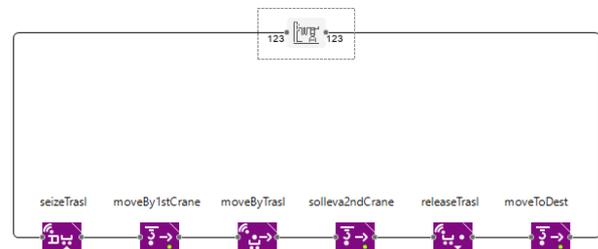


Figure 3: Loading area

The whole working process inside the automatic line has been realized following a different approach compared to the rest of the production process.

In fact, a *data-driven* method was followed, which will be suitably explained later in the paper, in order to increase the flexibility of the model (e.g. different plants configurations, change in the number of cranes, etc.) and the maintainability of the operating parameters by company personnel who do not necessarily have specific skills linked to the AnyLogic® software.

The process ends with the cleaning phase of the frames, which is necessary after the galvanic treatments. The return trolleys are therefore taken from the hoist and positioned in *Area B*. Frames are then washed and transported to the bindery, where they are once again available for use.

Data-driven model definition

Due to the complexity of the model in terms of, the model has been developed using a data-driven approach. Following this approach, the creation of the objects and their interaction between each other is entirely managed through external databases that provide the information for the construction of the model.

The data-driven approach gives the possibility to the user to easily change the characteristics of the plant without entering the model and consequently without the need to have knowledge about coding and simulation. On the other side, the effort needed in order to develop the model from scratch is higher in comparison with a traditional one. Moreover, the rules and the behavior of each element of the model has been described using the Java language, since this approach requires to define with custom action how entities move from one resource to the next one.

In detail, the custom objects that have been created are:

- *Galvanic Tanks*
- *Translator*.

Each of these is characterized by a process chart that describe the behavior of the entities within the object. The data regarding the processing time, transportation speed and how the entities move from one object to

another have been parametrized and stored in a database, that is read when the model starts.

Thanks to the functionalities made available by the AnyLogic software, it has been possible to connect the model to a database and parameters are stored within Microsoft Excel®.

Once the database had been defined, java functions able to execute queries which, at each step, determine the destination of the route to be followed and the operating parameters associated with it have been written.

A representation of the two objects is reported in Figure 4 and Figure 5.

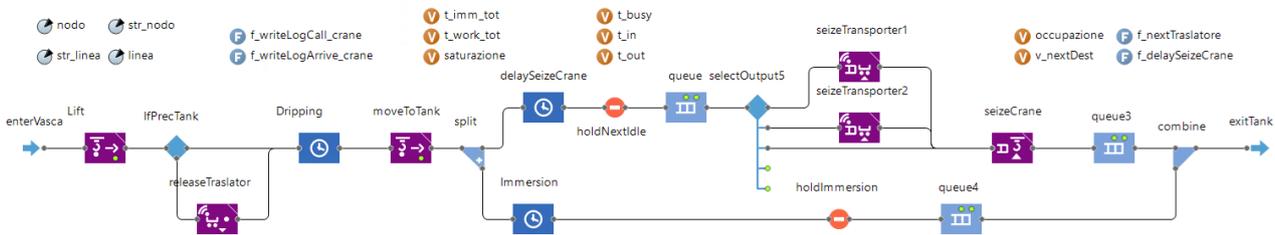


Figure 4: Galvanic Tank Object Process chart

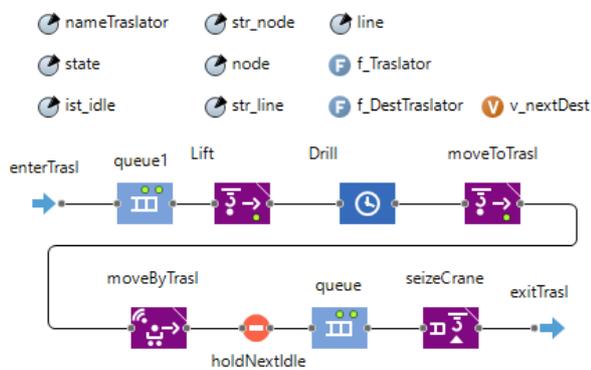


Figure 5: Translator Object Process chart

Due to the presence of cranes, no buffers have been set up between one resource and the following one, so a bar can be extracted and transported to its next step only when the second one is empty.

Moreover, it may happen that several different bars require simultaneously the use of a specific tank, consequently it is required a mechanism that defines the sequence of resources to be assigned to the different bars during the processing phase. The modeling of these aspects has been managed through the joint action of two functions.

The first function, launched every time a bar requires to be moved, executes a query that analyze the production cycles stored in the database and returns the possible next destination or destinations. The list of destinations is recorded, together with other parameters that characterize the bar, in an appropriate array. In each line are recorded the code and the processing cycle of the article, the type of resource and the name of the origin tank and the temporal instant of the movement request. This last parameter is essential for choosing the handling order of the various bars.

The second one is, instead, a *time-driven* function, activated cyclically at constant time intervals. This type

of function, in the simulation environment of AnyLogic, is managed through the Event functions.

This mechanism ensures a correct balance in the exploitation of parallel resources and allows a fair ordering mechanism of the movements of the bars that insist on shared resources. Once the appropriate destination has been defined for each of the bars to be handled, a further function executes a new query that reads the operating parameters of the subsequent destinations and assigns them to the entity that will be handled.

Process Charts and Crane Handling

As previously mentioned and reported in Figure 4 and Figure 5, flowcharts can be inserted inside every object, representing the activities that will be performed every time an entity crosses them.

The construction of a *data-driven* model also requires the realization of a mechanism that allows input, output and interfacing between each object. In AnyLogic this aspect is managed through the *Enter* and *Exit* functional blocks thanks to which each entity of allowed type can freely enter and exit each object.

To better clarify these concepts, it is useful to show the process charts of the two objects created (Figure 4 and Figure 5). In both cases functional blocks belonging to two libraries inside the program have been used: the *Process Modeling Library* and the *Material Handling Library*. In the specific case, the entities being handled are the bars to which the frames containing the semi-finished products to be treated are hooked. The means used for their handling along the four lines of the plant are bridge cranes and translators.

Animation of the agents is delegated to special tools known as *Space Markup*, which consist of special forms able to represent paths, movements or operations of the various objects within the system. As far as the realization of gantries is concerned, within the program there are dedicated *Space Markups* in which, with opportune parameters relative to their dimensions, to their shape, to the elevation height and to their

displacement speed, they allow to reproduce them graphically. For this reason, in order to faithfully reproduce the operation and movement times of the various gantries within their areas of competence, a graphical reproduction of the treatment tanks of the entire plant has been developed. Through AnyLogic, it is possible to create both a two-dimensional (2D) and a three-dimensional (3D) graphic representation, that have the double scope of reproducing the movements of the entities and generally constituting an excellent instrument useful for verifying the correct behavior of the model.

To such purpose, it has been realized a graphical reproduction in scale of all the resources inside the plant, as reported in Figure 6. In the figure are represented the tanks of treatment, disposed according to their order and layout, the mouths of loading and unloading, the warehouses of entry and exit, the translators, bound to execute movements exclusively along the lanes of their pertinence and the 12 bridge cranes opportunely distributed on the four lines. In this way, it was possible to realistically reproduce both the movements of the bars processed in the plant and the movements carried out by the *bridge cranes* and *translators*.

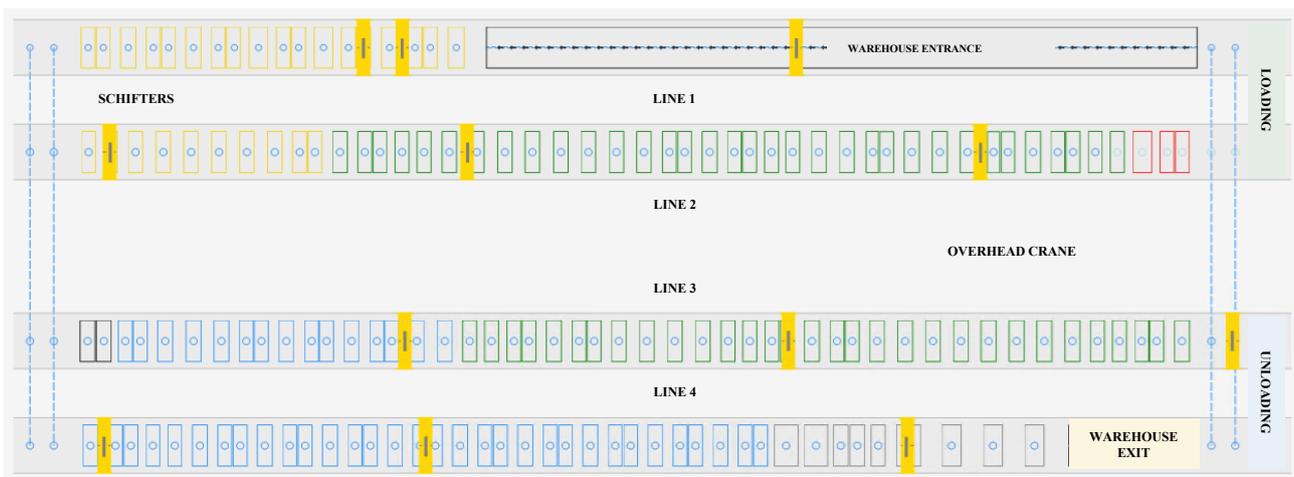


Figure 6: Two-dimensional representation of the automatic line

Once the immersion phase is over, the bar remains waiting for the next destination to be available, following the sorting and calling mechanism previously presented. In order to cope with a construction constraint linked to the process, a critical aspect of the automatic plant representation has been highlighted.

Every time the immersion time relative to a specific treatment end, it is necessary, in fact, that an overhead crane is immediately available to carry out the extraction of the bar from the tank. This constraint is due to requirements linked to the qualitative yield of galvanization, which requires that the treated product should not remain immersed for a longer time than allowed.

Although there are tolerances that allow a delay in the extraction of the bar from the galvanic bath, the permanence of the product inside the tank for too long risks compromising the success of the treatment.

In order to solve this problem, a split block has been added in order to generate a not real entity that, leaving the bath before the end of the treatment, is able to anticipate the request for transport by the bridge crane. With the split block, the *bar* entity is divided a fictitious copy is made. While, on the one hand, the real entity suffers a delay that is defined by the actual immersion time, the copy suffers a shorter delay given by the immersion time from which is subtracted the time that,

on average, the bridge crane takes to be physically present above the tank to begin extraction. Doing in this way, the entity copy can execute the seize of the resource with an opportune advance, guaranteeing the immediate availability to the conclusion of the treatment of the real entity. In the case in which the successive destination is, instead, a translator, the fictitious entity will supply to carry out also the seize of this last one, making also it available in advance regarding the conclusion of the galvanic treatment.

The functioning logic of the *Sideshifter* object foresees, instead, an initial lifting by means of a bridge crane and a dripping above the tank of origin, the transport inside the *Sideshifter* and the transfer towards the next line, from which it can subsequently perform the seize of the relevant bridge crane to transfer the bar towards the next treatment tank.

RESULTS

Once the model has been developed, a set of KPIs have been defined according to the company requirements and the model has been validated comparing them with the data coming from the actual physical plant.

Moreover, an optimization of the parameter combination to increase the global saturation have been determined, using the optimization tool OptQuest®.

Validation of the model

In order to validate this model, KPIs related to *productivity, immersion time, tanks saturations* and *queues* were examined.

The output data were extracted from the log files created by Anylogic®. Simulation has been replicated ten times and the simulation period chosen is three weeks, with five working days each. This time corresponds to the temporal horizon of the real production plan of the company. According to Figure 7, a warm up period of 4 days has been selected.

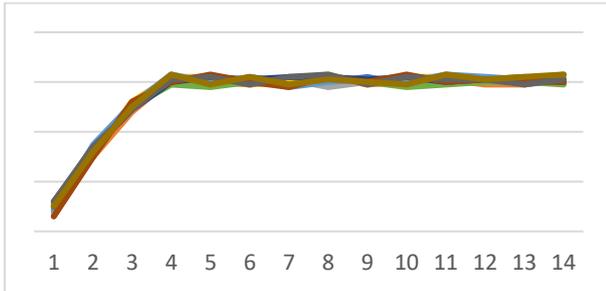


Figure 7: Warm-up

In order to compare simulation data and actual data were used the Minitab software, specifically, the Paired T-test. A graphical comparison between the simulated immersion time and the set tolerance is reported in Figure 8.

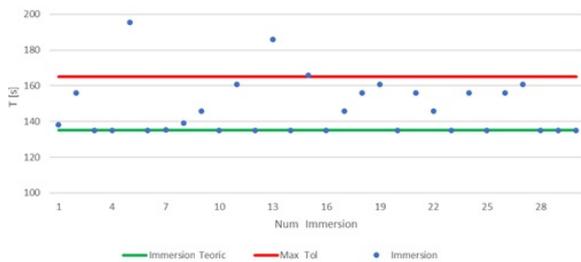


Figure 8: Simulated immersion time comparison

In blue there are the simulated immersions, in green there is the theoretical time and in red there is the max time tolerated. This graphic is related to the gold bath number one. The same graphic has been done for each bath, in order to validate that no bar entities have been stored a longer period than the target one plus the tolerance.

In order to provide a quantitative measure of the effectiveness of the model with respect to production constraints, an error measure was calculated for each tank.

This was obtained through the ratio, reported in terms of percentage, between the number of immersions with excess times and the number of total immersions within the same tank (i), as shown in the equation below:

$$\% error_i = \frac{\# immersion non ok_i}{\# immersion tot_i} \times 100$$

These data were used to determine the KPIs below:

- $\% error_{average} = \frac{\sum_i \% error_i}{N_{tanks}}$
- $\% error_{tot} = \frac{\# immersion non ok}{\# immersion tot} \times 100$

The first show the average accuracy of each tank, the second show global accuracy of the system. In both cases, a maximum limit of 5% was imposed.

Optimization experiment

A SME does not usually have the technical skills to be able to carry out scenario analyses. Then the optimization tool was used to perform the experiments automatically. In this way, the non-expert user only has to enter the input data, wait for the optimization results and obtain an optimal solution.

The optimization is based on OptQuest®, a proprietary software included into the Anylogic® software. The optimization process consists of repeated simulations of the model using different parameters each time. To do this a graphical user interface to set up and control the optimization process has been adopted. The final result provided by the program is the set of parameters that constitute the optimal solution related to the problem formulated. The interface is showed in the Figure 9.

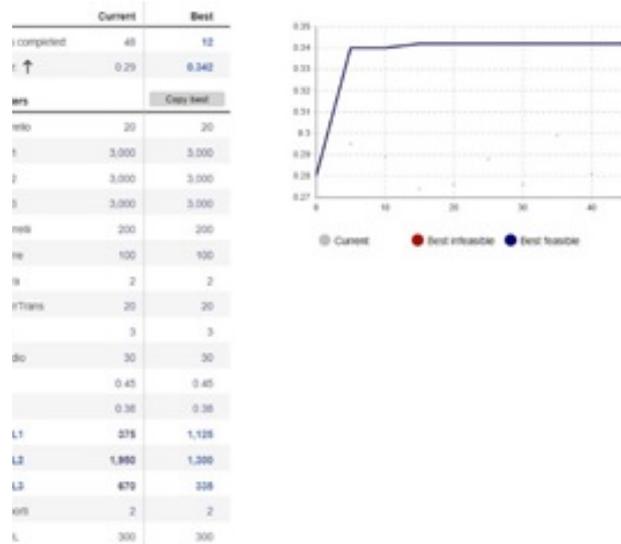


Figure 9: Optimization interface

Two different experiments were conducted: the first concerned the composition of the production mix and the objective function was the saturation of the production lines. The second case concerned the size of the production batches and their sequencing while the objective anointing was always the saturation of the production lines. The latter produced the best results. In Figure 10 the global saturation of each simulation of the first optimization is reported.

Finance and Economics and Social Science

Reflections on Assumptions for a Simulation Model of Dental Caries Prevention Planning in a Primary School

Maria Hajłasz and Bożena Mielczarek
Faculty of Management,
Wrocław University of Science and Technology,
ul. Ignacego Łukasiewicza 5, 50-371 Wrocław, Poland
E-mail: maria.hajlasz@pwr.edu.pl

KEYWORDS

simulation, dental caries, management.

ABSTRACT

The discrete event simulation method is commonly used to support decision-making in healthcare management. It is also used in planning the prevention of tooth decay in schools. Its usefulness largely depends on the concept of the model, which reproduces a fragment of reality along with the assumptions made. The aim of this paper is to discuss particular important modeling issues, which we faced, while developing a discrete event simulation model to support decision making in caries prevention planning in a sample primary school in one of the cities in the South-West Poland. We present reflections on the assumptions for the discrete event simulation model. The first stage of the simulation study confirms the relevance of the analysis of these assumptions and that their choice was appropriate. Therefore, the developed model may be the basis for further research and, as a result, be a tool to support management in planning the prevention of tooth decay in primary schools in Poland.

INTRODUCTION

The problem of tooth decay is considered a disease of the 21st century and affects more than half of the world's population (WHO, 2017). It is a relatively preventable disease, yet it is a big problem among the population. Therefore, prevention, properly planned and carried out from the earliest years of life, plays an important role. To take preventive action, a number of decisions related to their planning must be made. This requires financial, material and human resources, each of which is limited. So special care must be taken in planning the use of these resources to achieve the goal of preventing the spread of tooth decay. Managing processes in which people are involved and their decisions, behaviors, and health predispositions requires an approach that can take all these aspects into account. Thus, mapping health care systems is challenging and the reliability of the models developed can have a significant impact on management decisions.

One method that can be used to study preventive service delivery systems is simulation modeling. This method has been used for years to support management in health care, both in relation to medical issues and to support decision-making processes. One of the key steps in all simulation studies is *Model conceptualization*

(Banks Jerry et al. 2010). The assumptions formulated within every research are specific to the modeled system, its details, and the individual approach to the problem. Thus, it is not possible to write down universal rules for model conceptualization.

This paper focuses on the issue of formulating assumptions within the discrete event simulation (DES) approach to plan preventive care for dental caries in primary schools. The objective of this paper is to discuss particular important issues that we faced while developing a discrete event simulation model to support decision-making in caries prevention planning in a sample primary school in one of the cities in South-West Poland. The remainder of this paper is organized as follows. The next section describes the literature background. Subsequently, the main assumptions formulated in the conceptualization stage are presented. Then the simulation model and its verification and validation are discussed. Finally, conclusions and future research steps are presented.

LITERATURE BACKGROUND

Simulation methods are most often categorized according to 4 categories: discrete event simulation (DES), agent-based simulation (ABS), system dynamics (SD), or Monte Carlo (MC) (Brailsford et al. 2009). These methods are used in healthcare research (Katsaliaki and Mustafee, 2011). Different application areas may be distinguished; apart from health policy, simulation methods are being used in diagnosis and improvement, forecasting, medical decisions and threats (Mielczarek, 2014). In the field of dental caries prevention, examples of their use can also be found, but such studies are few and there is still a wide gap to fill.

In the area of simulation methods used in the context of research that addresses caries prevention management issues, papers using each of the four simulation approaches may be found. The ABS was a good method to model mechanisms affecting the occurrence of caries and to verify the effects of preventive interventions (Heaton et al. 2020). To determine the optimal combinations of staffing levels and sealant stations for school-based sealant programs, the DES method was chosen (Scherrer et al. 2007). The SD method was used to examine the relationships between dental caries status under different policy options (Urwannachotima et al. 2019) and to investigate the complex interrelationships among sugar-sweetened

beverage tax, sugar consumption and dental caries (Urwannachotima et al. 2020). Using MC, the lifelong costs of caries with and without fluoride use were modeled (Johnson et al. 2019) and the possible financial effects and impact on caries prevention of receiving fluoride varnish were estimated (Scherrer and Naavaal, 2019).

Thus, the utility of simulation methods in supporting management in caries prevention has already been noted. Each of the above-mentioned simulation methods is an advanced approach, the application of which requires knowledge of many elements from different areas. Undoubtedly, one of them is the real system modeled and the dependencies that enter into it.

PROCESS OF PROVIDING PREVENTIVE SERVICES

The process of providing preventive services from the beginning to the end of primary school was modelled according to the framework presented in Figure 1.

After starting school, the pupil is placed in a kindergarten ($n=0$). During the school year measured in days (d), preventive services are provided. When the school year ends, she or he is placed in the next grade to begin the new school year ($n=n+1$). At the end of the last grade ($n=8$), the pupil leaves the school. The watch symbols indicate the passage of time.

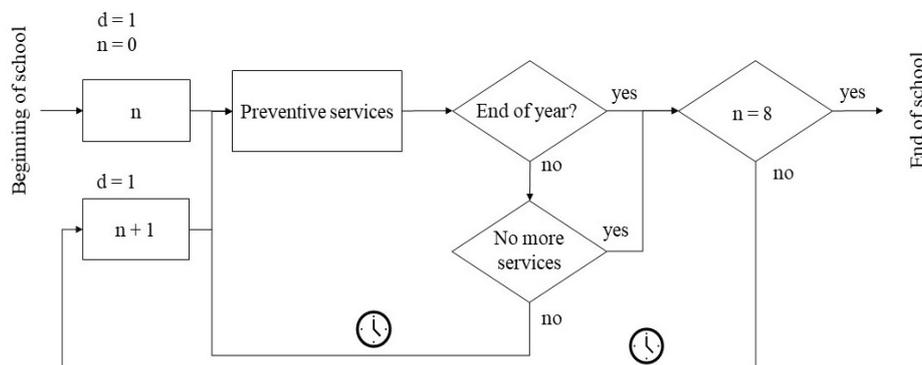


Figure 1: Flowchart describing preventive dental caries services from start of schooling through completion (d -day; n -grade)

Preventive services are provided to pupils during their education. Primary school education lasts 9 years, from kindergarten to grade 8. During this period, pupils are supposed to receive preventive care for caries. Preventive care in schools is financed by the state, but the way it is carried out is the responsibility of the directors of each educational institution. Preventive services include educational talk, fluoridation, dental check-up, and sealing of the first molars.

MAIN ASSUMPTIONS FORMULATED DURING THE CONCEPTUALIZATION PHASE

Assumptions relating to simulation parameters

Simulation parameters are a part of every simulation research. The main problems are related to the values that should be applied. How should a system be observed,

whether we can collect statistics right away, whether we need a warm-up period, or how many replications should be conducted?

In the case of the system of providing preventive services in primary school, observations can be carried out during the entire education or particular periods of education, for example, depending on the age of the pupils.

The system we present is observed for 9 years. This is a period of primary education in Polish schools. Before statistics can begin to be collected, all classes in the school must be filled with pupils. Therefore, the warm-up period is also equal to 9 years. Minutes are the base units of simulation because the duration of preventive services is usually a few minutes. The number of replications is 5.

Assumptions related to the inflow of pupils

Under the DES approach, pupils are represented in the model by entities that move through a system of caries prevention services provided at school. In the inflow of pupils section, we can consider how they are expected to flow into the model and which of their characteristics are relevant to the research. Moreover, we must determine whether to account for migration of pupils between classes or between schools and whether to assume absences, which may be due to a variety of reasons.

Should we model the individual inflow of pupils to a school or the inflow of whole classes? Modeling the inflow of pupils as individuals involves the subsequent need to group them into classes because in school, pupils attend the class with which they spend the most time and within which they are directed to the various services that take place in school. In contrast, modeling this inflow as a class involves separating classes to attribute the individual indicators changes and within certain services that are done individually.

This problem was solved by analyzing the use of two options. As described earlier, each has its pros and cons. It was decided to model the inflow of pupils within classes. Every year there are five incoming classes, the size of which is 20 which is determined by average number of pupils per class from Statistics Poland (SP). Furthermore, the model did not assume changes in the

number of pupils due to migration or absenteeism. It was assumed that all pupils were in a given school from the beginning to the end and that they were always present during preventive services.

Is it important to include the gender and age of the pupils in the model? Within the pupils' characteristics that can be modeled, the primary ones include gender and age. It is important to consider whether there are significant differences between genders that may be relevant for research. On the other hand, as for age, it is known that there are as many different birth dates as there are pupils in a school. In a given class, there may be differences in the years of the pupils. In addition, there are times when pupils start school at an earlier age. So, as with gender, is it necessary to account for such differences or is it necessary to assume equal age? We should also consider how detailed we want to analyze the demographics of the community. To decide whether to include gender, a dmft indicator values (*dmft: decayed, missing, and filled teeth*) were compared across the study region. These indicators are commonly used in dentistry to assess dental health. The symbol dmft is used for primary teeth and the DMFT for permanent teeth. The higher the index value, the more advanced the caries.

The dmft index for pupils at 6 years of age in the study region was 2.89 ± 2.90 and 4.32 ± 3.36 for women and men, respectively. At 7 years of age, the index values were 5.43 ± 3.28 and 5.41 ± 3.24 for women and men, respectively (Olczak-Kowalczyk et al. 2021).

By analyzing the data on the problem of dental caries and based on the reviewed literature, it was determined that gender would not be distinguished in the pupils. Although at age 6 the gender gap in indicators values was noticeable, by age 7 this difference had already disappeared.

Regarding the age of the pupils, it was assumed that the pupils start school at the age of 6 and are one year older in each subsequent grade until they reach the age of 14 in the last grade. We did not account for differences due to different birth dates.

Assumptions related to the structure of the school year

Within DES, we can see the dynamic changes that occur at specific points in time. Pupils start school, attend classes, and receive preventive services. But during the course of their education, there are days off, sometimes regular like weekends, and sometimes less regular like certain holidays or winter breaks. This raised the question how to model the structure of the school year so that it reflects reality as well as possible and is a universal one.

In Polish primary schools, pupils attend school from September to June, followed by a vacation period. During the school months, pupils have Saturdays and Sundays off, public holidays, two breaks for Easter and Christmas, and two weeks of winter holidays. The most detailed approach could be to write a calendar with school days and holidays for each simulated year. However, despite the differences that exist between each calendar, they are scheduled according to the same rules.

As part of making assumptions about the structure of the school year, we conducted detailed analyses of school calendars from 2011 to 2021. In these years, the number of school days ranged from 183 to 188, the winter break was always 2 weeks, the Christmas holiday break was 6 to 9 days, and the Easter holiday break was always 4 days. So, in the end, the model assumes the sample calendar corresponding to the 2021/2022 school year. It assumes 187 school days, a Christmas holiday break of 7 days, an Easter holiday break equal to 4 days, a two-week winter break, and 6 working days off due to public holidays. These components plus weekends add up to a school year that runs from September to June.

Assumptions relating to key attributes

Attributes in DES models are assigned to entities (e.g. pupils) and various characteristics can be stored in them. The question often arises as to which attributes are critical for the system being modeled and when to update them.

Regarding the process of providing preventive services, pupils may be assigned a number of attributes. Some of these may relate to basic pupils' characteristics such as age, class attended or year of school entry. Additionally, values for indicators, such as dmft and DMFT, related to oral health status can be recorded for a given pupil.

In our study, in addition to the attributes that characterize pupils, related to preventive services or defined for the correct performance of the model, among the key attributes are the dmft and DMFT indicators. The processes involved in changing oral health are continuous processes, but in the discrete model, the values of the relevant indicators are read once a year at the end of the school year. Each pupil is assigned an initial dmft and DMFT index value at the age of 6. The values of these indicators are updated at the end of each year according to the normal distributions assumed after reviewing the actual data (Olczak-Kowalczyk et al. 2021).

Depending on the sum of the dmft and DMFT indicators, one of three Dental Caries Status (DCS) can be assigned to each pupil: good, moderate, and bad (Table 1).

Table 1: Three states of DCS depending on the number of teeth with caries. The values in the table give the total number of primary and permanent teeth with caries

DCS	dmft +DMFT
Good	0
Moderate	1-3
Bad	4 and more

Caries disease progression is reflected in the model by random distributions that correspond to actual data (Table 2). For dmft, the index value increases between the ages of 6 and 7 and then decreases by the age of 12. In contrast, for DMFT, it increases with each additional year.

Table 2: Change in the dmft and DMFT indicators

Indicator	Initial condition	6-7 years change	7-10 years change	10-11* and 10-12** years change	12-15 years change
dmft	3.65 ±3.21	1.77 ±0.37	3.8 ±0.27	1.62 ±0.19	-
DMFT	0.09 ±0.47	0.52 ±0.08	1.27 ±0.18	1.72 ±0.22	1.82 ±0.40
* primary teeth					
** permanent teeth					

Source: Own elaboration based on real data (Olczak-Kowalczyk et al. 2021)

Assumptions related to the realization of preventive services

As part of preventive services, we include dental check-ups and seals of the first permanent molars in the model. The biggest challenge in this area was to model the impact of individual preventive services on pupils' oral health. A second challenge was to develop a level of detail that would be sufficient for management inference and medically correct. At this point, it is important to mention that the purpose of the model is not to predict the health status of pupils or to examine the impact of preventive services on this health status. The model is intended to support management related to preventive care planning. This stage examines how sealing results in a reduction in the average DMFT during subsequent school years.

The ideal situation would be to conduct clinical trials that focus on exactly the aspects needed in research. And then use the results obtained in further studies. Unfortunately, this is very difficult to do in practice, and other solutions should be sought. One of them may be to start cooperation with medical centers to conduct joint research. Or, if such collaboration is not possible, use already published research in the desired area.

In the present study, it is assumed that dental check-ups are performed twice a year. It is intended to provide caregivers for pupils with information on treatment needs, but in our study we do not address these aspects. Furthermore, the pupils' molars are sealed at the age of 6 years, when the first molars erupt. Based on the literature, we assumed that seal reduces the risk of caries by 79% (Wright et al. 2016).

SIMULATION MODEL

The purpose of the first stage of the conducted research was to check a simulation model for one type of preventive services provided in a sample primary school located in South-West Poland.

The authorial model based on the assumptions presented was built in Arena v16.1 (Rockwell Software), following the DES methodology. As preventive services

in the first stage of the study, we included dental check-ups and sealing of the first permanent molars. The output measure was the average DMFT index at the end of primary education.

The model consists of four main components: time control, inflow of pupils, preventive services, and update of indicators. The time control section is responsible for controlling the passage of time by counting days and consecutive years. Then, in the other elements of the model, the pupils are provided with preventive services or promoted to the next grade. In the inflow of pupils section, we generate five streams of entities, one for each of the five parallel classes that come into the school. The school should have 45 classes in total, five in each of the nine-year classes. In the section related to indicator updates, pupils have their dmft and DMFT values recalculated; this occurs at the end of each school year. Figure 2 shows dental examinations and the sealing section. Table 3 presents the key attributes.

Table 3: Key attributes used in the DES model

Name of attribute	Description
Beginning	It provides the year of the beginning of school
Age	It provides information on the age of a pupil.
dmft _n	It stores information on a number of decayed, missing and filled primary teeth of a pupil in the <i>n</i> year of education ($n = 1, \dots, 9$).
DMFT _n	It stores information on a number of decayed, missing and filled permanent teeth of a pupil in the <i>n</i> year of education ($n = 1, \dots, 9$).
Section ABCDE	It provides one of five school sections to which a pupil is assigned. We modeled a primary school that has five sections in one year. Each pupil is assigned letter A, B, C, D, or E.
Section number	It shows one of nine sections to which a pupil is assigned. We modeled a primary school where the pupils learn for 9 years. The number of sections changes according to the year of education.
School section	It is a combination of the <i>Section ABCDE</i> and the <i>Section number</i> . It stores all-encompassing information about the section that a pupil attends.
Section size	It stores information on a number of pupils in a school section.

MODEL VERIFICATION AND VALIDATION

The assumptions were consulted with a dentist who confirmed their accuracy. It was also checked whether the average values of the dmft and DMFT indicators obtained in the model correspond to the average values of these indicators in reality (Table 4).

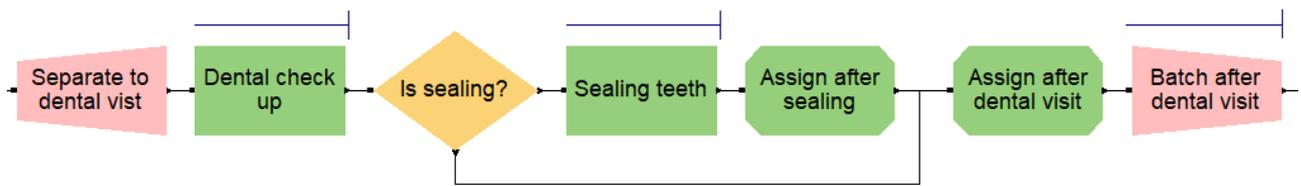


Figure 2: An excerpt from the model presenting the dental check-up and tooth sealing section

The averages obtained allowed for further verification to be performed for one type of preventive service, which is sealing the first permanent molars.

resources that can be used in its implementation or lack of proper planning. Caries prevention care planning is a comprehensive and complex issue influenced by many

Table 4: Comparison of average values of actual the dmft and DMFT indicators along with standard deviation (\pm) and those obtained in 5 replications of a simulation along with confidence interval (CI)

Indicator	Data	Age			
		6	7	10	12
dmft	Reality	3.65 ± 3.21	5.42 ± 3.25	1.62 ± 1.88	-
dmft	Simulation	3.52 (0.95 CI, 3.05-4)	5.22 (0.95 CI, 4.73-5.72)	1.7 (0.95 CI, 1.21-2.19)	-
DMFT	Reality	0.09 ± 0.47	0.61 ± 1.12	1.88 ± 1.63	3.6 ± 2.74
DMFT	Simulation	0.01 (0.95CI, -0.02,0.05)	0.22 (0.95 CI, 0.14-0.3)	1.39 (0.95 CI, 1.3-1.48)	3.10 (0.95 CI, 3.02-3.17)

We conducted two simulations. The first, in which pupils were not provided with sealants, and the second, in which sealants reduced the risk of caries by 79% (see Assumptions relating to the realization of preventive services). Five replications were performed. The length of one replication was 18 years, with the first 9 warming up the model and the next 9 collecting statistics. The results are presented for pupils who started primary school in the same year (Figure 3).

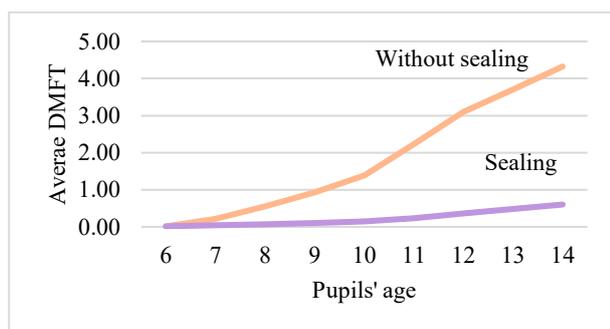


Figure 3: DMFT value in 100 pupils who started school at the same time, with and without sealing during primary education - average of 5 replications

In Figure 3, it may be notice that without sealing, the average number of DMFT for pupils increased with age until it reached a value greater than 4 at age 14. In contrast, the use of sealant resulted in the average number of DMFT less than 1 over the years.

CONCLUSIONS

Caries prevention care in Polish schools is not provided at the level it should be. This may be due to limited

factors and requires the use of advanced decision support tools such as DES. The aim of this paper was to discuss issues, which we faced, while developing the DES model to support decision making in caries prevention planning. The aim was achieved and a number of key assumptions were defined and carefully analyzed. To test the accuracy of the constructed assumptions, the DES model was built and checked for one type of preventive benefit. Our model enables simulating a system for providing caries prevention services in primary schools and its impact on pupils' oral health status. The model developed in this research is designed to support management related to the planning of preventive care. It is not a model to predict the health status of pupils. Depending on the specific focus of the research, it can be developed in a variety of ways, providing a universal tool for the development of adequate policy for the provision of caries prevention services.

In this paper, we have presented the key assumptions and formulation issues that we faced in the first stage of the research. The model has been validated and the results of that validation are presented for the potential effects of a single preventive service. Comprehensive verification and validation are planned for the next stage of the research. The results obtained are acceptable and promising for further research that will move in the direction of including more preventive services in different regions. In addition to dental check-ups and sealing, fluoridation and education will be included. By extending the research to different regions, it is planned to take into account the characteristics of the population living in these regions, for example: education level, poverty and access to medical services. The final model will be able to examine the effectiveness of

different configurations of more preventive services and compare them between different regions. Effectiveness will be understood as a reduction to 0 in the number of pupils who graduate with bad DCS. At this stage of the study, the validity of the assumptions has been checked first and foremost, making it possible to move on to the next stage of the study.

ACKNOWLEDGEMENTS

This project was financed by a grant *Hybrid modelling of the demand for specialist dental care in the field of dental caries prevention in children using computer simulation* from the National Science Centre, Poland that was awarded based on the decision 2021/41/N/HS4/03282.

REFERENCES

- Banks, J.; J.S. Carson II; B.L. Nelson; and D.M Nicol. 2010. *Discrete-event system simulation*. Fifth edition. Pearson Education, New Jersey.
- Brailsford, S.C.; P.R. Harper; and M. Pitt. 2009. "An analysis of the academic literature on simulation and modelling in health care." *Journal of Simulation*, 3(3), 130–140.
- Heaton, B.; S.T. Cherng; W. Sohn; R.I. Garcia; and S. Galea. 2020. "Complex Systems Model of Dynamic Mechanisms of Early Childhood Caries Development." *Journal of Dental Research*, 99(5), 537–543.
- Johnson, B.; N. Serban; P.M. Griffin; and S.L. Tomar. 2019. "Projecting the economic impact of silver diamine fluoride on caries treatment expenditures and outcomes in young U.S. children." *Journal of Public Health Dentistry*, 79(3), 215–221.
- Katsaliaki, K. and N. Mustafee. 2011. "Applications of simulation within the healthcare context." *Journal of the Operational Research Society*, 62(8), 1431–1451.
- Mielczarek, B. 2014. *Symulacja w zarządzaniu systemami ochrony zdrowia*, PWN, Warszawa.
- Olczak-Kowalczyk, D.; A. Mielczarek; U. Kaczmarek; A. Turska-Szybka; E. Rusyan; and K. Adamczyk. 2021. *Monitorowanie stanu zdrowia jamy ustnej populacji polskiej w latach 2016-2020: choroba próchnicowa i stan tkanek przyzębia populacji polskiej: podsumowanie wyników badań z lat 2016-2019*, red. Dorota Olczak-Kowalczyk, Dział Redakcji i Wydawnictw Warszawskiego Uniwersytetu Medycznego, Warszawa.
- Scherrer, C.R.; P.M. Griffin; and J.L. Swann. 2007. "Public health sealant delivery programs: Optimal delivery and the cost of practice acts." *Medical Decision Making*, 27(6), 762–771.
- Scherrer, C.R. and S. Naavaal. 2019. "Cost-Savings of Fluoride Varnish Application in Primary Care for Medicaid-Enrolled Children in Virginia." *Journal of Pediatrics*, 212, pp. 201-207.
- Urwannachotima, N.; P. Hanvoravongchai; J.P. Ansah; and P. Prasertsom. 2019. "System dynamics analysis of dental caries status among Thai adults and elderly." *Journal of Health Research*, 34(2), 134–146.
- Urwannachotima, N.; P. Hanvoravongchai; J.P. Ansah; P. Prasertsom; and VRY. Koh. 2020. "Impact of sugar-sweetened beverage tax on dental caries: a simulation analysis." *BMC oral health*, 20(1):76.

WHO. 2017. "Sugars and dental caries". Available at: [https://www.who.int/news-room/fact-sheets/detail/sugars-and-dental-caries](https://www.who.int/news-room/fact-sheets/detail/sugars-and-dental-carries) (accessed: 12 January 2022).

Wright, J.T.; M.P. Tampi; L. Graham; C. Estrich; J.J. Crall; M. Fontana; E.J. Gillette; B.B. Nový; V. Dhar; K. Donly K; E.R. Hewlett; R.B. Quinonez; J. Chaffin; M. Crespín; T. Iafolla; M.D. Siegal; and A. Carrasco-Labra. 2016. "Sealants for preventing and arresting pit-and-fissure occlusal caries in primary and permanent molars A systematic review of randomized controlled trials-a report of the American Dental Association and the American Academy of Pediatric Dentistry." *Journal of the American Dental Association*, 147(8), 631-645.

AUTHOR BIOGRAPHIES



MARIA HAJŁASZ was born in Poland and went to Wrocław University of Science and Technology, where she studied management science and obtained her degree in 2018. She is still associated with Wrocław University of Science and Technology. She works as an Assistant in the Department of Operations Research and Business Intelligence and she is a PhD student in Management and quality studies. Her research includes decision support in the management of preventive health care using simulation methods. Her e-mail address is: maria.hajlasz@pwr.edu.pl



BOŻENA MIELCZAREK is currently an Associate Professor in the Department of Operational Research and Business Intelligence, Wrocław University of Science and Technology (WUST), Poland. She received an MSc in Management Science, a PhD in Economics, and a D.Sc. in Economics from Wrocław University of Science and Technology. Her research interests include simulation modeling, health-service research, decision support, hybrid simulation, and financial risk analysis. She is the head of the MBA executive program at WUST. Her e-mail address is: bozena.mielczarek@pwr.edu.pl

EUROPEAN QUALITY OF LIFE IN RETIREMENT

Analyzing Personal Differences based on SHARE data

Sára Szanyi-Nagy

Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: szanyinagysara@gmail.com

Professor Dr. Erzsébet Kovács

Institute of Mathematics and Statistical Modelling
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: erzsebet.kovacs@uni-corvinus.hu

Ágnes Vaskövi

Institute of Finance, Accounting and Business Law
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: agnes.vaskovi@uni-corvinus.hu

KEYWORDS

Retirement, ageing, well-being, principal component analysis, SHARE

ABSTRACT

Background Population ageing is one of the greatest challenges of the 21st century. While in 1996 the number of retirees to the total population in the European Union was 14.97%, by 2020 this number had risen to 20.6%. Numerous studies talk about different aspects of ageing, however the European economic and demographic literature do not pay enough attention to the quality of pensioners' life. *Objective* In this paper, we provide a wide picture of their life exploring the individual differences. We used *data* from the 2017 wave of the multidisciplinary database Survey of Health, Aging and Retirement in Europe (SHARE), including personal data on 17,726 retired people from 24 European countries by demographics, education, health status, and their finances. *Method* We examined the differences with Principal component analysis and OneWay ANOVA evaluating the F-test significances. *Results* We found that (i) the health status of European pensioners depends mainly on their age and gender, (ii) investment habits are most significantly connected to education level and the region, (iii) happiness is particularly defined also by education and the region.

INTRODUCTION

2012 was the European Year for Active Ageing and Solidarity between Generations. Numerous initiatives have been launched all over the continent to improve the quality of life of older people and to preserve their activity. Physical and mental health programs and digital catch-ups were supported by smaller and larger communities. This initiative has put the ageing Europe in a more positive perspective and turned the focus to the active and healthy ageing.

At the research level, several scientific papers have been conducted examining the main aspects of ageing (active ageing, age-dependency, physical and mental health with ageing, sustainable pension systems etc.). However, global analysis of the European quality of life in retirement is not conducted in the last five years. We scanned the Web of Science database of articles with the keyword of "Quality of life" for the last 5 years and found that most of the papers analyzed mainly health concerns, less financial and overall happiness factors.

In our research, we present a comprehensive picture of the factors influencing the quality of life of retired citizens of the European Union, and provide a cross-country comparison of the most important factors influencing their well-being. To examine our research questions, we developed a linear factor model on the multidisciplinary database of Survey of Health, Ageing and Retirement in Europe (SHARE) to reduce the multivariate sample and to characterize the differences with a combination of some latent variables, and then to determine the differences between the groups of retirees.

DATA: SHARE WAVE 7

SHARE is a research database of citizens aged 50 or older from 28 European countries and Israel. The first wave of SHARE was conducted in 2004 with twelve countries, however the last wave took place in 2019 including a special covid-focus questionnaire. All surveys are taken personally with CAPI (computer-aided personal interview) and contain several modules, such as social network, health, housing, employment, and pension. Most of the waves create regular panel database, however some of them are supplemented SHARELIFE questionnaire, which focuses on respondents' life histories. SHARE research is harmonized with the US Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA) and has become exemplary panel database for ageing and

retirement studies in the world (Börsch-Supan et al., 2013).

Thus, SHARE database is a proper source for socio-economic and life-quality research, however its weaknesses derive from its strengths. Building a panel database is not straightforward, because new modules, questions are included in the upcoming waves, and based on the higher age of respondents, there is a certain level of attrition, as well (Börsch-Supan et al. 2013).

In our database, we used data of Wave 7 from 2017, because on one hand we wanted to include our home state, Hungary, and on the other hand we also wanted to analyze the most recent data available. (Before Wave 7, Hungary was included in Wave 4 (2011).) The first and most important challenge we faced during the compilation of the data set was the significant proportion of missing values. For example time-constant variables (e.g. nationality, marital status or number of children) are not asked in every wave only if it has changed since the previous waves, or not all questions or modules are answered by all respondents.

In order to minimize missing data, we built our database on a specific SHARE module included different kinds of multiple weight calibration and imputations, moreover we added financial criteria filtered from other financial SHARE modules, with a main focus on the biggest potential sample size. Thus, our database contains data on 17,726 retired persons from 24 European countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, Spain, Slovakia, Slovenia and Switzerland (retired means in SHARE: retired from own work, including semi-retired, partially retired, early retired, pre-retired, not including those retired who receive only survivor pensions and no pensions from own work (SHARE, 2017)). Wave 7 included also the Netherlands, Romania and Sweden but we left these three countries out from our research because of significant missing data. However, we aimed to form the widest possible sample, our dataset is not representative for none of the countries examined, therefore our findings are primarily valid for this sample, but may serve a basis for further research.

In our analyses, we worked with 38 variables that sufficiently describe the socio-economic conditions and well-being of 24 European countries' retirees. The distribution of our data by demographic criteria is shown in Table 1.

Table 1: Demographic Data Distribution

Variable	Values	Frequency (%)
Gender	Female	60,2
	Male	39,8
Currency	Euro-zone	69,2
	Non-euro-zone	30,8
Region	South-Europe	25,8
	North-Europe	24,2
	Western-Europe	26,1
	Eastern-Europe	23,9
Type of residency	Capital	15,3
	Agglomeration	6,8
	Big city	14,9
	Small city	23,4
	Village or other rural	39,3
	Missing or non-response	0,3
Marital status	Married, with spouse	51,9
	Married, separated	1,3
	Registered partnership	1,0
	Never married	5,5
	Divorced	10,3
	Widowed	30,1
ISCED (1997)	0 (less than primary)	3,9
	1 (primary)	13,9
	2 (lower secondary)	20,3
	3 (upper secondary)	37,5
	4 (post-secondary, non-degree)	4,8
	5 (degree)	18,8
6 (doctoral)	0,8	

METHODOLOGY

We used Principal Component Analysis (PCA) to reduce the dimension of 24 original (scale and categorical) variables. This method of factor extraction is used to form uncorrelated linear combinations of the original variables by eigenvector-eigenvalue decomposition. The first component has the highest eigenvalue, i.e. the maximum variance, however, the consecutive components explain decreasing variance. We applied the Kaiser rule to determine the optimal number of factors (where $\lambda > 1$) and sought to keep a minimum of two-thirds of the total variance. Last, we applied Varimax rotation on the orthogonal factors to minimize the number of variables with high loadings on each factor in order to simplify the factor interpretation. Similar research was conducted by Grané et al. (2021), who applied PCA on SHARE survey data.

The other 14 variables we did not include in our linear factor model were used as grouping variables in OneWay Analysis of Variance (ANOVA) to explore the differences of certain groups of pensioners. We created the F-test statistics and searched for the most significant variables visualized by using boxplots. These grouping variables were:

- age-groups: 41-60, 61-70, 71-80, 81-90, 91+
- gender: male or female
- income level: percentiles on a country-basis
- education: (i) 3 categories by rescaling ISCED-97 into primary, secondary and tertiary, (ii) 2 categories of non-degree (ISCED 0-4) / degree (ISCED 5-6) education
- region of residency: based on the real-time statistics of Worldometer (2022) North, South, West and East-European countries.

All our calculations are prepared using IBM SPSS Statistics 27.0.

RESULTS

In our linear factor model, we extracted 9 orthogonal components of the 24 original variables ($\lambda > 1$), which explain 69.088 percent of the original variance. The adequacy of PCA is measured with the determinant of the correlation matrix (≈ 0), the Kaiser-Meier-Olkin (KMO) measure of sampling adequacy (KMO = 0.792, meaning our data suitability is strong-average), and the Bartlett's Test of Sphericity ($\chi^2 = 284,951$, $df = 276$, $\alpha = 0.000$, meaning our data are not uncorrelated). The nine factors are shown in Table 2 (details are attached in Appendix).

Table 2: Components of Linear Factor Model

No	Factor	λ	Original variables
1	Partner	4.326	marital status, demographic indicators of partner
2	Health	2.864	number of chronic diseases, health problem or disability, self-reported health condition
3	Investments	1.839	investment in mutual funds, stocks or shares, life-insurance policy, retirement account
4	Hospital	1.499	nights spent in hospital, doctor visits
5	Pension	1.437	old-age, early retirement, and survivor pension, disability and sickness benefits
6	Residence	1.301	type of settlement and real estate
7	Happiness	1.248	self-reported life satisfaction and happiness
8	Expenses	1.067	food spending at home and outside
9	Education	1.000	years of education

In this paper, we want to focus on the three important factors, i.e. Health-, Investments-, and Happiness-factors. (Details of complete linear factor model are accessible by the authors.) Table 3 shows the F-values of OneWay ANOVA for the five grouping variables. The null-hypothesis of this F-test is that the means of groups do not differ, which could be rejected in all listed cases.

Table 3: F-values of ANOVA (all with $p=0.000$)

Grouping var	Health	Investments	Happiness
Age (10 years)	310.6	58.2	5.3
Gender (male / fem)	266.1	201.9	8.6
Income (percentile)	46.7	123.7	54.7
Education (degree / non)	125.2	742.4	132.6
Region (E, W, N, S)	72.0	499.1	134.6

The bold figures in Table 3 show the highest F-values of each component, meaning that by Health-factor the strongest grouping variable is the age, however by Investments-factor the education level, and by Happiness-factor the region of residence. In the next sections we visualize the most significant grouping variables by using box-plots.

Health-factor

Health-factor includes variables as the number of chronic diseases, health problem or disability, limitation with activities, self-reported health condition and the number of doctor visits in the previous year of the interview. Therefore, the higher Health-factor score is given to a person, the worse health condition s/he has. Figure 1 shows the relationship between the Health-factor and age (F=310.6, and $p=0.000$).

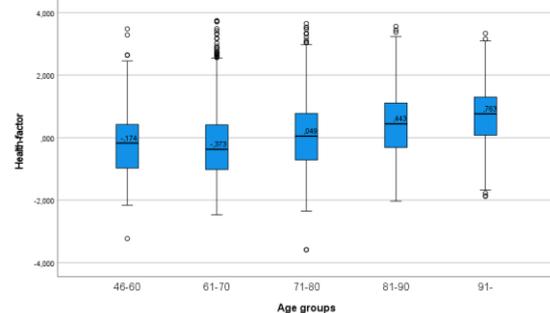


Figure 1: Health Differences by Age-groups

It is straightforward that there are the most health problems in the 91+ age group, however the first two age groups need more detailed explanation. When a person retires at the age of 50 (significantly before the official retirement age), there are most probably health issues in the background. This is the reason

why the youngest pensioners have more health problems than people in the second age group.

The same conclusion might also be found in Chatterji et al's (2015) meta-analysis, where they found rising number of health problems (morbidity) with age in all analyzed countries. However, they also found that elderly in Italy, Spain and Greece face this increase earlier (between 50 and 70 years) than those in the Netherlands, Sweden, or Switzerland.

Figure 2 shows the gender-dependency of Health-factor ($F = 266.1, p=0.000$). Based on gender variable, we concluded that rather women suffer from multi-morbidity (several chronic diseases, health and mobility problems) and consider their own health condition to be more serious.

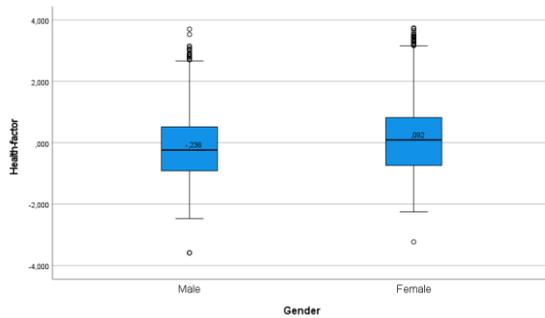


Figure 2: Health Differences on Gender

In line with our results, Grané et al (2020) found that depression and anxiety affect mostly older *women* causing serious health problems on the long run. Chatterji et al. (2015) also mention other longitudinal analysis besides SHARE (Health and Retirement Study, English Longitudinal Study of Ageing) that serve evidence for more health barriers of elder women.

Investments-factor

We obtained the highest F-values analyzing the Investments-factor, i.e. there are the most significant differences among European pensioners concerning their investment habits. In Table 3 the education grouping variable shows 742.4 F-value ($p=0.000$), which means the pensioners with high education have better financial literacy and they are more conscious about their investments. On Figure 3 we show the difference between the education groups.

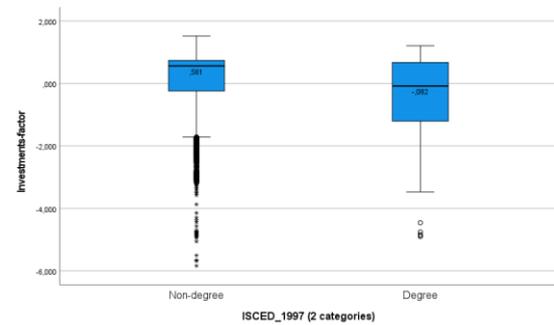


Figure 3: Investment Habit Differences between Higher and Less Educated People

Investments-factor consists of 4 original variables: if the person ever had investment in (i) mutual funds, (ii) stocks or shares, (iii) life-insurance policy, or (iv) retirement account. (The higher value of the variables show the person did not have that specific saving or investment form ever.) Pensioners with tertiary education had significantly more investments of this type, than those with no higher education. Table 4 shows the 95% confidence levels for median of investments-factor grouped by education.

Table 4: Confidence-level for Median (Investments-factor with Education)

	Median	95.0% Lower CL	95.0% Upper CL
Non-degree	0.561	0.549	0.572
Degree	-0.082	-0.138	-0.032

Garcia and Marques (2017) also used SHARE data (Wave 2 and 4) to analyze the influencing factors of the ownership of Individual Retirement Accounts (IRA) in eight EU countries. IRA is a special investment vehicle for long-term savings. They also came to similar results: they found that the years of education influence significantly and positively this kind of investments, however they did not find it statistically significant, if the person had degree.

Comparing the Investments-factor by the regions of Europe ($F=499.1, p=0.000$), we obtained that the Western countries' pensioners have had any of the four investment vehicles. However, pensioners in North-, East- and West-European countries do not show statistically significant differences. Figure 4 and Table 5 present the diverse Investments-factor by the European regions.

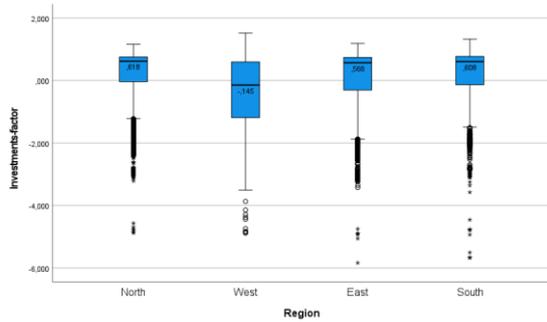


Figure 4: Investment Habit Differences between European Regions

Table 5: Confidence-level for Median (Investments-factor with Regions)

	Median	95.0% Lower CL	95.0% Upper CL
South	0.608	0.591	0.623
North	0.618	0.607	0.626
East	0.568	0.546	0.586
West	-0.145	-0.177	-0.096

By gender differences of Investments-factor ($F=201.9$, $p=0.000$) we found that men dominate the usage of this financial vehicles. With this result, we approved the mainstream literature inter alia Barber-Odean (2001) who proved men are more risk-taking and intend to invest in sophisticated financial vehicles than women. On the other hand, Garcia and Marques (2017) show that among SHARE Wave 2 respondents male people are less likely to have Individual Retirement Accounts. It seems contradictory to our result, however; women's low risk appetite could lay in the background of the use of low-risk savings products.

Happiness-factor

Considering the happiness of European pensioners, our prior assumption was that we would not find huge differences along with our grouping variables because the 24 countries are all developed countries and their happiness-indices (Helliweel et al., 2017) were all above the average 5.354 points in 2017 (except for Bulgaria with its 4.714). However, we found differences mostly by region ($F=134.6$, $p=0.000$) and education ($F=132.6$, $p=0.000$) (see Table 3). Age and gender differences are also significant but these grouping variables are less dominant.

Happiness-factor represents two manifest variables; how satisfied the person with his/her own life is (on a scale of 0 to 10) and how often s/he looks back on life with happiness (scale of 1 to 4). In Figure 5 and Figure 6 higher values show higher happiness of specific groups.

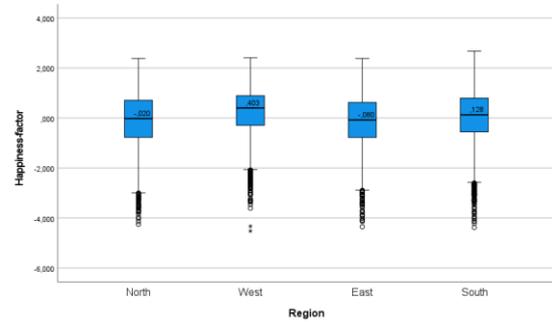


Figure 5: Happiness Differences by Regions

Our strongest grouping variable is the region with F-value of 134.6 ($p=0.000$). Figure 5 and Table 6 show pensioners in Western countries assess their life the happiest, however in Northern and Eastern countries people reported less happiness. (In our research, Northern countries include Denmark, Finland and also the 3 post-socialist Baltic countries, this could be the interpretation why our results are slightly different from e.g. the ranking of Global AgeWatch Index (GAWI, 2018).)

Table 6: Confidence-level for Median (Happiness-factor with Regions)

	Median	95.0% Lower CL	95.0% Upper CL
South	0.128	0.080	0.171
North	-0.020	-0.055	0.030
East	-0.080	-0.119	-0.044
West	0.403	0.376	0.435

The second strongest grouping variable was the education ($F=132.6$, $p=0.000$), as Figure 6 proves those retired people with degree self-reported happier than the others without degree.

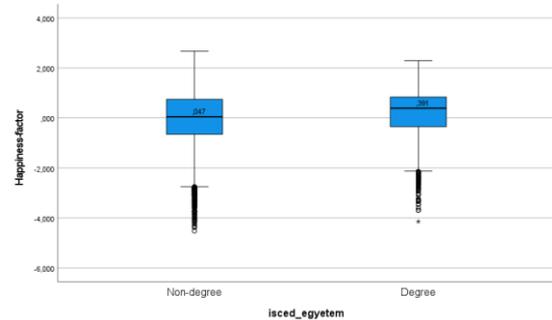


Figure 6: Happiness Differences by Education

Most of the well-being research projects report that education is a key factor of several aspects of life. Those people with higher education are happier, healthier, financially more literate, therefore they live in a higher level of well-being overall. For example, Becchetti et al. (2018) used SHARE data in their robust econometric model and found that people with higher ISCED level of education face lower risk of diabetes, hypertension and heart attack as those at lower ISCED levels. By including several other explanatory variables, their general conclusion

was that both objective and self-reported (subjective) health conditions of European highly educated pensioners are better deriving from their lifestyle differences.

CONCLUSIONS

Our empirical work investigates on a large sample of Europeans aged above 50 whether there are differences of health, investment habits, and happiness among their specific groups. We collected data from the Wave 7 (2017) of Survey of Health, Ageing and Retirement in Europe multi-disciplinary survey database and compiled responses of 17,726 retirees along 38 different variables. In order to reduce the dimension of these 38 variables, we built a linear factor model of 24 variables forming nine components and used the other 14 variables for comparing different aspects with ANOVA. This paper details three factors of our principal component analysis, however, the other six factors might be analyzed in further research papers.

Our analysis led to the following conclusions. We documented that age, gender, education, income, and region of residence are all statistically significant variables considering differences in health, investments and happiness of elderly Europeans. Investigating three factors of the factor model we found:

1. The health conditions are mostly dependent on age and gender, i.e. the older the person is, the more health issues s/he faces and women suffers more likely multi-morbidity than men.
2. Investment habits of elderly differs by the level of education, thus people with degree have more experience investing in sophisticated financial vehicles. We also found that pensioners living in Western-European countries are financially more literate than those in other regions of Europe.
3. Happiness level of highly educated pensioners is significantly higher than the less educated ones, and the ‘happiness-rank’ of European regions: west, south, north and east. (The Baltic countries belong to North-Europe in our analysis, along with the grouping of Worldometers, this could result in slightly different conclusions found in the literature.)

REFERENCES

- Barber, B., Odean, T. (2001): Boys will be boys. Gender, overconfidence and common stock investment. *Quarterly Journal of Economics*, Vol. 116. No. 1. pp. 261–292.
- Becchetti, L., Conzo, P., Pisani, F. (2018): Education and health in Europe, *Applied Economics*, Vol. 50. No. 12. pp. 1362–1377. <https://doi.org/10.1080/00036846.2017.1361013>
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmayer, J., Malter, F., Schaaf, B., Stuck, S. and

- Zuber, S. (2013): Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, Vol. 42, pp. 992–1001, <https://doi.org/10.1093/ije/dyt088>
- Chatterji, S., Byles, J., Cutler, D., Seeman, T., Verdes, E. (2015): Health, functioning, and disability in older adults – present status and future implications. *The Lancet*, Vol. 385. No. 9967. pp. 563–575. [https://doi.org/10.1016/S0140-6736\(14\)61462-8](https://doi.org/10.1016/S0140-6736(14)61462-8)
- De Luca G., Peracchi F., Börsch-Supan A., Jürges H. (2005): Survey participation in the first wave of SHARE, *The Survey of Health, Ageing and Retirement in Europe – Methodology*, Mannheim, Germany, Mannheim Research Institute for the Economics of Aging (MEA)
- Garcia, M. T. M., Marques, P. D. C. V. (2017): Ownership of individual retirement accounts—an empirical analysis based on SHARE. *International review of applied economics*, 31(1), pp. 69–82. <https://doi.org/10.1080/02692171.2016.1221389>
- GAWI (2018): Global AgeWatch Insights, 2018. HelpAge International, London, <https://www.helpage.org/global-agewatch/about/global-agewatch-index-version-20/>
- Grané, A., Albarrán, I., Lumley, R. (2020): Visualizing Inequality in Health and Socioeconomic Wellbeing in the EU. Findings from the SHARE Survey. International, *Journal of Environmental Research and Public Health*, Vol. 17. No. 21. 7747. pp. 1–18. <https://doi.org/10.3390/ijerph17217747>
- Grané, A., Albarrán, I., Guo, Q. (2021): Visualizing Health and Well-Being Inequalities Among Older Europeans. *Social Indicators Research*, Vol. 155. No. 2. pp. 1–25. <https://doi.org/10.1007/s11205-021-02621-x>
- Helliwell, J.F., Layard, R., Sachs, J.D. (2017): World Happiness Report 2017, New York: Sustainable Development Solutions Network
- Survey of Health, Ageing and Retirement in Europe (SHARE) (2017): Questionnaire Wave 7, access: <http://www.share-project.org/data-documentation/questionnaires/questionnaire-wave-7.html>
- SHARE (2019): Survey of Health, Ageing and Retirement in Europe Wave 7. Release version: 7.0.0. SHARE-ERIC. Data set (Börsch-Supan, 2019), <https://doi.org/10.6103/SHARE.w7.700>
- Tymowski, J. (2016): European Year for Active Ageing and Solidarity between Generations (2012) – European Implementation Assessment: in-depth analysis. European Parliament, Directorate-General for Parliamentary Research Services. <https://data.europa.eu/doi/10.2861/230093>
- Worldometers.info (2022) Population: Europe, <https://www.worldometers.info/population/europe/>

AUTHORS' BIOGRAPHIES

Prof. Dr. ERZSÉBET KOVÁCS is head of the Department of Operational Research and Actuarial Sciences. Her main fields of research are applications of multivariate statistical methods in international comparison of insurance markets, comparison and

modelling pension systems, mortality projections, risk analysis in student loan system, statistical analysis of the period of economic transition in Central-Eastern Europe. Her email address is erzsebet.kovacs@uni-corvinus.hu

SÁRA SZANYI-NAGY, MSc is a full-time retail analyst. She earned her master's degree in Finances from Corvinus University of Budapest, specializing in corporate finance. She wrote her Thesis about analyzing the dissimilarity of European pensioners, that is the basis of the current retirement research. She earned her BA in Applied Economics and wrote her Thesis about examining the motivations of people migrating to Germany. Her email address is szanyinagysara@gmail.com

ÁGNES VASKÖVI, MSc is a full time assistant professor of the Institute of Finance, Accounting and Business Law. She earned her master's degree in Economics from Corvinus University of Budapest, specializing in financial investment analysis. She gained professional experience in fields of project financing, venture capital and real estate investments. Currently, she teaches Finance, Corporate Finance, and Multivariate Data Analysis. On her main research agenda there are topics of behavioural finance, financial literacy, long term savings, longevity, and pension. Her email address is agnes.vaskovi@uni-corvinus.hu

Appendix: Total variance explained by the linear factor model

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.326	18.025	18.025	4.326	18.025	18.025	3.843	16.011	16.011
2	2.864	11.935	29.960	2.864	11.935	29.960	2.682	11.177	27.187
3	1.839	7.662	37.623	1.839	7.662	37.623	1.892	7.882	35.069
4	1.499	6.244	43.867	1.499	6.244	43.867	1.544	6.432	41.501
5	1.437	5.987	49.854	1.437	5.987	49.854	1.524	6.348	47.850
6	1.301	5.421	55.275	1.301	5.421	55.275	1.473	6.136	53.986
7	1.248	5.198	60.473	1.248	5.198	60.473	1.359	5.662	59.648
8	1.067	4.448	64.921	1.067	4.448	64.921	1.264	5.268	64.916
9	1.000	4.167	69.088	1.000	4.167	69.088	1.001	4.172	69.088
10	0.815	3.398	72.485						
11	0.782	3.260	75.745						
12	0.743	3.094	78.839						
13	0.727	3.027	81.867						
14	0.626	2.607	84.474						
15	0.589	2.453	86.927						
16	0.570	2.373	89.300						
17	0.522	2.176	91.476						
18	0.514	2.141	93.617						
19	0.477	1.989	95.606						
20	0.450	1.873	97.479						
21	0.446	1.859	99.339						
22	0.153	0.639	99.978						
23	0.004	0.018	99.996						
24	0.001	0.004	100.000						

Extraction method: Principal Component Analysis

REGIONAL MODELS OF CORPORATE SECTOR DEVELOPMENT IN RUSSIA: WHERE DOES FAMILY-FRIENDLY POLICY MATTER MOST? (A STUDY BASED ON CLUSTER ANALYSIS)

Anna Bagirova
Oksana Shubat
Ural Federal University
620002, Ekaterinburg, Russia
Email: a.p.bagirova@urfu.ru
Email: o.m.shubat@urfu.ru

KEYWORDS

Cluster analysis, modelling, corporate sector, demographic policy, Russian regions, family-friendly policy

ABSTRACT

Despite being popular across the world, the concept of family-friendly policy is not associated with business in Russia. Our research focuses on the corporate sector and its potential to change the demographic context in Russia. The aim of the study is to identify regional models of the Russian corporate sector development and suggest pilot regions to start family-friendly policies aimed at increasing the birth rate. We used variables which characterise the corporate sector development in Russian regions and applied the hierarchical cluster analysis. Regional values of the indicators for all 85 Russian regions were analysed. We revealed 5 clusters and then additionally profiled the clusters selected according to the Total Fertility Rate. The analysis allowed us to identify two clusters of Russian regions which may become pilot ones when promoting family-friendly policy practices in their enterprises. We argue that a family-friendly policy in these clusters' enterprises might become: relevant for employees; quite affordable for enterprises; a new effective tool of the demographic policy since the birth rates here are the lowest in Russia.

INTRODUCTION

Russia has extensively pursued a pro-natalist policy and introduced new measures to increase the birth rate and support to families with children; the number of people eligible for the support and the amount of benefits are increasing, too. The country has also started to implement a large national project "Demography", which includes a number of steps from the government to improve some demographic indicators. Despite the effort, the birth rate is not growing—in 2015–2020, the Total Fertility Rate (TFR) in Russia fell from 1.777 to 1.505, which is 15.4% (Total Fertility Rate data 2021).

We argue that one of the reasons behind the failure is that the demographic policy is poorly supported at the other levels of the socio-economic structure. Education, non-profit organisations, business, and other social institutes could contribute to the demographic policy

development since they are as concerned with the outcomes as the government. These institutes could both contribute to the demographic policy and reinforce it; thus, enhance its effectiveness.

Our research focuses on the corporate sector and its potential to change the demographic context in Russia. Despite being popular across the world, the concept of family-friendly policy is not associated with business in Russia. Instead, the corporate sector embraces the ambiguous notion of social responsibility, which rarely involves demographic aspects. Meanwhile, studies from different countries prove that pursuing a family-friendly policy has multiple effects on employees, employers, and society in general (Breugh and Frye 2007; Kim and Yeo 2019; Vysniauskiene and Braziene 2017).

Researchers agree that one of the aims of this policy is to mitigate the conflict between family and work, which is today a pressing issue for women with children (Jang and Ahn 2021; Feeney and Stritch 2019; Yu 2019). However, developing this kind of policy is especially challenging in transitional economies, including Russia. Nabergoj and Pahor claim that "the coordination of work and family life is complex and depends on the interplay between factors at three different levels: governmental, organisational, and individual. The relationship between these three levels is even more intertwined in economies that have undergone economic transition from socialism to capitalism" (Nabergoj and Pahor 2016).

We have to note, though, that family-friendly policies are mostly initiated and studied in countries with advanced economies. Researchers analyse how a family-friendly policy influences employees' satisfaction in different sectors and conclude that this influence is positive, but it is also determined by the demographic characteristics of employees (Kim and Wiggins 2011). Other studies explore how a family-friendly policy affects the image of a company and argue that a company pursuing a family-friendly policy is viewed as more attractive (Bourhis and Mekkaoui 2010). Lee and Hong evaluate how a family-friendly policy and its specific components influence turnover rates and companies' effectiveness and claim that some of the policy's programmes have a rather marked impact (Lee and Hong 2011). Callan reports on the relationship between a formal corporate family-friendly policy and an informal organisational culture (Callan 2007). In their analysis of

Slovenian business, Nabergoj and Pahor argue that some specific family-friendly practices have a profound—but with some limitations – influence on the organisational effects of companies in countries with economies in transition (Nabergoj and Pahor 2016).

While designing our research, we stemmed from the fact that 1) family-friendly policies are not widely developed by Russian companies; 2) the state demographic policy in Russia has not met its objectives yet; 3) Russian regions are highly differentiated by a number of socio-economic indicators, including those which deal with the corporate sector development. Thus, there are different models of the social and economic development within the country.

Therefore, the aim of our study is to identify regional models of the Russian corporate sector development and suggest pilot regions to start family-friendly policies aimed at increasing the birth rate.

DATA AND METHODS

For our research, we chose those variables which characterise the corporate sector development in Russian regions. They are traditionally used to characterise the economic development of the whole region. These variables are the following:

- Gross Regional Product (GRP) per capita, in roubles;
- Investment Activity in Russia (a share of organisations specialising in innovations among the total number of organisations, %);
- Share of Loss-making Organisations (% of the total number of organisations).

Also, it is important to analyse indicators which specify the actual or potential willingness of the corporate sector to meet social needs. To that end, we adopted the following indicators:

- Household Final Consumption Expenditure (families' expenditure on goods and services and the cost of consuming goods and services in kind);
- Retail Trade Turnover (per capita, in roubles).

To characterise the birth rate, we used Total Fertility Rate, which is most widely used in demographic studies.

We analysed regional values of the indicators mentioned above for all 85 Russian regions as of 2019, which is the most relevant data available. All data were extracted from an annual statistical report “Regions of Russia. Social and Economic Indicators 2021” issued by the Federal State Statistics Service (Regions of Russia. Social and Economic Indicators 2021).

To model the Russian economic space using indicators of the corporate sector, we applied the hierarchical cluster analysis. For modelling the regional economic space, we used various distance measures and distances between clusters. We compared clustering results obtained through different measures and chose those measures which allowed grouping the regions analysed most accurately.

To decide how many groups of regions to identify, we stemmed from the following:

- graphical representation of the clustering (we examined a dendrogram);
- evaluation of the between-group and within-group variability;
- cluster size (we controlled the number of regions that form a cluster to ensure that each group contained enough regions).

Variables used for clustering may be correlated, which can result in the distorted cluster structure; thus, we used the correlation analysis—based on Pearson correlation coefficients and Spearman's rank correlation—to assess whether variables are collinear. We used standard procedures to assess collinearity, which involve the correlation coefficient greater than 0.7.

As variables for clustering had different dimensions, we standardised them using the method of processing the initial data to the range of 0 to 1.

To ensure that clustering results are sound, we studied cluster centroids—medians of clustering variables. We avoided mean values, which are frequently used as cluster centroids, and referred to a non-parametric indicator because the values of clustering variables in the clusters identified were rather markedly different, which reduced the reliability of the mean value. The Median Test and Kruskal-Wallis Test were used to evaluate the statistical significance of differences between cluster centroids.

We used IBM SPSS Statistics 23.0 for our analysis.

RESULTS

We primarily focused on analysing how the indicators of the corporate sector development vary according to the region and found that these variations are either high or extremely high (Table 1). The lowest number of loss-making organisations was recorded in the Republic of Adygeya (22.6%); the highest in the Nenets Autonomous Area (59.6%). In this case, therefore, the Maximum-Minimum Ratio (MMR) accounts for 2.6. In the case of the investment activity, the MMR increases to 106. This huge regional differentiation is a precondition for forming groups of regions which differ radically by the corporate sector development; at the same time, these groups may incorporate regions with the similar values of the indicators analysed.

For clustering regions, it is important to note that some particular regions with minimax values for one indicator also showed those for some other indicator; it applies, for example, to the Retail Trade Turnover and Household Final Consumption Expenditure. Minimum values were recorded in the Republic of Chechnya, maximum ones in Moscow (Table 1). Therefore, we suggest that the indicators studied are collinear, which could result in the distorted cluster structure. To avoid negative effects from collinearity, we carried out a correlation analysis based on Pearson correlation coefficients and Spearman's rank correlation. Table 2 presents results of the analysis.

Table 1: Descriptive Statistics

Variable	Minimum		Maximum		MMR
	Value	Region	Value	Region	
Gross Regional Product, per capita, roubles	145723	Republic of Ingushetia	7530485	Nenets Autonomous Area	51.7
Household Final Consumption Expenditure, per capita, roubles	127351	Republic of Ingushetia	771175	Moscow ¹	6.1
Retail Trade Turnover, per capita, roubles	51702	Republic of Ingushetia	403426	Moscow	7.8
Investment Activity, %	0.2	Republic of Chechnya	21.2	Republic of Mordovia	106.0
Share of Loss-making Organisations, %	22.6	Republic of Adygeya	59.6	Nenets Autonomous Area	2.6

Table 2: Spearman's Rank Correlation

	Var 1	Var 2	Var 3	Var 4	Var 5
Var 1	1	0.814**	0.709**	0.199	0.006
Var 2	0.814**	1	0.930**	0.16	-0.006
Var 3	0.709**	0.930**	1	0.187	-0.092
Var 4	0.199	0.16	0.187	1	-0.397**
Var 5	0.006	-0.006	-0.092	-0.397**	1

In the Table:
 Var 1 - Gross Regional Product, per capita, roubles
 Var 2 - Household Final Consumption Expenditure, per capita, roubles
 Var 3 - Retail Trade Turnover, per capita, roubles
 Var 4 - Investment Activity, %
 Var 5 - Share of Loss-making Organisations, %

** Correlation is significant at the 0.01 level (2-tailed)

¹ According to the administrative division of Russian regions, Moscow is a separate region

The analysis showed that the Gross Regional Product (per capita), Retail Trade Turnover and Household Final Consumption Expenditure are highly correlated. To avoid possible clustering bias, we retained only one variable for further analysis – Household Final Consumption Expenditure. We argue that it characterises the regional economy's and corporate sector's focus towards meeting the social needs to the greatest extent.

Thus, we conducted the cluster analysis based on the following variables:

- Household Final Consumption Expenditure, per capita, roubles;
- Investment Activity in Russia, %;
- Share of Loss-making Organisations, %.

The clustering was based on Ward's method and the Euclidean distance – these measures showed the best differentiation power and allowed us to identify five clusters of regions. Graphically, the clustering is presented in Appendix 1 as a dendrogram; characteristics of the cluster centroids as median values are shown in Table 3. The statistical significance of differences in cluster centroids was tested with a non-parametric median test; results are presented in Table 4. According to our results, differences in the median values of the clustering variables in the groups of regions were statistically significant.

Table 3: Cluster Centroids: Median Values

Cluster	Household Final Consumption Expenditure per capita, roubles	Investment Activity, %	Share of Loss-making Organisations, %	Total Fertility Rate
1 (7 regions)	299 709	4.6	46	1.583
2 (19 regions)	494 911	7.2	35.7	1.572
3 (30 regions)	318 587	10.4	32	1.413
4 (10 regions)	361 235	14.9	27.5	1.352
5 (19 regions)	282 575	4.6	33.3	1.594

Table 4: Test Statistics (Median Test)

	Household Final Consumption Expenditure, per capita, roubles	Investment Activity,%	Share of Loss-making Organisations, %
Median	320128	8.1	32.9
Chi-Square	33.054	48.07	26.464
df	4	4	4
Asymp. Sig.	0	0	0

Proceeding to the most significant characteristics of the clusters identified, we believe that the most problematic clusters turned out to be Clusters 1 and 5. They include regions with the lowest level of the corporate sector development. The investment activity in these regions is equally low – and the lowest among other clusters. At the same time, regions in Cluster 1 have the highest share of loss-making enterprises; regions in Cluster 5 show the lowest Household Final Consumption Expenditure.

Cluster 4, on the other hand, includes regions with the highest level of the corporate sector development. The regions in this cluster show the highest investment activity and the lowest share of loss-making enterprises. The Household Final Consumption Expenditure here is effectively the highest among other clusters.

Clusters 2 and 3 are in between, having mid-level and statistically significant differences in the investment activity, and the share of loss-making organisations. Remarkably, Cluster 2 boasts the highest Household Final Consumption Expenditure.

Further, we additionally profiled the clusters selected according to the Total Fertility Rate. Its median values in each cluster are shown in Table 3. Conspicuously, two most challenging clusters have the highest TFR, and the most unproblematic Cluster 4 displays the lowest TFR.

We hypothesised a lag effect in the correlation between the level of the corporate sector development and the birth rate. Potentially, the analysis with the lag effect considered may reveal other peculiarities in the distribution of birth rates in the clusters identified. Clearly, to explore this effect, more specific and deeper research is needed. To a first approximation, we verified the hypothesis by analysing regional TFRs for adjoining years. Table 5 presents results of the correlation analysis of regional TFRs for 2018–2020. As is evident, all three variables are highly correlated; therefore, even with ± 1 -year lags, similar patterns may be found in the distribution of birth rates in the clusters identified.

Table 5: Spearman's Rank Correlation of Regional Birth Rates

	TFR 2018	TFR 2019	TFR 2020
TFR 2018	1	0.965**	0.945**
TFR 2019	0.965**	1	0.966**
TFR 2020	0.945**	0.966**	1

** Correlation is significant at the 0.01 level (2-tailed)

DISCUSSIONS

The analysis allowed us to identify two clusters of Russian regions – Cluster 3 and 4 – which may become pilot ones when promoting family-friendly policy practices in their enterprises. These clusters have a distinctly low birth rate and a relatively high level of the corporate sector development (i.e., high investment activity and low share of loss-making enterprises). Moreover, we observed a moderate level of the Household Final Consumption Expenditure within these clusters, which might indicate an average living standard for Russia. A family-friendly policy in the third and fourth clusters' enterprises might become, firstly, relevant for employees (since the final expenditure here is not the highest in Russia); secondly, quite affordable for enterprises (as the corporate sector here has a higher investment activity and makes fewer losses than organisations in other regions of Russia); thirdly, a new effective tool of the demographic policy since the birth rates here are the lowest in Russia.

A family-friendly policy traditionally consists of 4 programmes (Bourhis and Mekkaoui 2010):

- Measures to support employees' family members (children, parents);
- Measures to enable leave-taking (parental, personal, or family leave);
- Measures to introduce various family programmes for employees (counselling, leisure time organisation);
- Measures to introduce flexible work arrangements.

It appears that any of these measures could be used by the corporate sector in Russian regions from Clusters 3 and 4. However, when developing, implementing, and promoting these measures, it is crucial to emphasise their novelty and innovation for the Russian corporate sector and that employees should be seen not only as actors of qualified labour, but also as human beings with family responsibilities. In Russia, where the negative natural population increase has already exceeded 1 million people in 2021 – for the first time since 2000 – it is of high relevance to deliver a message to the population when introducing new instruments. It could be the following: the corporate sector introduces a family-friendly policy because it has realised that the top priority of their employees is their families – and this is what HR and social corporate policies rest on. Consequently, it is the corporate demographic policy that should be seen as a key element of the social responsibility policy.

Proceeding from at least two scientific theories, we suggest that this new instrument of the demographic

policy can be effective in Russia. The first theory is the neo-institutional sociology, which assumes that organisations are structured by phenomena in their environment and are usually isomorphic to that environment (Meyer and Rowan 1977). Organisations and the environment they operate in are interlinked and interchanging, which results in a certain mutual correspondence between organisations and the environment. This theory may explain why the corporate sector in the regions with the lowest birth rates may respond to demographic problems and introduce a new element in social responsibility policies (or initiate such policies from a demographic perspective).

The second theory is the social exchange theory, which suggests that people tend to avoid activities that would involve them in unfair exchanges; in doing so, they tend to perform activities that are rewarded by fairness. Moreover, a failure to perform these rewarding activities is considered a loss to them (Homans 1961). This theory explains why the introduction and promotion of family-friendly policies in the corporate sector can be an effective tool for stimulating birth rates in the regions of Russia.

Our study is pilot; we realise that, to estimate the potential effectiveness of a family-friendly policy in Russian regions, it is necessary to consider a variety of factors. For example, these factors may deal with regional demographic processes (the migration level, population ageing, etc.), the development of the infrastructure for families with children, existing regional support measures, people's religious beliefs, and others. Additionally, there may be revealed more valid indicators to characterise the level of the corporate sector development in the regions. In our research, we focused on the indicators which are used in Russian regions' economic studies most frequently.

Enhancing the state demographic policy by promoting a family-friendly policy in Russian companies is also affected by the regional demographic potential. In this context, a crucial indicator may be a structure of the region's population and a share of childbearing-age women.

CONCLUSIONS

Our study yielded the following conclusions:

1) When developing demographic, social, and economic policy measures, it is important to take into account that Russian regions are heterogeneous, which results from the high differentiation of many socio-economic and demographic indicators. Indeed, such heterogeneity means that it is impossible to develop universally effective measures of state support and birth rate stimulation. On the other hand, despite the high differentiation, there are still regions in Russia with a similar socio-economic and demographic situation and with similar models of demographic dynamics. Identifying and describing such models is necessary for enhancing the demographic policy.

2) We argue that the cluster analysis is effective for modelling the Russian demographic space to enhance state demographic policy measures. It involves the distribution of objects into homogeneous groups (i.e., clusters). More importantly, objects can be grouped using several variables, and clustering can be based on both quantitative and qualitative variables that have different dimensions.

3) In our study based on the cluster analysis, we identified groups of Russian regions where family-friendly policies may be the most relevant for the corporate sector personnel, quite affordable for enterprises, and effective as a new demographic policy instrument.

We argue that our results lay the foundation for more comprehensive and detailed research with aforementioned areas considered – that is, searching for and selecting the most valid indicators of the corporate sector development, identifying a range of determinants of the family-friendly policy effectiveness in Russian regions, exploring a family-friendly policy as a potential birth rate factor.

ACKNOWLEDGMENTS

The study was conducted as part of the project “Russian Pro-Natalist Policy Support Institutions: Potential and Prospects for Influencing Birth Rate Growth“, supported by the Council on grants of the President of the Russian Federation, project no. NSh-1327.2022.2.

REFERENCES

- Bourhis, A. and R. Mekkaoui. 2010. “Beyond Work-Family Balance: Are Family-Friendly Organizations More Attractive?”. *Relations Industrielles / Industrial Relations*, Vol 65 (1), 98–117. <http://www.jstor.org/stable/23078261>
- Breaugh, J. A. and N. K. Frye. 2007. “An Examination of the Antecedents and Consequences of the Use of Family-Friendly Benefits”. *Journal of Managerial Issues*, Vol 19 (1), 35-52.
- Callan, S. (2007). “Implications of family-friendly policies for organizational culture: findings from two case studies”. *Work, Employment & Society*, Vol 21 (4), 673–691. <http://www.jstor.org/stable/23748295>
- Feeney, M.K. and J.M. Stritch. 2019. “Family-Friendly Policies, Gender, and Work–Life Balance in the Public Sector”. *Review of Public Personnel Administration*, Vol 39 (3), 422-448.
- Homans, G. C. 1961. “*Social Behavior Its Elementary Forms*”. New York: Harcourt, Brace, and World.
- Jang, H. and H. Ahn. 2021. “Organizational responses to work-life balance issues: The adoption and use of family-friendly policies in Korean organizations”. *International Review of Public Administration*, Vol 26 (3), 238-253.
- Kim, J. and M.E. Wiggins. 2011. “Family-Friendly Human Resource Policy: Is It Still Working in the Public Sector?”. *Public Administration Review*, No 71, 728-739. <https://doi.org/10.1111/j.1540-6210.2011.02412.x>
- Kim, J.S. and Y.H. Yeo. 2019. “Family-friendly policy and childbirth intention: Mediating effect of family life satisfaction”. *Asia Life Sciences*, Vol 3, 75-84.

Lee, S.-Y., and J. H. Hong. (2011). “Does Family-Friendly Policy Matter? Testing Its Impact on Turnover and Performance”. *Public Administration Review*, Vol 71 (6), 870–879. <http://www.jstor.org/stable/41317386>

Meyer, J. W. and B. Rowan. 1977. “Institutionalized Organizations: Formal Structure as Myth and Ceremony”. *American Journal of Sociology*, Vol 83 (2), 340-363. <https://doi.org/10.1086/226550>

Nabergoj, A. S. and M. Pahor. 2016. “Family-friendly workplace: An analysis of organizational effects in the transition economy”. *Journal of East European Management Studies*, Vol 21 (3), 352–373. <http://www.jstor.org/stable/44111952>

Regions of Russia. Social and Economic Indicators 2021. Statistical Book. Rosstat, Moscow. URL: <https://rosstat.gov.ru/folder/210/document/13204> (access date 15.01.2022).

Total Fertility Rate data. Single inter-departmental information and statistical system (SIDIS). 2021. Rosstat, Moscow. URL: <https://fedstat.ru/indicator/31517> (access date 30.01.2022).

Vysniauskiene, S. and R. Braziene. 2017. “Evaluation of family friendly policy in Lithuania”. *Public Policy and Administration*, Vol 16 (3), 455-467.

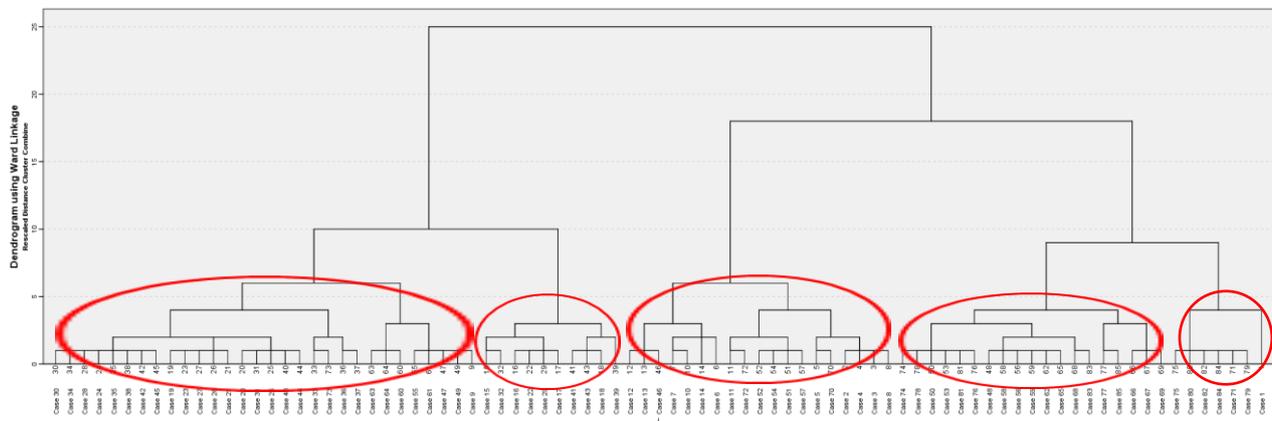
Yu, H.H. 2019. “Work-Life Balance: An Exploratory Analysis of Family-Friendly Policies for Reducing Turnover Intentions Among Women in U.S. Federal Law Enforcement”. *International Journal of Public Administration*, Vol 42 (4), 345-357.

AUTHOR BIOGRAPHIES

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 the sociology of labour. She is a doctoral supervisor and a member of the International Sociological Association. Her email address is a.p.bagirova@urfu.ru and her webpage can be found at <http://urfu.ru/ru/about/personal-pages/a.p.bagirova/>

OKSANA SHUBAT is an Associate Professor of Economics at Ural Federal University (Russia). She received her PhD in Accounting and Statistics in 2009. Her research interests include demographic processes, demographic dynamics and their 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 webpage can be found at <http://urfu.ru/ru/about/personal-pages/O.M.Shubat/>

Appendix 1: Clustering Dendrogram



RISING REGIONAL IMPORTANCE OF THE RENMINBI IN THE ASIA-PACIFIC AREA: A PANEL ANALYSIS

Eszter Boros and Gábor Sztanó
Magyar Nemzeti Bank (MNB), the Central Bank of Hungary
Krisztina krt. 55, HU-1013, Budapest, Hungary
Email: borosesz@mn.hu
sztanog@mn.hu

KEYWORDS

China, renminbi, exchange rate, Asia-Pacific region

ABSTRACT

In the past decade, China has taken several steps to make the renminbi (RMB) a globally important currency. Since the RMB was included in the IMF's SDR basket in 2016, several central banks have started using it as a reserve currency. Moreover, the exchange rate regime of the RMB has gradually been modified to add some flexibility, and Beijing has begun to promote cross-border RMB trade settlement. In this paper, we use panel data to analyze how changes in the RMB exchange rate might influence the movement of other currencies in the Asia-Pacific region. We find that the RMB's relative importance has significantly increased from 2015 onwards, i.e. since the date of the last reform of the RMB central parity quoting mechanism. Our results also suggest that the RMB's impact has remained broadly unchanged since the outbreak of the coronavirus pandemic. Contrary to some other findings, the COVID-19 has not reduced the relevance of the Chinese currency in the region.

INTRODUCTION

China has systematically strengthened its economic and political power in the past decades, although the internationalization of the renminbi was only started after the Global Financial Crisis. Beijing set the goal to turn the RMB into a globally used currency in order to extend economic and financial relations in Eurasia and beyond, especially with key trade partners. This plan involved the gradual opening of the capital account, making the RMB convertible in most of the cases. Several other proactive measures followed to increase the RMB's foreign turnover.

Statistical data reveal that cross-border renminbi flows and trade settlements keep rising for the past years, which could have changed the landscape of Asia-Pacific (AP) FX markets as well. So, in this paper, we analyze how changes in the RMB exchange rate might influence the movement of other currencies in the AP region. To be precise, we examine whether the exchange rate

movements of regional currencies are associated with fluctuations in the RMB exchange rate on a daily basis. Different periods are chosen to detect the possible shifts in this relationship. On the one hand, we examine the change which might have occurred after the announcement of the new quoting mechanism of the RMB/USD central parity in August 2015. On the other hand, it is also an interesting question whether there have been any changes since the onset of the pandemic and the resulting global economic turmoil.

Our analysis is structured as follows. The next section highlights some key trends and figures on how China has got more and more integrated in the global economy and how RMB transactions started to flourish. This section also summarizes the steps that China has made to internationalize its currency. After that, we introduce the empirical methodology and the results. The final section concludes.

THE INTERNATIONALIZATION OF THE RMB AND ITS EXCHANGE RATE REGIME

As China became the second largest economy after the U.S. a decade ago and it started to open up its financial markets, the "internationalization of the renminbi" appeared as a priority on its policy agenda. Beijing began to allow the "offshore" circulation of its currency in 2009, and subsequently, the RMB market outside of mainland China expanded to all key regions of the world. In the context of the "reform and opening-up" strategy and turning China into a modern, systemically important economy, the government decided to strengthen the role of the RMB in international trade settlements, investments and global reserves.

This endeavor proved to be vindicated by the rising footprint of China in the global economy. By 2020, the East Asian giant became the largest trade partner of the United States, the European Union as well as the Association of Southeast Asian Nations (ASEAN). China is also an important source of infrastructure finance in many regions, especially in many developing countries in Africa, Central and Southeast Asia. This role is underpinned by the Belt and Road Initiative (BRI) announced by Chinese president Xi Jinping in 2013 to

create an integrated Eurasian economic belt through upgrading physical and digital infrastructure.

The BRI and the connected institutions (such as the Asian Infrastructure Investment Bank, AIIB) are set to mobilize Chinese capital to finance the construction of ports, railways, roads, pipelines, power grids and enhance connectivity in financial markets as well. Although the rise in the cross-border use of the RMB has not been as fast as the increase in China's economic footprint worldwide, the amount of Chinese outward direct investment settled in RMB grew by 39.1% in 2020 (year on year; PBOC 2021). Moreover, the total amount of cross-border payments and receipts settled in the Chinese currency totaled RMB 28.39 trillion (approx. 4.5 trillion USD), which is an increase of 44.3% compared to 2019. This included a part of RMB 4.53 trillion settled between mainland China and the countries of the BRI, amounting to a year-on-year increase of 65.9% (PBOC 2021).

These facts illustrate the growing importance of RMB capital and cash flows for financial markets, especially in the Asia-Pacific region (including the ASEAN), which is closely linked to the Chinese economy. As a significant number of AP currencies is classified as floating or free floating (IMF 2021), it is fair to suppose that changes in the RMB exchange rate will increasingly affect them.

The evolution of the RMB exchange rate is largely determined by the onshore exchange rate regime. While no formal controls prevail over the exchange rate of the offshore RMB (also labelled as CNH), it is very strongly correlated to that of the onshore RMB (CNY). The general reason for this tight co-movement is the ongoing market liberalization and the growing RMB FX turnover (Erhart 2015).

According to the definition of the People's Bank of China (PBOC), the CNY is characterized by a managed floating exchange rate regime based on market supply and demand. The current regime was basically put in place in August 2015 when the PBOC changed the CNY/USD central parity quoting mechanism (Das 2019). This step was preceded by a decade of reforms when China had abolished a hard peg to the USD and gradually edged towards more exchange rate flexibility. Monetary authorities went on to widen the band around the central parity and ended up allowing +/-2% daily fluctuation from March 2014 onwards. The current mechanism of the central parity was clarified in early 2016 (with some fine-tuning since then). As of 2022, banks provide their daily central parity quotes with regard to two factors: the previous day's closing rate and the adjustment needed to account for the overnight changes in cross-rates of the CFETS currency basket (CFETS: China Foreign Exchange Trading System, the onshore interbank currency market). Based on the quotes, the PBOC announces the central parity on a daily basis, and the +/-2% band applies accordingly.

One of the first steps of RMB internationalization was to enable foreign importers to make their trades settled in RMB instead of a third currency. At the beginning of the 2010s, the outflow of currency was limited from China due to capital account restrictions. Therefore, the need for RMB liquidity was mostly connected to trade settlement. As a part of the internationalization policy, the PBOC has established a system of bilateral RMB currency swap agreements with other central banks, especially those that are part of the Belt and Road Initiative and other globally important central banks. This facilitates the provision of RMB liquidity for counterparties and promotes the RMB settlement of bilateral trade. Creating a renminbi hub has become one of the most important initiatives of BRI in the past decade (Song and Xi 2020).

The growing importance of the RMB paved the way for its inclusion into the IMF's SDR basket as of October 2016, acknowledging RMB as a globally important reserve currency. Several authors claimed that incorporating the RMB was making official the fact that many regional central banks had already been using the Chinese currency as an anchor in their execution of monetary policy operations (Uppal and Mudakkar 2020). The current share of the RMB is 10.92% in the SDR basket; the next re-evaluation of weights is due in mid-2022.

Several empirical papers found that the impact of the renminbi on other Asian currencies increased in the last decade. By employing a factor model, Fratzscher and Mehl (2011) found that the RMB had been a key driver of currency movements in Asia since as early as the mid-2000s, and claimed that the international monetary system is now tripolar: the RMB is the third most influential currency after the USD and the EUR. Shu et al. (2014) were more cautious with their findings: although they noted that both onshore and offshore renminbi had an impact on regional currency movements, they argued that the persistence of this influence depended on the progress in liberalizing the Chinese capital account. Recently, a network-based approach has become popular: two papers confirmed the recent increase of the RMB's regional dominance. Zhou et al. (2020) found that shocks originating from the RMB were significant in the relevant network of currencies, and the importance had increased over time. Besides, after examining SWIFT data, Liu et al. (2022) found that RMB shocks were indeed influential in the ASEAN region, but still insignificant in the global market. They noted that the RMB had become an influential regional currency; however, it was still far from global dominance.

All things considered, the growing importance of the RMB is undeniable and its regional impact is increasing gradually as a result of the internationalization efforts of the PBOC.

EMPIRICAL ANALYSIS OF THE IMPACT OF RMB EXCHANGE RATE CHANGES

In the empirical part of the paper, we examine how the relative importance of the RMB changed over time in the AP region. Our work broadly follows the methodology used by Marconi (2017).

Data and Variables

To establish the impact of the CNY exchange rate on AP exchange rate changes, 8 regional currencies are selected: the Australian dollar (AUD), the Indonesian rupiah (IDR), the Indian rupee (INR), the South Korean won (KRW), the Malaysian ringgit (MYR), the New Zealand dollar (NZD), the Philippine peso (PHP) and the Thai baht (THB). These currencies are all classified by the IMF as floating or free floating, so their exchange rates are supposed to be largely market-determined (IMF 2021). All 8 exchange rates, as well as the CNY exchange rate are measured in terms of the USD as a numeraire.

We also use a set of control variables that aim to capture all other factors that may influence the given exchange rate. As a proxy for global risk appetite and market volatility, we include the VIX index (*VIX*), i.e. the Chicago Board Options Exchange's CBOE Volatility Index as it is commonly used in the literature. In order to capture overall emerging market movements, we use EMBI Global and Fred's EM USD index. The first (*EMBI*) is a well-known emerging bond composite that moves along with the risk perception of emerging countries. The second (*DTWEXEMEGS*) is a USD nominal effective exchange index against emerging trade partners, published by the St. Louis Fed. We decided to apply two different variables to control non-Chinese, but emerging market-specific factors: while the exchange rate index directly reflects the overall performance of emerging currencies, the bond index is related to general risk perception in this market.

As not all currencies in our dependent variable are issued by an emerging country, we also incorporate a USD index by Bloomberg that grabs the movement of the USD vis-à-vis the 10 largest currencies (except the USD itself) (*BUSDIN*). We also consider a commodity index published by Bloomberg (*COMM*). Finally, we use the Chinese one-week repo rate to control for Chinese monetary policy shocks and liquidity conditions (*REPO*).

All data are daily closing data and were published by Bloomberg, except the Fed's index that was downloaded from the Fred database by the St. Louis Fed.

Model

In a way similar to Marconi (2017), we estimate a linear model for our panel dataset, which is given by *Equation (1)* as its basic specification:

$$\Delta \log(E_{i,t}) = \alpha + \beta_1 \Delta \log(CNY_t) + \beta_2 X_t + \varepsilon \quad (1)$$

where $E_{i,t}$ denotes the exchange rate of currency i at time t ($i = \text{AUD, IDR, INR, KRW, MYR, NZD, PHP, THB}$). CNY_t is the CNY/USD daily exchange rate – thus, our primary interest in this research is coefficient β_1 . For the sake of brevity, X refers to our control variables introduced under the previous subtitle, using the first differences of logarithms in all cases. Intercept α is common and time-invariant. (Chow tests showed that currency-specific fixed effects were not needed in the model.) Finally, ε denotes the error term.

By obtaining coefficient β_1 for variable CNY, we can establish the significance of the CNY exchange rate regarding the fluctuation of AP currencies. Control variables are intended to separate the RMB-specific impacts from other relevant factors in the global and emerging financial markets. Limitations of this approach are discussed later.

Results and Discussion

To capture the possible changes in the impact of the CNY exchange rate over time, we estimate *Equation (1)* for three periods. First, we consider the whole period since China started to allow the RMB to fluctuate in a slightly wider band. The first complete year in which the band was +/-0.5% happened to be 2008, so our initial dataset runs from 1 January 2008 to 7 January 2022 ($t = 3,659$). Note that during this period, the band was further extended in April 2012 (+/-1%) and in March 2014 (+/-2%), plus the quoting mechanism of the central parity was changed in August 2015 (with further clarifications in the subsequent months). Thus, this long period involves quite meaningful policy changes, and as such, it serves mainly as a reference point in our research.

In the second case, the estimation is confined to the period since the announcement of the new quoting mechanism in August 2015 (1 September 2015 – 7 January 2022, $t = 1,659$). For the third estimation, we only look at the period of the pandemic, starting from the mass infections in China (1 January 2020, $t = 528$). The latter allows us to investigate whether the fallout of the pandemic instigated any changes in the relationship between AP currencies and the CNY. Our results for the different periods are reported in *Tables 1-3*.

Table 1: Pooled model results for 2008-2022

	coeff.	std.err	t-ratio	p-value
Constant	0.000	0.000	0.054	0.957
CNY	0.051	0.020	2.600	0.009***
VIX	0.003	0.001	6.722	0.000***
EMBI	0.018	0.002	9.049	0.000***
DTWEXEMEGS	0.492	0.013	36.92	0.000***
BUSDIN	0.174	0.007	23.42	0.000***
COMM	-0.025	0.004	-6.811	0.000***
REPO	-0.001	0.000	-1.953	0.051
F-statistic: 974.3 p-value: 0.000*** Adj. R-squared: 0.203				

Table 2: Pooled model results for August 2015-2022

	coeff.	std.err	t-ratio	p-value
Constant	0.000	0.000	-0.763	0.446
CNY	0.107	0.019	5.600	0.000***
VIX	0.002	0.001	4.206	0.000***
EMBI	0.026	0.003	8.947	0.000***
DTWEXEMEGS	0.318	0.016	20.18	0.000***
BUSDIN	0.207	0.010	20.18	0.000***
COMM	-0.011	0.005	-2.192	0.028**
REPO	-0.001	0.000	-3.273	0.001***
F-statistic: 460.8 p-value: 0.000*** Adj. R-squared: 0.211				

Table 3: Pooled model results for 2020-2022

	coeff.	std.err	t-ratio	p-value
Constant	0.000	0.000	1.192	0.233
CNY	0.103	0.035	2.944	0.003***
VIX	0.001	0.001	1.165	0.244
EMBI	0.045	0.005	9.923	0.000***
DTWEXEMEGS	0.185	0.028	6.561	0.000***
BUSDIN	0.302	0.020	15.43	0.000***
COMM	-0.019	0.008	-2.484	0.013**
REPO	-0.001	0.000	-2.550	0.011**
F-statistic: 186.7 p-value: 0.000*** Adj. R-squared: 0.251				

Source of *Table 1-3*: Own estimations. Variables are used as first differences of logarithms.

Significance codes: *** $p < 0.01$, ** $p < 0.05$

The results show that in the full sample (*Table 1*), the Chinese currency has a significant impact on the daily exchange rate changes of the 8 regional currencies. The estimated coefficient suggests that ceteris paribus, a 1% change in the CNY exchange rate is expected to be coupled with a 0.05% change in the AP currencies' exchange rates. The positive sign shows that changes tend to happen in the same directions. It is important to note that all control variables, except the one-week Chinese repo rate, are significant. The signs of the β_2 coefficients show reasonable relationships (with all other things unchanged). Investors' expectations of higher volatility (higher *VIX*) seem to come along with the depreciation of AP currencies (due to risk aversion). The same is true for the emerging market bond composite (*EMBI*): higher emerging market yields also point to increased risk perception, and thus, the devaluation of our currencies. As for the coefficients of the USD indexes (*DTWEXEMEGS* and *BUSDIN*), their higher value indicates the appreciation of the USD in both cases, which is reasonably coupled with the depreciation of the selected AP currencies. At the same time, higher commodity prices (*COMM*) might involve an appreciation in the AP region as these economies usually benefit from price increases of raw materials. (The significance and the signs of the β_2 coefficients remain almost the same in our models over time. One notable exception is *REPO*, which we discuss below.)

As we noted earlier, the exchange rate mechanism of the CNY changed substantially in August 2015 (before the SDR inclusion in 2016). Our sample with observations between September 2015 and January 2022 (*Table 2*) shows that the coefficient of the *CNY* was not only significant, but it also increased, even doubled. This outcome is similar to the results of Marconi (2017) although the time series in her sample were certainly shorter. Thus, we can establish that the new central parity quoting mechanism, which was a step closer to market demand and supply, helped to extend the regional impact of the Chinese currency.

Note that in this period, *REPO* became significant. In other words, the one-week repo interest rate set by the PBOC directly appeared as a relevant factor in the Asia-Pacific FX markets. This hints at the growing international importance of Beijing's monetary policy. (From September 2015 onwards, higher repo rates tend to be accompanied by the appreciation of AP currencies against the USD.) This finding is generally in line with the strengthened role of China in the global economy and the process of RMB internationalization.

Regarding the third subsample covering the pandemic period (*Table 3*), the coefficient of the *CNY* is still significant and does not change substantially. This result is relatively positive from the viewpoint of the renminbi: the outbreak of the coronavirus crisis initially threatened with a severe economic slowdown and a fall in international financial transactions; however, these obstacles did not prevent the renminbi from maintaining its role in the Asia-Pacific region. This finding is remarkable because for instance, Fang and Cao (2021) found that the influence of the Chinese currency weakened in the countries of the Belt and Road Initiative during the pandemic. In light of their result, we might conclude that South and Southeast Asia, along with Australia and New Zealand, represent the core area of the growing economic power of China. Strengthening trade, investment and financial ties with these countries will be a key interest of Beijing, and the region will continue to be the mainstay of the BRI and the connected Maritime Silk Road in the future.

Overall, our results suggest that the importance of the RMB has increased in the Asia-Pacific FX markets since 2015, and the pandemic has not broken its influence, either.

Limitations

Explaining changes in the foreign exchange market is undoubtedly a difficult task as several unobservable factors may influence price movements. The CNY exchange rate is also dependent on the explanatory variables that we used, so the question of endogeneity may arise. Through comparing the model specifications with and without the *CNY* variable, we could ensure that the explanatory power increased, and the significance of

the CNY also confirms that it plays a significant role in the exchange rate changes of the selected AP currencies.

CONCLUSIONS AND OUTLOOK

China has grown at an enormous pace during the past decades and its financial reforms have paved the way to becoming one of the most important superpowers in the world. China's regional impact is indisputable in several areas of international relations, most notably as a result of the Belt and Road Initiative. However, the academic literature about its relative importance in global financial markets is still relatively scarce as the country is still on the way to opening-up its financial markets. The gradual introduction of RMB convertibility, flexibility and its inclusion into the IMF's SDR basket all contributed to strengthening the global and regional importance of the Chinese currency.

In our paper, we presented a simple panel regression framework that aimed to uncover the relationship between daily exchange rate changes of the RMB and those of 8 Asia-Pacific currencies (with floating or free-floating regimes according to IMF classification). Our results show that the renminbi became more relevant in the region after 2015, compared to earlier years. This suggests that the reform of the central parity quoting mechanism in the onshore FX market was a significant step by China to better reflect market valuations. This, in turn, helped the RMB to take a more pronounced role in influencing exchange rate changes in the region. While the pandemic has posed severe challenges to global and regional financial integration, the Chinese currency succeeded in maintaining its significance, at least in the neighborhood of the East Asian giant, i.e. Southeast Asia and the Pacific.

Looking ahead, announcements and policy frameworks show that China will stick to the policy of "reform and opening up" and accelerate the measures which are necessary for a further increase in cross-border RMB transactions. These policies are quite complex, ranging from the reduction of investment restrictions (the so-called "negative lists") to the creation of a new macroprudential policy framework for outbound RMB investments. As a result of such policies, onshore investors gain more opportunities to tap foreign financial markets, through initiatives like the Shenzhen-Hong Kong Stock Connect or the Shanghai-London Stock Connect Scheme. International investment banks have recently been granted licenses to start operations in mainland China. Last but not least, one should not forget about the substantial progress China has made in terms of digitalizing its currency and adapting it to the requirements of the digital age. Quite large-scale tests of the Chinese central bank digital currency (labelled as DC/EP or e-CNY) have been taking place since 2020. This project has the potential to create some favorable features of the RMB which might be necessary for its global acceptance and use in the long run.

REFERENCES

- Das, S. 2019. "China's Evolving Exchange Rate Regime." *IMF Working Paper*, WP/19/50.
- Erhart, S. 2015. "Liberalisation of the Renminbi Exchange Rate Regime and Foreign Currency Regulations." *Budapest Renminbi Initiative Papers*, 2.
- Fang, X. and W. Cao. 2021. "The Impact of COVID-19 on the Status of RMB as an Anchor Currency." *Asian Economics Letters*, 2(1).
- Fratzcher, M. and A. Mehl. 2011. "China's Dominance Hypothesis and the Emergence of a Tri-polar Global Currency System." *ECB Working Papers Series*, No. 1392/2011.
- International Monetary Fund. 2021. "Annual Report on Exchange Arrangements and Exchange Restrictions 2020."
- Liu, T; X. Wang; and W. Woo. 2022. "The Rise of Renminbi in Asia: Evidence from Network Analysis and SWIFT Dataset." *Journal of Asian Economics*, 78 (2022).
- Marconi, D. 2017. "Currency Co-Movements in Asia-Pacific: The Regional Role of the Renminbi." *BOFIT Discussion Papers*, No. 10/2017.
- People's Bank of China. 2021. "RMB Internationalization Report 2020."
- Shu, C.; D. He; and X. Cheng. 2014. "One Currency, Two Markets: the Renminbi's Growing influence in Asia-Pacific." *BIS Working Papers Series*, No. 446.
- Song, K. and L. Xia. 2020. "Bilateral Swap Agreement and Renminbi Settlement in Cross-Border Trade." *Economic and Financial Studies*, 8:3, 355-373.
- Uppal, J. and S. Mudakkar. 2020. "China's Belt and Road Initiative and the Rise of Yuan - Evidence from Pakistan" *The Lahore Journal of Economics*, 25:1 (Spring), 1-26.
- Zhou, Y.; X. Cheng; and Y. Wang. 2020. "Measuring the Importance of RMB in the Exchange Rate Spill-over Networks: New Indices of RMB Internationalization." *Economical and Political Studies*, DOI: 10.1080/20954816.2020.1775374

DATA SOURCES

- Bloomberg. Tickers are available upon request. Downloaded from the terminal on 08.01.2022
- Fred database. Tickers are available upon request. Downloaded from the website on 08.01.2022

AUTHOR BIOGRAPHIES

Eszter BOROS, PhD is an International Expert at the International Relations Directorate of the Central Bank of Hungary. She joined the MNB in 2016 as a Methodology Expert in banking supervision. In 2020, she became an International Expert specializing in economic analysis with a focus on China and the RMB, as well as EU-China relations and Eurasian cooperation. She earned her PhD from Corvinus University of Budapest in 2021. Her email address is borosesz@mnbb.hu

Gábor SZTANÓ has been an Economist at the Monetary Policy and Financial Market Analysis Directorate of the Central Bank of Hungary since 2014 and he is a member of the financial market monitoring team. He earned his master's degree in Economics from Corvinus University of Budapest, specializing in Bank and Public Finance and currently is a PhD candidate at CUB. His main field of research is monetary policy in emerging countries. His email address is sztanog@mnb.hu

The paper contains the opinion of the authors which is not necessarily in line with that of the Central Bank of Hungary.

Implied volatility based margin calculation on cryptocurrency markets

Balázs Králik, Nóra Felföldi-Szűcs, Kata Váradi
Department of Finance
Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest Hungary

KEYWORDS

Crypto Markets; Implied Volatility; Margin

ABSTRACT

The focus of our research is the use of implied volatility in the context of margin calculations on BTC positions. We will compare the standard deviation estimated from historical data to the implied volatility values estimated from ATM option prices. As our main result, we show that the implied volatility is a superior basis for margining purposes. Not only does it perform better in a back-test, it also requires lower average margin levels to provide the same risk profile. Our results also show that the logreturns of BTC cannot be assumed to be normally distributed.

INTRODUCTION

In our paper, we test the performance of implied volatility based margin calculation for Bitcoin (BTC) positions. Our research connects to a large body of already existing literature; it relates to the following areas:

First, the broadest context for our work is provided by the literature on the margining of clearinghouses, more generally the risk management of clearinghouses. Berndsen (2021) provides a good overview on the latter topic, while Murphy, Vasios, and Vause (2016) describes how margins should be defined to meet actual regulations and to cover possible losses coming from counterparty risk. Usually, to open a position on a centralized market, agents

need to buy the asset outright or deposit margin (called the initial margin) to signal their ability to settle their obligations and to cover possible future losses incurred in case of failure to pay. The losses and gains from the daily settlement drive the balance of the margin account. Once the level of the margin account falls below the maintenance margin, the trader receives a margin call and is obliged to restore the margin to the required level. If the price movements of the asset decrease the trader's margin to the extent that he will not be able to meet the margin call and his positions have to be closed by the clearinghouse, the losses occurring when closing the position should be covered by the margin account of the trader.

Since the volatility of the given asset is the basis of the margin calculation, measuring and forecasting volatility of financial assets is the second important part of the related literature for our paper. The volatility of logreturns for the asset, i.e. the second moment of the probability distribution, is a risk measure. It quantifies possible deviations of ex-post returns from the expected level. It serves as an input for optimization problems (see e.g. portfolio selection in Markowitz 1952 and Sharpe 1964), risk management models and pricing (see e.g. Black and Scholes 1973). Historical and implied volatility models are already distinguished and compared in the early literature review of Mayhew (1965). There is already strong evidence that returns are not normally distributed (e.g. Fama 1965) and volatility is not constant over time. Several models are applied to assess the empirical characteristics of volatility. Granger and Poon compare the performance of the following three model classes based on their previous work surveying 93 arti-

cles (Granger and Poon 2002): time-series models (constant historical volatility, autoregressive conditional heteroscedasticity (ARCH) models), stochastic volatility models and implied volatility models. The results of Granger and Poon (2005) show that for forecasting, the implied volatility outperforms time-series models since it incorporates both current information and expectations for the future. Historical ARCH models do not outperform implied volatility models, but they show better results than stochastic volatility forecasts. Granger and Poon (2005) also report that the forecasting power of volatility models can be improved by using data of higher frequency (e.g. 5 minutes-data for developed markets), and short-horizon forecasts are more successful than those for long horizon. Our comparison historical volatility to implied volatility is based on relatively short, 8-hour periods.

Next we provide a short overview of implied volatility estimation. The most important milestone of derivative pricing is the Black-Scholes-Merton model (see Black and Scholes 1973; Merton 1973) where the value of a European call option can be calculated as follows:

$$C(S, T - t) = N(d_1)S - N(d_2)Ke^{-r(T-t)} \quad (1)$$

$$d_1 = \frac{1}{\sigma\sqrt{T-t}} \left[\ln\left(\frac{S}{K}\right) + (T-t)\left(r + \frac{\sigma^2}{2}\right) \right] \quad (2)$$

$$d_2 = \frac{1}{\sigma\sqrt{T-t}} \left[\ln\left(\frac{S}{K}\right) + (T-t)\left(r - \frac{\sigma^2}{2}\right) \right] \quad (3)$$

where

- S: spot price of the underlying product,
- K: strike price,
- σ : standard deviation of the underlying
- r: risk free rate
- T-t: time to expiration
- N: normal cumulative distribution function

Several authors report that the distribution of return has fat tails instead of following normal distribution which allows the so-called volatility smile

to occur. (See for a recent detailed overview of the topic: Zulfiqar and Gulzar 2021). When observing the prices of options traded at different exchanges and in different asset classes, traders usually experience a difference between the current market price and the theoretical fair value of the derivative. Implied volatility is the volatility level of the underlying product which, used as input for Black-Scholes-Merton model gives the observed market price as a result of the theoretical model. Since the level of implied volatility reflects the expectations of the market as well, using it as an estimate for future volatility is a widespread and - as cited in the above literature overview - an efficient method.

For margining purposes, there is one further step from volatility estimation to margin calculation. On the basis of the volatility forecasts, a downside risk measure like Value-at-Risk (VaR) or expected shortfall (ES) needs to be calculated (for details see e.g. Jorion 2001). Based on Bams, Blanchard, and Lehnert (2017), in case of VaR calculation, the implied volatility based VaR does not outperform the historical volatility based ones: they arrive at the opposite result to the one we cited earlier in this article. These contradictory, non-obvious results led us to pose our research questions. We compare an implied volatility based margin calculation method to one where the margin level is specified by the actual regulations in the European Union (EMIR 2012, RTS 2013). For simplicity, we will disregard the handling of procyclicality within the initial margin calculation, nor will we apply any additions to the calculated margin requirements. Based on the regulation, the margin will be calculated to be sufficient at a 99% significance level, with a 12 months look-back period. As a further simplification, we have modified the required 2-day liquidation period to an 8-hourly liquidation period, which is more in keeping with the global nature of the crypto markets and the prevailing 8-hour margining regime customary there. This margin value will be the basis of comparison to the other margin calculations described later in this paper in the Back-testing section.

The most recent area within the literature is the application and testing of already existing results on the dynamically emerging crypto markets. Hence the focus of our research is how the use of implied volatility can contribute to margin calculation on BTC positions; the most closely related

antecedent of our work is the contribution of (Zulfiqar and Gulzar 2021) who are the first to have tested for the volatility smile and implied volatility for BTC options. They use 14-day maturity Bitcoin options of two time periods traded on Deribit Bitcoin Futures and Options Exchange. Root-finding iterative techniques, namely the Newton Raphson and the Bisection methods proved to effectively estimate the implied volatility of BTC options. The results show similarities with that of assets from the commodity class, therefore Zulfiqar and Gulzar classify BTC as a commodity. More generally, all recent works on crypto markets provide a background for our paper. The cryptocurrency markets have shown extreme development over the last years, therefore they received considerable attention as a research topic as well. We can distinguish the following directions in the dynamically evolving research: price discovery, market efficiency and optimal trading strategies (especially hedging) or the use of crypto products in risk management (Alexander et al. 2020; Deng et al. 2019). Pichl and Kaizoji (2017) defines the fundamental importance of Bitcoin and its security aspects also as an existing direction in the literature. The behaviour of volatility on crypto markets is an important question of price discovery and risk management strategies as well. While Pichl and Kaizoji (2017) used a Heterogenous Autoregressive model for Realized Volatility for BTCUSD data, the most widespread methodology is provided by the GARCH models. See e.g. Katsiampa et al. (2017) who compared the forecasting power of several GARCH models for the volatility of BTC. According to them, the ARCGARCH model achieved the highest goodness-of-fit on their dataset.

SIMULATION DATA

We chose to use the free tier of market data integrator coinapi.com that has downloadable historical data for both underlying and derivative crypto data. This tier allows the daily download of up to 100 klines (open/close/high/low candle data) for up to 100 products (specific option contracts). Given the very large number of option contracts (on 19th January 2022, *coinapi* listed 25,843 options contracts on BTC expiring after 1st January 2021, across all exchanges), we had to settle for a subset

of them. *Binance* (binance.com) lists vanilla European options on BTC/USDT with regular weekly and monthly expiration. On 19th January 2022 there were 1,348 contracts on BTC/USDT that had expiration after 1st January 2021. We further reduce this number by considering only those contracts whose strike price was close to being ATM over its liquid (30 day) lifetime. In this context, a contract (strike price K , underlying price S , expiration time T expressed as days) is close to being ATM if:

$$S^{max} = \max_{T-30 \leq t \leq T} S_t \quad (4)$$

$$S^{min} = \min_{T-30 \leq t \leq T} S_t \quad (5)$$

$$S^{min} - (S^{max} - S^{min}) < K < S^{max} + (S^{max} - S^{min}) \quad (6)$$

This restriction yields 767 contracts for this study.

Each of the contracts has an associated time series of 8-hour kline data (open/high/low/close price, period volume), maximally 100 periods long. Contracts become liquid when they are ATM, otherwise there are often no transactions. There are altogether 7,942 non-empty klines in this data set, which is less than 10% of the maximum possible.

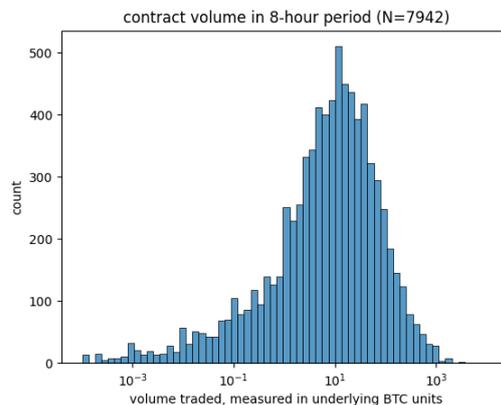


Figure 1: Available volume in BTC options on Binance in 2021

In addition to the *Binance* option data, we also collected information on the underlying security. For *Binance*, the underlying security is the BTC/USDT pair that is very liquid. $S =$

BTC/USDT is the price of BTC in units of USDT, the latter being a stablecoin, i.e. a coin whose value is closely tethered to the fiat USD currency. The data for S is available for all 8-hour periods. We save them between 2020-01-01 and 2022-01-19.

Implied Volatility Calculation

The concept of implied volatility was described above. In our study, we used `scipy.optimize.scalar_root` (Jones, Oliphant, Peterson, et al. 2001–) to find the implied volatility, which converges in all cases except for prices implying negative time-value for options – these prices were filtered out.

For each 8-hour period, multiple contracts would typically have a non-empty kline: puts, calls at various strike prices and expiration times trade at the same time. In order to estimate a single implied volatility σ_t for each period, we use a weighted average defined below in Equation 7 and 8. In using the Black-Scholes formula to find $\sigma_t^{(i)}$, we assume no dividends and a risk-free rate of 1.85%. This was a typical 2021 US T-Bill rate. The results below are not sensitive to this choice if r is within reasonable bounds, because the time to maturity of the liquid options is so short, typically less than a week.

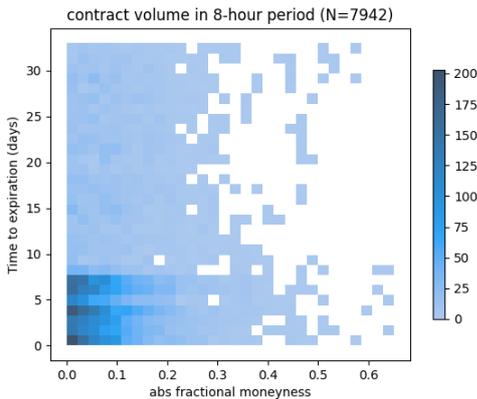


Figure 2: Available volume in BTC options on Binance in 2021 vs time to expiration and moneyness

$$\sigma_t = \frac{\sum_{i \in \mathcal{C}} w_{i,t} \sigma_t^{(i)}}{\sum_{i \in \mathcal{C}} w_{i,t}} \quad (7)$$

$$w_{i,t} = \frac{V_{i,t}}{1 + (S_t - K_i)^2} \quad (8)$$

where

- \mathcal{C} is the set of all contracts,
- $V_{i,t}$ is the volume of contract i in time period t

As seen in Figure 2, the majority of the weight in this average comes from contracts in the last week of trading, with strike prices within 20% of the underlying.

BACKTESTING

For each 8-hour period in 2021 we calculate the forward 8-hour maximum loss L_{8h}^{max} by using the underlying minimum and maximum over that period. L_{8h}^{max} is the larger of the absolute difference of the next 8-hour maximum (or minimum) of the underlying and the current close price. The goal of a margining policy is to make sure that in any period the maximum loss is covered by the margin posted by the trader at a given significance level. Margin in this simplified simulation will be taken to be a percentage M_t of the trader’s net position in the underlying at the end of each period. The trader will be required to post sufficient margin in order to be allowed to increase her position in the next period. In this simulation we examine 3 policies:

1. Realized Trailing Sigma Policy
2. Implied Volatility Policy
3. Normal VaR Policy

Margining regimes will be compared according to their average margin level. Clearly, a CCP will be more competitive in the market if it tends to require smaller margins, while maintaining the same risk profile. We will find, through simulation, that the above methods do differ in how much margin they require on average to reach the same fraction C incidence of insufficient margin. C is usually 1%,

based on the 99% significance level requirement of the EMIR 2012.

For a particular margin policy, at time t , the forward incidence rate will be denoted F_t^{fwd} . For each t we determine F_t^{fwd} , and the distribution of $\{F_t^{fwd}|t \in 2021\}$ will be compared to the margin M_t in effect at time t . The total failure rate is $F^{fwd} = E(\{F_t^{fwd} > M_t|t \in 2021\})$.

Realized Trailing Sigma Policy

In this study, the Realized Trailing Sigma Policies involve calculating the population standard deviation (SD) of the realized 8-hour returns of the underlying over a 250-day trailing time interval. The policy parameter m^{hist} is used to multiply the historical sigma to obtain the margin $M_t = m^{hist} \sigma_t^{trail}$, where σ_t^{trail} is the trailing historical sigma at time t . BTC/USDT data is available for dates prior to 2021, so we are able to use 2020 data to get trailing standard deviation estimates from the beginning of 2021. We look at F^{fwd} as a function of m^{hist} . In this study we do not attempt to cross-validate m^{hist} on an outsample period, mainly due to the scarcity of options data.

Implied Volatility Policy

In contrast to the Realized Trailing Sigma Policy, the Implied Volatility Policy does not rely on a long trailing historical value of the standard deviation of the underlying. Instead, it uses a smoothed function of the implied volatility calculated in the manner described earlier.

For each time t , we compute the implied volatility σ_t^{impl} , and $M_t = m^{impl} \sigma_t^{impl}$. We look at F^{fwd} as a function of m^{impl} .

EMA of Maximum Variation Policy

Instead of looking at implied volatilities, we can also consider the recent volatility of the market to suggest an optimal margin. As a short-term estimator of volatility, we use the difference of the minimum and maximum underlying prices (maximum variation) within a kline. An exponential moving average (EMA) of these observations is then taken as the statistic on which to base this policy.

For each time t , we compute the maximum variation σ_t^{var} , and $M_t = m^{var} \sigma_t^{var}$. We look at F^{fwd} as a function of m^{var} .

Normal VaR Policy

The Normal VaR Policy is based on the results of the previous two policies, namely the VaR is calculated both from the Realized Trailing Sigma, and from the Implied Volatility as well. In the Normal VaR method, we assume that the periodic profit of a position is normally distributed. If the expected logreturns were normally distributed as $N(0, \sigma)$, the m parameter would be fixed at 2.33 in order that $E(F^{trail}) = 1\%$. In this study, the Realized Trailing Sigma Policy, and the Implied Volatility Policy is the generalization of the VaR policy involving actual calibration of the m parameter on different significance levels (99% and 97%). However the Normal VaR Policy based on the normal assumption is still interesting to consider, since it has been a commonly applied regulatory approach in the past.

SIMULATION RESULTS

Figure 3 shows characteristics of the underlying returns over the past 2 years. What we see is that the 1-year trailing standard deviation of the 8-hour returns is a poor proxy for the possible maximum loss values of a position, as it is approximately constant. Maximum loss L is defined for each period using the kline data: $L_t = \max_{X \in \{high, low\}} (|S_{t-1}^{close} - S^X|)$

Option data is only available from February of 2021. Figure 4 compares the historical standard deviation of returns to standard deviations implied by options prices. The smoothed implied volatilities in Figure 4 are given by the 5-day rolling maximum of the exponential moving average (half-life = 1 day) of the implied volatility. It is visually clear that the implied volatilities follow the evolution of the actual maximum loss quite well. The 1 month EMA of the maximum variation performs similarly to the implied volatility: it is high when the realized maximum loss is high.

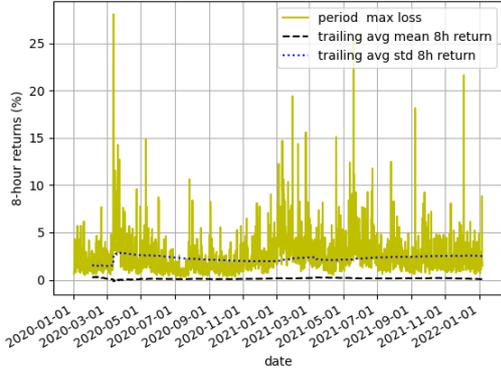


Figure 3: characterization of BTC/USDT 8hr returns

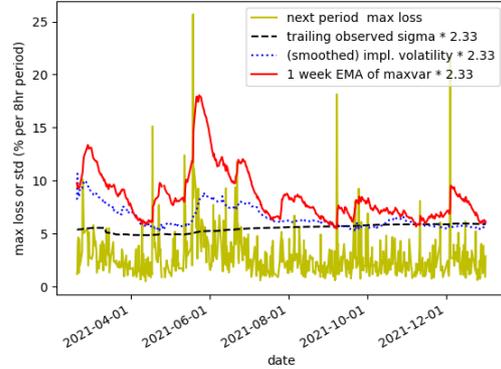


Figure 4: Historical and implied standard deviations compared to next period maximum loss. See text for motivation of the factor 2.33

Comparing Policies

An Implied Volatility Policy and a Realized Trailing Sigma Policy are both defined by a multiplier m^{policy} . In this study we look at what multiplier is necessary to achieve a certain failure rate in 2021. In fact we try a number of multipliers, and for each multiplier we look at what the simulated failure rate is, as well as what the average margin rate is. Due to the scarcity of options data, we are effectively doing an in-sample optimization of m^{policy} . A lower margin rate for a given failure rate signals a better margining policy.

If returns were normally distributed, the multiplier required for a particular desired failure rate would be trivially given by the inverse normal CDF: For example a desired 1% failure rate would imply $m^{policy} = 2.33$ and 3% would correspond to $m^{policy} = 1.88$. The average (over t) required margin M^{avg} would be given by the markers on Fig 5. It is immediately clear that the returns are not normally distributed. For example, the Realized Trailing Sigma VaR Policy with $m = 2.33$ has a failure rate of more than 8%, instead of the intended 1%.

The Implied Volatility Policy with $m = 2.33$ does much better than the Realized Trailing Sigma Policy with $m = 2.33$, as it only fails in about 3% of the time. It is clear that the normal assumption is faulty, and calibrated m -s are necessary.

As the lines on Figure 5 show, even if we calibrate m for the Realized Trailing Sigma and Implied Volatility policies, the latter does better: at

all failure rates, a calibrated Realized Trailing Policy gives a higher average margin, than a calibrated Implied Volatility Policy. Moreover if we calculate the VaR based on the Implied Volatility Policy, that backtest performs significantly better.

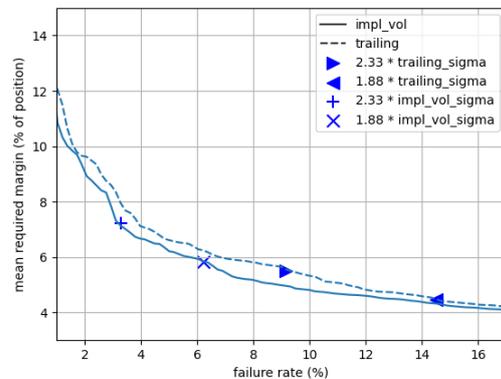


Figure 5: Comparison of Realized Trailing Sigma, Implied Volatility, and EMA of Maximum Variation Policies. For any desired failure rate, the mean margin is obtained by calibrating m , then simulating on 2021

Figure 6 shows a more detailed view of how the Implied Volatility and the EMA of Maximum Variation perform with respect to the baseline Realized Trailing Sigma policy. Over the full range

of targeted failure rates, Implied Volatility and EMA policies require lower mean margins, meaning cheaper trading for the investor with the same risk for the exchange.

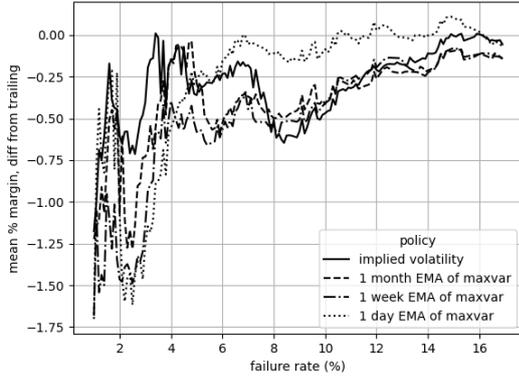


Figure 6: Comparison of Implied Volatility, and EMA of Maximum Variation Policies. We show the mean required margin gain with respect to the Realized Trailing Sigma policy. The greatest gains are to be had for the most relevant (low) failure rates.

CONCLUSIONS

Our paper shows that using implied volatilities in BTC/USDT margining regimes is superior to using historical standard deviations. First of all, implied volatilities predict market volatility much better than historical standard deviations. This allows the CCP to calculate a lower required margin on average, while obtaining the same failure rate. This remains the case whether we assume normally distributed logreturns or not. For normally distributed logreturns, the Implied Volatility Policy gives, over 2021, a 3% failure rate compared to 8% for the Realized Trailing Sigma Policy. If we drop the normality assumption, and calibrate m to obtain a desired failure rate, it is still the case that a calibrated Implied Volatility Policy yields a lower average margin than the Realized Trailing Sigma policy.

Overall, our results are consistent with the commonly held view that BTC/USDT returns follow very fat tailed distributions.

APPENDIX

The raw data as it is downloaded from coinbase consists of csv files for each symbol. For example, for BINANCEOPTV_OPT_BTC_USDT_211231_48000_P (vanilla put option on BTC/USDT expiring 2021.12.31, strike 48000), the kline data looks like the sample in . The option and underlying data are in the same format, but it is revealing comparing the two, especially in terms of the volumes.

Option Attribute	Value
time_period_end	2021-12-31T08:00:00.000000Z
time_period_start	2021-12-31T00:00:00.000000Z
time_open	2021-12-31T00:38:56.277000Z
time_close	2021-12-31T03:25:34.610000Z
price_open	3935.01
price_high	4155.18
price_low	206.67
price_close	989.04
volume_traded	91.1301
trades_count	2500
Underlying Attribute	Value
time_period_end	2021-12-31T08:00:00.000000Z
time_period_start	2021-12-31T00:00:00.000000Z
time_open	2021-12-31T00:00:00.000000Z
time_close	2021-12-31T07:59:59.999000Z
price_open	47120.88
price_high	47550.0
price_low	46825.38
price_close	47191.09
volume_traded	6896.44835
trades_count	248930

Figure 7: Coinapi option data format

We have 7942 relevant options data points, distributed over 579 (8-hour) time periods, 46 expirations (one per week). For each of these time periods, as well as for all of 2020, we have underlying data as well, one kline per 8-hour period.

REFERENCES

- Alexander, Carol et al. (2020). “BitMEX bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness”. In: *Journal of Futures Markets* 40.1, pp. 23–43.
- Bams, Dennis, Gildas Blanchard, and Thorsten Lehnert (2017). “Volatility measures and Value-at-Risk”. In: *International Journal of Forecasting* 33.4, pp. 848–863.
- Berndsen, Ron (2021). “Fundamental questions on central counterparties: A review of the literature”. In: *Journal of Futures Markets* 41.12, pp. 2009–2022.
- Black, Fischer and Myron Scholes (1973). “The Pricing of Options and Corporate Liabilities”. In: *Journal of Political Economy* 81.3, pp. 637–654. DOI: 10.1086/260062.
- Deng, Jun et al. (2019). “Optimal Bitcoin Trading with Inverse Futures”. In: *Available at SSRN 3441913*.
- EMIR (2012). *Regulation (EU) No. 648/2012 on OTC derivatives, central counterparties and trade repositories*. Data retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R0648&from=EN>.
- Fama, E. (1965). “The Behaviour of Stock Market Prices”. In: *Journal of Business* 64, pp. 34–105. DOI: <http://dx.doi.org/10.1086/294632>.
- Granger, Clive W. and Ser-Huang Poon (2002). “Forecasting Volatility in Financial Markets: A Review (Revised Edition)”. In: DOI: <http://dx.doi.org/10.2139/ssrn.331800>.
- (2005). “Practical Issues in Forecasting Volatility”. In: URL: <https://ssrn.com/abstract=661423>.
- Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001–). *SciPy: Open source scientific tools for Python*. URL: <http://www.scipy.org/>.
- Jorion, P. (2001). *Value at Risk: The New Benchmark for Managing Financial Risk*. 2nd. United States of America: McGraw-Hill.
- Katsiampa, Paraskevi et al. (2017). “Volatility estimation for Bitcoin: A comparison of GARCH models”. In: *Economics Letters* 158.C, pp. 3–6.
- Markowitz, Harry (1952). “Portfolio Selection in The Journal of Finance Vol. 7”. In.
- Mayhew, S. (1965). “Implied Volatility”. In: *Financial Analysts Journal* 51.4, pp. 8–20. URL: <http://www.jstor.org/stable/4479853>.
- Merton, Robert C (1973). “Theory of rational option pricing”. In: *The Bell Journal of economics and management science*, pp. 141–183.
- Murphy, David, Michalis Vasios, and Nicholas Vause (2016). “A comparative analysis of tools to limit the procyclicality of initial margin requirements”. In.
- Pichl, Lukáš and Taisei Kaizoji (2017). “Volatility analysis of bitcoin”. In: *Quantitative Finance and Economics* 1.4, pp. 474–485.
- RTS (2013). *Regulation (EU) No. 153/2013 supplementing Regulation (EU) No. 648/2012 with regard to regulatory technical standards on requirements for central counterparties*. Data retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0153&from=EN>.
- Sharpe, William F (1964). “Capital asset prices: A theory of market equilibrium under conditions of risk”. In: *The journal of finance* 19.3, pp. 425–442.
- Zulfiqar, Noshaba and Saqib Gulzar (2021). “Implied volatility estimation of bitcoin options and the stylized facts of option pricing”. In: *Financial Innovation* 7.1, pp. 1–30. DOI: 10.1186/s40854-021-00280-.

AUTHOR BIOGRAPHIES

BALÁZS KRÁLIK (balazs.kralik@uni-corvinus.hu) is a Senior Lecturer at Corvinus University Budapest. He joined the Finance Department after a career in quantitative trading. His research interests include pricing and risk management of novel financial instruments.

NÓRA SZÜCS (nora.felfoldi-szucs@uni-corvinus.hu) has served as a lecturer at Corvinus University since 2006, where she obtained her PhD in 2013. Between 2015-2020, she worked as a researcher at John von Neumann University. Her field of interest is microfinance, credit risk, and contract theory.

KATA VÁRADI (kata.varadi@uni-corvinus.hu) is an Associate Professor at Corvinus University of Budapest, at the Department of Finance. She graduated from CUB in 2009, and then obtained her PhD in 2012. Her main research areas are market liquidity, central counterparties, capital structure and risk management.

IDENTIFYING REGIONAL MODELS OF ACTIVE GRANDPARENTING IN RUSSIA BASED ON CLUSTER ANALYSIS

Oksana Shubat
Irina Shmarova
Ural Federal University
620002, Ekaterinburg, Russia
Email: o.m.shubat@urfu.ru
Email: i.v.shmarova@urfu.ru

KEYWORDS

Cluster analysis, modelling, grandparenting, grandmothers, Russian regions.

ABSTRACT

One of the important social roles the elderly perform is that of grandparents. Our study aims to identify groups of Russian regions with similar models of grandparental activity. The research focuses only on grandmothers. To determine these models, we applied the hierarchical cluster analysis. We used indicators that characterize potential (based on the age criterion) and active (based on intensive involvement in caring for grandchildren) grandparenting in Russian regions. In the process of clustering, we use the growth rates of active grandmothers in the total number of potential grandmothers in 2011-2014, 2014-2016, 2016-2018. The analysis based on the Ward method and the Euclidean distance allowed us to identify 4 models of grandparental activity (regarding grandmothers) in Russian regions. The models differ significantly in the specifics of changes in the degree of grandmothers' involvement in caring for their grandchildren. These models provide the framework for developing specific demographic policy measures with the regional heterogeneity in mind.

INTRODUCTION

In recent decades, population ageing has become one of the most crucial global trends, which is now becoming more evident in many countries. The share of the elderly is steadily increasing in Russia, too. For example, over the past twenty years, the share of those aged 65 and above has grown from 12.4% to 15.8% (Demographic indicators 2022). The consistent trend results in the more intense economic, social, political, and demographic involvement of the elderly population in the social life.

One of the important social roles the elderly perform is that of grandparents while taking part in childcare is becoming increasingly popular among the older generation. Their activity – actual or potential – generates interest in new research. The last decade has seen a rapid development of the grandparenthood demography (Arpino et al. 2018; Margolis and Arpino

2019). An increasing number of grandparenthood demography studies found, inter alia, some positive psychological, social, demographic, and economic effects from grandparents' involvement in their grandchildren's lives (Arpino and Bordone 2014; Mahne and Huxhold 2015; Hilbrand et al 2017; Shubat and Shubat 2021).

One of the positive demographic effects is the influence of proactive grandparenting on increasing the birth rate, which is particularly important for Russia. In fact, the country witnesses negative demographic trends; the birth rate is falling annually, the natural population decline is growing whereas the total population is decreasing. According to a medium variant of demographic projection (i.e., the most likely one), the population will be declining until 2035 (Demographic indicators 2022; Demographic projections 2021). To address demographic problems, the government has extensively designed and implemented various demographic policy measures; however, according to statistics, they have not produced desired results yet – the birth rate keeps declining. Clearly, it inspires new initiatives, which could enable more effective demographic policy measures. One of these initiatives is to support proactive grandparenting in the country.

Researchers from different countries explore a positive impact of proactive grandparenting on increasing the birth rate. Kaptijn R. et al. (2010) demonstrated that in the Netherlands, childcare support from grandparents increases the probability that parents have additional children in the next 8 to 10 years. The conclusion is based on three-generation longitudinal data anchored in a sample of grandparents aged 55 and over. Similar results were obtained by Thomese and Liebroer (2013), who used a survey of 898 Dutch men and women. In other countries, researchers also found positive implications of grandparenting for birth rates (Chapman S.N. et al. 2021; Hejun Gu et al. 2021; Hank and Kreyenfeld 2003) or the reverse effect, when the absence of grandparents (or their death) results in lower birth rates (Okun and Stecklov 2021).

The grandparents' childcare may be regionally and nationally specific. For example, Buchanan and Rotkirch (2018) showed differences between childcare patterns in the Northern and the Southern Europe. Similarly, Russia may also have its own specificities of

how grandparenting affects the birth rate. However, a critical issue now is to identify these specificities. Although the active involvement of grandparents into taking care of and developing their grandchildren is traditional in Russia, the country lacks national-level studies which could help soundly estimate characteristics of the grandparents' activity. Today, Russian statistics does not collect data on grandparents as people with grandchildren. Making evaluations is possible only through secondary data and indirect recalculations.

One of the peculiarities of grandparenting in Russia may be considered its high regional specificity. Historically, Russian regions differ greatly in a number of social, economic, and demographic indicators. For example, Table 1 presents minimum and maximum regional values of indicators which are frequently used for exploring the socio-economic development of the country or a region (Regions of Russia 2021). The Table proves that these values differ sharply.

Table 1: Regional Differences of Some Socio-Economic Indicators in Russia

Variable	Minimum		Maximum		MMR*
	Value	Region	Value	Region	
Total Fertility Rate	1.061	Leningrad region	2.971	Tyva Republic	2.8
Unemployment Rate, %	2.4	Yamal-Nenets Autonomous Area	29.8	Republic of Ingushetia	12.4
Gross Regional Product, per capita, roubles	145723	Republic of Ingushetia	7530485	Nenets Autonomous Area	51.7

* MMR is Maximum-Minimum Ratio

We argue that the statistical method of the cluster analysis may be appropriate to meet the goal. Interestingly, the cluster analysis is not widely used today for designing differentiated demographic policy measures. Although researchers perform modelling of space and other structures based on the cluster analysis quite often, demographic studies use this method of modelling infrequently; it is supported by our analysis of papers indexed in the Web of Science global citation database. We chose those publications which used "cluster analysis" as a keyword and compared the number of such publications in different research areas from the Web of Science classification. Results of the

analysis are presented in Table 2. They reveal that only 26 publications out of nearly 18 thousand which used cluster analysis deal with demography.

Table 2: Number of Publications with "Cluster Analysis" as a Keyword (indexed in Web of Science as of 7 February 2022)

Web of Science Categories	Number of Papers
Demography	26
Sociology	111
Economics	787
Total	17,972

Due to positive implications from the active involvement of grandparents into taking care of grandchildren, which were found by researchers from different countries, and the uneven demographic and socio-economic development of Russia, exploring regional models of active grandparenting in the country and tendencies for their changes is of great relevance. Thus, our study aims to identify groups of Russian regions with similar models of grandparental activity. We hypothesise that the degree of grandmothers' and grandfathers' involvement in caring for grandchildren differs and thus argue that these groups should be studied separately; this research focuses only on grandmothers.

DATA AND METHODS

From the methodological point of view, it is difficult to identify the socio-demographic group of grandparents. As was mentioned, Russia does not collect statistical data on the number of grandparents. We can estimate it only through secondary data – based on the age criterion. To that end, we have to define the age when men and women in Russia become grandparents. In our previous research, we presented the methodology for it (Shubat and Bagirova 2020, Shubat and Shubat 2021); to define the age of becoming grandmothers, the age when a woman gives birth to her first child is added for two adjacent generations of mothers.

On defining the age of becoming grandparents, we can evaluate the number of grandparents based on the age criterion; however, this method allows us to evaluate only the number of potential grandmothers. Not all of them will be active grandmothers, that is actively involved in taking care of and developing their grandchildren. To determine them, we used "Comprehensive monitoring of the living conditions of the population" – a survey by the Russian Federal State Statistics Service held biannually (Comprehensive monitoring 2018). At the time of the study, data were available for 2011, 2014, 2016, and 2018. Some questions in this survey give a preliminary idea of how active grandmothers are in taking care of grandchildren. In particular, the following question may be a good identifier: "Is taking care of children – yours or someone else's – a part of your daily routine (without

being paid for it)?" A positive answer to this question allowed us to categorise a potential (by age) grandparent as an active one.

Based on these data, we calculated three variables for each region of Russia and used them throughout the research:

- Var 1: the growth rate of active grandmothers in the total number of potential grandmothers in 2011-2014;
- Var 2: the growth rate of active grandmothers in the total number of potential grandmothers in 2014-2016;
- Var 3: the growth rate of active grandmothers in the total number of potential grandmothers in 2016-2018.

To determine regional models of active grandparenting, we applied the hierarchical cluster analysis. We tested different distance measures and various ways to group regions in clusters. We chose those measures which showed the best differentiation potential and allowed dividing regions into clusters as clearly as possibly. To decide on the number of clusters, we used a dendrogram and coefficients of the agglomeration schedule. Clusters were profiled through analysing cluster centroids. For each group of regions, we calculated mean and median values of clustering variables. Medians are known to be a non-parametric measure of central tendency insensitive to outliers in the population analysed – as opposed to means. They can be used when analysing the mean value is inadequate due to outliers or high variation of data.

RESULTS

We obtained the following results.

1. Based on the methodology for assessing Russian women's age of becoming grandmothers, we concluded that, on average, a woman in Russia becomes a grandmother at the age of 47-48 years (Table 3). These numbers are growing due to the increasing age of first births – a tendency not only in Russia but also in many other developed and developing countries.

Table 3: Russian Women's age of Becoming Grandparents

Year	2011	2014	2016	2018
Age	47.86	48.03	48.28	48.51

We used these data to form groups of potential grandmothers for each year. Then we selected active grandmothers – those who participate in taking care of and developing grandchildren daily. Further, we calculated variables characterising the dynamic of the number of active grandmothers in Russian regions.

2. The analysis showed that the variables characterising the dynamic of the number of active grandmothers differ significantly in Russian regions – the minimax ratio of their values varies from 7.8 to 16.8 times, as shown in Table 4.

Table 4: Minimum and Maximum Values of Clustering Variables

	Var 1 2014/2011	Var 2 2016/2014	Var 3 2018/2016
Minimum	0.262	0.000	0.213
Maximum	4.423	2.032	1.661
Maximum- Minimum Ratio	16.882	-	7.798

Such heterogeneity is a precondition for performing the cluster analysis that may identify typical region groups, that is, regional models of grandparental activity.

3. The hierarchical cluster analysis based on the Ward method and the Euclidean distance allowed us to identify 4 clusters of Russian regions that differ significantly in the levels of the indicators studied. Thus, we determined 4 models of grandparental activity (regarding grandmothers) in Russian regions.

The clustering dendrogram is shown in Figure 1; the mean and median values of the variables studied are presented in Table 5. We applied the cluster analysis based on these distance measures because it provides the clearest division of all regions into clusters.

Table 5: Mean and Median Values of Variables in Clusters

Model (cluster)	Statistics	Var 1 2014/2011	Var 2 2016/2014	Var 3 2018/2016
1	Mean	0.76	1.21	0.75
	Median	0.76	1.16	0.76
2	Mean	0.98	0.81	1.01
	Median	1.06	0.82	0.91
3	Mean	1.95	1.12	0.95
	Median	1.91	1.11	0.98
4	Mean	3.45	0.89	0.57
	Median	3.23	0.79	0.58

The study of cluster centroids based on median values revealed a non-linear dynamic of the number of active grandmothers in Russian regions. We did not observe a continuous growth in any model of grandparental activity. Crucially, all models have shown a decrease in the number of active grandmothers in recent years – from 9% in Model 2 to 42% in Model 4.

At the same time, there are obvious differences in the models identified. For example, in Model 1, the number of active grandmothers changed non-linearly – it decreased in the beginning, then increased, and decreased again; this is the only regional model where in 2011–2014 the number of active grandmothers decreased.

Model 4 is characterised by the highest volatility. With the largest increase in the number of active grandmothers in 2011–2014 (by more than 220%), there was also the greatest decrease in subsequent periods – by 21% and 42%.

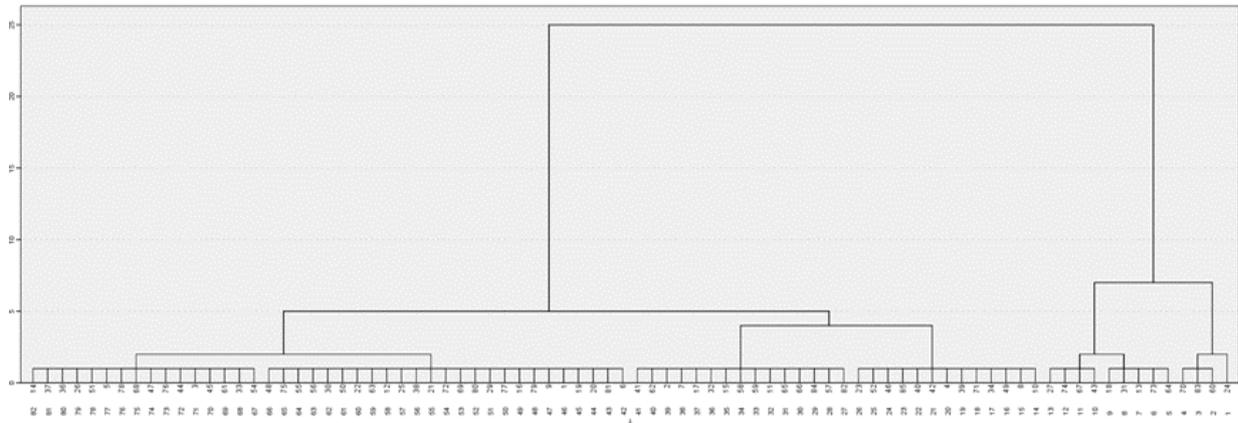


Figure 1: Clustering Dendrogram

Model 3 is the most stable and unproblematic one. It is characterised by an increase in the number of active grandmothers in 2011–2016 and the smallest decrease among other models in 2016–2018.

As an additional variable for profiling the clusters found, we can consider the growth rate of active grandmothers in the total number of potential grandmothers in 2011–2018 – that is, how it changed relative to the base year. We argue that using this variable for clustering is irrelevant because it undermines our initial idea – to identify models of the Russian grandparents’ activity according to dynamics of specific time periods. Considering changes in the more long-term period will clearly conceal important specificities of changes, which occurred within the period. At the same time, the analysis of this variable can definitely reveal crucial characteristics of active grandparenting models. Table 6 shows mean and median values of the additional variable.

Table 6: The Growth Rate of Active Grandmothers in the Total Number of Potential Grandmothers in 2011–2018

Model (cluster)	Statistics	Value
1	Mean	0.69
	Median	0.67
2	Mean	0.80
	Median	0.79
3	Mean	2.07
	Median	2.08
4	Mean	1.75
	Median	1.48

To some extent, additional profiling proved our results. Model 3 indeed turned out to be more positive. The share of active grandmothers here more than doubled in 2011–2018. The total growth in the share of active grandmothers in Model 4, though, should not be regarded as positive because it results from a dramatic increase in 2011–2014, which did not subside even though the share of active grandmothers markedly decreased in the following years. Between 2011 and

2018, Models 2 and 3 saw a decline in the number of active grandmothers. However, the dynamic of this indicator within the period studied differed in two models, which allows us to conclude that these are two different models of the Russian grandmothers’ activity.

DISCUSSIONS

Our results raise some questions, and we see two courses of discussion – how these results can be used to improve the demographic policy and whether it is even possible to use them.

First of all, in 2016–2018, the number of active grandparents decreased in all models identified. At the same time, since 2016, the total fertility rate in Russia has also begun to decline. To confirm the influence of grandparenting on the birth rate in Russia, specialised studies are needed, but our results allow us to hypothesise this influence. A number of studies which showed a positive influence of active grandparenting on the birth rate increase in different countries also prove the adequacy of our hypothesis. In particular, these positive effects were revealed by Kaptijn R. et al (2010), Thomese and Liefbroer (2013), Chapman S.N. et al (2021), Hejun Gu et al (2021). Analogous positive effects in Russia should be explored in a dedicated study, which cannot be pursued at present due to the lack of necessary data and statistics.

Models of active grandparenting provide the framework for developing specific state measures with the regional heterogeneity in mind, which would be aimed at supporting and stimulating fertility considering possible positive implications from active grandparenting. The potential of active grandparenting needs governmental support; it can be provided, for example, in the form of compensation payments to grandparents who spend time with their grandchildren instead of their parents. It is also necessary to support active grandparenting in the media. Earlier, we showed some ways to support and stimulate the potential of active grandparenting (Shubat & Bagirova, 2020). We argue that the most active measures are demanded in Model 4 – the most dramatic one.

Further research is required to identify reasons for the specific dynamic of the number of active grandparents in different regional models. Our research may be considered pilot, but at the same time it inspires further and deeper research in this area. Besides, it is relevant to examine regional characteristics of the grandfathers' activity. Current trends for feminisation and active fatherhood may result in fundamentally different patterns of the grandparental activity.

We argue that it will be of importance to explore the duration of grandparenting and its effectiveness in further studies. It is critical that Russia has recently experienced an annual increase in the life expectancy; in 2011-2020, it grew by more than 4% and, in 2020, equaled to 72.91 years (Life expectancy, 2022). At the same time, our studies show that the age of Russian women's entering grandparenting accounts for 47-48 years; the median age of active grandmothers in 2018 was 60 years, whereas 90% of these grandmothers were up to 71 years inclusive. Therefore, grandparenting effectiveness will be mostly affected by an active life position and grandparents' willingness to take part in raising grandchildren.

The second course of discussion concerning our results is related to the possibility of using the cluster analysis to enhance demographic policy measures.

The cluster analysis is a method for modelling the structure of data based on the so-called person-oriented approach to data analysis (Bergman and Magnusson 1997). It is the opposite of a variable-oriented approach, which is much more often used in social, demographic studies and is oriented at studying the interaction between variables using linear statistical models. The person-oriented approach considers all the variables simultaneously as a complex system that cannot be divided and which functions and develops as a totality (Bergman and Trost 2006). The approach is aimed at studying intra-individual dynamics and variation, which is clearly its methodological advantage compared to the variable-oriented approach.

However, we should be prudent in considering the use of the cluster analysis for demographic research and the development of differentiated state measures of the demographic policy. One of the problems of the cluster analysis is that different clustering methods can give different results, which can result both in a different number of clusters and in a different composition of clusters. Aldenderfer and Blashfield (1984, 1988) presented a broad overview of using various clustering methods and described results of testing various metrics both on real data and on data modelled by the Monte Carlo method. The results of empirical studies showed some advantages of the Ward method.

In our study, this method also gave better results and segmented Russian regions into homogeneous groups more clearly. But we argue that the development of effective state demographic policy measures, differentiated according to regional peculiarities, requires further research and further modelling of the Russian demographic space performed both on real data

and the Monte Carlo simulation. It is necessary to compare clustering results and to search for reasons behind the differences identified in our research.

CONCLUSIONS

Our study yielded the following conclusions.

1. When developing demographic policy measures, it is important to consider the heterogeneity of Russian regions. The dynamic of demographic indicators in the regions is not similar and often multidirectional; therefore, designing universal state support measures for stimulating fertility equally effective in all regions cannot be possible.

2. The cluster analysis may be an effective analytical tool for identifying regional models in demography. Its results may help suggest regionally specific demographic policy measures. However, it should be first tested with different methods of clustering in a number of studies.

3. The grandparents' activity may be viewed as a resource for addressing demographic problems in Russia. The models of active grandparenting we identified allow determining regions with the most problematic situation, which require the most urgent support from the government. These models may be also used for more profound research in the field. Our study raises another issue for Russian authorities – a need to collect national-level statistical data on grandparenting. Without these data, research conducted cannot build a solid foundation for enhancing the demographic policy.

Future research on the topic can concentrate on additional studies intended to develop an information pool for finding cause-effect relationships between the birth rate and an active contribution of grandparents in Russia. Important aspects for future studies also include exploring the grandparenting duration in Russia for grandmothers and grandfathers and modelling the Russian demographic space based on both actual data and Monte Carlo simulation.

ACKNOWLEDGMENTS

The reported study was funded by RFBR, project number 20-011-00280.

REFERENCES

- Aldenderfer, M. S., Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage Publications.
- Arpino B., Guma J., Julia A. (2018). Family histories and the demography of grandparenthood, *Demographic research*, No. 39, pp. 1105-115 .
- Arpino, B., Bordone, V. (2014). Does Grandparenting Pay Off? The Effect of Child Care on Grandparents' Cognitive Functioning. *Journal of Marriage and Family*, 76(2), 337-351. doi: <https://doi.org/10.1111/jomf.12096>
- Bergman, L. R., Magnusson, D. (1997). A person-oriented approach in research on developmental

- psychopathology. *Development & Psychopathology*, 9, 291-319.
- Bergman, L. R., Trost, K. (2006). The person oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds?, *Merrill-Palmer Quarterly*, 3, 601–632.
- Blashfield, R. K., Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselrode, R. B. Cattell (Eds.). *International handbook of multivariate experimental psychology*. New York: Plenum Press.
- Buchanan A., Rotkirch A. (2018). Twenty-first century grandparents: global perspectives on changing roles and consequences. *Contemporary Social Science*, 13(2), 131-144.
- Chapman S.N., Lahdenperä M., Pettay J.E., Lynch R.F., Lummaa V. (2021) Offspring fertility and grandchild survival enhanced by maternal grandmothers in a pre-industrial human society. *Scientific Reports*, Feb 11(1), 36-52. doi: 10.1038/s41598-021-83353-3.
- Comprehensive monitoring of the living conditions of the population, conducted by the Federal State Statistics Service in 2018. Retrieved from https://www.gks.ru/free_doc/new_site/KOUZ18/index.html (access date 15.12.2021)
- Demographic indicators of the Federal State Statistics Service. Retrieved from <https://rosstat.gov.ru/folder/12781> (access date 01.02.2022).
- Demographic projections of the Federal State Statistics Service. Changes in population size by variants of projections (2021-2035). Retrieved from <https://rosstat.gov.ru/storage/mediabank/progn1.xls> (access date 13.01.2022).
- Hank, K., Kreyenfeld, M. (2003). A Multilevel Analysis of Child Care and Women's Fertility Decisions in Western Germany. *Journal of Marriage and Family*, 65(3), 584-596. <https://doi.org/10.1111/j.1741-3737.2003.00584.x>
- Hejun, Gu., Fengqin, Bian., Ehsan, Elahi. (2021). Impact of availability of grandparents' care on birth in working women: An empirical analysis based on data of Chinese dynamic labour force. *Children and Youth Services Review*, 121, (105859). <https://doi.org/10.1016/j.childyouth.2020.105859>
- Hilbrand, S., Coall, D. A., Gerstorf, D., & Hertwig, R. (2017). Caregiving within and beyond the family is associated with lower mortality for the caregiver: A prospective study. *Evolution and Human Behavior*, 38(3), 397–403. doi: <https://doi.org/10.1016/j.evolhumbehav.2016.11.010>
- Kaptijn R, Thomese F, van Tilburg TG, Liefbroer AC. How Grandparents Matter: Support for the Cooperative Breeding Hypothesis in a Contemporary Dutch Population. *Human Nature* 2010 Dec;21(4):393-405. doi: 10.1007/s12110-010-9098-9.
- Life expectancy at birth (2022). Federal State Statistics Service. Retrieved from <https://www.fedstat.ru/indicator/31293> (access date 29.03.2022).
- Mahne, K. and Huxhold, O. (2015). Grandparenthood and subjective well-being: Moderating effects of educational level. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 70(5), 782-792. <https://doi.org/10.1093/geronb/gbu147>
- Margolis R., Arpino B. (2019). The demography of grandparenthood in 16 European countries and two North American countries / Timonen V. (ed.) *Grandparenting practices around the world: Reshaping family*. Bristol: Bristol University Press, Policy Press. Pp. 23-41.
- Okun, B.S., Stecklov, G. (2021). The Impact of Grandparental Death on the Fertility of Adult Children. *Demography*, 58(3), 847-870. doi: <https://doi.org/10.1215/00703370-9015536>
- Regions of Russia. Social and Economic Indicators 2021. Statistical Book. Rosstat, Moscow. URL: <https://rosstat.gov.ru/folder/210/document/13204> (access date 15.01.2022).
- Shubat O., Bagirova A. (2020). Russian Grandparenting: Demographic and Statistical Modelling Experience, *Communications of the ECMS*, no. 34(1), pp. 78-83. DOI: <http://doi.org/10.7148/2020>
- Shubat, O., Shubat, M. (2021). Demographic and statistical modelling of grandfatherhood in Russia. *Communications of the ECMS*, 35(1), 57-62.
- Thomese, F., Liefbroer, A. (2013). Child Care and Child Births: The Role of Grandparents in the Netherlands. *Journal of Marriage and Family*, 75(2), 403-421. doi: <https://doi.org/10.1111/jomf.12005>

AUTHOR BIOGRAPHIES

OXSANA SHUBAT is an Associate Professor of Economics at Ural Federal University (Russia). She received her PhD in Accounting and Statistics in 2009. Her research interests include demographic processes, demographic dynamics and their 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 webpage can be found at <http://urfu.ru/ru/about/personal-pages/O.M.Shubat/>

IRINA SHMAROVA is an Associate Professor of Economics at Ural Federal University (Russia). She received her PhD in Economics in 2020. Her research interests include demographic processes and their determinants, development of human capital of children and spending on population reproduction. Her email address is i.v.shmarova@urfu.ru and her webpage can be found at <https://urfu.ru/ru/about/personal-pages/personal/person/i.v.shmarova/>

COMPARISON OF SEPARATED FAMILIES' STANDARD OF LIVING IN GERMANY

Analyzing the Equalised Incomes in Simulated Families after Child Support and Child Benefit Paid

Erzsébet Teréz Varga, PhD (erzsebet.varga@uni-corvinus.hu)
Department of Banking and Monetary Finance

Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest, Hungary

KEYWORDS

Child support, risk of poverty, one-parent household, equivalence scales, child penalty, inequality

ABSTRACT

In this paper, I describe the inequality in the standard of living in Germany after divorce and compare their risk of poverty. The one-parent families have the highest poverty risk everywhere in the world. In Germany, a directive is available for anybody to determine the child support geared to the non-custodial parent's disposable income. Assuming that the non-custodial parent pays child support following this directive of *düsseldorfer* tables I found deep differences in the equalised incomes of the divorced households in simulated cases. Equalised incomes were determined by two types of the OECD scales to make comparable the different composed families' incomes. Both methods result in fewer life standards for one-parent households in more than 83 % of the cases, however, the risk of poverty is not higher for the custodial parent's household. This indicates some modification in the directive: the respect of the custodial parent's income and/or correction of the amounts in the tables mainly on the higher income categories.

INTRODUCTION

The risk of poverty in lone parent households is the highest in the world as well as in the European Union. First of all, the market is responsible for this phenomenon, so the states usually help these families by subventions in the form of financial or institutional assets. But what can the non-custodial parents do for their children to avoid this risk? There are judicial recommendations for child support payment in some countries, e.g. in Germany the so-called '*Düsseldorfer Tabelle*'. This paper investigates if this guideline is enough to prevent social sinking or only for conscience's sake? Could the non-custodial parent pay more without endangering her/his own well-being more than it needs? Simulating 1000 families regarding the children's effect on the wages and the state subventions, I calculate the difference in the standard of living of the two new households after their divorce. I supposed that the non-custodial parent pays child support according to the official guideline. I investigate if the recommendation is enough or if any modification is needed

to make a fair system. Two (or more) different composed households' disposable incomes are comparable by OECD equivalence scales. Two alternatives are popular nowadays: the modified and the square-root methods. I wondered also if they give the same or different results, which is better for economic analysis. I choose Germany because there exists a guideline that is available for anybody so it can be a standard for other countries in the European region (especially for similar social market economies). It is the most populated country in the EU meanwhile one of the most developed although the one-parent families' risk of poverty here is higher than the average in both the EU and Euro area. Henceforward there are many studies about poverty among households regarding their composition but none of these papers analyses the effect of the child support system. Obviously many aspects are responsible for the well-being of a household from the market to the state but it is time to see the role of a child support guideline.

BACKGROUND

Risk of Poverty

The Measurement Method

The at-risk-of-poverty rate is the share of people whose equivalised disposable income (after social transfer) is below the at-risk-of-poverty threshold. This poverty line is set at 60 % of the national median equivalised disposable income (after social transfers) in the Eurostat statistics (Eurostat. Glossary 2022) but it can also be addressed at 50 % of the median value. In the samples of Hansen et al. (2005) level of the threshold does not disturb the results, I have also seen it in my calculation as well.

Determining and comparing the standard of living in different households is a reasonably hard task. The necessities of a younger or older adult, either a baby or a pupil diverged. Near to this two adults of the same age usually don't need twice the disposable income as one adult to achieve the same well-being due to the fixed costs of a household, the economies of scale in consumption.

Consequently, the income data of households determined per capita can be elusive. To avoid this problem OECD worked out many methods in the past decades. Nowadays both the Eurostat and the OECD statistics supply the concept of equivalised income.

“The equivalised disposable income is the total income of a household, after tax and other deductions, that is available for spending or saving, divided by the number of household members converted into equalised adults.” (Eurostat, 2022a)

Eurostat calculates the equivalised adults as the size of the household according to the so-called “OECD-modified” (or OECD2) equivalence scales. In this method, the weight of the first adult is 1, and any other member of the family aged over 14 years is 0.5, meanwhile, a child (aged under 14) weighs 0.3. Practically it means that: if one adult needs 100 units then one adult plus one child has 130 units for the same standard of living. It comes from that the equivalised adults in the second household are $1+0.3 = 1.3$ (one adult plus one child), so their income has to be divided by this amount: $130/1.3 = 100$.

Nowadays OECD publications apply a later suggested method, the so-called square-root scale. According to it the age of any member of the household doesn’t matter only the size of the household. The total income of a household is divided by the square root of the number of members. (E.g. in the case of four members in the household the income is divided by two: $\sqrt{4}=2$). Practically it means that if one adult has 100 units then one adult plus one child need 141 units for the same standard of living. It comes from that the equivalised adults in the second household are $\sqrt{2} \approx 1.41$ (two persons disregarding their age), so their income has to be divided by this amount: $141/1.41 = 100$. (OECD, 2022)

Table 1 compares the two methods if there is only 1 adult in the household and 0 to 4 children. Hansen et al. (2005) used the modified OECD scale I calculated both methods to compare the results.

Table 1: Equalised adults in one-adult households (own calculation based on OECD, 2022)

Household size	„OECD-modified” scale	Square root scale
1 adult	1	1
1 adult, 1 child	1.3	1.4
1 adult, 2 children	1.6	1.7
1 adult, 3 children	1.9	2.0
1 adult, 4 children	2.2	2.2

Statistics

In the European Union and the Euro area, every fifth citizen lives at risk of poverty or social exclusion (see Total row in Table 2). Seeing the details according to the composition of the households the most endangered type is where one adult lives with one or more children. (Unfortunately, there is no detailed data for one, two, three, or more children.) Their rate is the highest, more than 40 % of the people who live in one-parent families lie under this risk. In Germany, which is the most populated country in the EU, the same rate is closer to 50 %! In the next

subsection, I present the phenomenon of child penalty which is mainly responsible for it.

Table 2: People at risk of poverty or social exclusion by different household types in 2020 (Eurostat, 2022b) (%)

Household composed of ...	European Union - 27 countries	Euro area - 19 countries	Germany
... 1 adult	33.2	32.2	34.8
... 1 adult with dependent children	42.1	43	46.7
... 2 adults	16.4	16.2	16.3
... 2 adults with 1 dependent child	15.7	16.6	15.8
... 2 adults with 2 dependent children	16.7	17	17.6
... 2 adults with 3 or more dependent children	29.6	29.8	30.9
... 3 or more adults	18.1	18.8	15.4
... 3 or more adults with dependent children	25.2	25.7	19.7
Total	21.9	22	22.5

As can be seen from Table 1, the household composed of 1 adult with dependent children has the highest chance to live at risk of poverty or social exclusion both in the EU and the Euro area (not only in 2020). In 2020 that data was worse in Germany by 4.6 (3.7) percentage points than in the EU (Euro area) on average. Almost every second person who lives in a one-parent family was at risk of poverty. Almost half of these people have less equalised income than half of the median income. They are in the poorest quartile of society. At the same time, a household with only one adult without children has the second-highest chance to live at risk of poverty or social exclusion. Practically third of these people have less equalised income than half of the median income. Divorce prognosticates both the exes-partner and their child(ren) increasing risk of poverty if they do not find a new partner.

Impact of having a child

Child Penalty: the Market

The ‘motherhood pay gap’ or in more general form the ‘child penalty’ is a well-known and researched social problem. (see Grimshaw and Rubery, 2009) Over the gender gap, the mothers earn less than other women without children. Many studies investigate this phenomenon and try to determine its extent. Gangl – Ziefle (2009) found that the German wage penalty for motherhood was 16%–18% per child which was the highest result among the investigated countries (USA 9-16 %, Britain 13 %) in the 50th and 60th decades. Kleven et al. (2019) determined 61 % as the long-run child penalty in Germany. In their paper, Germany had the highest rate

among the investigated six countries. E.g. in the case of Denmark, it was ‘only’ 21 % for women who birthed their first child between 1985-2003. The penalties can come from three deviations: employment, working hours, and the wage rate.

Child Benefit: the State

Since the phenomenon of child penalty and to avoid children’s high risk of poverty governments usually gave some subvention for parents. In Germany, its extent depends on the number of children as Table 3 shows it.

Table 3: Child benefits in Germany (Bundesregierung, 2022)

Child benefit per capita in euro	
1st and 2nd child	219
3rd child	225
from 4th child	250

Child Support: the Ex-Spouse

The most important financial asset which can moderate (if not solve fulfilled) the risk of poverty is a well-defined (and paid) child support from the non-custodial parent. (Monostori, 2019). In Germany, the recommendation for the amount of child support is contained by the so-called Düsseldorf Tabelle (see Table 3 and Appendix Table 4 and 5). There is also a recommendation for the amount of spousal support in the directive but this paper does not deal with the theme.

Table 3: Child support (euro) payment for the first and second child per capita in 2022

Chargeable net income (euro)		Child support depends on the age of the child (years)			
from	to	0-5	6-11	12-17	from 18
1	1900	286.5	345.5	423.5	350
1901	2300	306.5	368.5	450.5	379
2301	2700	326.5	391.5	477.5	407
2701	3100	346.5	414.5	503.5	436
3101	3500	366.5	436.5	530.5	464
3501	3900	397.5	473.5	573.5	510
3901	4300	429.5	509.5	615.5	555
4301	4700	461.5	546.5	658.5	601
4701	5100	492.5	582.5	701.5	646
5101	5500	524.5	618.5	743.5	692
5501	6200	556.5	655.5	786.5	737
6201	7000	587.5	691.5	829.5	783

7001	8000	619.5	728.5	871.5	828
8001	9500	651.5	764.5	914.5	874
9501	11000	682.5	800.5	956.5	919

According to Table 3 when the non-custodial parent’s chargeable net income before paying child support is e.g. 4500 euros a month he/she should pay 601 euros for a child aged 18 and 658.5 euros for a child aged 13 if there are no more children. If there is a third (younger) child in this family the non-custodial parent has to pay according to a lower income bracket: 555 euros for the child aged 18 and 615.5 euros for the child aged 13. The child support payment for the third (youngest) child is determined by Table 4 (see the Appendix). In the case of four children, the discount is -2, so the non-custodial parent has to pay according to the lower income bracket by 2. From the example above if the non-custodial parent should pay 510 euros for the child aged 18 and 573.5 euros for the child aged 13 if the two siblings are younger. (The first child is always the oldest.)

Near to it, there is a maximum level of child support depending on the net chargeable income. If this couldn’t cover the sum of the calculated support and the minimum living expenses (in 2022: 1.160 euros), he/she only has to pay the difference between the chargeable net income and the minimum living expenses. In the case when this difference is negative no payment is obligated.

METHODS

Dataset by simulation

Size of the families

In the beginning, 1000 sample families were simulated. They could have 0, 1, 2, 3 or 4 children altogether but at the most only one in every different age bracket (0-5, 6-11, 12-17 and from 18) which is signed in Table 3. It means that the numbers of the child in each category are independent and identically distributed random variables and for every bracket, it could be 0 or 1. The adult (age 18+) child was supposed to be a pupil in the analysis (otherwise the amount of child support would diverge from Tables 3, 4 and 5). Among the simulated families, there were 0 children in 62 cases in each category. Those data were eliminated so the cleaned sample had 938 cases.

The further calculation supposed that every family has one custodial and one non-custodial parent, all children in the family live with the custodial parent and the non-custodial parent lives alone. A new marriage or new children would modify both the disposable income and the number of equalised adults in their households. Accordingly, their quotient, the equalised income, could be higher and less as well so the analysis disregarded these opportunities and assumed one adult and zero children in the non-custodial parent’s household.

Disposable income of the households

The non-custodial parents:

1. step: net income from employment was set as independent and identically distributed random variables between 960 and 11.000 euros. The aim was to test the effect of the whole range of the German guideline.
2. step: for determining the child support the chargeable net income came from the net income by reducing it by 5 %. (This is the countable cost related to work which is deductible according to the guideline.) The child support was determined from the chargeable net income according to Tables 3, 4 and 5 depending on the number of children suggested by the reviewed guideline, regarding the mentioned maximum level of child support.
3. step: the non-custodial parent's net income from employment (see step 1) decreased by the child support (see step 2) to determine the disposable income.

The custodial parents:

1. step: net income from employment was set as independent and identically distributed random variables between 960 and 11.000 euros as in the case of the non-custodial parent
2. step: the data set was modified by the so-called child penalty. I chose the result of Kleven et al. (2019) since it is recent and gender-neutral so I discounted the random variable income data from step 1 by 61 %.
3. step: the custodial parent's income from step 2 is increased by the child support from non-custodial's step 2 and the child benefit according to Table 3.

Methodological strategy

Since the non-custodial parents were assumed to live alone, their equalised income is simply their disposable income. In the case of custodial parents' households, the equalised income was determined by both the OECD-modified (OECD2) and Square root scales. The latter was simpler because it doesn't deal with the age of a person. In the case of the OECD-modified scale, the age categories in the düsseldorf tabelle did not match it as a whole. In the first two categories (age 0-5, 6-11) the weight of a child is 0.3, and the fourth category (from 18 years) gets the weight of 0.5. But in the third bracket (age 12-17) 0.4 was chosen since the method gives 0.3 under age 14 and 0.5 above it.

According to Table 1, I supposed significant difference in the results along with the two scales but there were not demonstrable only some differences as you can see in the next section.

RESULTS

In 141 cases (15 %) both of the indicators signed the non-custodial had less equalised income than in the single parent's household. In 17 cases (1.8%) the OECD2 and

the square root method gave a different signs. In the remaining 780 (83.2 %) cases according to both of the methods, the lone parent's household has less equalised income than the non-custodial parent. Henceforth the extent of the difference is measured by the quotient of the equalised incomes of the households.

'Negative' cases

First I investigated those 158 'negative' cases when at least one method give a worse result (less equalised income) for the non-custodial parent. The average number of children in this sample was 2.37 meanwhile in the original sample it was 2.15. Not surprising significantly less.

In 116 cases the custodial parent's net income from employment was higher than the non-custodial's one. On average the custodial parent had a higher original income of 829 euros. Consequently, the higher life standard comes from this higher salary mainly not from the high child support. In detail: in 21 cases the child support was 0, in 57 cases was less than would come from the Tabelle and paid only amount above the 1160 limit.

Among the 158 cases in the most extreme situation, the custodial parent's household has 2.8 times higher equalised income than the non-custodial.

The differences can be seen in Figure 1.

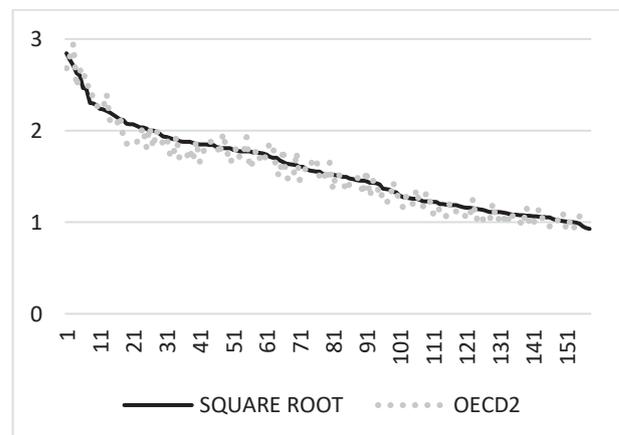


Figure 1: The quotient of the equalised income of the custodial and non-custodial parent's household in the cases where at least one scale signed less equalised income for the non-custodial parent (own calculation)

'Positive' cases

In the 780 'positive' cases (83.2 %) both of the indicators signed the non-custodial has more equalised income than in the single parent's household. At the highest difference, the custodial parent has 8.8 times higher equalised income than his/her child(ren). Figure 2 shows the quotients of equalised income for these 780 cases.

The average number of children in this sample was 2.11. Only in 6 cases, the custodial parent's original income was higher than the non-custodial's one. On average the custodial parent has less net income from employment by 4.453 euros.

In 12 cases the child support was 0 even though the life standard was higher for the custodial parent. In 19 cases the support was less than would come from the Tabelle and paid only the amount above the 1160 limit. In 211 cases the non-custodial parent's equalised income was more than 2.8 times higher than his/her children. (2.8 was the highest value in the 'negative' cases.)

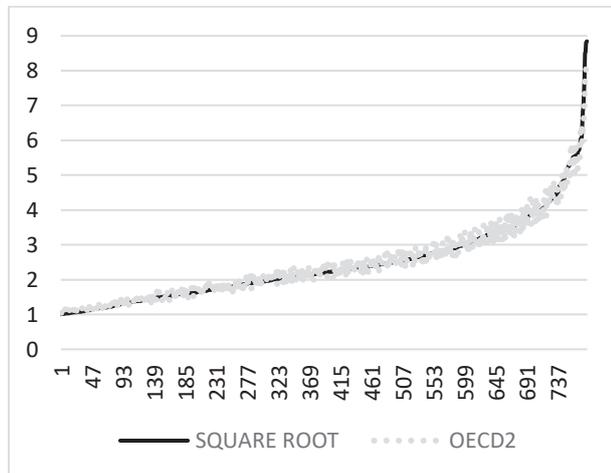


Figure 2: The quotient of the equalised income of the non-custodial and custodial parent's household in those 780 cases where both scales signed higher equalised income for the non-custodial parent (own calculation)

The whole sample

For the whole sample Figure 3 shows the differences between non-custodial and custodial parent's households. The average difference in equalised income (the broken line in the figure) is more than twice (2.2). The line at 1 unit shows the equal case when both households had the same equalised income.

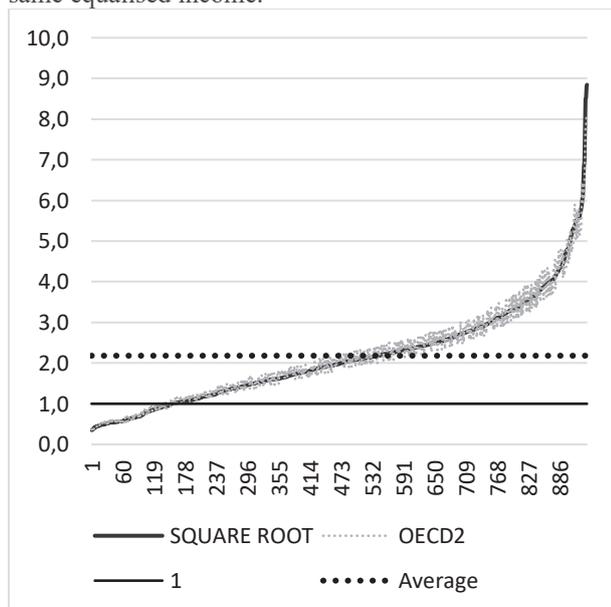


Figure 3: The quotient of the equalised income of the non-custodial and custodial parent's household (own calculation)

The risk of poverty in the sample for the households at different levels of median income and on different OECD scales are summarized in Table 4. It gave us a less extreme situation. In the sample, the risk of poverty is not too far in the two types of households.

Table 4: Households at risk of poverty or social exclusion in the sample (own calculation)

	OECD 2		Square root	
	50 % median	60 % median	50 % median	60 % median
Custodial household	5.9%	11.5%	5.5%	11.0%
Non-custodial household	6.2%	7.8%	6.3%	7.9%

DISCUSSION

The paper investigated the German child support guideline considering the child benefit system as well. It used the OECD scales to determine the equalised income to compare the standard of living in separated households. Regarding 938 cases with a different number of children, it found an unfair distribution of income after divorce.

In general, if we see the differences between the living standards the sum of the child support should become at least twice the current values. As we saw in the details there are some cases where it is not necessary but in other cases, the amount could be either 8 times higher.

At the same time, if we see only the risk of poverty the German guideline is enough.

However, the construction of a fair system needs further research, this paper wished to draw attention to the current unfair system.

REFERENCES

- Bundesministerium für Arbeit und Soziales (2020): Mindestlohn steigt auf 10.45 Euro im Jahr 2022 <https://www.bmas.de/DE/Service/Presse/Pressemitteilung/n/2020/mindestlohn-ab-2022-erhoeht.html>
- Bundesregierung (2022): Child benefit to rise. <https://www.bundesregierung.de/breg-en/news/kindergeld-steigt-1772690>
- Düsseldorfer Tabelle (2022): https://www.olg-duesseldorf.nrw.de/infos/Duesseldorfer_Tabelle/Tabelle-2022/Duesseldorfer-Tabelle-2022.pdf
- Eurostat (2022a): *Glossary*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:At-risk-of-poverty_rate
- Eurostat (2022b): People at risk of poverty or social exclusion by income quintile and household type - new definition https://ec.europa.eu/eurostat/databrowser/view/ILC_PEPS_03N_custom_1451128/bookmark/table?lang=en&bookmarkId=dd86fb07-acc6-4813-87e7-d0b8df824b33&page=time:2020

Gangl, M. – A. Ziefle (2009). Motherhood, Labor Force Behavior, and Women’s Careers: An Empirical Assessment of the Wage Penalty for Motherhood in Britain, Germany, and the United States. *Demography* 46 (2), pp. 341-369.

Grimshaw, D. and J. Rubery (2015). *The Motherhood Pay Gap: A Review of the Issue, Theory and International Evidence*, Volume 57. Geneva: International Labour Office, Inclusive Labour Markets, Labour Relations and Working Conditions Branch Conditions of Work and Employment Series.

Hansen, H.-T. – O. Jürgens – A. H. H. Stand – W. Voges (2006): Poverty among households with children: a comparative study of Norway and Germany. *International Journal of Social Welfare*. (15) pp. 269-279.
<https://doi.org/10.1111/j.1468-2397.2006.00469.x>

Kleven, H. – C. Landais – J. Posch – A. Steinhauer – J. Zweimüller (2019): Child Penalties across Countries: Evidence and Explanations. *AEA Papers and Proceedings* Vol. 109, May 2019 pp. 122-12.
<https://pubs.aeaweb.org/doi/pdfplus/10.1257/pandp.20191078>

Monostori J. (2019): Egyszülős családok és politikák Magyarországon és Európában. *Demográfia* 66 (1) pp 5-41.

OECD (2022): What are equivalence scales?
<https://www.oecd.org/economy/growth/OECD-Note-EquivalenceScales.pdf>

AUTHOR BIOGRAPHIES

Erzsébet Teréz VARGA, PhD is an assistant professor at the Department of Banking and Monetary Finance at the Corvinus University of Budapest. Her main research areas are tax theory and public finance. Her email address is erzsebet.varga@uni-corvinus.hu.

Appendix

Table 4: Child support (euro) for a third child in 2022

Chargeable net income (euro)		Child support depends on the age of the child (years)			
from	to	0-5	6-11	12-17	from 18
1	1900	283.5	342.5	420.5	344
1901	2300	303.5	365.5	447.5	373
2301	2700	323.5	388.5	474.5	401
2701	3100	343.5	411.5	500.5	430
3101	3500	363.5	433.5	527.5	458
3501	3900	394.5	470.5	570.5	504
3901	4300	426.5	506.5	612.5	549
4301	4700	458.5	543.5	655.5	595
4701	5100	489.5	579.5	698.5	640
5101	5500	521.5	615.5	740.5	686
5501	6200	553.5	652.5	783.5	731
6201	7000	584.5	688.5	826.5	777
7001	8000	616.5	725.5	868.5	822
8001	9500	648.5	761.5	911.5	868
9501	11000	679.5	797.5	953.5	913

Table 5: Child support (euro) from fourth child per capita in 2022

Chargeable net income (euro)		Child support depends on the age of the child (years)			
from	to	0-5	6-11	12-17	from 18
1	1900	271	330	408	319
1901	2300	291	353	435	348
2301	2700	311	376	462	376
2701	3100	331	399	488	405
3101	3500	351	421	515	433
3501	3900	382	458	558	479
3901	4300	414	494	600	524
4301	4700	446	531	643	570
4701	5100	477	567	686	615
5101	5500	509	603	728	661
5501	6200	541	640	771	706
6201	7000	572	676	814	752
7001	8000	604	713	856	797
8001	9500	636	749	899	843
9501	11000	679.5	797.5	953.5	913

FORECASTING MODELS FOR FIRST YEAR PREMIUM OF LIFE INSURANCE

Somsri Banditvilai
Choojai Kuharattanachai
Department of Statistics
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
E-mail: somsri.ba@kmitl.ac.th

KEYWORDS

Forecasting, Ordinary Life Insurance, Holt-Winters method, Box-Jenkins method, ANN

ABSTRACT

The objective of this research is to study forecasting models for the first year premium of life insurance. The premium data are gathered from the Office of Insurance Commission (OIC) during January 2003 to November, 2021. The data are divided into 2 sets. The first set from January, 2003 to December 2020 is used for constructing and selection the forecasting models. The second one from January 2021 to November 2021 is used for computing the accuracy of the forecasting model. The forecasting models are chosen by considering the minimum Root Mean Square Error (RMSE). The Mean Absolute Percentage Error (MAPE) is used to measure the accuracy of the model. The results showed that the multiplicative model with initial values from 18 years Decomposition method give the appropriate model for the first year premium of life insurance and yields the MAPE = 17.29%

INTRODUCTION

The Office of Insurance Commission (OIC) of Thailand defines life insurance as it is a way for a group of people to share the dangers of death, dismemberment, disability and loss of income in old age. When any person has to face those dangers and received an average amount of help to alleviate suffering for themselves and their families. The life insurance company will act as the center for bringing such sums to pay to the victims.

First year premium is the amount of an insured person must pay to the life insurance company in the first year to purchase the coverage that they will receive from life insurance.

Life insurance is an essential tool in financial risk services for those who wish to prepare for unforeseen emergencies and retirement that will occur in the future. There are also many other benefits in terms of saving money with discipline and continuity. Most types of life insurance are insurance with long-term commitments. This will allow the insured to use this money for future needs. Both as scholarship money for children or as money saving for old age. It is a guarantee for one's life

and family in helping the insured person feel more secure.

Life insurance therefore plays a direct role in the well-being of the people. In terms of financial planning to create stability for yourself not to be a burden on others and also helps to drive the economy. It is a source of long-term fund raising in the form of life insurance premium that can be invested for profits as government bonds causing money to circulate in the economy.

Forecasting is a powerful tool in helping life insurance companies make predictions the growth trend of life insurance premium in the future for the decision of the executives in driving the business to continue well.

At present, the most popular technique for forecasting is time series analysis. It is a statistical method that uses historical data collected in continuous chronological order to study the relationship patterns of the data by creating equations to guide future forecasts. There are several methods used in forecasting. This research will find a model for forecasting ordinary life insurance premium. This is because it is the most popular type of insurance in the market. which accounted for 79.76% (Thai Life Insurance Association 2019)

Artificial Neural Networks was used for forecasting model of insurance premium revenue. The actual and forecast value was almost the same. Therefore, the results confirmed to be strong and useful to deploy it for forecasting the insurance premium revenue. (Bahia 2013) Box-Jenkins method was employed for modeling the forecasting of life insurance premium and insurance penetration rate and the results were good and degree of accuracy were over 80%. (Namaweje and Geoffrey 2020) Safyar (2010) studied the forecasting life insurance premium by ARIMA and Neural Networks and it was found that Neural Networks is the best model to predict life insurance premium. The Holt-Winters method is suitable for time series forecasts with both linear trends and seasonal influences. It is simple and easy and it is one of the most widely used time series forecasting methods. This has led to studies of different trend defaults in different studies. Changing the trend initiation affects the Holt-Winters forecasting performance. Different initiation configurations provide significantly different MAPE values (Booranawong and Booranawong 2018). Holt-Winters and Extended Additive Holt-Winters (EAHW) method, with 4 different initiation values were used for modeling crude palm oil yield and price

forecasting in Thailand and the results showed that different initiation values in both forecasting methods gave a significant different MAPE values. (Suppalakpanya et al. 2019) Therefore, this research employs the Holt-Winters Exponential Smoothing method with different initiation, Box-Jenkins, and Artificial Neural Networks in forecasting the first year premium of ordinary life insurance of Thailand.

DATA COLLECTION AND METHODOLOGY

The first year premium of ordinary life insurance are gathered from the Office of Insurance Commission (OIC) of Thailand. The monthly data are collected from January, 2003 to November 2021 and it is divided into 2 sets. The first set from January, 2003 to December 2020 is used for constructing the models and employed the minimum Root Mean Square Error (RMSE) for model selection. The second one from January 2021 to November 2021, it is used to compute the accuracy of forecasting models by using the Mean Absolute Percentage Error (MAPE). This research employs three forecasting methods which are the Holt-Winters Exponential Smoothing methods with different initiation, Box-Jenkins method, and Artificial Neural Networks to construct the forecasting models.

Holt-Winters method with different initial values

The Holt-Winters method of exponential smoothing involve linear trend and seasonal variation which are based on three smoothing equations: for level, for trend and for seasonal variation. The decision regarding which model to use depends on time series characteristics: the additive model is used when the seasonal component is constant. The multiplicative model is used when the size of the seasonal component is proportion to trend level (Chatfield 1996)

Additive Holt-Winters Model

If a time series has a linear trend with a fixed growth rate, β_1 , and a fixed seasonal pattern, S_t with constant variation, then the time series may be described by the model.

$$Y_t = (\beta_0 + \beta_1 t) + S_t + \varepsilon_t$$

For this model, the level of the time series at time t-1 is $T_{t-1} = \beta_0 + \beta_1(t-1)$ and at time t is $T_t = \beta_0 + \beta_1 t$. Hence, the growth rate in the level from one time period to the next is β_1 . Y_t is the observed data at time t and ε_t is an error at time t.

The estimate \hat{T}_t for the level at time t, the estimate b_t for the growth rate at time t, and the estimate \hat{S}_t for the seasonal factor at time t are given by the smoothing equations (Bowerman et al. 2005)

$$\begin{aligned}\hat{T}_t &= \hat{T}_{t-1} + b_{t-1} + \alpha e_t \\ b_t &= b_{t-1} + \alpha \gamma e_t\end{aligned}$$

$$\hat{S}_t = \hat{S}_{t-L} + (1 - \alpha) \delta e_t$$

The estimate e_t for the error of the time series in time period t where α, γ, δ are smoothing constants between 0 and 1. \hat{T}_{t-1} and b_{t-1} are estimate in time period t-1 for the level and growth rate, \hat{S}_{t-L} is the estimate in time period t-L for the seasonal factor, and L denotes the number of seasons per year.

Multiplicative Holt-Winters Model

If a time series has a linear trend with a fixed growth rate, β_1 , and a fixed seasonal pattern, S_t with increasing variation, then the time series may be described by the model.

$$Y_t = (\beta_0 + \beta_1 t) \times S_t \times \varepsilon_t$$

The smoothing equations for the multiplicative Holt-Winters model are:

$$\begin{aligned}\hat{T}_t &= \hat{T}_{t-1} + b_{t-1} + \frac{\alpha e_t}{S_{t-L}} \\ b_t &= b_{t-1} + \frac{\alpha \gamma e_t}{S_{t-L}} \\ \hat{S}_t &= \hat{S}_{t-L} + \frac{(1 - \alpha) \delta e_t}{T_t}\end{aligned}$$

The Holt-Winters method requires starting values in which this research separate starting values into 2 methods. The first method used the first year of data to calculate \hat{T}_t, \hat{S}_t and b_t which separated into 5 different patterns which level component, and seasonal factor for additive and multiplicative model defined as follows:

$$\begin{aligned}\hat{T}_L &= \frac{(Y_1 + Y_2 + \dots + Y_L)}{L} \\ \hat{S}_i &= Y_i - \hat{T}_L \quad ; i = 1, \dots, L \\ \hat{S}_i &= \frac{Y_i}{\hat{T}_L} \quad ; i = 1, \dots, L\end{aligned}$$

The growth rate of each pattern is defined in Table 1. The second one employed Decomposition method which spend the data from the first set 2-18 years ($T=2, \dots, 18$) in calculating trend and seasonal variation by employed ratio to moving average method to decompose. Then, we get b_0, b_1 and \hat{S}_i from linear trend and seasonal factor, then applied.

$$\begin{aligned}\hat{T}_L &= b_0 + b_1 \times L \\ b_L &= b_1\end{aligned}$$

This research estimated the smoothing parameters α, γ and δ to minimize the root mean square error (RMSE) by using Solver module in Microsoft Excel.

Table 1: The growth rate for each patterns

Pattern	Growth rate
1	$b_L = \frac{1}{L} \sum_{i=1}^L \left(\frac{Y_{i+L} - Y_i}{L} \right)$ (Hyndman and Athannasopoulos 2018)
2	$b_L = Y_2 - Y_1$ (Kalekas 2020; Suppalakpanya et al. 2019)
3	$b_L = \frac{(Y_2 - Y_1) + (Y_3 - Y_2) + (Y_4 - Y_3)}{3}$ (Kalekas 2020; Suppalakpanya et al. 2019)
4	$b_L = \frac{Y_L - Y_1}{L - 1}$ (Kalekas 2020; Suppalakpanya et al. 2019)
5	$b_L = 0$ (Kalekas 2020; Suppalakpanya et al. 2019)

Box-Jenkins method

The Box-Jenkins method was developed in 1974 and it is widely used in modelling and forecasting the number of airport passengers. Box-Jenkins method uses Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify the model under the stationary condition. Box-Jenkins method is a four-step process (Bowerman et al. 2005).

Step 1: Tentative Identification: historical data are used to identify an appropriate Box-Jenkins model.

Step 2: Estimation: historical data are used to estimate the parameters of the identified model.

Step 3: Diagnostic checking: various diagnostics are used to check the adequacy of the identified model. In some cases may need to suggest an improved model, which is then regarded as a new identified model.

Step 4: Forecasting: once a final model is obtained, it is used to forecast future time series values.

It is possible that several models may be identified, and the selection of an optimum model is necessary. Such selection of models is usually based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) defined respectively as follows:

$$AIC = 2 \ln L + 2k \quad \text{and} \quad BIC = 2 \ln L + \ln(n)k$$

Where L represents the likelihood function, k is the number of estimated parameters from the model and n is the number of residuals that can be computed for the time series. The optimum model gives the minimum AIC and SBC.

Neural Networks

Artificial Neural Networks (ANN) are designed to mimic the characteristics of the biological neurons in the human brain and nervous system (Jiang et al. 2011). In the case of modelling first year premium of life insurance time series, the historical incidence are sent into the input neurons, and corresponding forecasting incidence is generated from the output neurons after the network is adequately trained. The network learns the information

contained in the first year premium time series by adjusting the interconnections between layers. The structure and neural networks can only be viewed in terms of the input, output and transfer characteristics. The specific interconnections cannot be seen even after the training process. There is no easy way to interpret the specific meaning of the parameters and interconnections within networks trained using the first year premium time series data. There are two advantages of employing neural networks for forecasting time series data. First, they can fully extract the complex nonlinear relationships hidden in the time series. Second, they have no assumption of the underlining distribution for the collected data (Zhang et al. 1998).

Back Propagation Neural Networks are a type of feed forward artificial neural networks. In feed-forward neural networks, the data flow is in one direction and the answer is obtained solely based on the current set of inputs. Back Propagation Neural Networks consist of an input layer, a hidden layer, and an output layer. Each layer is formed by a number of nodes, and each node represents a neuron. The upper-layer and lower-layer nodes are connected by the weights.

The first year premium of life insurance was divided into three subsets. First year premium from January, 2003 to December, 2015 was employed as the training set used for training the network. The first year premium from January, 2021 to November, 2021 was employed as the validation set. The remaining set of the series was used as the test set.

This research employed Back Propagation Neural Networks training. The number of inputs of the neural networks was determined by trend and seasonal period of the time series. In this study, the number of input nodes was selected to be 13, 26, 39 and 52. The output layer of artificial neural networks contains only one neuron representing the forecast value of the first year premium for the next month. This research employed Weka 3.8.6 in running artificial neural networks. The different learning rate was examined from 0.005 to 0.1 with 0.005 increments. The different momentum was examined from 0.1 to 0.8 with 0.05 increments. The number of hidden neurons were varied from 2 to 26 at an increment of 1.

MODEL SELECTION CRITERION

The forecasting models were chosen by considering the smallest root mean square error (RMSE) and the mean absolute percentage error (MAPE) were used to measure the accuracy of the model.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{Y_t} \right| \times 100$$

RESULTS

The first set of data from January 2003 to December 2020 is employed to build the Holt-Winters models with different initial values, Box-Jenkins models, and artificial neural networks models.

The Holt-Winters method

The results from Holt-Winters models with different initial values are shown in Table 2 and 3.

Table 2: RMSE of Holt-Winters models with five different initial values

Pattern	Additive model	Multiplicative model
1	863,466.60	837,216.40
2	856,548.71	861,502.89
3	870,418.17	851,926.00
4	856,548.71	829,798.23
5	860,619.49	832,226.44

From Table 2, the Holt-Winters models with initial values from pattern 4 obtained the minimum RMSE for both additive and multiplicative models. Where multiplicative model yields the minimum RMSE (829,798.23).

Table 3: RMSE of Holt-Winters models with initial values from Decomposition method.

Number of years (T)	Additive model	Multiplicative model
2	836,940.92	844,950.91
3	828,661.13	834,436.60
4	824,061.29	831,282.64
5	820,765.35	824,791.16
6	819,956.77	821,765.97
7	818,993.45	813,264.40
8	816,004.33	809,525.52
9	814,967.62	809,131.77
10	814,439.14	803,938.39
11	813,530.57	793,130.86
12	815,220.53	773,647.63
13	815,438.55	773,636.17
14	814,388.20	775,593.93
15	812,695.82	773,822.60
16	810,527.82	773,210.66
17	810,018.92	765,074.48
18	807,140.72	756,342.63

From Table 3, The RMSE of additive and multiplicative models are decreased while using more data in calculating trend and seasonal factor. The multiplicative Holt-Winters model with initial values from 18 years decomposition method yields minimum RMSE (756,342.63).

Box-Jenkins method

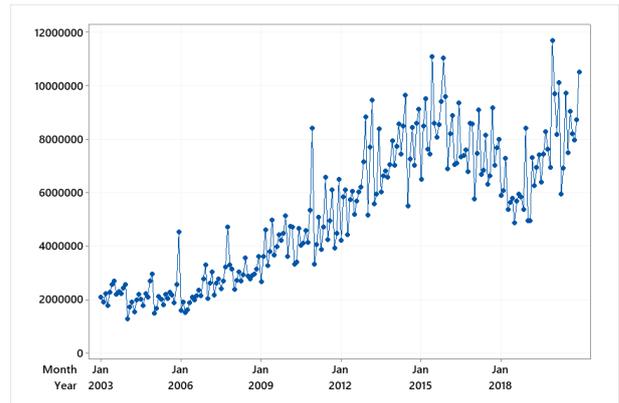


Figure 1: First year premium of life insurance from January 2003 to December 2020.

From Figure 1, the non-stationary is visible since the first year premium have non-linear trend and seasonal variation. One regular difference ($d=1$) and one seasonal difference ($D=1$) are taken to make time series stationary. Based on Autocorrelation Function (ACF) as shown in Figure 2 and Partial Autocorrelation Function (PACF) as shown in Figure 3 ACF are exponentially decayed, and PACF are significant spike at lag 3 then it suggests ARIMA(3,1,0). Additionally, PACF are significant spike at lag 12, and 48 suggest a seasonal SARIMA(4,1,0) However, ARIMA(3,1,0)xSARIMA(4,1,0)₁₂ model do not pass the diagnostic checking. Therefore, the improved model ARIMA(4,1,0)xSARIMA(4,1,0)₁₂ is suggested.

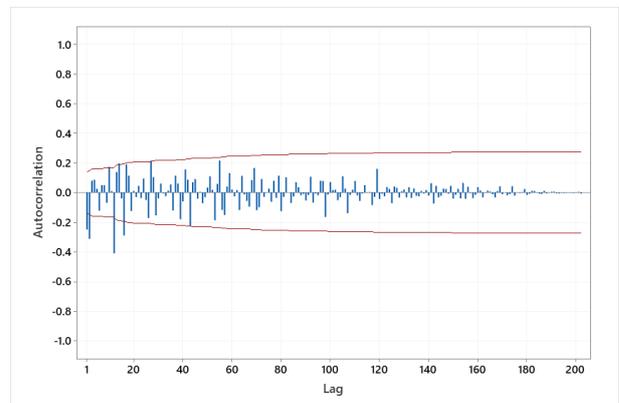


Figure 2: Autocorrelation Function of first year premium of life insurance with one regular difference and one seasonal difference.

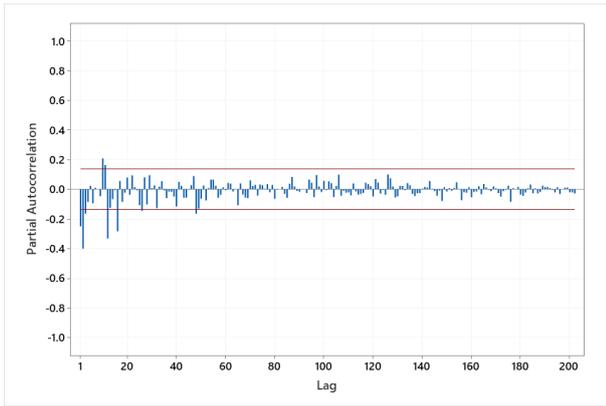


Figure 3: Partial Autocorrelation Function of first year premium of life insurance with one regular difference and one seasonal difference.

Table 4: Minitab output of Box-Jenkins model of ARIMA(4,1,0)xSARIMA(4,1,0)₁₂

Type	Coef	SE Coef	t-value	p-value
AR1	-0.5238	0.0720	-7.28	0.000
AR2	-0.5947	0.0789	-7.54	0.000
AR3	-0.2187	0.0793	-2.76	0.006
AR4	-0.1881	0.0720	-2.61	0.010
SAR12	-0.6719	0.0728	-9.23	0.000
SAR24	-0.3970	0.0865	-4.59	0.000
SAR36	-0.4860	0.0927	-5.24	0.000
SAR48	-0.4705	0.0858	-5.49	0.000
Modified Box-Pierce(Box-Ljung)Chi-Square statistic				
Lag	12	24	36	48
Chi-Square	7.39	22.15	36.71	45.62
DF	4	16	28	40
p-value	0.117	0.138	0.125	0.250

Table 4 shows all parameters ($\phi_1, \phi_2, \phi_3, \phi_4, \Phi_{12}, \Phi_{24}, \Phi_{36}, \Phi_{48}$) from ARIMA(4,1,0)xSARIMA(4,1,0)₁₂ model are statistically significant from zero (p-value is less than 0.05). From Box-Ljung test, residuals from the model are independent (p-value is greater than 0.05). Therefore, the model ARIMA(4,1,0)xSARIMA(4,1,0)₁₂ fits with first year premium of life insurance. In addition, ARIMA(0,1,3)xSARIMA(5,1,0)₁₂ and ARIMA(2,1,3)xSARIMA(5,1,0)₁₂ also pass the diagnostic checking. Table 3 shows all Box-Jenkins models with RMSE, AIC and SBC.

Table 5: Box-Jenkins models with RMSE, AIC and SBC

RMSE	AIC	SBC
ARIMA(0,1,3)xSARIMA(5,1,0) ₁₂ model		
783,809.29	70.29	96.47
ARIMA(4,1,0)xSARIMA(4,1,0) ₁₂ model		
794,947.80	70.34	96.53
ARIMA(2,1,3)xSARIMA(5,1,0) ₁₂ model		
779,229.75	74.26	106.89

From Table 5, based on the minimum AIC and SBC values, the ARIMA(0,1,3)xSARIMA(5,1,0)₁₂ model are the optimum model for first year premium of life insurance.

Artificial Neural Networks

Increasing the input nodes, the RMSE of training set is decreasing significantly. However, the RMSE of testing set is increasing. To avoid the overfitting, the model which has closest training and testing RMSE is chosen. The optimal model for ANN is 13-4-1 with learning rate of 0.01, momentum of 0.6 and 500 iterations. The RMSE of the model are shown in Table 6.

Table 6: The RMSE of training set, test set and validation set.

Model	RMSE		
	Training set	Test set	Validation set
13-4-1	737,179.1	1,060,518.2	1,408,953.9

Table 7: The RMSE of three forecasting methods.

Forecasting model	RMSE
Multiplicative model with initial value pattern 4	829,798.23
Multiplicative model with initial values from 18 years Decomposition method	756,342.63
ARIMA(0,1,3)xSARIMA(5,1,0) ₁₂	783,809.29
Artificial Neural Networks	1,060,518.23

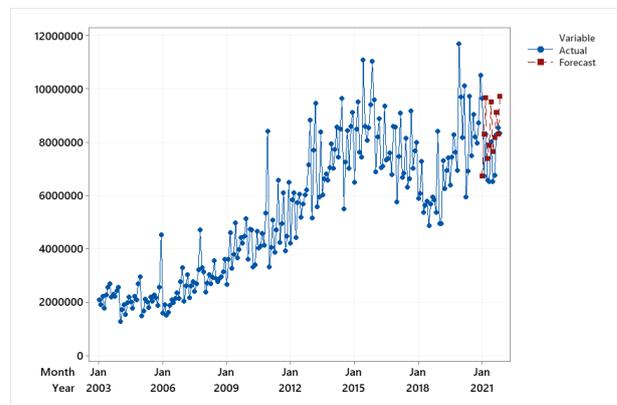


Figure 4. Actual and forecast by the multiplicative model with initial values from 18 years Decomposition method.

Model Selection and Performance Measure

From Table 7, the multiplicative model with in initial values from 18 years Decomposition method obtains the smallest RMSE. This research confirms that the different initial value can improve the efficient of Holt-Winters method significantly. Although the Holt-Winters method is suitable for time series that have linear trend and seasonal variation. This research shows that the different initial value can increase the efficiency of Holt-Winters

method even though the time series have non-linear trend and seasonal variation as first year premium of life insurance.

CONCLUSION

This research presents three different forecasting methods which are Holt-Winters method, Box-Jenkins method and Artificial Neural Networks to model the first year premium of life insurance. The results showed that the multiplicative model with initial values from 18 years Decomposition method give the appropriate model for the first year premium of life insurance and yields the MAPE = 17.29%. The Holt-Winters method with different initial value is simpler and require less data to estimate than Box-Jenkins method and Artificial Neural Networks. However, its performance can be outperformed both Box-Jenkins method and Artificial Neural Network. The increased complexity of Box-Jenkins method and Artificial Neural Networks do not guarantee the better results. As shown in this research, the simpler Holt-Winters method with different initial value can give higher accuracy in some cases. When the researcher models time series data with trend and seasonal factor, Holt-Winters method with different initial value is a good choice.

ACKNOWLEDGMENT

This work was supported by School of Science, King Mongkut's Institute of Technology Ladkrabang Research Fund. Grant no: 2564-02-05-002.

REFERENCES

- Booranawong T. and Booranawong A. 2018. "Double exponential smoothing and Holt-Winters methods with optimal initial values and weighting factors for forecasting lime, Thai chili and lemongrass prices in Thailand". *Engineering and Applied Science Research*, 45(1), 32-38.
- Bahia, Itedal Sabri Hashim 2013. "Using Artificial Neural Netork Modeling in Forecasting Revenue: Case Study in National Insurance Company/Iraq". *International Journal of Intelligence Science*, Vol 3, 136-143.
- Bowerman, B. L.; Richard T. O'; Connell and Anne B. Koehler. 2005. "Forecasting, Time Series, and Regression: An Applied Approach". 4th ed., Thomson Brooks/Cole. USA.
- Chatfield C. 1996. "The Analysis of Time Series". 5th ed., Chapman & Hall, New York.
- Hyndman R. J. and Athanasopoulos G. 2018. "Forecasting Principle and Practice". 2nd ed., OTexts: Melbourne, Australia.
- Jiang L-H; Wang A-G; Tian N-Y; Zhang W-C; Fan Q-L. 2011. "BP Neural Network of Continuous Casting Technological Parameter and Secondary Dendrite Arm Spacing of Spring Steel". *Journal of Iron and Steel Research*, Vol 18, 25-29.
- Kalekar, P.S. 2004. "Time Series Forecasting using Holt-Winters Exponential Smoothing"
- Namaweje, H. and Geoffrey, K. 2020. "Forecasting Life Insurance Premiums and Insurance Penetration Rate in Uganda". *International Journal of Scientific and Research Publications*, Vol 10, Issue 5, 708-718.
- Safyar A. 2010. "Forecasting Life Insurance Premium by Arima Model s and Neural Network." *SANAAT-E-BIMEH*. Vol 25, no 2(98),121-138.
- Suppalakpanya, K.; Nikhom, R.; Booranawong A.; Booranawong T. 2019. "An Evaluation of Holt-Winters Methods with Different Initial Trend Values for Forecasting Crude Palm Oil Production and Prices in Thailand". *Suranaree Journal of Science and Technology*, 26(1), 13-22.
- Thai Life Assurance Association, 2019. "Overview of Thai Life Insurance Business 2018-2019 and Trends of 2020" Annual Report 2019, 78-110.
- Zhang G.; Eddy Patuwo B.; Y. Hu M. 1998. "The State of the Art" *International Journal of Forecasting*, Vol 14, 35-62.

Simulation and Optimization

A SIMPLE ALGORITHM SELECTOR FOR CONTINUOUS OPTIMISATION PROBLEMS

Tarek A. El-Mihoub¹, Christoph Tholen^{1,2}, Lars Nolle^{1,2},

¹Department of Engineering Science,
Jade University of Applied Sciences,
26389 Wilhelmshaven, Germany

²German Research Center for Artificial Intelligence (DFKI),
Marine Perception,
26129 Oldenburg, Germany
{tarek.el-mihoub | lars.nolle | ✉@jade-hs.de
christoph.tholen@dfki.de

KEYWORDS

Algorithm Selection Problem, Evolution Strategy, Covariance Matrix Adaptation, Linear Search, STEP, Continuous Optimisation, Nelder-Mead Algorithm, One-Fifth Success Rule, (1+1)-CMA-ES, COCO Framework, BBOB-2009 Testbed.

ABSTRACT

A large number of algorithms has been proposed for solving continuous optimisation problems. However, there is limited theoretical understanding of the strengths and weaknesses of most algorithms and their individual applicability. Furthermore, the performance of these algorithms is highly dependent on their control parameters, which need to be configured to achieve a peak performance. Automating the processes of selecting the most suitable algorithm and the right control parameters can help in solving continuous optimisation problems effectively and efficiently. In this paper, a simple online algorithm selector is proposed. It decides on selecting the right algorithm based on the current state of the search process to solve a given problem. Each algorithm in the portfolio of the algorithm selector competes with others and utilises the results of other algorithms to locate the global optimum. The proposed algorithm selector and the algorithms of the portfolio as stand-alone algorithms were benchmarked on the noise-free BBOB-2009 testbed. The results show that the performance of the simple algorithm selector is better than the performances of the individual algorithms in general. It was also able to solve eleven out of twenty-four functions of the test suite to the ultimate accuracy of 10^{-8} .

INTRODUCTION

A large number of search algorithms for solving optimisation problem has been developed (Cuevas, et al., 2020; Whitley, 2019). In most cases, the performance of these algorithms can be further improved by introducing small modifications through adjusting their control (hyper) parameters (Vermetten, et al., 2020; El-Mihoub, et al., 2014). However, the no free-lunch theorem

(Wolpert & Macready, 1997) states that no single algorithm can outperform all other algorithms on all optimisation problems. Based on this theorem, solving any optimisation problem requires selecting a suitable algorithm with a suitable configuration (Huang, et al., 2019). Practitioners usually face the Algorithm Selection (AS) problem and the Algorithm Configuration (AC) problem when applying optimisation algorithms (Kerschke, et al., 2018). The AS and the AC problems have gained increasing attention in the last decades (Huang, et al., 2019; Kerschke, et al., 2018).

Different methodologies have been followed to automate deciding on the right algorithm and the right configuration. Some approaches adopt Rice's framework (Rice, 1976) for solving the AS problem to learn from the algorithms experience in solving different problems (Kerschke, et al., 2018). Most of these approaches use Rice's features-based model to solve the AS (Cruz-Reyes, et al., 2012) and the AC (Belkhir, et al., 2017) problems. Most of these approaches depict the AS and the AC problems as two separated problems in spite of the strong relations between them (Janković & Doerr, 2019). In most cases, the AS and AC problems are addressed sequentially. This separation can affect the algorithm efficiency or narrow the range of problems, on which an approach can be applied. Few researches have dealt with the problem as a Combined Algorithm Selection and Hyper-parameter optimisation (CASH) problem (Vermetten, et al., 2020).

Hyper-heuristics methodology (Drake, et al., 2019) solves the AS and the AC problems using different approaches. The generative hyper-heuristics explore the space of the algorithmic components of algorithms to design a customised search algorithm. This methodology can be used to generate tailor-made and well-configured optimisation algorithms to solve the CASH problem. These approaches are also suitable for solving dynamic problems. However, they can be extremely costly (Miranda, et al., 2017).

On the other hand, the selective hyper-heuristics approach deals with the AS problem as an optimisation problem. A hyper-heuristic is employed to discriminate between different search algorithms based on their

performance on a given problem. Following this approach, a suitable algorithm is selected for each search iteration. Off-line selective hyper-heuristics approaches require featuring different states of different optimisation problems. Whereas, online selective hyper-heuristics approaches incur an additional high cost on the optimisation process.

A simple online algorithm selector with minimum overhead learning costs might be efficient in solving the basic AS problem. Most state-of-the-art optimisation algorithms utilise restart to resample good potential solutions (Hansen, et al., 2021). Instead of restarting, a simple selector can decide to resample a new solution or select the right algorithm for the current search state based on the most recent performances of a set of optimisation algorithms. Based on this idea, a simple algorithm selector is proposed in this paper.

A set of optimisation algorithms was chosen as a portfolio of the algorithm selector. This set consists of one multi-point and three single-point optimisation algorithms. The multi-point algorithm is the Nelder-Mead downhill simplex (Nelder & Mead, 1965). The single-point algorithms are the one plus one Covariance Matrix Adaptation Evolutionary Strategy, i.e. (1+1)-CMA-ES (Igel, et al., 2006), the one plus one Evolution Strategy with one-fifth success rule, i.e. (1+1)-ES (Auger, 2009), and the Line Search with the STEP (Swarzberg, et al., 1994), LSSStep algorithm. This set of single- and multi-point algorithms was selected to show that different types of algorithms can be unified easily in the proposed algorithm selector.

In the next section, the simple algorithm selector is introduced. Then, the algorithms that constitute the selector portfolio are briefly described. Before presenting and discussing the results of the benchmarking, the experimental setup is presented. The paper ends with the conclusion and future work.

THE SIMPLE ALGORITHM SELECTOR

Most local search algorithms and even global search algorithms rely on probabilistic restart to achieve globalisation (Pošik & Huyer, 2012). However, sharing the optimisation resources by more than one algorithm can reduce the risk of failure in solving a range of optimisation problems.

The concept of the proposed algorithm selector is to rely on a portfolio of optimisation algorithms instead of relying on a single algorithm with probabilistic restart to reach the global optimum. An algorithm selector with a simple discriminating mechanism based on the current state of the search can solve the AS problem with a minimum overhead cost. The main aim of the algorithm selector is to locate the global optimum and not to find the best algorithm for a given problem.

The algorithm selector should have the ability to unify single point and population-based search algorithms within its framework. The algorithm selector should be

capable of distributing the search resources in an effective way. It should be able to select a suitable algorithm based on the current state of the search process and to build on search results of previously selected algorithms. The algorithm selector should enable different algorithms to compete with each other and cooperate to reach the global optimum.

We consider an objective function $f: \mathbb{R}^D \rightarrow \mathbb{R}$, where $x \rightarrow f(x)$ to be minimised. To minimise this function, the proposed algorithm selector follows the algorithm shown in Algorithm 1. The algorithm selector starts by initialising the search environment, line 1 in Algorithm 1. The initialisation process includes selecting an initial point as a potential solution. Then, the selector chooses randomly an algorithm from the portfolio and applies it to the optimisation problem using the potential solution. The selector checks whether the global optimum is reached. In this case, the selector stops. Otherwise, it assigns the selected algorithm a score, which is equal to the change in the fitness of the potential solution. The selector repeats the previous steps until reaching the global optimum or all the algorithms of the portfolio are applied.

In the case of not reaching the global optimum, two algorithms from the portfolio are selected randomly. The algorithm with the best score out of these two algorithms is chosen for utilisation. If the last algorithm applied is selected again, and that algorithm has probably reached a local optimum, the algorithm selector resamples a new potential solution to replace the local reached optimum. Next, the new selected algorithm is applied and its score is updated. The previous steps (12-19 in algorithm 1) are repeated until reaching the target optimum or the search resources are consumed.

Algorithm 1 : The Simple Algorithm Selector

```

1  initiate search environment
2  while not all algorithms of the portfolio selected
3    current algorithm = select an algorithm from the
      algorithms portfolio randomly
4    apply current algorithm on the optimisation
      problem
5    if target value reached
6      print results and exit
7    end if
8    current algorithm score = change in fitness
9    last selected = current algorithm
10 end while
11 while global optimum not reached
12   selected algorithms = select randomly two
     algorithms
13   current algorithm = algorithm with best score of
     selected algorithms
14   if last selected == current algorithm
15     resample new solution
16   end if
17   apply current algorithm on the optimisation
     problem
18   current algorithm score = change in fitness
19   last selected = current algorithm
20 end while

```

The proposed algorithm selector can use a portfolio that consists of a number of single- and multi-point search algorithms. In this paper, a portfolio of four algorithms is implemented. These algorithms are classified as local search algorithms. These algorithms are briefly described in the following sections.

THE NELDER-MEAD DOWNHILL SIMPLEX

The Nelder–Mead algorithm (Nelder & Mead, 1965) is also known as the downhill simplex method. It is an optimisation algorithm for real-value problems. The Nelder-Mead method starts with a set of $D + 1$ initial solutions, a simplex, where D is dimension of the search space. A new solution is generated through reflecting the worst solution on the centroid of the remaining D solutions. Other operations are conducted to either further improve the new generated solution or to focus on the most promising region of the search space.

A pseudocode of Nelder-Mead method for function minimisation is shown in algorithm 2. The simplex method modifies the vertices of the simplex using four operations to generate better solution based on the fitness of the vertices. These operations are reflection, expansion, contraction and shrinking. The coefficients of these operations are χ , γ , ρ and σ , respectively. Table 1 shows the formulas for executing these operations in addition to the formula for calculating the centroid. The reflection, expansion and the contraction operations are applied to the worst vertex. Meanwhile, the shrinking operation is applied to all vertices except the best one. The standard values for the operations coefficients are $\rho=0.5$, $\chi=2$, $\gamma=0.5$, and $\sigma=0.5$ (McKinnon, 1998).

Table 1: Operations of Nelder-Mead Method

Operations on simplex	Formula
centroid calculation	$x_c = \frac{1}{n} \sum_{i=2}^{n+1} x_i$
reflection	$x_r = x_c + \rho(x_c - x_{n+1})$
expansion	$x_e = (1 + \rho\chi)x_c - \rho\chi x_{n+1}$
outside contraction	$x_{out_c} = (1 + \rho\gamma)x_c - \rho\gamma x_{n+1}$
inside contraction	$x_{in_c} = (1 - \gamma)x_c - \gamma x_{n+1}$
shrinking	$x_{i=2,n+1} = x_i + \sigma(x_i - x_1)$

Starting with an initial solution, a simplex can be generated around this solution. This initial solution together with D generated points can be used as starting points, which represent the vertex of the simplex. The simplex can be generated by adding a small value (delta) to each component of the initial solution. In this paper, the value of delta is set to 0.00025 for components with values of zero and a delta of 0.05 for other component values. These values are used by Matlab in the *fminsearch* function (Hansen, 2009).

EVOLUTION STRATEGIES

Evolution strategies (ES) for real-valued optimisation usually rely on Gaussian random variations. They assume

that the space around the global optimum can be represented by a multi-variant distribution and the global optimum is at the distribution's centre. Variant ES algorithms have been proposed to locate the global optimum and determine the multi-variant distribution around it by sampling points in the search space.

The (1+1)-ES algorithms start with an initial solution (x_i , $i = 0$) and assumes this is the mean of the distribution. They resample a new solution according to the adopted distribution. Once a better solution is found, this solution becomes the new mean of the distribution. The algorithm keeps a track of the changes in the objective function values and the change in solutions' locations in the search space. The algorithm uses these changes to amend the shape of the distribution to improve the quality of new sampled solutions. In this section, two variants of the (1+1)-ES are concisely described.

Algorithm 2 : The Nelder-Mead Algorithm

```

1  Input: (D+1) points
2  while not terminated do
3    order the vertices according to their fitness
4     $f(x_1) \leq f(x_2) \leq \dots \leq f(x_{D+1})$ 
5    calculate the centroid point of all vertices except
6     $x_{D+1}, x_c$ 
7    calculate the reflection point  $x_r$ 
8    if  $f(x_1) \leq f(x_r) < f(x_n)$ 
9       $x_{D+1} = x_r$ 
10   else
11     if  $f(x_r) < f(x_1)$ 
12       calculate the expansion point  $x_e$ 
13       if  $f(x_e) < f(x_r)$ 
14          $x_{D+1} = x_e$ 
15       else
16          $x_{D+1} = x_r$ 
17       end
18     else
19       if  $f(x_D) \leq f(x_r) < f(x_{D+1})$ 
20         calculate the outside contraction point  $x_{out\_c}$ 
21         if  $f(x_{out\_c}) < f(x_r)$ 
22            $x_{n+1} = x_{out\_c}$ 
23         else
24           shrink the simplex
25         end
26       else
27         if  $f(x_r) \geq f(x_{D+1})$ 
28           calculate the inside contraction point  $x_{in\_c}$ 
29           if  $f(x_{in\_c}) < f(x_{D+1})$ 
30              $x_{n+1} = x_{in\_c}$ 
31           else
32             shrink the simplex
33           end
34         end
35       end
36     end
37   end
38 end
39 end

```

The (1+1)-ES with One-Fifth Success Rule Algorithm

This (1+1)-ES algorithm is based on the idea that the search step from the current solution should increase in a case of successive successful steps and should decrease otherwise. Many successful steps indicates that the

search can be improved by taking a larger step. On the other hand, very few successful steps indicates the search step might be too large and need to be reduced. According to the one-fifth success rule (Schumer & Steiglitz, 1968), the step-size should not change if the success probability of the sampled solutions is about one-fifth, increase if the success probability is larger than one-fifth and decrease otherwise.

The factors of 1.5 and $1.5^{-1/4}$ for increasing and decreasing the step-size can implement the idea of the one-fifth success rule (Auger, 2009). Pseudocode of the (1+1)-ES with one-fifth success rule is shown in Algorithm 3. A sample from the standard multivariate normal distribution is selected randomly (z_i , line 5). This sample is multiplied by the step size σ and is added to the current mean of the distribution to generate a new solution. The algorithm assumes the distribution is symmetric around the global optimum.

Algorithm 3 : (1+1)-ES with One-Fifth Success Rule

1	Input: x_0, σ_0
2	$x_c = x_0$
3	$\sigma = \sigma_0$
4	while <i>not terminated</i> do
5	$z_i = N(0, I)$
6	$x_i = x_c + \sigma z_i$
7	if $f(x_i) \leq f(x_c)$
8	$x_c = x_i$
9	$\sigma = 1.5 \sigma$
10	else
11	$\sigma = 1.5^{-1/4} \sigma$
12	end if
13	end while

The (1+1)-CMA-ES Algorithm

CMA-ES algorithm (Hansen, 2006) add a covariance matrix, $C \in \mathbb{R}^{n \times n}$, component to the multi-variant distribution, which is used to generate new solutions. By appropriate adaptation of the covariance matrix, a more accurate representation of the space around the global optimum can be achieved. This help in locating the global optimum, especially, for ill-conditioned problems.

The (1+1)-CMA-ES algorithm starts with an initial solution (x_0), or initial mean, as in any (1+1)-ES algorithm. It also starts with an initial covariance matrix (C_0) of the distribution. Usually, the algorithm starts with a standard multi-variate normal distribution with an initial global step size or sigma (σ_0). As in the case with (1+1)-ES with one-fifth success rule, the changes in the new sampled solution and its cost value are used to adapt the mean the global step size. However, in the (1+1)-CMA-ES, they are also used to adapt the covariance matrix of the distribution.

The covariance matrix C_i needs to be decomposed into Cholesky factors in order to sample a general multivariate normal distribution (i.e. $C_i = A_i A_i^T$). A new solution is generated using the multi-variant distribution as shown in line 10 of Algorithm 4, which shows pseudocode of the (1+1)-CMA-ES algorithm. The global

step size is then updated based on the average success rate $p_{succ} \in [0, 1]$. The covariance matrix C is updated in the case of a decrease in the cost values of new solutions compared the cost of the current mean of the distribution. The update is done based on the values of the average success rate p_{succ} , the global step size σ and the evolution path p_c . Table 2 shows the rules for updating these parameters and Table 3 shows the default parameter values (Igel, et al., 2006).

Table 2: Updating rules of (1+1)-CMA-ES

Parameter Name	Updating rules
average success rate	if $f(x_{i+1}) < f(x_i)$ $p_{succ_{i+1}} = (1 - c_p)p_{succ_i} + c_p$ else $p_{succ_{i+1}} = (1 - c_p)p_{succ_i}$
step size	$\sigma_{i+1} = \sigma_i \times \exp\left(\frac{1}{d} \left(\frac{p_{succ_i} - p_{succ}^{target}}{1 - p_{succ}^{target}} \right)\right)$
evolution path	if $p_{succ} < p_{thresh}$ $p_{c_{i+1}} = (1 - c_p)p_{c_i} + \sqrt{c_c(2 - c_c)}Az_i$ else $p_{c_{i+1}} = (1 - c_p)p_{c_i}$
Covariance matrix	if $p_{succ} < p_{thresh}$ $C_{i+1} = (1 - c_{cov})C_i + c_{cov} \cdot p_c p_c^T$ else $C_{i+1} = (1 - c_{cov})C_i + c_{cov} \cdot (p_c p_c^T + c_c(2 - c_c)C_i)$

Table 3: Parameters of the (1+1)-CMA-ES and their Default Values

Step size control	Covariance matrix adaptation
d : the damping parameter controls the rate of the step size adaptation $d = 1 + \frac{n}{2}$	c_c : the learning rate for the evolution path $c_c = \frac{2}{n+2}$
p_{succ}^{target} : the target success rate $p_{succ}^{target} = \frac{2}{11}$	c_{cov} : the learning rate for the covariance matrix $c_{cov} = \frac{2}{n^2+6}$
c_p : the learning rate for the step size $c_p = \frac{1}{12}$	p_{thresh} : the threshold of the success rate to prevent fast increase of C matrix axes with small step sizes $p_{thresh} = 0.44$

Algorithm 4 : (1+1)-CMA-ES

1	Input: x_0, σ_0
2	$x_c = x_0, \sigma = \sigma_0$
3	$p_{succ} = p_{succ}^{target}, p_c = 0, C = I$
4	while <i>not terminated</i> do
5	decompose C_i such that $C_i = AA^T$
6	$z_i = r_{nadam} \text{ sample of } N(0, I)$
7	$x_{i+1} = x_c + \sigma_i Az_i$
8	$p_{succ_{i+1}} = \text{updated_average_success_rate}$
9	$\sigma_{i+1} = \text{updated_step_size}$
10	if $f(x_i) \leq f(x_c)$
11	$x_c = x_{i+1}$
12	$C_{i+1} = \text{updated_covariance_matrix}$
13	end if
14	end while

LINE SEARCH WITH STEP

Line search algorithms are simple optimisation algorithms. They are effective and efficient in solving separable optimisation problems. They can be used to discriminate between separable and non-separable optimisation test functions (Pošík & Huyer, 2012). The algorithm starts with a randomly selected solution and iterates through individual directions. It optimises the function with respect to the chosen direction while keeping the other components of the solution fixed. Once the optimum in one direction is found, it switches to another direction starting from the best-found solution. No change in the objective value of the solution after going through all the directions indicates a local optimum and the algorithm stops.

The STEP method (Swarzberg, et al., 1994) is a univariate global search algorithm with interval division. The basic idea of the STEP is to sample a new solution with the greatest chance of exceeding the best-found solution.

Therefore, it starts with an initial interval as shown in Algorithm 5. It calculates the objective values of the endpoints of that interval. Then, it divides the interval into two halves by sampling the middle point of the interval. It determines the interval, which has the greatest chance to include a solution, which is better than the current best solution. It repeats the division process for the most promising interval until reaching the stopping criteria. The interval difficulty is used as criteria for selecting the most promising interval. The algorithm selects the interval that enables sampling a solution, which is better than the current best solution.

To determine the difficulty of an interval, STEP assumes a quadratic function $y = ax^2 + bx + c$ that goes through the interval boundaries and $y = f_{best} - \epsilon$, where f_{best} is the objective function of the best found solution and ϵ is a small positive number. The value of ϵ determines the value by which f_{best} is to be at least exceeded. In other words, it determines the tolerance in objective function values. In this paper, the value of ϵ is set to 10^{-8} (Pošík & Huyer, 2012). STEP uses the value of the coefficient a of this quadratic function to measure the interval difficulty. The interval with the smallest value of a is more likely to enable sampling better solutions (Swarzberg, et al., 1994).

Algorithm 5 : The STEP algorithm

1	Input: boundaries of the initial interval
2	evaluate the boundaries of the initial interval
3	sample the middle point of the interval
4	evaluate the middle point
5	add the intervals to the interval list
6	while <i>not terminated</i> do
7	determine the interval with the lowest difficulty
8	sample the middle point of this interval
9	evaluate the middle point
10	add the new intervals to the interval list
11	end while

EXPERIMENTS

This work tries to answer the question ‘Can a simple algorithm selector outperform the individual algorithms, those constitute its portfolio, on a range of optimisation problems?’. To answer this question, the performance of the algorithm selector needs to be compared with the performances of the individual algorithms. Experiments were conducted to evaluate the performance of different algorithms on a range of optimisation problems.

Experimental Framework Description

The experiments were carried out using the Comparing Continuous Optimisers (COCO) framework (Hansen, et al., 2021). This framework was also used for the Black-box Optimisation Benchmarking workshop at the GECCO-2009 and 2010 conferences.

The experiments were conducted using the BOBB test suite, which consists of 24 test functions (Hansen, et al., 2009). The functions are classified based on their properties as multimodality, ill-conditioning, global structure and separability. All functions are scalable in terms of dimensionality. The search domain is $[-5; 5]$ for each dimension. Different instances of the same function can be produced by rotating and shifting each function.

Each algorithm was tested for 15 trials on different instances of each function for different dimensions [2, 3, 5, 10, 20, 40].

Algorithm and Experiment Parameter Settings

For fair comparison between the different algorithms, the maximum number of function evaluations for the different algorithms was set to $D \times 10^4$, where D is the dimensionality of the problem. The algorithms were benchmarked using the BBOB2009 settings, i.e. the algorithms were run on the 24 benchmark functions, 5 instances each, 3 trials per instance.

No specific parameter tuning has been done during the experiments. The settings were identical for all functions such that the crafting effort is zero (Hansen, et al., 2010). The control parameters of algorithms within the simple algorithm selector are the same as that of the individual algorithms. The values of the control parameters for the individual algorithms were set as defined in research papers that benchmarked these algorithms using the COCO framework (Auger, 2009; Auger & Hansen, 2009; Posík, 2009). To avoid the impact of the implementation details on the evaluation process, all algorithms have been implemented by the authors and the comparison was done based on results of the implemented algorithms, not on the archived data of the COCO framework. The implementation process was done to accurately replicate the algorithms as described in the papers (Auger, 2009; Auger & Hansen, 2009; Posík, 2009; McKinnon, 1998).

RESULTS AND DISCUSSION

The results of benchmarking the individual algorithms, as implemented by the authors, were compared with the

archived data on <https://numbbo.github.io/data-archive/bbob/>. The comparison shows a big difference in the results of the (1+1)-CMA-ES compared with the archived results on f_5 . It also shows a significant difference on f_7 . For the (1+1)-ES with one-fifth rule, the comparison shows that there is a significant difference in the performances on f_5 , f_8 , f_9 , f_{10} and f_{11} for the different dimensions. The comparison also shows a significant difference in the results of the STEP algorithm, on f_6 , f_8 , f_9 , f_{10} , f_{15} , f_{21} , and f_{22} for small dimensions (i.e. $d=[2, 3, 5]$). In most cases, the archived results are better than the results of the conducted experiments. The comparison results are not shown in this paper due to limitation on the number of pages.

The results of the experiments that compares the simple algorithm selector with individual algorithms on the test functions for different dimensions are shown in Figures 1-6. The post-processing tools of the COCO platform were used to generate these plots. In these plots, the best algorithm is the algorithm, which is able to solve the highest fraction of test functions for different target values. The best 2009 line shown in the figures corresponds to the algorithms from BBOB-2009 with the best expected run time for each of the targets considered. Whereas, the Selector, CMA, OneFifth, Simplex and LSSStep lines correspond to the algorithms the simple algorithm selector, (1+1)-CMA-ES, (1+1)-ES with fifth rule, the Nelder-Mead and the line search STEP algorithms respectively.

The results show that in general the simple selector algorithm performs better than or at least as good as the best individual algorithm of the algorithms portfolio. The performance in terms of the fraction of function-target pairs of the simple algorithm selector that is better than the individual algorithms for the dimensions of 2, 5, 10, 20 and 40 as shown in Figures 1, 3, 4, 5 and 6. However, for three-dimensional test functions, the Selector and CMA show similar performance as depicted in Figure 2. For larger dimensions as illustrated in Figures 3-6, the difference in the performances between the simple algorithm selector and other algorithms becomes more significant.

The simple algorithm selector has solved the functions f_1 , f_2 , f_3 , f_4 , f_5 , f_8 , f_9 , f_{10} , f_{12} , and f_{14} . It was able to reach a target of 10^{-8} for all dimensions and in all the experiments for these test functions. As in any learning mechanisms, an additional learning cost was expected, which can affect its expected run-time to reach different targets. However, the algorithm selector shows a performance, which is better than that of the best 2009 algorithm on f_4 as shown in Figure 7. It was also able to locate the global optimum of f_4 in 15 experiments compared only 6 experiments out of 15 for the best 2009 algorithm.

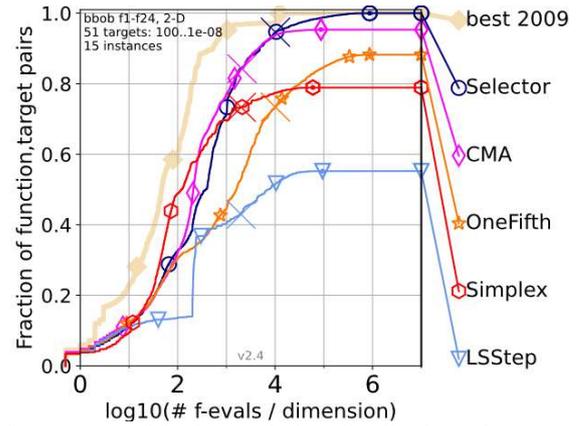


Figure 1: Empirical cumulative distribution of expected run time (ERT) over dimension for 51 targets in $10^{[2..8]}$ for all functions in 2-D.

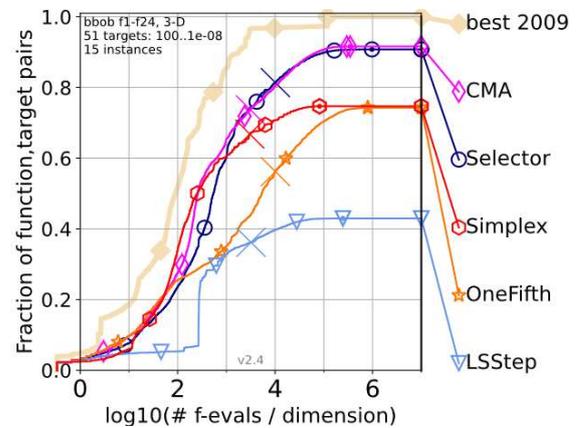


Figure 2: Empirical cumulative distribution of ERT over dimension for all functions in 3-D.

The results show that the simple algorithm selector is able to reach a target of 10^{-8} for the different test functions and for the tested dimensions, more than any of the individual algorithms. Table 4 compares the success rate of the different algorithms for reaching this target for different dimensions. The table clearly shows the superiority of the simple selector over the individual algorithms on all test functions for different dimensions. The success rate is calculated as the ratio of the number of experiments that have reached the target of 10^{-8} to the total number of conducted experiments for a specific dimension on all test functions.

Table 4: The success rate of the algorithms in finding the ultimate precision of 10^{-8} in the 24 function for different dimensions.

Algorithm	Dimension (D)					
	2	3	5	10	20	40
Selector	0.89	0.73	0.62	0.59	0.54	0.49
CMA	0.77	0.63	0.53	0.46	0.42	0.40
Simplex	0.65	0.58	0.48	0.37	0.25	0.07
OneFifth	0.53	0.33	0.21	0.15	0.14	0.09
LSSStep	0.23	0.21	0.21	0.20	0.19	0.21

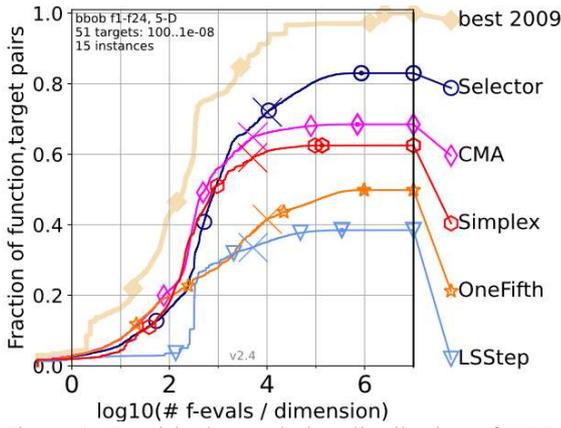


Figure 3: Empirical cumulative distribution of ERT over dimension for all functions in 5-D.

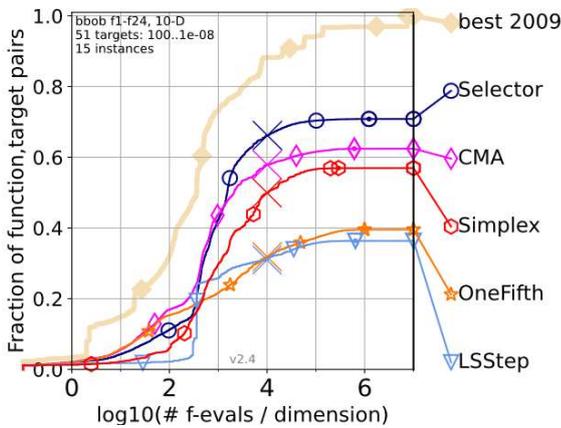


Figure 4: Empirical cumulative distribution of ERT over dimension for all functions in 10-D.

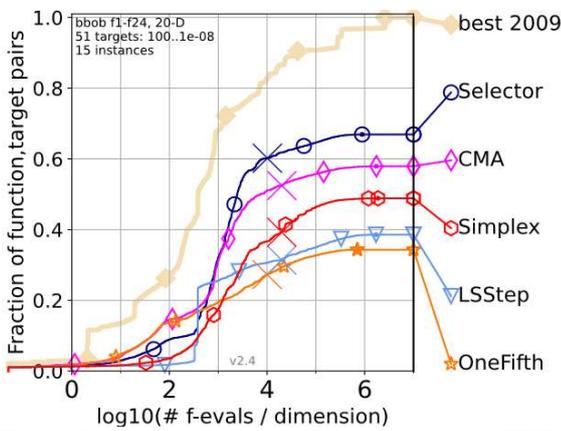


Figure 5: Empirical cumulative distribution of ERT over dimension for all functions in 20-D.

In Table 5, the success rate of the different algorithms in reaching the precision of 10^{-8} for different function groups and for all dimensions is shown. The table shows that the simple algorithm selector has a success rate of 100% on the separable functions, i.e. f_1 - f_5 for all dimensions. It also demonstrates that the simple selector has a success rate better than the success rate of the individual algorithms on the unimodal functions with moderate conditioning, i.e. f_6 - f_9 , and the multimodal functions, i.e. f_{15} - f_{19} . However, the (1+1)-CMA-ES algorithm has a slightly better success rate than the

simple algorithm selector on the unimodal ill-conditioned functions, i.e. f_{10} - f_{14} , and the multimodal functions with weak structure, i.e. f_{20} - f_{24} . The performance on the group of multimodal functions with weak structure can be explained with the shape of the fitness landscapes of this group. They do not enable the algorithm selector to benefit from utilising previous experience for deciding on the best algorithm for the current state. For the unimodal ill conditioned functions, selecting the (1+1)-ES with one fifth rule or the LSStep, which have a success rate of approximately 0, can lead to waste a high fraction of the search resources.

Table 5: The success rate of the algorithms in finding the ultimate precision of 10^{-8} for different function groups.

Algorithm	Function group				
	f_1 - f_5	f_6 - f_9	f_{10} - f_{14}	f_{15} - f_{19}	f_{20} - f_{24}
Selector	1.00	0.78	0.92	0.13	0.41
CMA	0.53	0.70	0.94	0.08	0.44
Simplex	0.56	0.50	0.59	0.03	0.35
OneFifth	0.39	0.31	0.00	0.10	0.41
LSStep	0.97	0.00	0.01	0.01	0.01

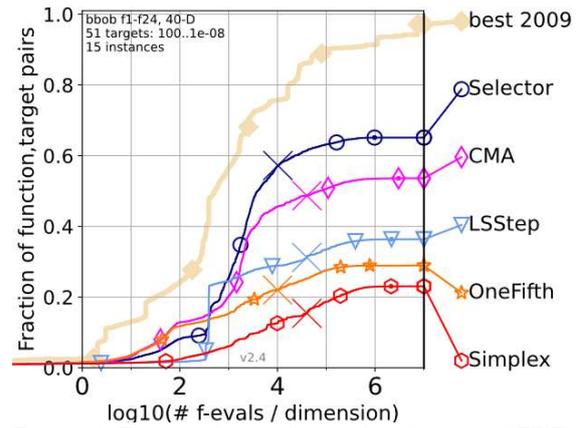


Figure 6: Empirical cumulative distribution of ERT over dimension for all functions in 40-D.

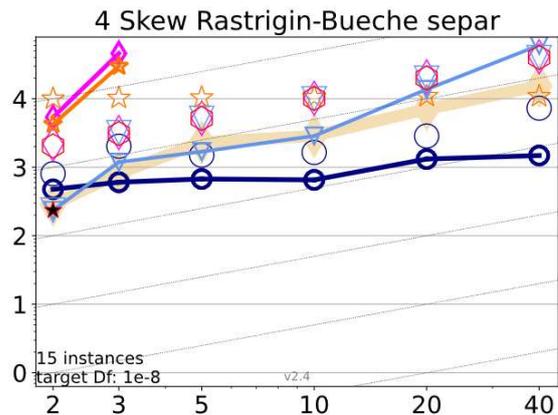


Figure 7: Expected running time (ERT in number of f -evaluations as \log_{10} value), divided by dimension for target function value 10^{-8} versus dimension. Legend: \circ : Selector, \diamond : CMA, \star : OneFifth, \blacktriangledown : LSStep, \circ : Simplex

CONCLUSION AND FUTURE WORK

A simple algorithm selector that differentiate between the performances on a set of four search algorithms is proposed. A portfolio, which unifies single point and multi-point search algorithms to enable utilising them within an algorithm selector, is constructed. A simple discrimination method is used to decide on a suitable search algorithm based on the search experience. The simple algorithm selector outperforms the stand-alone search algorithms on the noise-free BBOB-2009 test suite. It was able to solve $f_1, f_2, f_3, f_4, f_5, f_8, f_9, f_{10}, f_{12}$, and f_{14} to the ultimate precision of 10^{-8} for dimensions of 2, 3, 5, 10, 20 and 40.

The next step in this research is to investigate reinforcement learning techniques, such as q-learning, to decide on the right algorithm based on the accumulated search experience

REFERENCES

- Auger, A., 2009. Benchmarking the (1+1) evolution strategy with one-fifth success rule on the BBOB-2009 function testbed. Montréal Québec, Canada, s.n.
- Auger, A. & Hansen, N., 2009. Benchmarking the (1+1)-CMA-ES on the BBOB-2009 Functions testbed. Montreal, Canada, s.n.
- Belkhir, N., Dréo, J., Savéant, P. & Schoenauer, M., 2017. Per Instance Algorithm Configuration of CMA-ES with Limited Budget. Berlin, Germany, s.n.
- Cruz-Reyes, L. et al., 2012. Algorithm Selection: From Meta-Learning to Hyper-Heuristics. In: Intelligent Systems. s.l.:InTech, pp. 77-102.
- Cuevas, E., Fausto, F. & González, A., 2020. Metaheuristics and Swarm Methods: A Discussion on Their Performance and Applications. In: New Advancements in Swarm Algorithms: Operators and Applications. s.l.:Springer, Cham, pp. 43-67.
- Drake, J., Kheiri, A., Özcan, E. & Burk, E., 2019. Recent advances in selection hyper-heuristics. European Journal of Operational research, pp. 1-24.
- El-Mihoub, T., Hopgood, A. A. & Aref, I., 2014. Self-adaptive Hybrid Genetic Algorithm using an Ant-based Algorithm. Kuala Lumpur, IEEE, pp. 166-170.
- Hansen, N., 2006. The CMA Evolution Strategy: A Comparing Review. In: J. Lozano, P. Larrañaga, I. Inza & E. Bengoetxea, eds. Towards a New Evolutionary Computation. Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer, pp. 75-102.
- Hansen, N., 2009. Benchmarking the Nelder-Mead Downhill Simplex Algorithm With Many Local Restarts. Montreal, Canada, ACM, p. 2403-2408.
- Hansen, N., Auger, A. & Finck, S., R., 2010. Real-parameter black-box optimization benchmarking BBOB-2010: Experimental setup, s.l.: INRIA.
- Hansen, N. et al., 2021. COCO: a platform for comparing continuous optimizers in a black-box setting. Optimization Methods and Software, pp. 1-36.
- Hansen, N., Finck, S., Ros, R. & Auger, A., 2009. Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions Technical Report RR-6829, s.l.: INRIA.
- Huang, C., Li, Y. & Yao, X., 2019. A Survey of Automatic Parameter Tuning Methods for Metaheuristics. IEEE Transactions on Evolutionary Computation, pp. 1-16.
- Igel, C., Sutton, T. & Hansen, N., 2006. A Computational Efficient Covariance Matrix Update and a (1+1)-CMA for Evolution Strategies. Seattle, Washington, USA, ACM.
- Janković, A. & Doerr, C., 2019. Adaptive Landscape Analysis. Prague, Czech Republic, s.n., p. 2032-2035.
- Kerschke, P., Hoos, H. H., Neumann, F. & Trautmann, H., 2018. Automated Algorithm Selection: Survey and Perspectives. Evolutionary Computation, 27(1), p. 3-45.
- McKinnon, K., 1998. Convergence of the Nelder-Mead Simplex Method to a Nonstationary Point. SIAM Journal on Optimization, p. 148-158.
- Miranda, P., Prudêncio, R. B. & Pappa, G. L., 2017. H3AD: A hybrid hyper-heuristic for algorithm design. Information Sciences, Volume 414, pp. 340-354.
- Nelder, J. & Mead, R., 1965. A simplex method for function minimization. The Computer Journal, p. 308-313.
- Pošik, P. & Huyer, W., 2012. Restarted Local Search Algorithms for Continuous Black-Box Optimization. Evolutionary Computation, 20(4), pp. 575-607.
- Posik, P., 2009. BBOB-benchmarking two variants of the line-search algorithm. Montréal Québec, Canada, s.n.
- Rice, J. R., 1976. The algorithm selection problem. Advances in Computers, Volume 15, pp. 65-118.
- Schumer, M. & Steiglitz, K., 1968. Adaptive step size random search. IEEE Transactions on Automatic Control, 13(3), pp. 270-276.
- Swarzberg, S., Seront, G. & Bersini, H., 1994. S.T.E.P.: the easiest way to optimize a function. Orlando, FL, USA, IEEE, pp. 519-524.
- Vermetten, D., Wang, H., Doerr, C. & Bäck, T., 2020. Integrated vs. sequential approaches for selecting and tuning CMA-ES variants. Cancun, Mexico, s.n.
- Whitley, D., 2019. Next Generation Genetic Algorithms: A User's Guide and Tutorial. In: Handbook of Metaheuristics. International Series in Operations Research & Management Science. s.l.:Springer, Cham, pp. 245-274.
- Wolpert, D. H. & Macready, W. G., 1997. No Free Lunch Theorems for Optimization. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, pp. 67-82.

OPTIMISED BUMBLEBEE PATHS AS SEARCH STRATEGY FOR AUTONOMOUS UNDERWATER VEHICLES

Christoph Tholen

Department of Engineering Science
Jade University of Applied Science
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
and

German Research Centre for Artificial Intelligence
Marie-Curie-Straße 1
26129 Oldenburg, Germany
christoph.tholen@dfki.de

Tarek A. El-Mihoub

Department of Engineering Science
Jade University of Applied Science
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
tarek.el-mihoub@jade-hs.de

Lars Nolle

Department of Engineering Science
Jade University of Applied Science
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
and

German Research Centre for Artificial Intelligence
Marie-Curie-Straße 1
26129 Oldenburg, Germany
lars.nolle@jade-hs.de

Oliver Zielinski

Institute for Chemistry and Biology of the Marine
Environment

Carl von Ossietzky University of Oldenburg
Schleußenstraße 1
26382 Wilhelmshaven, Germany
and

German Research Centre for Artificial Intelligence
Marie-Curie-Straße 1
26129 Oldenburg, Germany
oliver.zielinski@uol.de

KEYWORDS

Autonomous Underwater Vehicles, Path Planning Algorithms, Travelling Salesman Problem, Optimisation, k-opt

ABSTRACT

In this paper, the concept of optimised bumblebee (BB) patterns as a search strategy for autonomous underwater vehicles (AUV) is presented. Here, an AUV is used to detect submarine groundwater discharge (SGD) in coastal areas. The optimisation of the BB paths is achieved utilising k-opt optimisation. In this research, 2-opt, 3-opt and 4-opt is used for the optimisation of the BB paths. It is shown using computer simulations that all three optimisation strategies are able to improve the search capabilities of the BB search strategy. The optimisation of the BB path shortens the length of the path to visit the waypoints generated. The saved energy can be used for exploring the search space in more detail, allowing the visit of waypoint the unoptimized BB was not able to reach. The median saved path length is 33.8 m, 43.5 m and 52.6 m for the 2-opt, 3-opt and 4-opt, respectively. The median error over 1,000 experiments of the not-optimised BB is 76.26, while the median error of the optimised BB are 71.63, 72.02 and 72.23 for the 2-opt, 3-opt and 4-opt, respectively.

INTRODUCTION

The long-term goal of this research is to develop a flexible and low-cost autonomous multi-sensor platform

for submarine exploration. Such a platform could be used for the localisation and investigation of submarine sources of interest like dumped waste, lost harmful cargo or submarine groundwater discharge (SGD) (Burnett et al. 2006). The term SGD covers any flow of water across the seabed regardless of the composition and the driving forces (Burnett et al. 2006; Moore 2010). Hence, SGD includes the discharge of fresh groundwater as well as the discharge of recirculating seawater (Figure 1). Due to the higher load of nutrients, SGD inflow can have an influence on the marine environment (Luijendijk et al. 2020).

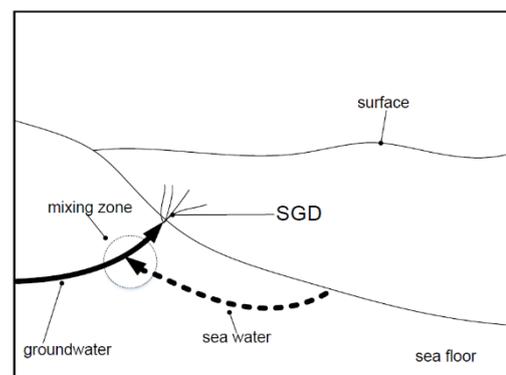


Figure 1: Submarine groundwater discharge (SGD) consisting of fresh groundwater and recirculating seawater

Different methods, such as seepage meters (Lee 1977; Taniguchi et al. 2003; Seibert et al. 2020), tracer studies

(Burnett et al. 2006), remote sensing (Mallast and Siebert 2019), or seismic surveys (Smith et al. 2003; Stieglitz and Ridd 2000; Taniguchi et al. 2019), have been utilised for SGD investigation. More recently, small unmanned underwater vehicles (UUVs) were used as a tool for SGD site investigation (Tholen et al. 2021).

UUVs can either be remotely controlled by a human pilot over a tether, i.e. remotely operated vehicle (ROV) or, without a tether, by an algorithm running on an onboard computer, i.e. autonomous underwater vehicle (AUV) (Christ and Wernli 2011).

During their missions, AUVs usually follow a pre-defined path, for instance a series of different transects defined by waypoints (Wynn et al. 2014; Marouchos et al. 2015), or an adaptive sampling strategy (Hwang et al. 2019; Mo-Bjorkelund et al. 2020) to achieve a given goal. The search capabilities could be potentially improved by incorporating artificial intelligence (AI) into the strategy. Often, AI strategies, for instance particle swarm optimisation (PSO) (Kennedy and Eberhart 1995) or ant colony optimisation (ACO) (Dorigo et al. 2006; Nolle 2008), mimic the behaviour of social entities, like schools of fish, flocks of birds, or colonies of ants, and hence are population-based.

The bumblebee (BB) search strategy has been used as search strategy to guide a small swarm of AUVs during the search for SGD sites (Tholen et al. 2022). However, in this research, only a single AUV was used during the simulations.

Bumblebee

The BB search strategy is inspired by bumblebee flight paths (Dukas and Real 1993; Philippides et al. 2013) and was developed by Hwang et al. (2020). The search strategy applies a combination of zigzag and double loops within the search space. In the first step, a specific number of waypoints is generated randomly and a path to visit all waypoints is computed. Upon arrival at a waypoint, the AUV undertakes a bow-tie shaped path with two loops. The radius of the loops r and the offset between the loops O are chosen by the operator prior to the search. The execution of the loops adds local exploitation capabilities to the search algorithm.

Due to the limited energy storage onboard of an AUV, a maximum travel distance for each AUV is defined. Therefore the maximum number of waypoints, utilising the given travel distance, are generated to maximize the search capabilities. All waypoints are chosen during a planning phase prior the search take place. During this planning phase, new waypoints are added iteratively to the paths until the expected path length is longer than the maximum travel distance of the AUV. The waypoint generation can be viewed as a travelling salesman problem (TSP) (Lawler 1995). Algorithm 1 shows pseudocode for the waypoint-planning algorithm.

Figure 2 shows an example trajectory of a single AUV utilising the BB algorithm as search strategy. It can be observed that the generated waypoints are spread over the whole search area, allowing the exploration of the entire area under investigation. However, the search strategy

does not utilise the information gained during the search to adapt the search path to exploit promising regions in more detail.

Algorithm 1: Pseudocode for waypoint creation of the bumblebee (BB) algorithm

```

1  WP = generate two random position
2  max_dist_reached = false
3  while not max_dist_reached do
4    distance = calc_travel_dist()
5    if distance < max_distance do
6      WP = [WP , random position]
7    else
8      max_dist_reached = true
9    end if
10 end while
11
12 function dist = calc_travel_dist()
13   N = number of WP
14   dist = N*4*π*r+O
15   current_point = startpoint
16   open_list = WP
17   for I = 1 : N do
18     td = vector of distances between
        current_point and all elements of
        open_list
19     n = index of min(td)
20     if td(n) > threshold do
21       dist = dist + td (n)
22       current_point = open_list(n)
23       delete open_list(n)
24     else do
25       %Drop WP, too near to other WP
26       delete open_list(n)
27     end if
28   end for
29 end function

```

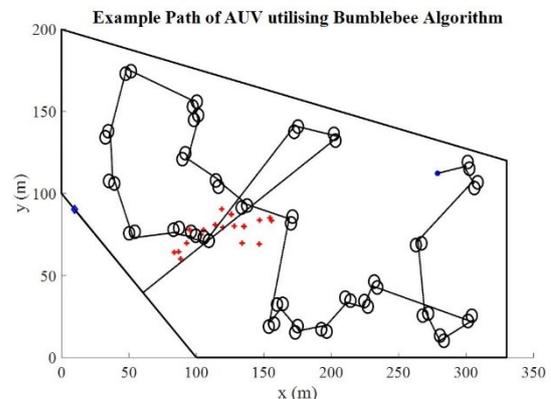


Figure 2: Example path of the BB search algorithm, start and end point of the autonomous underwater vehicle (AUV) are marked by the blue circle and diamond, respectively; Position of submarine groundwater discharges (SGDs) are marked by red crosses

During the generation of the waypoints, direct travelling in a straight line between two waypoints is assumed. However, due to self-localisation errors, the AUV is not travelling in a straight line. It rather produces zigzag paths resulting in a waste of energy between two waypoints. Therefore, not all waypoints generated can be visited (Tholen et al. 2022). Hence, in most cases, the BB strategy is not able to guide the AUV to investigate all parts of the search area, resulting in a poor search performance of the BB strategy, compared to other strategies (Tholen et al. 2022).

Optimising the path, i.e. find a shorter path to visit all waypoints, might be a suitable solution to tackle the problem described above. A shorter path will save energy which then can be used to visit more waypoints from the list of generated waypoints. This will increase the coverage of the search area and therefore potentially increase the search performance of the BB algorithm. The following research hypothesis will be addressed in this paper: “For an AUV, which utilises BB search strategy, optimising, i.e. minimizing, its path length can increase the search performance”. A positive correlation between the saved path length and the performance of the search strategy is assumed.

Different approaches to optimise TSP problems, like ACO (Dorigo et al. 2006), simulated annealing (Linhares and Torreão 2011) or k-opt heuristic (Chandra et al. 1999) were proposed in the past.

K-opt Optimisation

In this research, k-opt optimisation was used, due to the simple implementation. Other optimisation strategies, for instance simulated annealing would require additional afford for parameter tuning. To answer the research questions of this paper a simple optimisation strategy is sufficient. During the optimisation process, k points from the list of waypoints are randomly chosen. In the next step, all possible permutations of the k points are calculated. For each of the permutations, the travel length for visiting all points is calculated. The permutation of the points is kept, if this calculated travel length is shorter than the travel length without the permutation. Otherwise, the permutation is rejected. The optimisation process is repeated n times. Algorithm 2 shows pseudocode of the optimisation process described.

Algorithm 2: Pseudocode of the k-opt strategy used

```

1  for I = 1:n do
2    k_WP = select k WP randomly
3    perm_k = all permutations of k_WP
4    for j = 1:k! do
5      temp_WP = WP using perm_k(j)
6      temp_tl = calculate travel length
       with temp_WP
7      if temp_tl < best_tl do
8        best_tl = temp_tl
9        WP = temp_WP
10     end if
11   end for
12 end for

```

It can be observed from Algorithm 2 that the computational costs of the optimisation process depend on the chosen values for k and n . The number of optimisation steps can be calculated as follows:

$$S = n \cdot k!. \quad (1)$$

Where S represents the total number of optimisation steps, k represents the number of WP chosen for optimisation and n is the number or repetitions.

Simulation environment

In this research, a dynamic simulation based on a real harbour environment was used. As shown in Figure 2, the dimensions of the simulated environment are 330 m x 200 m. The environment contains 20 SGDs. The number, position, strength and composition of the SGDs are randomly selected. The simulated environment used in this research is described in detail in Tholen et al. (2022).

To measure the success of the search strategy, the following error calculation was used:

$$E = \frac{1}{n} \sum_{i=1}^n \frac{1}{\beta_i} \cdot \min(d_{1:t,i}). \quad (2)$$

Where E represents the error of the search run, n denotes the number of SGDs in the environment, β represents the flowrate coefficient of the SGD and $d_{1:t,i}$ denotes the distances between the AUV and the SGD i for all time steps $\{1 \dots t\}$ of the simulation.

This measure reflects which search strategy best fulfils the intended aims of the search. In this work, the environment contained n SGDs with different flowrates. In the best case, the search strategy would be able to guide the AUV to visit all SGDs within the given maximum travel path length. Hence, the minimum distance between the AUV and all SGDs is used for fitness evaluation. In addition, the flowrate of the different SGDs is kept into account. That means it is more important to investigate SGDs with higher inflow, rather than visiting SGDs with lower inflow.

EXPERIMENTS

To answer the research question of this work, a set of 1,000 experiments was conducted. The number of experiments was chosen as a trade-off between computing time and number of results. In each experiment, different values of $k \in \{2,3,4\}$ were investigated. In each experiment, the optimisation process was repeated for $n = 1,000$ times.

For a fair comparison, all four different options, i.e. not-optimised-BB, 2-opt-BB, 3-opt-BB and 4-opt-BB, were evaluated at the same time within the same simulated environment. Each search option is assigned to a single AUV. Here, in the first step of the setup, a set of waypoints is generated according to Algorithm 1.

The set of waypoints is directly used by the not-optimised-BB and used as starting point for the 2-opt-BB, 3-opt-BB and 4-opt-BB.

After the optimisation is finished, all four solutions are executed simultaneously within the same environment. The simulation is iteration based, while the time-lapse of each iteration step is 1 second. In each iteration, the four AUVs are moved and the performance, following equation (2) is updated. At the end of each experiment, the performance of the four options is stored for later use. This will allow for a fair comparison between the four different options.

For the experiments, the maximum travel length of the AUVs was set to 2,700 m. The radius r was set to 6 m and the offset between the circles was 3 m. The self-localisation of the AUV was error affected using a Gaussian error model with a standard deviation of 0.5 m. These values were chosen according to the findings presented in Tholen et al. (2022).

RESULTS AND DISCUSSION

The length of the optimised paths should be shorter than the length of the not-optimised path. Figure 3 shows a histogram of the path length saved by the three optimised BB compared to the not-optimised BB. Statistical parameters for the different optimised BB are summarised in Table 1. It can be observed that the average amount of path length saved is positive correlated with the value chosen for k . Hence, the chosen k -opt optimisation strategy is able to optimise the path generated by the BB algorithm presented in Algorithm 1.

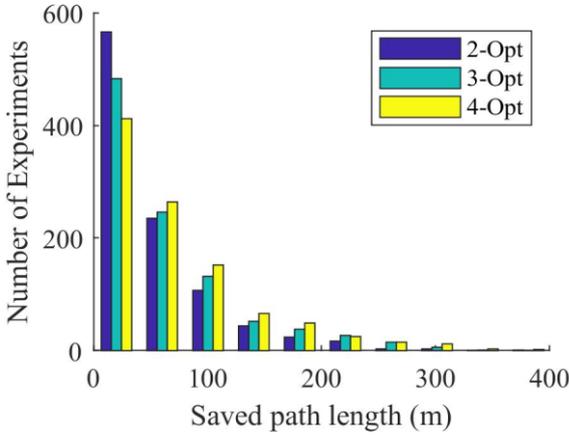


Figure 3: Histogram of path length saved by the three optimised BB compared to the not-optimised BB

Table 1: Summarised statistical parameters of the path length saved by the three optimised BB

	Strategy		
	2-Opt	3-Opt	4-Opt
Median	33.8 m	43.5 m	52.6 m
Mean	49.3 m	61.4 m	70.4 m
Standard deviation	52.8 m	62.2 m	67.5 m
Minimum	0 m	0 m	0 m
Maximum	402.0 m	328.5 m	402.0 m

The length of the paths are calculated prior the search took place. Direct movement is assumed between the waypoints. However, as mentioned above, the movement of the AUV is affected by a self-localisation error of the AUV. Therefore, in most cases, the AUVs are not able to visit all waypoints generated before the energy of the AUV is consumed. Figure 4 shows a histogram summarising the percentage of remaining waypoints for the four different BBs. The remaining waypoints, are the waypoints that cannot be visited by the AUV due to energy restrictions. It can be seen from the figure that the

optimisation process is capable of reducing the number of waypoints that the AUV was not able to visit.

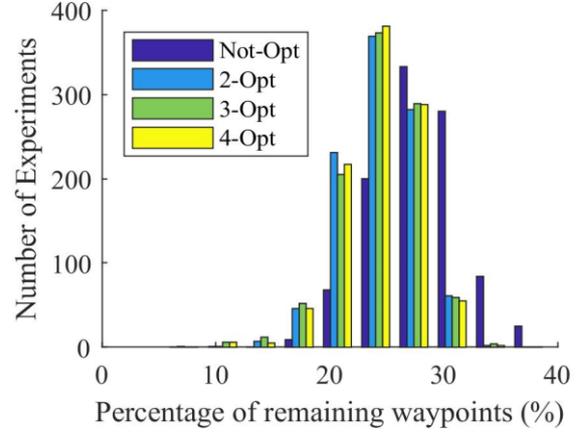


Figure 4: Histogram of the remaining waypoints for the not-optimised BB and the three optimised BB

Figure 5 shows the histogram of the error, calculated following equation (2). In Table 2 statistical parameters of the error scored by the different options are summarised. It can be seen from the figure and the table that all optimised versions of BB gave better results than the not-optimised BB. However, the results for all three optimised BB are in the same order of magnitude.

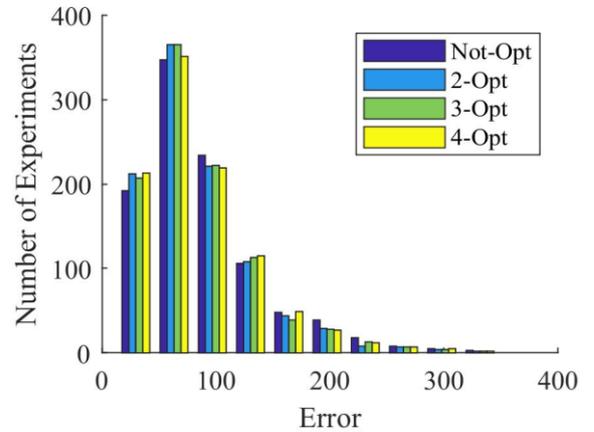


Figure 5: Histogram of error values for the not-optimised BB and the three optimised BB

Table 2: Summarised statistical parameters of the error for the not-optimised BB and the three optimised BB

	Strategy			
	Not-Opt	2-Opt	3-Opt	4-Opt
Median	76.26	71.63	72.02	72.23
Mean	88.48	83.40	83.99	84.68
Standard deviation	52.23	48.20	48.74	49.76
Minimum	13.62	14.35	14.76	12.92
Maximum	346.61	333.95	348.53	346.93

The optimised BB used the same waypoints as the not-optimised BB. Therefore, the difference in the error for each experiment ΔE_i can be calculated as follows:

$$\Delta E_i = \frac{E_{not-opt,i} - E_{opt,i}}{E_{not-opt,i}} \cdot 100. \quad (3)$$

Where $E_{opt,i}$ denotes the error value of the k-opt optimised BB and $E_{not-opt,i}$ denotes the error value of the not-optimised BB in the specific experiment i . Positive values for ΔE_i indicate a better result achieved by the optimised BB, while negative values indicate a worse result for the optimised BB compared to the not-optimised BB. Figure 6 shows the histogram of ΔE_i for all three optimised BB strategies investigated. Statistical parameters of ΔE_i are summarised in Table 3.

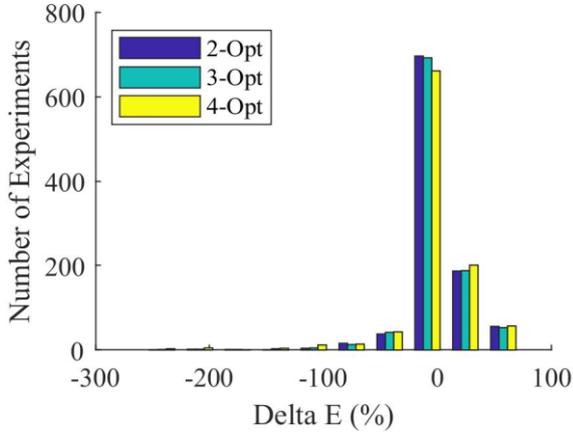


Figure 6: Histogram of ΔE_i ; positive values are representing a decrease in the error, i.e. improvement compared to not-optimised, while negative ones represent an increase respectively

Table 3: Summarised statistical parameters of ΔE_i

	Strategy		
	2-Opt	3-Opt	4-Opt
Median	1.1 %	1.0 %	1.0 %
Mean	2.2 %	1.2 %	-0.14 %
Standard deviation	24.4 %	26.6 %	32.5 %
Minimum	-219.2 %	-227.0 %	-258.2 %
Maximum	74.3 %	73.3 %	74.5 %
Better	590	550	561
Worse	410	450	439

It can be seen from the table that, based on the median, all three optimised BB performed better than the not optimised BB. However, in 41.0 %, 45.0 % and 43.9 % of the experiments, the performance of the 2-opt, 3-opt and 4-opt BB strategy is worse compared to the not optimised BB. A possible explanation for this worse performance is shown in Figure 7. In some cases, the optimisation process may change the order of the

waypoints in such a way that the new path does not cross the area of the SGD, even if the not-optimised path did.

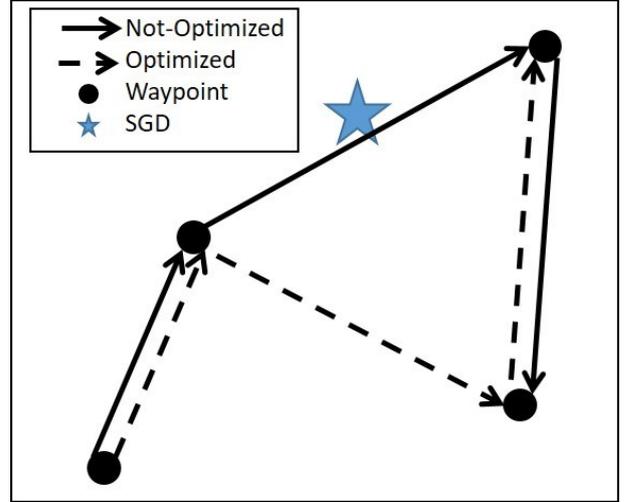


Figure 7: Possible explanation for the decrease in performance caused by optimisation; waypoints are marked by circles

If the postulated research hypothesis would be true, a positive correlation between ΔE_i and the saved path length would be expected. Figure 8 shows a scatter plot of ΔE_i over the saved path length for the three different optimised BBs. No positive correlation between the two variables can be observed for any of the optimised BB. In addition, in some experiments with a high saved path length, the achieved performance is bad compared to the performance of the not-optimised BB.

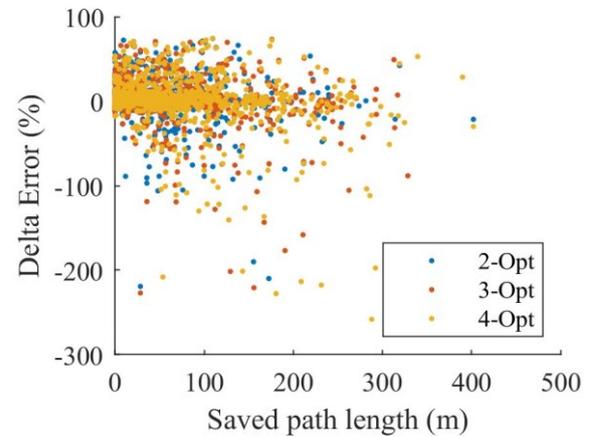


Figure 8: Scatter plot of ΔE_i over the saved path length

CONCLUSIONS AND FUTURE WORK

In this research a simple optimisation strategy, i.e. the k-opt strategy, was used to improve the search performance of an AUV utilising the BB search strategy to search for SGDs. In this research, 2-opt, 3-opt and 4-opt were used. On average, all three optimised BB outperformed the not-optimised BB. However, the achieved reward of all three optimised BB was in the same order of magnitude. Therefore, the 2-opt strategy is the best option, due to the

much lower computational costs, compared to the 3-opt and 4-opt strategy.

Only in 56.7 % of all conducted experiments the optimised BB performed better than the not-optimised BB. Therefore, nearly in every second run, the optimised BB performed worse compared to the not-optimised BB. In the worst case, the error of the 4-opt optimised path was 258.2 % inferior to the not-optimised BB. Thus, optimising the path of the AUV does not guarantee an increasing performance of the search. Another possible solution to improve the performance of the BB search would be the incorporation of feedback, gained from the environment, to guide the search towards more promising regions of the search space.

In future research, different ways for the online adaptation of the list of waypoints, based on the current state of the search, will be evaluated.

REFERENCES

- Burnett, W. C.; Aggarwal, P. K.; Aureli, A.; Bokuniewicz, H.; Cable, J. E.; Charette, M. A. et al. (2006): Quantifying submarine groundwater discharge in the coastal zone via multiple methods. In *Science of The Total Environment* 367 (2-3), pp. 498–543. DOI: 10.1016/j.scitotenv.2006.05.009.
- Chandra, Barun; Karloff, Howard; Tovey, Craig (1999): New Results on the Old k-opt Algorithm for the Traveling Salesman Problem. In *SIAM J. Comput.* 28 (6), pp. 1998–2029. DOI: 10.1137/S0097539793251244.
- Christ, Robert D.; Wernli, Robert L. (Eds.) (2011): The ROV manual. A user guide to observation-class remotely operated vehicles. 1st ed. Amsterdam, Boston, London: Elsevier Butterworth-Heinemann.
- Dorigo, Marco; Birattari, Mauro; Stutzle, Thomas (2006): Ant colony optimization. In *IEEE Comput. Intell. Mag.* 1 (4), pp. 28–39. DOI: 10.1109/MCI.2006.329691.
- Dukas, Reuven; Real, Leslie A. (1993): Effects of recent experience on foraging decisions by bumble bees. In *Oecologia* 94 (2), pp. 244–246. DOI: 10.1007/BF00341323.
- Hwang, Jimin; Bose, Neil; Fan, Shuangshuang (2019): AUV Adaptive Sampling Methods. A Review. In *Applied Sciences* 9 (15), p. 3145. DOI: 10.3390/app9153145.
- Hwang, Jimin; Bose, Neil; Nguyen, Hung Duc; Williams, Guy (2020): Acoustic Search and Detection of Oil Plumes Using an Autonomous Underwater Vehicle. In *JMSE* 8 (8), p. 618. DOI: 10.3390/jmse8080618.
- Kennedy, J.; Eberhart, R. (1995): Particle swarm optimization. In : 1995 IEEE International Conference on Neural Networks. Proceedings, the University of Western Australia, Perth, Western Australia, 27 November-1 December 1995. ICNN'95 - International Conference on Neural Networks. Perth, WA, Australia, 27 Nov.-1 Dec. 1995. IEEE International Conference on Neural Networks; IEEE Neural Networks Council. New York, Piscataway, NJ: Institute of Electrical and Electronics Engineers, pp. 1942–1948.
- Lawler, Eugene L. (Ed.) (1995): The traveling salesman problem. A guided tour of combinatorial optimization. Reprint. Chichester: Wiley (Wiley-interscience series in discrete mathematics and optimization).
- Lee, David Robert (1977): A device for measuring seepage flux in lakes and estuaries. In *Limnol Oceanogr* 22 (1), pp. 140–147. DOI: 10.4319/lo.1977.22.1.0140.
- Linhares, Alexandre; Torreão, José R. A. (2011): Microcanonical Optimization Applied to the Traveling Salesman Problem. In *Int. J. Mod. Phys. C* 09 (01), pp. 133–146. DOI: 10.1142/S012918319800011X.
- Luijendijk, E.; Gleeson, T.; Moosdorf, N. (2020): Fresh groundwater discharge insignificant for the world's oceans but important for coastal ecosystems. In *Nature communications* 11 (1), p. 1260. DOI: 10.1038/s41467-020-15064-8
- Mallast, Ulf; Siebert, Christian (2019): Combining continuous spatial and temporal scales for SGD investigations using UAV-based thermal infrared measurements. In *Hydrol. Earth Syst. Sci.* 23 (3), pp. 1375–1392. DOI: 10.5194/hess-23-1375-2019.
- Marouchos, Andreas; Muir, Brett; Babcock, Russ; Dunbabin, Matthew (2015): A shallow water AUV for benthic and water column observations. In : OCEANS 2015 - Genova: IEEE.
- Mo-Bjorkelund, Tore; Fossum, Trygve O.; Norgren, Petter; Ludvigsen, Martin (2020): Hexagonal Grid Graph as a Basis for Adaptive Sampling of Ocean Gradients using AUVs. In : Global Oceans 2020: Singapore – U.S. Gulf Coast. Global Oceans 2020: Singapore - U.S. Gulf Coast. Biloxi, MS, USA: IEEE, pp. 1–5.
- Moore, Willard S. (2010): The effect of submarine groundwater discharge on the ocean. In *Annual review of marine science* 2, pp. 59–88. DOI: 10.1146/annurev-marine-120308-081019.
- Nolle, Lars (2008): On a Novel ACO-Estimator and its Application to the Target Motion Analysis Problem. In Richard Ellis, Tony Allen, Miltos Petridis (Eds.): Applications and Innovations in Intelligent Systems XV. London: Springer London, pp. 3–16.
- Philippides, Andrew; Ibarra, Natalie Hempel de; Riabinina, Olena; Collett, Thomas S. (2013): Bumblebee calligraphy. The design and control of flight motifs in the learning and return flights of *Bombus terrestris*. In *Journal of Experimental Biology* 216 (6), pp. 1093–1104. DOI: 10.1242/jeb.081455.
- Siebert, Stephan Ludger; Degenhardt, Julius; Ahrens, Janis; Reckhardt, Anja; Schwalfenberg, Kai; Waska, Hannelore (2020): Investigating the Land-Sea Transition Zone. In Simon Jungblut, Viola Liebich, Maya Bode-Dalby (Eds.): YOUNG MARES 9 - The Oceans: Our Research, Our Future. Proceedings of the 2018 conference for YOUNG MARINE RESEARCHER in Oldenburg, Germany. 1st ed. 2020. Cham: Springer International Publishing, pp. 225–242.
- Smith, A. J.; Herne, D. E.; Hick, W. P.; Turner, J. V. (2003): Quantifying submarine groundwater discharge and nutrient discharge into Cockburn Sound Western Australia. A technical report to the Coast and Clean Seas Project WA9911 : quantifying submarine groundwater discharge and demonstrating innovative clean-up to protect Cockburn Sound from nutrient discharge. Wembley, Western Australia: CSIRO (CSIRO Land and Water technical report, no. 01/03).
- Stieglitz, T.; Ridd, P. (2000): Submarine Groundwater Discharge from Paleochannels?: “Wonky Holes” on the Inner Shelf of the Great Barrier Reef. In Australia Institution of Engineers (Ed.): Hydro 2000: Interactive Hydrology; Proceedings. 3rd International Hydrology and Water Resources Symposium of the Institution of Engineers, Australia ; 20 - 23 November 2000, Sheraton Perth Hotel, Perth, Western Australia. International Hydrology and Water Resources

Symposium (3rd : 2000 : Perth, W.A.). Perth. Institution of Engineers, Australia. Barton, A.C.T.: Institution of Engineers, Australia, pp. 189–194.

Taniguchi, Makoto; Burnett, William C.; Smith, Christopher F.; Paulsen, Ronald J.; O'Rourke, Daniel; Krupa, Steve L.; Christoff, Jamie L. (2003): Spatial and temporal distributions of submarine groundwater discharge rates obtained from various types of seepage meters at a site in the Northeastern Gulf of Mexico. In *Biogeochemistry* 66 (1/2), pp. 35–53. DOI: 10.1023/B:BIOG.0000006090.25949.8d.

Taniguchi, Makoto; Dulai, Henrietta; Burnett, Kimberly M.; Santos, Isaac R.; Sugimoto, Ryo; Stieglitz, Thomas et al. (2019): Submarine Groundwater Discharge. Updates on Its Measurement Techniques, Geophysical Drivers, Magnitudes, and Effects. In *Front. Environ. Sci.* 7, p. 335. DOI: 10.3389/fenvs.2019.00141.

Tholen, Christoph; El-Mihoub, Tarek A.; Nolle, Lars; Zielinski, Oliver (2022): Artificial Intelligence Search Strategies for Autonomous Underwater Vehicles Applied for Submarine Groundwater Discharge Site Investigation. In *JMSE* 10 (1), p. 7. DOI: 10.3390/jmse10010007.

Tholen, Christoph; Parnum, Iain; Rofallski, Robin; Nolle, Lars; Zielinski, Oliver (2021): Investigation of the Spatio-Temporal Behaviour of Submarine Groundwater Discharge Using a Low-Cost Multi-Sensor-Platform. In *JMSE* 9 (8), p. 802. DOI: 10.3390/jmse9080802.

Wynn, Russell B.; Huvenne, Veerle A.I.; Le Bas, Timothy P.; Murton, Bramley J.; Connelly, Douglas P.; Bett, Brian J. et al. (2014): Autonomous Underwater Vehicles (AUVs). Their past, present and future contributions to the advancement of marine geoscience. In *Marine Geology* 352, pp. 451–468. DOI: 10.1016/j.margeo.2014.03.012.

AUTHOR BIOGRAPHIES

CHRISTOPH THOLEN graduated from Jade University of Applied Science in Wilhelmshaven, Germany, with a MEng in mechanical engineering in 2015. He received his doctoral degree in 2022 from the Carl von Ossietzky University of Oldenburg. From 2016 to 2022, he worked on a joint project between the Jade University of Applied Science and the Institute for Chemistry and Biology of the Marine Environment (ICBM), at the Carl von Ossietzky University of Oldenburg for the development of a low cost and intelligent environmental observatory. Since 2022 he is a researcher with Marine Perception research department at the German Research Centre for Artificial Intelligence (DFKI). His current research interests including the application of marine robotics

for SGD investigation and Artificial Intelligence applied to the maritime context.

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. He is also affiliated with the Marine Perception research group of the German Research Centre for Artificial Intelligence (DFKI). His main research interests are AI and computational optimisation methods for real-world scientific and engineering applications.

TAREK A. EL-MIHOUB graduated with a BSc in computer engineering from University of Tripoli, Tripoli, Libya. He obtained his MSc in engineering multimedia and his PhD in computational intelligence from Nottingham Trent University in the UK. He was an assistant professor at the Department of Computer Engineering, University of Tripoli. He is currently a postdoctoral researcher with Jade University of Applied Science. His current research is in the fields of applied computational intelligence

OLIVER ZIELINSKI is head of the research group “Marine Sensor Systems” at the Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University of Oldenburg. He is also heading the research department Marine Perception at the German Research Centre for Artificial Intelligence (DFKI). After receiving his Ph.D. degree in Physics in 1999 from University of Oldenburg, he moved to industry where he became scientific director and CEO of “Optimare Group,” an international supplier of marine sensor systems. In 2005, he was appointed Professor at the University of Applied Science in Bremerhaven, Germany. He returned to the Carl von Ossietzky University of Oldenburg in 2011. His area of research covers marine optics and marine physics, with a special focus on coastal systems, marine sensors, and operational observatories involving different user groups and stakeholders.

EMISSION REDUCTION THROUGH PRODUCTION SCHEDULING BY PRIORITY RULES AND ENERGY ONSITE GENERATION

Hajo Terbrack
Thorsten Claus
Technische Universität Dresden
International Institute (IHI) Zittau
Markt 23, 02763 Zittau, Germany
Email: hajo.terbrack@mailbox.tu-dresden.de

Frank Herrmann
Ostbayerische Technische Hochschule Regensburg
Innovation and Competence Centre for Production
Logistics and Factory Planning (IPF)
P.O. Box 12 03 27, 93025 Regensburg, Germany

KEYWORDS

Sustainable Production Planning, Energy-related Emissions, Job Shop Scheduling, Simulation

ABSTRACT

This article describes primary findings of various simulation runs on job shop scheduling dealing with energy consumption and emission pollution. By two combinations of priority rules, production is linked to the generation output of a renewable energy source installed on-site. The resulting schedules show a reduction of energy-related emissions and makespan compared to several conventional priority rules often used in industrial practice.

INTRODUCTION

Climate change, resource scarcity and the associated costs are causing companies to pay more attention to sustainability. Along with this, also increasing social interest as well as legal and structural framework conditions are reinforcing the need for companies to act more sustainably.

For manufacturing companies, enormous potential for improvement in terms of sustainability is given in production. This potential can be addressed through sustainable production planning and control, for example by planning methods associated with the concept of hierarchical production planning (regarding this, see Herrmann and Manitz (2021)). In addition to classic economic objectives, ecological and social indicators are increasingly being taken into account in production planning (Trost et al. (2019)). A comprehensive listing of a large number of articles in the context of sustainable production planning can be found in the online literature database on sustainable production planning, developed by the authors' research group (see Terbrack et al. (2020), Terbrack et al. (2021c)).

In the field of ecologically oriented production planning, the integration of energy in particular has attracted enormous interest in recent years. Numerous scientific papers in the context of production planning already address energy issues in the form of different objectives and restrictions. Especially the integration of different energy sources and energy storage systems within production planning is increasing. Some of these articles address energy-related emissions as well, although the emphasis is mainly on reducing energy costs and energy consumption so far (Bänsch et al. (2021), Terbrack et al. (2021b)).

However, current developments show that the reduction of emissions is becoming increasingly important as well. Germany is pursuing the goals of reducing greenhouse gas emissions by 65 % by 2030 in relation to the year 1990 and achieving climate neutrality by 2045

(BKSG (2021)). Moreover, the Paris Agreement aims to achieve global greenhouse gas neutrality in the second half of the 21st century (UN (2015)). A need for action can be inferred from these goals, also for the manufacturing industry.

For these reasons, the article at hand aims to make a contribution to the consideration of emissions, in specific energy-related emissions, in the context of production scheduling.

By extending our preliminary work in Terbrack et al. (2021a), the application of commonly used priority rules in a job shop environment combined with renewable energy onsite generation is investigated by means of a simulation study. Two different combinations of priority rules are discussed which allow production to be adjusted to self-generated energy and thus a reduction in energy-related emissions. The resulting production schedules are analyzed in terms of makespan, consumption of electrical energy and emissions.

The remainder of the paper is structured as follows. The next chapter presents the literature review and the problem definition. In chapter 3, the design of the simulation study is outlined, followed by the results and their discussion. The article ends with a conclusion and an outlook on further proposed research.

LITERATURE REVIEW AND PROBLEM DEFINITION

In general, production scheduling aims at an optimal allocation of jobs to the respective machines required for processing. Therefore, it is determined at which time, on which machine and in which order each job is processed. Common objective criteria are for example the minimization of completion time or tardiness (Herrmann (2009)).

Nonetheless, some scientific articles already take into account the minimization of energy-related emissions within production scheduling. For example Wang et al. (2019) consider emissions in a single machine scheduling and vehicle routing problem. Guo et al. (2020) present a flow shop scheduling approach to minimize energy-related emissions, makespan and noise. In Foumani and Smith-Miles (2019), a flow shop scheduling problem is introduced to minimize both, emission quantity and costs, as well as makespan. For a flexible job shop environment, Coca et al. (2019) address the minimization of emissions and energy consumption costs along with several economic and social indicators as total completion time, water consumption, penalties for waste and others.

Several approaches from the literature argue that a significant share of electrical energy is generated by fossil fuels as coal or gas and the efficient utilization of energy offers enormous potential for emission reduction. Following that, the reduction of energy-related emissions in

short-term production planning is achieved by minimizing total energy consumption (e.g. Ding et al. (2016)). However, the decrease of energy-related emissions does not necessarily correlate with a reduction of total energy consumption. Thus, another approach lies in the utilization of renewable energy – generated onsite and with zero emissions, as for example through photovoltaic systems and wind turbines (Wu et al. (2018)).

By using renewable energy sources like photovoltaic (PV) systems, a certain proportion of total electricity consumption in production can be covered by emission-free electricity. Yet the amount of solar power generated depends strongly on different factors such as the time of day, the hours of sunshine and the season. Therefore, companies with photovoltaic systems have different amounts of solar electricity available every day. Due to this correlation between time and the amount of solar electricity generated, it is important to consider that the generation of emission-free electricity can fluctuate over the course of the day. In this manner, for instance Liu (2016) presents two optimization models for single machine scheduling, both taking such renewable energy uncertainty into account while addressing total weighted flow time and energy-related emissions.

Besides optimization, metaheuristics and simple heuristics like priority rules are used for production scheduling. For the latter, basically, the jobs released for processing are queued in buffers in front of the individual machines. As soon as a machine is free and ready for operation, the job with the highest priority from the corresponding queue is assigned to the machine for processing (Herrmann (2011)).

Due to the fact that especially heuristics such as priority rules are used in industrial practice, the paper at hand further analyses the application of conventional priority rules for job shop scheduling combined with a photovoltaic system and macrogrid procurement. For this, we

present two combinations of conventional priority rules that address energy onsite generation. To the best of the authors' knowledge, such an approach has not yet been further discussed in research.

SIMULATION STUDY

We continue the work described in Terbrack et al. (2021a) as presented in the following. The simulation study is performed in Plant Simulation, version 2201. The simulation layout is shown in figure 1. A job shop is considered as the underlying shopfloor layout, consisting of five different production machines ("M1" to "M5"). In front of each machine, a buffer is placed representing the queue. For each of those machines, the machine states "working", "setting up", "operational", "standby" and "off" as well as the corresponding state transitions "off → operational", "operational → off", "operational → standby", "standby → operational" and "standby → off" are modelled. Power demand takes a value between 0 and 7 kW for each machine state and follows this relation:

$$P_{\text{working}} > P_{\text{setting up}} > P_{\text{operational}} > P_{\text{standby}} \geq P_{\text{off}}$$

These power values are assumed to be constant for each state. In following the EnergyBlocks methodology as introduced in Weinert et al. (2011), this means that each machine state equals one energy block.

In the job shop, three different products (P1, P2, P3) are processed whereby P1 and P3 undergo seven processing steps and P2 passes eight operations, each in different order. The product- and machine-specific setup and processing times range between 5 and 17 minutes. In total, 9 jobs with production quantities between 1 and 4 units are processed and are released at the beginning of the work day (06.00 am).

In performing production scheduling by the application of priority rules, the jobs in every queue are sorted according to the respective priority indicator every time

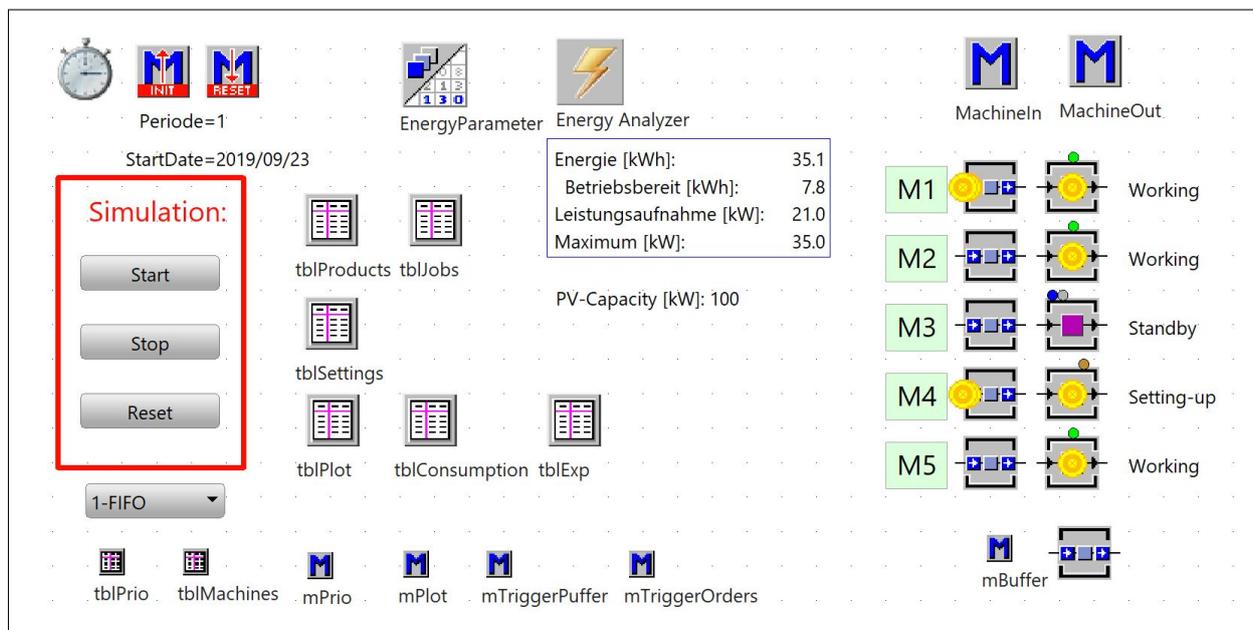


Figure 1: Layout of the simulation model (see Terbrack et al. (2021a)).

a production machine has finished an operation and the job just processed is passed on to the next machine in the shopfloor. Then, the job with the highest priority score is processed next. In every simulation run, job scheduling is carried out through the following six priority rules and by two combinations of these as described in more detail in the next chapter.

- FIFO, LIFO: First in, first out (FIFO) and Last in, first out (LIFO) priority rules. The jobs in a queue are sorted according to the length of the waiting time in descending (FIFO) or ascending (LIFO) order.
- KOZ, LOZ: shortest processing time (KOZ) and longest processing time (LOZ) priority rules. The priority of a job is determined by the length of the processing time on the machine. The highest priority is given to the job with the shortest processing time (KOZ, equiv. SPT) or the longest processing time (LOZ, equiv. LPT).
- KRB, GRB: shortest remaining processing time (KRB) and largest remaining processing time (GRB) priority rules. The priority of a job is determined depending on the remaining processing time of the outstanding operations. According to the GRB rule, the job with the largest remaining processing time receives the highest priority – according to the KRB rule, the job with the shortest remaining processing time.
- Two combinations: LOZ-KOZ-s and GRB-KRB-s – combinations of the LOZ and KOZ rules respectively GRB and KRB priority rules. Both combinations depend on a threshold value s , which relates on the generated power of a renewable energy source.

To determine the energy demand and energy consumption in the shopfloor, the PlantSimulation tool "Energy-Analyzer" is used. Note that in our study, solely production machines demand energy while indirect energy consumption for example caused by transportation between the machines and HVAC (heating, ventilation, air conditioning) is neglected since it is out of the scope of our current research.

Energy is provided by two sources: a renewable energy source and the macrogrid. A photovoltaic system supplies electricity as renewable and thus emission-free energy. Based on surveys conducted by our industry partner, a capacity of 100 kW and 30 degrees south orientation are assumed for this PV system. To analyze the influence of variable solar energy supply, five different generation profiles of the PV system are considered as weather scenarios and each production schedule is assessed with each weather scenario. For this, the data for five typical days in September in Regensburg (Germany) are taken from PV*SOL®. A graphical illustration of the weather scenarios is shown as figure 2.

As long as the PV system can meet the energy demand caused in production, zero energy is supplied from the macrogrid. However, as soon as the shopfloor's energy demand exceeds the generation of the photovoltaic system, energy is procured from the macrogrid. While the surplus power of the PV system is not further addressed in this study (e.g. by means of energy storage systems or feed-in possibilities), the excess in energy demand above the generated solar energy results in energy-related emissions. In that sense, the procurement of energy from the macrogrid is assumed to cause energy-related emissions as further discussed in the following.

For each kWh supplied by the macrogrid, a constant conversion factor equal to 401 gCO₂/kWh is used to calculate the associated emissions. This value is based on a study by Icha and Kuhs (2019) and represents the average emission factor for the German energy mix in 2019. Although we are aware of the fact that in reality, the conversion factor depends on the current energy mix and therefore varies over time and depending on location, a large share of research approaches that include energy-related emissions in production scheduling consider a constant emission factor in their studies, as for example Jiang et al. (2017), Piroozfard et al. (2018) and Zheng and Wang (2018). Therefore, we conclude that the considered constant conversion factor equal to 401 gCO₂/kWh is sufficient for our study.

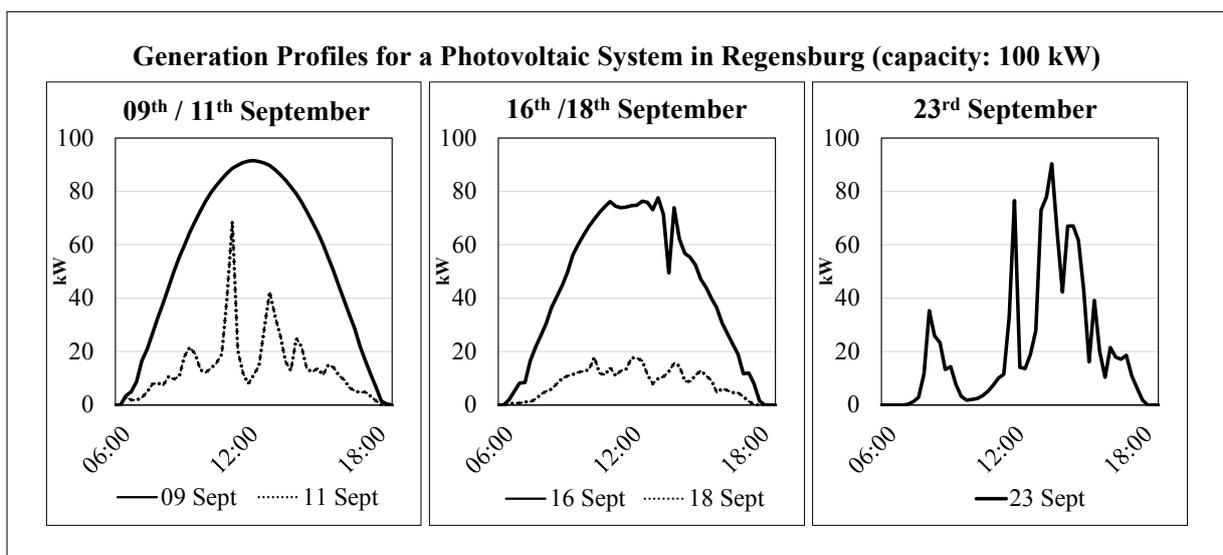


Figure 2: Generation profiles for the photovoltaic system used in the simulation study: data for five September days in Regensburg, Germany.

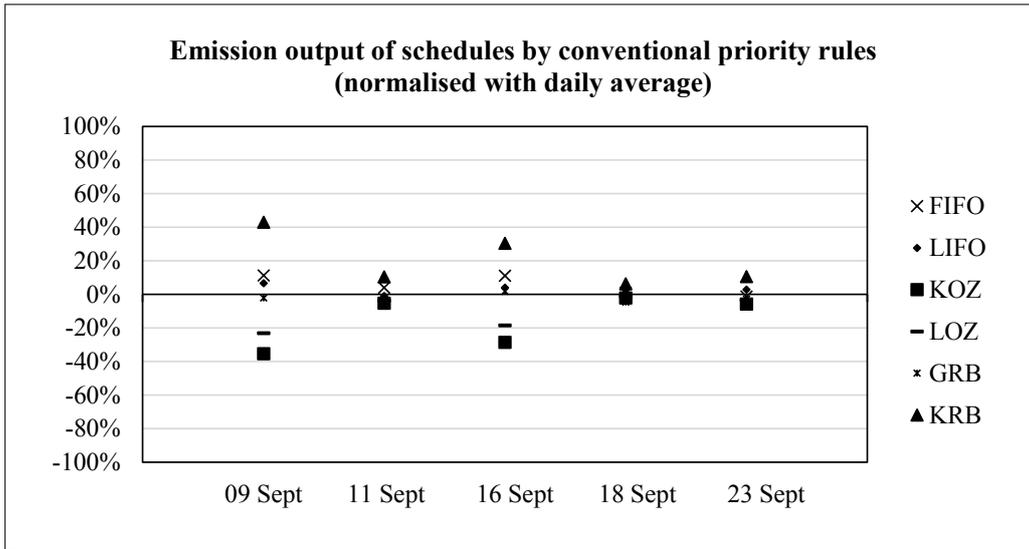


Figure 3: Variation in the amount of energy-related emissions associated with the production schedules based on conventional priority rules.

RESULTS

Previous results

We were able to show in Terbrack et al. (2021a) that the amount of energy-related emissions associated with production can relate to the production schedule. Besides differences in makespan and energy consumption of the production schedules depending on the respective priority rule, the emission output in production differed in respect to the choice of the priority rule as well. This is summarized in figure 3, which shows the emission output of the different production schedules for each weather scenario in relation to the day-specific average of the six schedules. As can be seen, there are weather scenarios in which the emission output differs strongly based on the chosen priority rule. For example on September 9th, the emission output by the KRB priority rule is 43% higher than the day average of 7.44 kgCO₂ while the production schedule received through the KOZ priority rule leads to an amount of energy-related emissions 35% less than the average. For some of the weather scenarios illustrated in figure 3, the differences between the production schedules' emission output might not seem large. However, it should be noted that even small savings in production-related emissions are beneficial in respect to ecological sustainability.

It was further recognized that in four of five weather scenarios, even the production based on the schedules with the lowest emissions still does not utilize the entire amount of generated solar energy, representing an improvement potential by increasing the utilisation level of the photovoltaic system.

While the presented conventional priority rules do not take into account the amount of onsite generated energy in production scheduling, a combination of two priority rules – KOZ and LOZ – was introduced in the above cited article that considers the renewable energy source to some extent. This combination relies on the choice of the KOZ priority rule as long as the average generation power of the photovoltaic system within the next hour is less than a specific threshold value s multiplied with the nominal capacity of the PV system (100 kW). Otherwise, in case the average generation of the next hour is equal to or higher than the product of s and the nominal capacity of the PV system, the LOZ priority rule is chosen. Note that the generation output of the PV system for the next hour is assumed to be known.

With this heuristic, production schedules with reduced emissions compared to the conventional priority rules were achieved for each of the five weather scenarios. In fact, also the makespan was reduced by this combination. These findings are reflected in table 1.

Table 1: Results of the first combination of priority rules compared to the conventional priority rules (PR) with lowest makespan and lowest emissions in each weather scenario: makespan (MS), emission quantity (E) and threshold value (s) are listed. Note that in two weather scenarios, an interval for s is given.

Weather Scenario	PR with lowest makespan			PR with lowest emissions			LOZ-KOZ- s		
	Rule	MS [h]	E [kg]	Rule	MS [h]	E [kg]	s	MS [h]	E [kg]
09 Sept	LOZ	12.78	5.71	KOZ	13.02	4.81	[0.64; 0.76]	12.42	2.07
11 Sept	LOZ	12.78	33.95	KOZ	13.02	33.42	0.16	12.20	28.90
16 Sept	LOZ	12.78	8.04	KOZ	13.02	7.04	[0.50; 0.60]	12.42	4.23
18 Sept	LOZ	12.78	50.51	GRB	12.87	48.71	0.13	12.73	47.60
23 Sept	LOZ	12.78	32.28	KOZ	13.02	31.49	0.33	12.73	28.78

Scheduling by a second combination of priority rules: GRB-KRB-s

As described above, we were able to reduce energy-related emissions and makespan in every weather scenario by applying the LOZ-KOZ-s heuristic. However, we concluded further optimization potential due to the fact that electricity was procured from the macrogrid in every weather scenario although the photovoltaic system's output exceeded the production's consumption of solar energy in four of five weather scenarios. Therefore, we extended our research approach as described next.

Based on a similar idea as for the LOZ-KOZ-s heuristic – namely that a longer processing duration results in higher energy consumption since the energy demand of a production machine remains the same for every job – we repeated our simulation runs with a new combination of priority rules. The combination of the GRB priority rule and the KRB priority rule was tested, whereby the choice of priority rule depends on a threshold value s and the generation output of the PV system.

In specific, the jobs within the queue in front of a production machine are sorted in ascending order of remaining processing time as long as the average PV output within the next hour is less than the threshold value s multiplied with the nominal capacity of the photovoltaic system. So, when a machine finished an operation and is available for processing the next job, the one with the lowest remaining processing time is next. In reverse, as soon as the the average generation power of the PV system of the next hour is greater than or equal to the threshold value s multiplied with the photovoltaic generation capacity, the job with the longest remaining processing time is processed next. To summarize our attempt, scheduling by the new combination GRB-KRB-s follows this logic:

if $\overline{PV}_{Power}^{next\ hour} < s \cdot PV_{Capacity}$ then

Apply KRB priority rule

else Apply GRB priority rule

We conducted 101 simulation runs for each weather scenario while increasing the threshold value s iteratively by 0.01, starting at $s = 0$ and ending at $s = 1$. By this, different s values were derived that lead to a reduction in makespan and emissions compared to the conventional priority rules. Out of these 505 runs, the results of the simulation runs with the most suitable s values

per weather scenario are stated in table 2. In relation to the schedules of the conventional priority rules with the lowest makespan and the lowest emissions, this second combination of priority rules achieves improvements in terms of economic and ecological manner as well. In every weather scenario, the makespan of the schedules by the GRB-KRB-s heuristic is lower than the minimal makespan of the conventional priority rules, equal to 12.78 h. Moreover, the emission quantity is reduced in each scenario, between 3.1% (18 Sept) and 60.3% (09 Sept).

However, as can be inferred from table 2, the value of s is essential to obtain these favourable results. For example for the weather scenario September 11th, the stated results are only achieved for a threshold value s equal to 0.14. Similarly, the intervals of the s values in the other scenarios are rather small. To some extent, the same holds true for our first heuristic LOZ-KOZ-s. Based on the data presented in table 3, we discuss both heuristics, LOZ-KOZ-s and GRB-KRB-s, in the following.

DISCUSSION

According to our results, both heuristics provide schedules with lower makespan and energy-related emissions than the considered conventional priority rules. Since the two combinations LOZ-KOZ-s and GRB-KRB-s take into account onsite generated solar energy to some extent, the amount of energy procured from the macrogrid is reduced.

In comparing both heuristics, it can be observed from the results in table 3 that the two heuristics yield to different values for makespan and emission quantity. The first combination, LOZ-KOZ-s, achieves a lower makespan in the first three weather scenarios and a lower emission quantity in the second and fifth weather scenario. Consequently, the GRB-KRB-s heuristic delivers better results for makespan in the fourth and fifth weather scenario as well as lower emission quantity in the first, third and fourth weather scenario.

These reduced values in emission output are based either on differences in total energy consumption or due to differences in consumption of generated solar energy and the amount of procured energy from the macrogrid. The former is the case for the GRB-KRB-s combination on September 16th for example. Although the consumption of solar energy is higher for the LOZ-KOZ-s schedule and therefore, the amount of unused solar energy is lower, the GRB-KRB-s schedule results in less energy-related emissions for that day because it causes a lower

Table 2: Results of the second combination of priority rules compared to the conventional priority rules (PR) with lowest makespan and lowest emissions in each weather scenario: makespan (MS), emission quantity (E) and threshold value (s) are listed. Note that in four weather scenarios, an interval for s is given.

Weather Scenario	PR with lowest makespan			PR with lowest emissions			GRB-KRB-s		
	Rule	MS [h]	E [kg]	Rule	MS [h]	E [kg]	s	MS [h]	E [kg]
09 Sept	LOZ	12.78	5.71	KOZ	13.02	4.81	[0.53; 0.59]	12.45	1.91
11 Sept	LOZ	12.78	33.95	KOZ	13.02	33.42	0.14	12.47	29.55
16 Sept	LOZ	12.78	8.04	KOZ	13.02	7.04	[0.48; 0.50]	12.47	3.12
18 Sept	LOZ	12.78	50.51	GRB	12.87	48.71	[0.09; 0.10]	12.45	47.20
23 Sept	LOZ	12.78	32.28	KOZ	13.02	31.49	[0.08; 0.12]	12.25	29.38

Table 3: Comparison of the two outlined combinations LOZ-KOS-s and GRB-KRB-s: threshold value (s), makespan (MS), total energy consumption (TEC), consumption of generated solar energy (SEC), amount of unused solar energy (USE), energy procurement from the macrogrid (EPG) and emission quantity (E) are stated.

Weather Scenario	Heuristic	s	MS [h]	TEC [kWh]	SEC [kWh]	USE [kWh]	EPG [kWh]	E [kg]
09 Sept	LOZ-KOZ-s	[0.64; 0.76]	12.42	225.50	220.33	472.85	5.17	2.07
	GRB-KRB-s	[0.53; 0.59]	12.45	222.75	217.98	475.21	4.77	1.91
11 Sept	LOZ-KOZ-s	0.16	12.20	222.00	149.92	28.01	72.08	28.90
	GRB-KRB-s	0.14	12.47	220.75	147.05	30.88	73.70	29.55
16 Sept	LOZ-KOZ-s	[0.50; 0.60]	12.42	225.50	214.95	341.62	10.55	4.23
	GRB-KRB-s	[0.48; 0.50]	12.47	220.75	212.98	343.59	7.77	3.12
18 Sept	LOZ-KOZ-s	0.13	12.73	223.75	105.04	0	118.71	47.60
	GRB-KRB-s	[0.09; 0.10]	12.45	222.75	105.04	0	117.71	47.20
23 Sept	LOZ-KOZ-s	0.33	12.73	225.75	153.98	129.88	71.77	28.78
	GRB-KRB-s	[0.08; 0.12]	12.25	226.75	153.47	130.39	73.28	29.38

total energy consumption. Conversely, on September 11th for instance, scheduling by the GRB-KRB-s combination achieves a higher amount of emissions than LOZ-KOZ-s even though the total energy consumption is lower than that of the first combination. Rather, the reason for this lies in the higher utilization of the generated solar energy by the LOZ-KOZ-s heuristic which is reflected by the corresponding values for SEC and USE in table 3.

Based on the present data, it appears that the first heuristic uses the onsite generated energy to a higher proportion in most cases, while the second heuristic achieves lower total energy consumption. Moreover, the comparison of the two combinations of priority rules shows that there is a very high dependence on the weather data and the threshold values when choosing the appropriate heuristic.

Nonetheless, the two heuristics only provide good results under certain conditions. As such, none of the outlined combination of priority rules leads to superior results in all weather scenarios. Consequently, further adjustments are necessary to achieve more sufficient results. For instance, this could be realised in terms of an optimization model combined with the simulation of weather scenarios.

CONCLUSION AND OUTLOOK

Throughout this article, we presented different production scheduling approaches with regard to energy-related emissions. For a job shop environment with a renewable energy source onsite, several simulation runs were performed and two combinations of conventional priority rules were analyzed in terms of makespan, energy consumption and emission output. By addressing the varying energy generation to some extent, both heuristics were able to reduce emissions in production. The outlined research expands previous work and provides additional insights on the relation between total energy consumption, the utilization of renewable energy and the associated emissions in production scheduling.

Regarding the objective of reducing energy-related emissions by production scheduling, further research is planned for the future. For a large share of the production schedules stated, there was still a high proportion of solar energy that remained unused. An increased flexibility in production, e.g. by heterogeneous production ma-

chines or speed scaling strategies, could lead to a higher utilisation of such renewable energy sources. By taking into account variations in the energy demand of parallel machines for instance, production scheduling could align jobs to the production machine with higher energy demand especially in periods of high solar energy supply. Vice versa, less jobs would be scheduled for processing by this machine in periods with lower power output of the photovoltaic system. Moreover, the adjustment of the machine speed and thus the energy demand depending on the renewable energy supply could serve as a further improvement option. In this context, modifying the prediction horizon for the PV generation output may bring further benefits. Besides that, a consideration of time-varying emission factors could help to align research closer to reality.

Since these extensions are unlikely to be representable by simple heuristics such as priority rules, these issues should be addressed by means of a multi-criteria optimization model. With this, a sufficient balance between economic and ecological objectives could be derived and the indicated trade-off between reduced total energy consumption and increased consumption of renewable energy could be further investigated. Furthermore, a practical approach might lie in single machine scheduling, as soon as individual larger energy consumers occur in the shopfloor.

REFERENCES

- Bänsch, K., Busse, J., Meisel, F., Rieck, J., Scholz, S., Volling, T., and Wichmann, M. G. (2021). Energy-aware decision support models in production environments: A systematic literature review. *Computers & Industrial Engineering*, 159:107456.
- BKSG (2021). Bundes-Klimaschutzgesetz vom 12. Dezember 2019 (BGBl. I S. 2513), das durch Artikel 1 des Gesetzes vom 18. August 2021 (BGBl. I S. 3905) geändert worden ist [German Climate Protection Act as changed in 2021].
- Coca, G., Castrillón, O. D., Ruiz, S., Mateo-Sanz, J. M., and Jiménez, L. (2019). Sustainable evaluation of environmental and occupational risks scheduling flexible job shop manufacturing systems. *Journal of Cleaner Production*, 209:146–168.
- Ding, J.-Y., Song, S., and Wu, C. (2016). Carbon-efficient scheduling of flow shops by multi-objective optimization.

- European Journal of Operational Research*, 248(3):758–771.
- Foumani, M. and Smith-Miles, K. (2019). The impact of various carbon reduction policies on green flowshop scheduling. *Applied Energy*, 249:300–315.
- Guo, H., Li, J., Yang, B., Mao, X., and Zhou, Q. (2020). Green scheduling optimization of ship plane block flow line considering carbon emission and noise. *Computers & Industrial Engineering*, 148:106680.
- Herrmann, F. (2009). *Logik der Produktionslogistik*. Oldenbourg, München.
- Herrmann, F. (2011). *Operative Planung in IT-Systemen für die Produktionsplanung und -steuerung: Wirkung, Auswahl und Einstellhinweise*. Vieweg & Teubner, Wiesbaden.
- Herrmann, F. and Manitz, M. (2021). Ein hierarchisches Planungskonzept zur operativen Produktionsplanung und -steuerung. In *Claus, T., Herrmann, F. and Manitz, M. (Eds.): Produktionsplanung und -steuerung*, pages 9–25. Springer.
- Icha, P. and Kuhs, G. (2019). Entwicklung der spezifischen Kohlendioxid-Emissionen des deutschen Strommix in den Jahren 1990 - 2019, Umweltbundesamt. *Climate Change*, 13/2020.
- Jiang, E., Wang, L., and Lu, J. (2017). Modified multiobjective evolutionary algorithm based on decomposition for low-carbon scheduling of distributed permutation flowshop. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.
- Liu, C.-H. (2016). Mathematical programming formulations for single-machine scheduling problems while considering renewable energy uncertainty. *International Journal of Production Research*, 54(4):1122–1133.
- Piroozfard, H., Wong, K. Y., and Wong, W. P. (2018). Minimizing total carbon footprint and total late work criterion in flexible job shop scheduling by using an improved multi-objective genetic algorithm. *Resources, Conservation and Recycling*, 128:267–283.
- Terbrack, H., Claus, T., Götz, M., Herrmann, F., and Selmair, M. (2021a). Analyse von konventionellen Prioritätsregeln zur Reduktion von CO₂-Emissionen durch den Einsatz von Photovoltaikanlagen. In *Franke, Jörg and Schuderer, Peter (Eds.): Simulation in Produktion und Logistik 2021: Erlangen, 15.-17. September 2021*, pages 75–84. Cuvillier Verlag.
- Terbrack, H., Claus, T., and Herrmann, F. (2021b). Energy-oriented Production Planning in Industry: A Systematic Literature Review and Classification Scheme. *Sustainability*, 13(23):13317.
- Terbrack, H., Frank, I., Herrmann, F., Claus, T., and Trost, M. (2021c). Eine Literaturlistenbank zur Nachhaltigkeit in der hierarchischen Produktionsplanung. In *Claus, T., Herrmann, F. and Manitz, M. (Eds.): Produktionsplanung und -steuerung*, pages 27–35. Springer.
- Terbrack, H., Frank, I., Herrmann, F., Claus, T., Trost, M., and Götz, M. (2020). A literature database on ecological sustainability in industrial production planning. *Anwendungen und Konzepte der Wirtschaftsinformatik*, 12:36–40.
- Trost, M., Forstner, R., Claus, T., Herrmann, F., Frank, I., and Terbrack, H. (2019). Sustainable Production Planning and Control: A Systematic Literature Review. *Proceedings of the 33th ECMS International Conference on Modelling and Simulation. Caserta, Italy*, pages 303–309.
- UN (2015). Paris Agreement, United Nations Treaty Collection, Chapter XXVII 7. d.
- Wang, J., Yao, S., Sheng, J., and Yang, H. (2019). Minimizing total carbon emissions in an integrated machine scheduling and vehicle routing problem. *Journal of Cleaner Production*, 229:1004–1017.
- Weinert, N., Chiotellis, S., and Seliger, G. (2011). Methodology for planning and operating energy-efficient production systems. *CIRP annals*, 60(1):41–44.
- Wu, X., Shen, X., and Cui, Q. (2018). Multi-objective flexible flow shop scheduling problem considering variable processing time due to renewable energy. *Sustainability*, 10(3):841.
- Zheng, X.-L. and Wang, L. (2018). A collaborative multi-objective fruit fly optimization algorithm for the resource constrained unrelated parallel machine green scheduling problem. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(5):790–800.

AUTHOR BIOGRAPHIES

HAJO TERBRACK is a doctoral student at the Chair of Production Economy and Information Technology at the International Institute (IHI) Zittau, a central academic unit of Technische Universität Dresden. His e-mail address is: Hajo.Terbrack@mailbox.tu-dresden.de.

PROFESSOR DR. THORSTEN CLAUS holds the Chair of Production Economy and Information Technology at the International Institute (IHI) Zittau, a central academic unit of Technische Universität Dresden, and he is the director of the International Institute (IHI) Zittau. His e-mail address is: Thorsten.Claus@tu-dresden.de.

PROFESSOR DR. FRANK HERRMANN is Professor for Operative Production Planning and Control at the Ostbayerische Technische Hochschule 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.

SUPPLY CHAIN RESILIENCE MANAGEMENT USING PROCESS MINING

Frank Schätter

Frank Morelli

Florian Haas

Business School

Pforzheim University

75175, Pforzheim, Germany

E-mail: frank.schaetter@hs-pforzheim.de, frank.morelli@hs-pforzheim.de, florian.haas@hs-pforzheim.de

KEYWORDS

Supply chain resilience, data-based supply chain model, key resilience areas, process mining.

ABSTRACT

Recent events such as the Coronavirus Pandemic or the disruption of the Suez Canal have shown how vulnerable supply chains can be and have led to an increased focus on resilience analysis by companies. We believe that all the data needed to understand the resilience status of a supply chain and identify opportunities for improvement already exist within companies. Therefore, we provide an approach to guide decision makers in this regard. We propose to first perform a rough resilience analysis using a limited set of transactional data. This analysis is based on key resilience areas to identify vulnerable elements of the supply chain that should be further investigated in terms of specific entities, transport relations, and materials. Based on these elements, process mining becomes a promising approach to understand the underlying actions, problems, and possible bottlenecks and to reveal improvement strategies.

INTRODUCTION

The goal of this article is to focus on systematic, performance-based risk and resilience management for supply chains. Therefore, our research deals with the questions of (1) how to provide an easy-to-use process for decision makers to identify the relevant data required to monitor the resilience status from a strategic perspective, and (2) how to understand the processes and implications of their supply chain risk management efforts.

Accordingly, performance measurement should explicitly be used in the sense of resilience. Therefore, critical disruptive factors and problem sources for relevant supply chain processes are to be identified and prioritized. Both (inter-)dependencies and the relevance of risks are to be evaluated and summarized into profiles. Related analyses enable proactive decisions and actions. Optimizing processes in terms of resilience implies using new knowledge and capabilities and applying different process-oriented performance measurements. This article examines the extent to which process mining is suitable for ensuring associated supply chain risk management. The remainder of this paper is organized as follows: in the next section, the concept of process mining is

discussed from a general perspective. We subsequently highlight how to analyze the triggers of supply chain disruptions on the basis of transaction data. The resulting vulnerabilities in the supply chain are the starting point for a detailed analysis in terms of process mining. Finally, a conclusion is drawn, and the tasks of future research are outlined.

GENERAL CONSIDERATIONS

Supply chains can be viewed as complex systems in which different business processes exist within and across companies and whose efficiency depends on how well they are interlinked or coordinated. The process mining approach offers an option to analyze and optimize supply chain processes. The question of whether the associated use of this technology proves meaningful is the starting point for the following considerations.

Process mining is an interface technology that links data mining methods with the use of process management (van der Aalst et al. 2012; Ramesh et al. 2020). Key figures can be used to identify relevant or critical business processes and the associated process optimization potential.

The state-of-the art literature differentiates three process mining approaches: discovery, conformance checking, and enhancement. In the course of discovery, information is read unchanged from an event log and displayed. This allows companies to gain transparency about the actual processes. During the conformance check, the expected process flow is compared with the actual process flow. For this purpose, a process mining tool analyzes the correspondences or deviations between an existing target process model and the variants in the actual. Process mining enhancement aims to extend or optimize the existing process structure.

Information about the as-is process flow is extracted from the event log and added to the existing process model. In comparison to the conformance check, the extension is not limited to the comparison between actual and target process but integrates the identified deviations directly into the process model (van der Aalst et al. 2012). In addition to the aforementioned approaches, a further manifestation of process mining can be identified in practice: it involves the operational support of IT-based systems. Here, insights gained from process mining applications are used to support the process execution of operational systems in real time. For example, decisions

can be made based on empirical values from past process execution (Peters and Nauroth 2019). The permanent maintenance of the event log enables real-time analysis. The evaluation criteria described below can serve as a basis for decisions and provide statements on whether process mining is applicable to supply chain processes. For the identification of the central building blocks in the sense of a framework, three different perspectives are taken: the first perspective includes requirements that are relevant for the goal setting process of the involved companies with regard to process mining. From the second perspective, the analysis of the supply chain process structure is performed. The third perspective deals with the data basis of the supply chain processes under consideration.

Before using process mining, suitable goals or questions should be defined. In this case, it is a matter of identifying vulnerable process areas of a supply chain with the help of resilience performance checks and taking optimization measures on the basis of the analyzed weak points. In addition to generating actual process models, it is also possible to compare this with an existing target process model (e.g., within the Supply Chain Operations Reference (SCOR) framework) or to extend the existing process model (van der Aalst et al. 2012). In addition, process mining can be used for real-time-based decision making (Peters and Nauroth 2019).

In order to specify how to use process mining in the company, it should be stated in the goal setting which process mining approaches are relevant for the company. An existing process model has to be available both for the process mining conformance test and for the process mining extension.

Process mining focuses on process analysis. However, not all business processes have the same value for the supply chain. Rather, they can be distinguished and differentiated from one another on the basis of different characteristics in terms of resilience in the sense of a preselection. Therefore, an approach is required that ensures that relevant elements of the supply chain are identified to be further processed via process mining.

A process characteristic which has a direct impact on the application of process mining methods, for example, is the degree of structuring of a process. This determines how precisely and in detail a business process has been defined and how often it deviates from its process flow chart (Allweyer 2005). However, if so-called concept shifts are incorrectly identified as process deviations, this leads to an erroneous interpretation of the process mining model. Identified concept shifts must therefore be taken into account when considering the model (Hierzer 2017). In this case, it is a matter of changing from an efficiency orientation to a resilience orientation, or rather an adequate balance.

Furthermore, it proves useful to examine the process-related capabilities of the parties involved. Process mining benefits from a solid, digitally oriented process and data infrastructure (van der Aalst et al. 2012). In the present case, it is assumed that this already exists in

supply chain management. Otherwise, corresponding capabilities should be built up first.

Maturity models can be used to evaluate the suitability of a business process for process mining (Becker et al. 2009). To determine the process maturity level, a suitable process maturity model has to be identified and this must be transferred into an evaluation framework. The evaluation should be carried out on the basis of uniformly formulated criteria, for example on the basis of a decision matrix. This is typically an extension of the Capability Maturity Model Integration (CMMI). For this purpose, business processes are classified into defined levels on the basis of specific process objectives and characteristics (Bürgin 2007). The process objectives and process characteristics of the individual maturity levels can be a benchmark for the process maturity level.

Another prerequisite for the efficient use of process mining is to have consistent access to all essential information. In this case, the relevant data must first be identified, merged, and structured in a uniform granularity (Hierzer 2017). Thus, the structure of the event data must be reviewed if the process is handled by several systems, which can be assumed in the context of supply chain management. Criteria such as the recording frequency of the process data, the respective data granularity as well as the process reference of the data should be used for the analysis. If an associated uniform structure is missing, this must be created. Otherwise, the use of a process mining tool proves inadequate.

Data from application systems can be interpreted as "raw material" for a process mining tool. To generate a process model that reflects reality as accurately as possible, the data basis must be of the highest possible quality. Accordingly, the data to be used later as event data must be analyzed in advance – which is in particular true when dealing with complex supply chain networks. A suitable evaluation method for this purpose is the maturity model developed by Wil van der Aalst: similar to the principle of process maturity models, the event log maturity model classifies the event logs under consideration into different levels based on certain criteria and characteristics (van der Aalst et al. 2012). To generate a robust process model, the event log under consideration should have a maturity level of at least three (Peters and Nauroth 2019; van der Aalst et al. 2012).

Only if a suitable data basis is available, it proves useful to analyze the attributes of the event logs. As a general rule, the more attributes there are in the event log, the more details can be represented in the process model. The analysis of the attributes is necessary to check whether the event log contains all relevant data for the goal fulfillment. If, for example, the objective requires that incidents be evaluated for each customer, it is necessary to check whether the data basis of the event log permits an assignment between customer and incident.

SUPPLY CHAIN RESILIENCE ANALYSIS

Recent events such as the coronavirus pandemic or the disruption of the Suez Canal have shown how fragile

supply chains can be. These events have meant that reliable supplies of ordered goods to customers at various stages of the value chains could no longer be guaranteed. As in other socially relevant areas, there has been an increased discussion about the *resilience* of supply chains. This discussion is directly related to the availability and application of suitable supply chain risk management approaches in companies.

Fundamentally, supply chain risk management has been on the corporate agenda for many years. As part of their efforts, companies can turn to one of the many commercial IT tools that typically suggest and promise assistance in developing reactive emergency response. They enable operational business continuity by providing tailored decision-relevant information that can be used to initiate ad hoc measures such as changing the transportation mode from sea freight to emergency air to deal with a port strike.

We confirm that the use of such tools is indeed very valuable to manage supply chain risk on a daily basis as they deal with high probability low impact events. Therefore, we highlight the vendors' claim that their tools increase operational resilience in the supply chain. However, the limitations of these tools become apparent when it comes to the overall functionality of supply chains during an incident, as it was the case with the recent events mentioned above.

Events such as the current pandemic are characterized by low probability and (potentially) severe impact. Therefore, the focus of our research is on ways to monitor the overall functioning of the networks during disruptive events and thus understand supply chain resilience from a strategic perspective. This is of great importance because only if this holistic resilience status is transparent, in-depth analysis are useful to identify proactive decision options that can be made in advance to safeguard against disruptive events, e.g., adjustments to the current supply chain design.

From a theoretical point of view, the resilience of a supply chain can be defined by its ability to recover from a disruptive event and even reach a higher level of performance in the long run (Anbumozhi et al. 2020). Several authors have provided a classification of concrete actions that can be selected by decision makers to improve the resilience of their networks. For example, Melnyk et al. (2014) present a set of "investments" to improve resilience, such as investments in discovery, information, operational flexibility, and buffers. Although these investments are useful to achieve improvements, they are valuable only if decision makers are informed about the true state of resilience of their networks. We believe that it is precisely this transparency that is extremely important as an upstream step, as it allows decision makers not to make decisions "in the dark" but to know exactly where and what vulnerabilities exist in their supply chains.

Therefore, our research focuses on the questions of (1) how to provide an easy-to-use process for decision makers to monitor resilience status from a strategic perspective, and, which subsequently applies resilience-

related data to (2) understand, evaluate, and improve the processes regarding their supply chain risk management efforts.

We believe that these goals can be achieved if companies perform resilience analysis that is clearly based on data. A supply chain is always characterized by the physical flows that take place between different entities. These can be factories and warehouses (intra supply chain) or suppliers and customers (b2b and b2c) (inter supply chain). The physical flows can occur in any relationship between these entities and each flow is therefore defined as a material-specific delivery between a sender and a receiver location. The core idea of our research regarding (1) is to define a limited set of data which is sufficient to get a first idea regarding the status of strategic resilience. Based on the result, elements of the supply chain (e.g., specific materials, entities, relations) can be revealed whose disruptions would lead to major turbulence in the supply chain and, thus, would have a negative effect on the supply chain's overall resilience.

Regarding the first research objective, this means that our proposal is to avoid elaborated and detailed models of the supply chain with sophisticated analyses, but to first perform a rough analysis of the weaknesses and strengths of the network to get a first indication of the strategic resilience status. Therefore, a data-based supply chain model is to be developed on the basis of a limited amount of transaction data that a company has anyway (Schätter and Morelli 2021). In order to translate this data into insights about the strategic resilience status of the company, a limited number of transaction data can be considered sufficient, as highlighted in Table 1.

Table 1: Data sets for supply chain resilience analysis

Data set	Description
Sender ID	Source of physical flow
Sender City	City in which sender is located
Receiver ID	Sink of physical flow
Receiver City	City in which receiver is located
Material ID	Unique ID of delivered material
Sending date and time	Date at which delivery has started
Receiving date and time	Date at which delivery has finished
Distance of delivery	Distance between sending and receiving location
Duration of delivery	Duration between sending and receiving location
Volume of delivery	Volume of the delivery [m3]

The limited transaction data highlighted in Table 1 can be easily captured from corporate data warehouses, ERP systems or SCM systems. Nevertheless, the data contains all the information needed to strategically analyze the effects of potential disruptions and vulnerabilities within the supply chain. Thus, we offer a cost-efficient method for an initial resilience analysis. Based on the data, we aim at uncovering vulnerabilities within the supply chain

that could affect smooth functioning during a disruptive event. They thus provide information on the state of resilience which are in the following understood as *key resilience areas (KRA)*:

- *KRA1 - geographic distribution of entities*: visibility into the locations and distribution of entities (factories, warehouses, suppliers, customers) provide an initial indication of supply chain resilience. For example, large aggregations of suppliers in certain areas increase the risk of large-scale disruptions within the network if a risk event affects the entire area. The geographic distribution (e.g., the number of customers per city) can be captured directly from the geographic datasets (sender and receiver cities).
- *KRA2 - sourcing strategy of materials*: using a single-sourcing approach for certain materials provides cost benefits but is one of the most important aspects of managing the impact of supply disruptions. There are no redundancies for these materials, so a supplier failure can lead to a shortage of the material (unless other inventory is available). Through data analysis, it is possible to directly determine what proportion of materials is not purchased on the basis of a multi-sourcing strategy. by analyzing which material IDs are supplied by only one sender ID.
- *KRA3 - warehouse materials*: one of the most positive impacts on supply chain resilience can be buffer stocks of certain materials in warehouses. This allows temporary outages of certain suppliers to be bridged without negatively impacting supply chain performance. The dataset allows for an analysis of material IDs and quantities delivered to warehouses during the period under consideration. We believe that this information is sufficient for a rough first check. Of course, further data can be optionally included to the data such as “inventory snapshots”, indicating inventories per material and month.
- *KRA4 - average storage time*: a further indication of the resilience of the network in terms of buffers, linked to point 3, refers to the average time materials are stored in the warehouse. The data described show all physical upstream flows into the warehouses as well as all physical downstream flows to the next entity. Based on this data, the average time can be estimated as the total quantity of a stocked material in relation to demand. This allows critical materials to be identified.
- *KRA 5 - transport delays*: for each relation, the distance and average delivery duration is available. By comparing these target data with the actual delivery durations, which can be derived by comparing the sending and receiving times, delays in the supply chain become visible. In this way, critical transport relations can be identified and both sending or receiving entities in the supply chain,

delay-prone material IDs, and the corresponding delivery volumes [m3] can be analyzed.

- *KRA6 - consolidation of deliveries*: based on the data, consolidation of material deliveries between shipping and receiving locations can be estimated. Deliveries should generally be consolidated, e.g., different materials delivered in the same relation on the same day should be combined. This leads to better utilization of shipments with positive cost effects. In addition, consolidation can have a positive impact on resilience, as the relation is used by fewer individual deliveries, which reduces the risk of disruptions (e.g., due to congestion). Consolidation can be read directly from the data by counting the delivery days and delivery quantities per relation.
- *KRA7 - transport distance*: the data allows an analysis of the transport distances between the used transport relations. Thus, the share of regional deliveries in inbound and/or outbound transports compared to long-distance transports can be determined. Furthermore, the data for each transport distance can be used to read off the actual volumes delivered in the reference period. In principle, more national and regional networks reduce the risk of large-scale, global disruptions.
- *KRA8 - intra-logistics processes*: the data enables transparency regarding the processes from an intra-logistical perspective. Thus, it can be analyzed which materials are delivered to which warehouses and in which factories they are further processed. From a resilience perspective, for example, it is possible to assess which elements of the company's own supply chain are particularly dependent on critical materials and can therefore be affected by disruptions. Insights regarding the intra-logistics processes might be already included in the further KRAs. However, we suggest considering those processes as explicit category because weaknesses within the internal supply chain can be revealed.

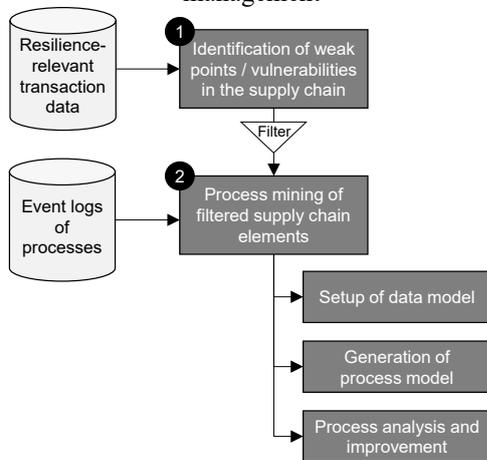
Hence, rather than focusing on overly analytical approaches to assess the status quo of resilience, our research focuses on ways to assist logistics managers in an easily applicable and pragmatic manner. We believe that the data highlighted above, and the guidance provided in the eight KRAs of analysis, are sufficient to provide an initial understanding of the status of the supply chain from a strategic perspective. To extend and enhance this analysis in terms of monitoring the actions taken by decision makers leading to the physical flows shown by the data, we believe *process mining* is a promising approach. Therefore, the next chapter focuses on the basic considerations for using process mining in supply chain resilience management.

SUPPLY CHAIN RESILIENCE MANAGEMENT BASED ON PROCESS MINING

Regarding our research objective (2), we aim at improving supply chain resilience by analyzing critical elements of the supply chain and underlying processes

and actions in order to identify concrete improvement opportunities. The basic principle of our approach is shown in Figure 1. The starting point refers to the data set and the analysis of the KRAs to identify vulnerabilities and susceptibilities within the supply chain (Step 1). We assume that to gain deeper insight into these vulnerabilities, it is promising to analyze the processes behind the vulnerabilities in more detail. At this point, we propose to focus on process mining (Step 2). In fact, the identified vulnerable supply chain elements are used as filters to cut the focus area of the supply chain and the process areas to be considered, such as single-source suppliers, relationships prone to transportation delays, or relationships characterized by weak consolidation in the past. Finally, improvement opportunities should emerge from the process mining.

Figure 1: Approach of supply chain resilience management



Step 1: Identification of weak points / vulnerabilities in the supply chain

As mentioned earlier, the supply chain is defined by the physical flows in the form of material-specific deliveries that are included in the transaction data set. Step 1 can therefore be understood as a filter to identify the relevant elements of the supply chain on which the subsequent process mining should focus. Hence, the KRA analysis has to create transparency regarding vulnerable entities, transport relations, and specific materials:

- *Vulnerable entities* can be identified primarily by KRA1, KRA2, KRA6, and KRA8. The geographic distribution of entities (KRA1) highlights specific suppliers, customers, warehouses, and factories that are in an unusual environment, for example, because they are highly clustered or located in areas of political risk. KRA2 considers suppliers that are linked to a single sourcing strategy and are therefore vulnerable to disruption. KRA6 suggests generally considering shippers that are characterized by a high mix of materials delivered on different shipping days; this could indicate poor consolidation of supplies and, thus, increasing vulnerability to disruption. Finally, KRA8 should point to

intralogistics operations such as factories and warehouses that may be vulnerable to critical materials due to high volumes, for example.

- *Vulnerable transport relations* can be identified primarily by KRA5 and KRA7. The dataset shows transport delays in physical flows (KRA5), which occur when there is a significant discrepancy between the expected duration of delivery and the difference between the timestamps of the send and receive dates and times. KRA7 is based on a static analysis of transport relations that may be at risk because they run over long distances. These relations can be selected directly from the data set.
- *Vulnerable materials* can be identified primarily by KRA3 and KRA4. First, the most important materials in terms of volume that are temporarily stored in warehouses should be considered in more detail. This can be done by looking at the high input volume flows into the warehouses (KRA3). In addition, those materials that are stored for only a limited or no time are also considered critical, as they are of great importance for maintaining the supply chain. In this context, an analysis of the time stamps between the input and output flows to and from the warehouses is required (KRA4).

All elements related to the above endangered entities, transport relations, and materials must be filtered. For example, by following the vulnerable transport relations, the corresponding entities and materials should be part of the further analyses. The filter reduces complexity as it highlights the supply chain's elements that are important for resilience management using process mining.

Step 2: Process mining of filtered supply chain elements

Data is needed as "raw material" for a process mining tool. Instead of trying to gain insights into resilience-relevant actions of decision makers by processing all available transaction data of deliveries for a specific reference point in time, our approach proposes to first perform a rough analysis of potentially critical elements in the supply chain in order to reduce complexity. This is in line with the assumptions of process mining, since the event data need to be analyzed in advance. We suggest that process mining for supply chain resilience management should follow a practical application of the general steps of discovery, conformance checking, and enhancement (see section "General Considerations"). From a practical perspective this means that first a data model should be built, second a process model should be generated based on process mining algorithms followed by the identification of optimization potential.

Setup of data model

Data extraction into an event log represents an integral part of process mining activities. The event log has to contain the relevant data for the relevant process under consideration. Then, based on the traces of the underlying

IT systems (i.e., database entries), process mining algorithms can reconstruct the as-is process flows by connecting events to activities.

An event log can be regarded as a particular view of the event data available. The assumption about event logs is that a process consists of cases, which comprise events, and the events within a case are ordered (van der Aalst 2018a). Therefore, three components for business processes are required: a date stamp, a characterization of the event (e.g., “goods receipt” activity), and a key (“case id”) to this operation (e.g., “purchase order item”). Each case consists of a sequence of events carried out within a process instance. Each unique sequence of events from the beginning to the end of a process instance is referred to as a variant, and each case/trace belongs to exactly one (Suriadi et al. 2017).

Regarding resilience process analysis, it is necessary to store additional data elements (attributes) to use information about resources and the organizational perspective (entities, transport relations, and materials) for later discoveries and / or enhancements. As supply chain management comprises several business processes, multi-event logs have to be considered as an adequate data source. The critical elements of the supply chain identified in the previous analysis are now the subject of process mining. Thus, all actions related to these elements of the supply chain should be further processed. This might be, for example, all actions taken for transport scheduling or order processing. These are basically all steps in the purchase-to-pay process starting with demand planning, through supplier selection, disposition, approval and monitoring to goods receipt.

Taking the example of KRA2, we might have identified a material that has been exclusively ordered from only one supplier for the given time period. All actions taken by the planner leading to the resulting single-sourcing are part of the event log and its capturing is therefore the first step of process mining.

The quality of the data (both form and content) reveals to be critical for the overall success. On the one hand, the preparation of event logs for our strategic application case has to focus on the relevant data following Occam’s razor in the sense of simplicity and granularity. On the other hand, it has to minimize information loss so that the event log is valid in the context of the resilience domain. From a technical perspective, XES (“extensible event stream”) can be used as a future data format standard, instead of the MXML format, as it is suitable for exchanging event logs between process mining and simulation tools (van der Aalst 2018b).

Generation of process model

Process mining algorithms create a process flow out of the traces from the event log. This is the basis for further discoveries, conformance checks, and /or enhancements. The algorithm used needs to generalize the behavior contained in the event log to show the most likely underlying model that is not invalidated by the next set of observations. Especially the balance between “overfitting” (creating a model too specific) and

“underfitting” (generating a model too general) reveals as a challenge for this phase (van der Aalst 2018a).

Taking again the example of single-sourcing, process mining algorithms allow to plot all actions included in the event log. Thus, the standard process of ordering material becomes transparent. In addition also deviations from the standard are revealed such as buying from only one source although the standard specifies a split between two suppliers that goes along with a reduced risk.

The control flow perspective as well as the organizational perspective must be considered regarding the target group and its willingness for acceptance. The process model has to provide transparency, allowing to trace process flows, analyze delays, loops and to identify process complexity drivers (Reinkemeyer 2020). Furthermore, resilience performance measures have to be extracted or calculated from the event log data.

Process analysis and improvement

The analysis takes place on the created process model and the provided performance measures to enable data-driven decision making and to discover as well as to monitor. For the supply chain management, the logistics orchestration becomes transparent. The KRAs form the basis for value proposition considerations regarding resilience, due to the holistic approach.

In the example of single sourcing, it might become obvious that although there is a second supplier established, only one supplier received the orders. This gives a clear indication for possible improvement in terms of activating the second available supplier.

Besides this, conformance checks for different process variants are possible to find communalities as well as discrepancies. Furthermore, process mining can be combined with business process management efforts to optimize the as-is model by creating a to-be model (e.g., based on the SCOR framework). This enhancement approach changes or extends the a-priori model. There is also a benchmarking opportunity, based on the key performing indicators created (e.g., costs, throughput time or rework).

CONCLUSION AND OUTLOOK

It seems that the aim of uncovering vulnerabilities in the supply chain by using transaction data and KRAs as a starting point can be advanced by a subsequent process mining: based on the existing IT infrastructure like data warehouses, ERP software and/or SCM systems, it is possible to compose data (and meta data) and convert them into an event log. The challenges are limited, as the required data is structured in contrast to the domain of social media. However, it has to be ensured that additional data (entities, transport relations, and the warehouse material flow) are adequately mapped.

The target of creating an end-to-end supply chain process promises to be manageable from a strategic point of view: it is the core focus of process mining to create as-is variants out of the event log. The challenge is to avoid “overfitting” and “underfitting”. Discovery as a type of

process mining offers an adequate platform for analysis. It can be enhanced by other procedures like conformance checking, enhancement, and bench-marking.

As an outlook, event logs created for process mining reasons also can be used as a fundament for constructing predictive models (van der Aalst 2018a). However, to switch from backward-looking by the process mining approach to design alternatives and to anticipate the future simulation is a promising choice: corresponding work may employ different “what if” issues to be answered and alternatives with respect to the resilience indicators are able to be evaluated. Various scenarios to combine process mining and simulation can be used therefore (van der Aalst 2018b).

We believe that the approach described in this contribution is promising to support decision makers in understanding their supply chains with respect to resilience – a requirement that has been significantly revealed by past supply chain disruptions. Our suggestions ensure that the relevant elements of the supply chain that provide insights regarding the resilience status are identified in a pragmatic and cost-efficient manner. In this regard, an important next step of our research will be to evaluate the proposed KRAs: are they sufficient to understand the network under consideration in terms of resilience? Are there overlaps in the KRAs or should additional categories be included? Therefore, a case study based on real corporate transaction data is currently underway.

The initial resilience analysis sets the basis to apply a process mining approach in order to understand the actions of the decision makers on the different management levels and to identify improvements. We have exemplarily highlighted how this would look like for a typical purchasing process. Future research should provide a prioritization of the specific disposition processes (e.g., order and transport scheduling) that should be analyzed in this regard. The concrete augmentation of the underlying specifications (e.g., event log) is work in progress as well as the verification and illustration of the approach with real company data and therefore an essential part of our future research.

REFERENCES

- Allweyer, T. 2005. *Geschäftsprozessmanagement, Strategie, Entwurf, Implementierung, Controlling*. W3L-Verlag, Herdecke, Bochum.
- Anbumozhi, V.; F. Kimura; S. Mugan Thangavelu. 2020. *Supply chain resilience, Reducing vulnerability to economic shocks, financial crises, and natural disasters*. Springer, Singapore.
- Becker, J.; R. Knackstedt; J. Pöppelbuß. 2009. “Entwicklung von Reifegradmodellen für das IT-Management.” *Wirtschaftsinformatik* 51, No.3, 249–260.
- Bürgin, C. 2007. *Reifegradmodell zur Kontrolle des Innovationssystemes von Unternehmen*. Doctoral Thesis, ETH Zurich.
- Hierzer, R. 2017. *Prozessoptimierung 4.0, Den digitalen Wandel als Chance nutzen*. Haufe Gruppe, Freiburg, München, Stuttgart.
- Melnyk, S.; D.J. Closs; S. Griffis; C. Zobel; J. Macdonald. 2014. “Understanding supply chain resilience.” *Supply Chain Management Review* 18, 34–41.

- Peters, R. and M. Nauroth. 2019. *Process-Mining, Geschäftsprozesse: smart, schnell und einfach*. Springer Fachmedien, Wiesbaden.
- Ramesh, G.S.; T.V. Ranjini Kanth; D. Vasumathi. 2020. “A Comparative Study of Data Mining Tools and Techniques for Business Intelligence”. In *Performance Management of Integrated Systems and its Applications in Software Engineering 2020*, M. Pant, T.K. Sharma, S. Basterrech, C. Banerjee (Eds.), Springer, Singapore, 163–173.
- Reinkemeyer, L. 2020. *Process mining in action, Principles, use cases and outlook*. Springer, Cham.
- Schätter, F. and F. Morelli. 2021. “Business Process Simulation Focusing Supply Chain Risk Management Aspects”. In *Special Track: Simulation and Modelling in Supply Chains, along with SIMUL 2021*, 38–43.
- Suriadi, S.; R. Andrews; A. ter Hofstede; M.T. Wynn. 2017. “Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs.” *Information Systems* 64, 132–150.
- van der Aalst, W. 2018a. *Process Mining, Data Science in Action*. Springer, Berlin.
- van der Aalst, W. 2018b. “Process Mining and Simulation: A Match Made in Heaven!” In *Proceedings of the 50th Computer Simulation Conference: Society for Modeling and Simulation International (SCS)*.
- van der Aalst, W.; A. Adriansyah; A.K.A. de Medeiros; F. Arcieri; T. Baier; T. Blickle et al. 2012. “Process Mining Manifesto”. In *Business Process Management Workshops. BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011*, F. Daniel, K. Barkaoui, S. Dustdar (Eds.). Springer, Berlin, 169–194.

AUTHOR BIOGRAPHIES



Frank Schätter has been a Professor of Supply Chain Processes Management at Pforzheim University. He earned his doctoral degree in 2016 at the Institute for Industrial Production (IIP) at the Karlsruhe Institute of Technology (KIT) in the field of supply chain risk management. His teaching and research focus is on modeling, analysis, and optimization of supply chain processes.



Frank Morelli is Professor of Information Systems - Management & IT at Pforzheim University. In addition to his teaching activities, he is involved in practical and research projects from the fields of business process management, business intelligence, SAP ERP, project management and IT organization. A close cooperation also exists with the Celonis Academic Alliance on the topic of “Process Mining Education”.



Florian Haas is Professor of Purchasing and Supply Management and leads the bachelor’s degree program Purchasing and Logistics at Pforzheim University. He has over 15 years of practical experience in several management positions within logistics at Robert Bosch GmbH. Most recently he was responsible for the implementation of a central sea and air freight transport management. His research focuses on the evaluation of purchasing processes.

STRATIFICATION OF TIMED PETRI NETS AT THE EXAMPLE OF A PRODUCTION PROCESS

Carlo Simon and Stefan Haag and Lara Zakfeld
Fachbereich Informatik
Hochschule Worms
Erenburgerstr. 19, 67549 Worms, Germany
E-Mail: {simon,haag,zakfeld}@hs-worms.de

KEYWORDS

Stratified Modeling and Simulation, Petri Net Folding, Timed Dynamic Systems, Process Management

ABSTRACT

Timed dynamic systems can be modeled and simulated with Petri nets using very different approaches. In Clock Pulse Models (CPM), one marking represents exactly one moment in time and all enabled transitions fire simultaneously according to a global clock. This allows for real-time observation of the modeled system during a simulation run. Since simulation time is proportional to observed real-time, such an approach barely scales concerning the observed time horizon. If not the observable behavior over time is of interest but the final simulation result, Event Triggered Models (ETM) can overcome this limitation as has been shown in former publications. A time-lapse model accelerates simulation runs by focusing on the moments state changes occur.

CPM and ETM, however, share one disadvantage: the models' sizes tend to be proportional to the number of events that may occur simultaneously in the real world. Hence, they scale barely with the number of modeled entities. This unsolved scaling problem is addressed in this paper and pursues the idea of folding similar subnets in temporal layers called strata. The result is a Stratified Simulation Model (SSM), a small, flexible model that scales well concerning the number of modeled entities. SSM are derivatives of CPM, since they still allow for nearly-real-time observation of the real world. Moreover, these models ease the visualization of the simulation in dashboards. The approach is explained at the example of a small production process.

INTRODUCTION

It is the nature of simulation that it is performed for systems that are too complex for a human mind to understand them entirely. Computational complexity considers the two dimensions time and space (Arona and Barak 2009). Time complexity classifies the time needed to execute an algorithm on an input for a given parameter such as the number of elements in an array. Space complexity classifies the memory amount needed.

Batty and Torrens (2001) apply this idea to the term modeling complexity which links it to simulation. In simulation research, we must handle both kinds of complexity to make simulation methods applicable.

The two parameters that influence the complexity of a simulation run are the complexity of the modeled system and the observation time. Popovics and Monostori (2016) distinguish between structural and software complexity measures. The structural complexity is derived from the number of elements a modeled system consists of. Both structural and software complexity influence the memory space complexity of a simulation problem.

Also, the model complexity must be considered which consists of the model's size and the connections within the model. This complexity is influenced by the problem but also by the modeling approach used. If, for example, modelers try to build simulation models of production lines with the aid of basic Petri net concepts, they will fail soon because of the model complexity. But even Clock Pulse Models (CPM) and Event Triggered Models (ETM), high-level Petri nets which have been proven to solve real-world modeling and simulation problems in production and logistics, must handle the burden of complexity. This paper discusses Stratified Simulation Models (SSM), a derivative of CPM, which is intended to solve this problem partially.

Figures 1 and 2 illustrate time complexity with respect to simulated time horizon and model complexity with respect to number of simulated components that CPM, ETM, and the newly introduced SSM tend to. Since CPM and, as a derivative, SSM are used to observe systems' behavior over time, they scale barely with the considered time horizon. And since SSM may stratify a simulated moment over several subsequent Petri net states, SSM may even perform inferior to CPM. In ETM, on the other side, only state changes and not the progress in time is simulated which enables a time-lapse simulation.

In CPM and ETM, real world components are modeled and simulated with subnets, and their composition models and simulates the entire system. In SSM, these subnets are folded into one if they are equally structured. The behavior of the formerly different subnets is then expressed with the aid of token attributes: an SSM can be interpreted as a simulation model of its CPM.

The folding technique that leads to SSM has two major advantages: First, SSM is superior regarding model complexity compared with CPM and ETM as shown in figure 2. Second, the overall system state is no longer distributed over a larger number of places and their marking but is concentrated in small number of places. This makes it much easier to observe the system and visualize the system state in a dashboard.

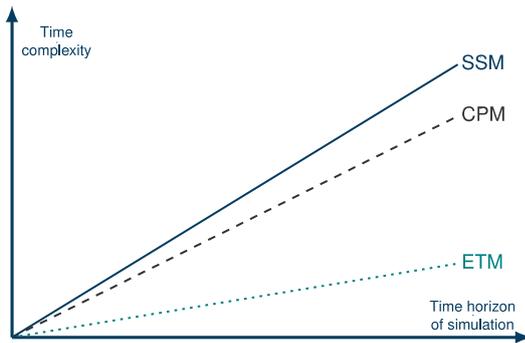


Figure 1: Time Complexity Trends for CPM, ETM, and SSM

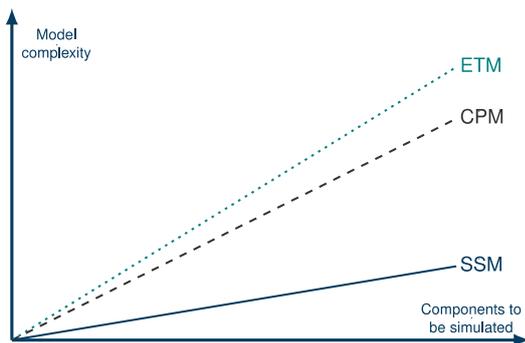


Figure 2: Model Complexity Trends for CPM, ETM, and SSM

The presented research is experimental, as the authors do not have an algorithm that takes a CPM and generates an SSM automatically. Experiments need a laboratory and ingredients. The authors' laboratory is the Process-Simulation.Center (P-S.C), a modeling and simulation environment for high-level Petri nets which is free to use for academic purposes. The ingredient is a simple production process that may happen all over the world.

The remainder of the paper is structured as follows: after this introduction, related work on Petri net folding techniques is presented which leads to the SSM approach. Next, the example process is explained. Using a pattern from literature, a CPM is developed first. In a non-algorithmic but structured way, an SSM is derived from this step-by-step afterwards. Moreover, a dashboard is designed to present the simulation results. For a discussion of the results and an outlook to the future work, the role of ETM is explained in more detail.

RELATED WORK

Reisig (2013) provides a fine introduction to Petri nets which are well suited to model, simulate, and analyze dynamic systems. Even large systems can be examined by expanding and connecting local components; different Petri net concepts allow for apposite modeling (Recalde et al. 2004). Their mathematical power is a blessing and a challenge at the same time since the same problem can be solved with the aid of Petri nets in very different ways. It is difficult to teach the required creativity and for some people it is hard to follow this approach, especially if they do not have access to a suitable Petri net tool.

CPM, ETM, and the SSM approach introduced with this paper deliver modeling and simulation patterns for processes in production and logistic. They make use of high-level nets in the form of Predicate/Transition nets (Genrich and Lautenbach 1981). Tokens carry individual information, and a firing rule decides on which tokens to use for the next transition.

Since time concepts are highly relevant in practice, many different extensions have been proposed to include time information to otherwise timeless basic Petri nets. In the modeling approaches discussed here, time data is included beside any other system relevant information in the high-level net structure.

However, Petri nets per se exhibit an issue most graphical presentation methods ail from: models of real systems become very large quickly, i.e., the model complexity of Petri nets raises fast. Beside the already mentioned high-level Petri net concepts, other methods have been proposed to address this problem and squeeze large Petri net models:

Modularization is a decomposition approach. For example, Righini (1993) modularizes larger nets using Petri subnets as an implementation of hierarchical nets (Fehling 1993). Christensen and Petrucci (2000) modularize even without explicit subnets.

Clustering and Folding morphisms are further approaches (Keller 2002). These two methods make it possible to establish smaller models, however clusters are no Petri net concepts which causes semantical problems.

A special form of clustering and folding results from avoiding **redundant transitions** and **implicit places** (Garcia-Valles and Colom 1999; Simon 2008).

As a shortcoming of these methods, the scaling problems remain partly unsolved or are only relocated. For example, subnets or modules externalize portions of one model into other models that need to be integrated. This possibly leads to increased system complexity as interfaces need to be established and a necessity for redundant handling of data may arise.

A practical solution would be a small Petri net model that retains all information from a larger one while minimizing the footprint. To this end, folding seems to be a fitting technique. This paper examines the folding of a CPM and the challenges this poses to modelers.

A CPM FOR THE EXAMPLE

To present the SSM approach, a small real-world production process is introduced. It originates from the region the authors' university is located: viniculture in Rhenish Hesse in Rhineland-Palatinate, Germany. The example is used to examine timed dynamic systems in logistics and production. The following model is derived from a model pattern introduced in Simon et al. (2021b).

The process illustrated in figure 3 is the handling of personalized wine gifts in a winery. *Unlabeled bottles* of wine are taken from the storage and delivered to the input inventory of a workbench where they receive a personalized label according to customer wishes. In parallel, *cartons* are dispatched in form of *sheets* to another workplace to be assembled there.

Both items are taken to the packing station where bottle and newly created supporting material like greeting cards are put into the carton. Afterwards, the *packed gift box* (for short: *box*) is taken to another workplace where the shipping label is attached, and the box is sealed. The *completed box* is deposited in the outgoing goods area.

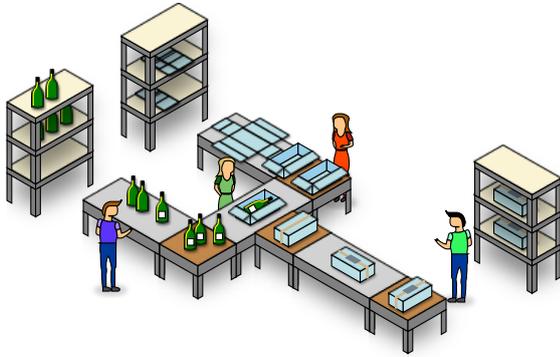


Figure 3: Layout Illustration of the Sample Process

Despite its relative simplicity, the process enables the examination of many different problems encountered in larger scale production processes. The concept can be modified as needed: different production strategies can be simulated, bottlenecks or constraints may be added, or process steps can be inserted, omitted, or temporally altered. Thus, it serves both as an easy-to-understand example and as testing grounds for bigger systems. The main point of this sample process is its transferability and scalability to general work and assembly workplaces.

In the following, customer orders of 75 gift boxes are examined. Thus, there are 75 unlabeled bottles and 75 unfolded cartons in the material storage. Implemented as a push process, items are put into the assembly-line as fast as possible. The process times are estimated: building a carton takes roughly 9 seconds while labeling a bottle requires 18 seconds. This time difference means that the downstream inventory of the carton production fills up faster than the other one. The most time-consuming activity with 25 seconds is the packaging step. Thus, both already mentioned inventories build up over time. The completing step accounts for 7 seconds.

Figure 4 shows the CPM. In short, the upper part depicts the bottles' labeling on the left side, while the right side presents the folding of the associated cartons. The subnet starting with the transition *startP* is associated with the merging of the two lines where bottle and additional items get packed into the carton, creating the gift box. Afterwards, the package is finalized and cleared.

Four types of places must be distinguished:

- Places modeling inventories like *inM* for the raw material or *inU* for unlabeled bottles.
- Places modeling activities to express that a specific task is currently conducted like *labeling* or *packing*.
- *feeding**-places that provide the first inner inventories with their raw materials.
- *idle**-places serving as semaphores to prevent more than one item to be processed at once.

The *idle**-places may additionally track the workplaces' idle times to measure utilization rates. *start**- and *stop**-transitions frame the activity-places.

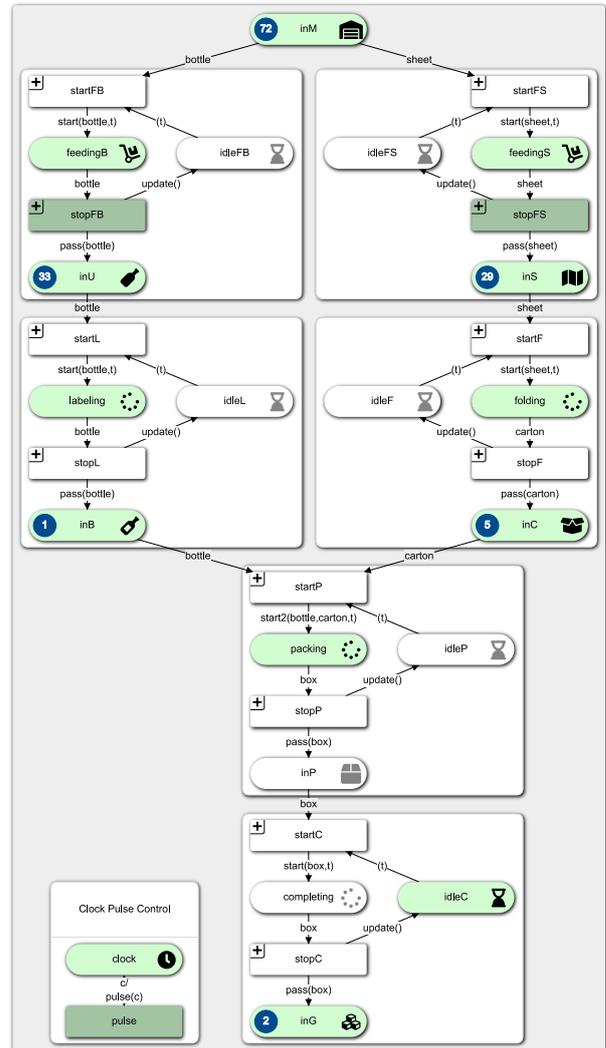


Figure 4: CPM of the Personalized Wine Gift's Production after 77 Clock Pulses

A further place type is *clock*, marked with the current simulation time. Firing the transition *pulse* increases this value by one time unit, here, a second. All other enabled transitions fire in unison, establishing a new system state. The *P-S.C* allows for exporting all system states. Using this data, the grounding system can be analyzed deeply.

This model can be simulated using different input data on place *inM*. The net's current state is represented as follows: blue dots show the number of tokens on high-level places that can be marked with records comparable to a table in a database. This is omitted for the *idle**-places and the *clock* which carry at most one token.

Simon et al. (2021b) have demonstrated how to integrate a simple dashboard into the given net which counts the number of items of each inventory in every moment in time. The SSM developed next is not only more compact than this CPM but also establishes new possibilities to design a simulation dashboard.

TRANSFORMATION FROM CPM TO SSM

Frames in figure 4 show repeating structures. *start**- and *stop**-transitions frame their activity places. The *start**-transitions take "material" from incoming inventory and, in addition, are controlled by an *idle**-place. The processed items are put on an outgoing inventory that also serves as incoming inventory for the following structure. A deviation can be observed for *startP* which has two incoming arcs and, correspondingly, two incoming inventories.

Since the corresponding places are type equivalent and the arcs and transitions are compatible, differing only concerning the included time information, they can be folded to the structure shown in figure 5: transition *start* takes an object from inventory *in* and puts it on *perform* indicating that an activity takes place. At the same time, the availability marker is removed from *idle* which avoids that a second object is "performed". When the performance ends, transition *stop* takes the object from *perform* and puts it on inventory *in* again. Marking place *idle* unblocks the *start-stop* subnet.

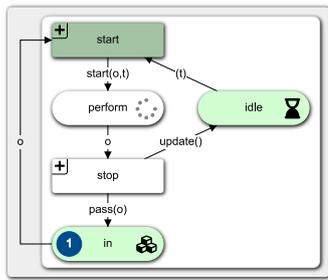


Figure 5: Folded subnet for repetitions in the CPM of figure 4

In the folded net information gets lost during the folding process: the particular place an item is located on. This information must be integrated into the item itself. To understand this important step, the high-level structure of the CPM in figure 4 must be considered.

Table 1 (left) shows a part of the initial marking of the place *inM* for the CPM. All other inventory places of the CPM are of the same structure. In a first step, this can be changed to the right structure where each token is augmented by the name of the place it lays on. Available simplifications are explained later.

Table 1: Initial Sample Allocation for Items in (left) the CPM and (right) the SSM

id	type	stamp	id	type	stamp	place
1	bottle	0:00	1	bottle	0:00	inM
⋮	⋮	⋮	⋮	⋮	⋮	⋮
76	carton	0:00	76	carton	0:00	inM
⋮	⋮	⋮	⋮	⋮	⋮	⋮

The time-typed *idle**-places of the original CPM must be enriched, too, as depicted in table 2: an attribute *place* refers to the former place and attribute *type* to the kind of item that lies on the place. The Boolean attribute *idle* symbolizes the availability of the workplace.

Table 2: Initial Sample Allocation for Workplace Availability in (left) the CPM and (right) the SSM

stamp	type	place	idle	stamp
0:00	bottle	inM	true	0:00
⋮	⋮	⋮	⋮	⋮
0:00	carton	inM	true	0:00
⋮	⋮	⋮	⋮	⋮

Transitions select tokens they may use for firing. The existence of selection rules is indicated by the plus symbol in the upper left corner in the depicted model. Outgoing arcs from transitions allow for computations to be made. Functions on these arcs are parametrized with variables taken from the incoming arcs.

The newly introduced additional information and the time information must be updated while they circle in the folded net. For the time information, the *clock/pulse* subnet remains in the SSM. The *pulse* transition fires whenever no other transition is enabled. This means that this stratum – the concept of all activities to be performed in a given time period – has been simulated, completely.

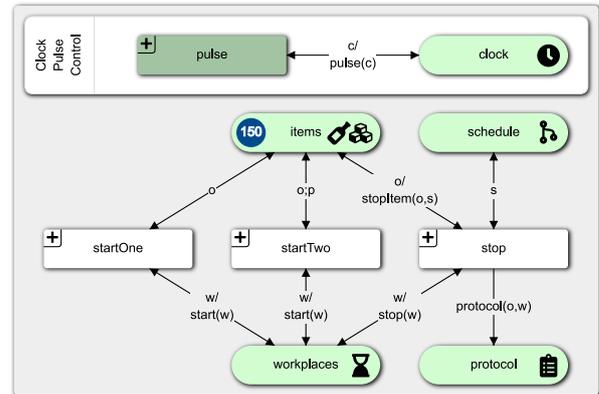


Figure 6: SSM of the Personalized Wine Gift's Production

The SSM in figure 6 takes the considerations made above into account and is a derivate of the CPM in figure 4. The savings that can be achieved concerning model complexity are tremendous as the number of places for the actual process is reduced from 19 to 4, and the number of transitions is reduced from 12 to 3. Actually, the savings concerning the places is even better since place *protocol* has only been introduced to ease the observation of the simulation.

Two aspects of the model in figure 6 have not been explained in advance: the necessity of two start transitions and the role of the place *schedule*.

Since in Petri nets the number of tokens that may be transferred over an arc simultaneously is fixed, two transitions are needed: One to model taking a single item, and a second to model taking two items at once. The selection criteria on these places guarantee that *startOne* cannot "steal" tokens that *startTwo* needs for firing.

During the process of folding the CPM manually, a characteristic of the P-S.C that seemed to be limiting at first glance led to the introduction of the place *schedule*.

This place supplies information about the processing times of the workstations and the state changes that occur for these items. While performing a specific work step on an item is time consuming, accessing an item out of an inventory is regarded as timeless. Figure 7 shows the chain of states the items traverse while they are processed by the Petri net, and which is implemented as the marking of place *schedule*.

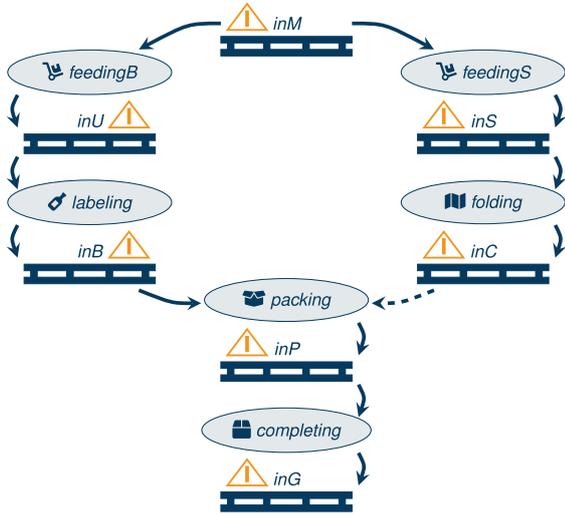


Figure 7: States of the items in the CPM and possible state changes

These considerations led to the final simplification: it is not necessary to present the current time for items on the place *items*. It is sufficient to code this information on *workplaces* as its data combined with that from *items* establish new record sets for *protocol* at those moments an inventory receives an item, or an activity starts.

As has been pointed out before, all folding steps have been conducted in a structured way but without an algorithm. The authors, however, see a chance and the potential of further automation of these steps.

A DASHBOARD FOR SIMULATION RESULTS

The CPM of the exemplary process allows to observe the raising and sinking of the inventory stocks while the Petri net simulation runs. Since these stock quantities and the utilization of the working places are folded into few places, this is not possible for SSM anymore.

As an alternative, the place *protocol* was introduced which summarizes the simulation and its results over time. For each item, all reached states are stored in attributes *id* (to identify the item), *type* for the kind of item and *place* with the information where the item is currently located. An attribute *busy* indicates the moment the state was reached and *idle* the moment a following item is allowed to reach the same state.

Every information about the model, its dynamics and its results can be extracted from this protocol easily. Table 3 shows an excerpt of the protocol relevant to produce the first gift box.

Table 3: Protocol Excerpt for the First Giftbox

id	type	place	busy	idle
1	bottle	feedingB	0:00	0:02
76	carton	feedingS	0:00	0:02
1	bottle	inU	0:02	0:02
76	carton	inS	0:02	0:02
76	carton	folding	0:02	0:11
76	carton	inC	0:11	0:11
1	bottle	labeling	0:02	0:20
1	bottle	inB	0:20	0:20
1	bottle	packing	0:20	0:45
76	carton	packing	0:20	0:45
1	box	inP	0:45	0:45
1	box	completing	0:45	0:52
1	box	inG	0:52	0:52

The advantage of SSM concerning model complexity is paid for with a higher time complexity compared to CPM. This can also be seen in the marking of place *protocol*. It is possible that no action at all takes place in one second, i.e., one stratum. On the other side, table 4 shows all state changes that can be observed for second 20. And since each of these changes are written by transition *stop*, the strata span in the simulation for this moment in real time becomes obvious. Moreover, also *start*-transitions may fire for one second.

Table 4: Protocol Excerpt for second 20

id	type	place	busy	idle
1	bottle	labeling	0:02	0:20
10	bottle	feedingB	0:18	0:20
1	bottle	inB	0:20	0:20
77	carton	building	0:11	0:20
10	bottle	inU	0:20	0:20
85	carton	feedingS	0:18	0:20
77	carton	inC	0:20	0:20
85	carton	inS	0:20	0:20
11	bottle	feedingB	0:20	0:22
86	carton	feedingS	0:20	0:22
78	carton	building	0:20	0:29
2	bottle	labeling	0:20	0:38
1	bottle	packing	0:20	0:45
76	carton	packing	0:20	0:45

After a stratum has been simulated completely, the clock increments by one and the next stratum can be simulated. This special form of simulating a timed system led to the name for this modeling technique: *Stratified Simulation Model*.

The final step for the discussed approach is how to represent the simulation results in a dashboard, a major problem identified by Simon et al. (2021a).

Dashboards serve as a communicator to the user, the business community, or to attract the attention of other researchers. On the other hand, they serve as a checkpoint to ensure that the results in the newly developed concepts are still correct.

For the generation of a dashboard out of the simulation results of a Petri net, the data must be prepared first. In business intelligence, this is conducted in the ETL-phases Extract, Transform, and Load (van der Lans 2012).

For CPM, the information concerning reachable system states must be collected from the different places all over the net. This is different for the presented SSM and its place *protocol* which focusses the observation of the entire system on a single place. This makes the ETL process much easier. This become obvious when the amount of information generated during the simulation is considered: for the CPM model, a CSV file is generated with a size of 3.1 Mbyte, while the SSM model only generates a CSV file of 38 Kbyte, i.e., only 1% of the original size. However, this file still contains all relevant information concerning stock utilization and idle times of working places.

Figure 8 demonstrates how to prepare the data for further use. Assume that the Winery in Worms wants to analyze the productivity of the individual areas for the purpose of overhead cost accounting. Based on the time stamps of place *protocol* it is possible to examine exactly which box was where and when, and when it was processed. Interactive filters in dropdown menus exemplarily help to find the shortest and longest waiting time per inventory and many other information.

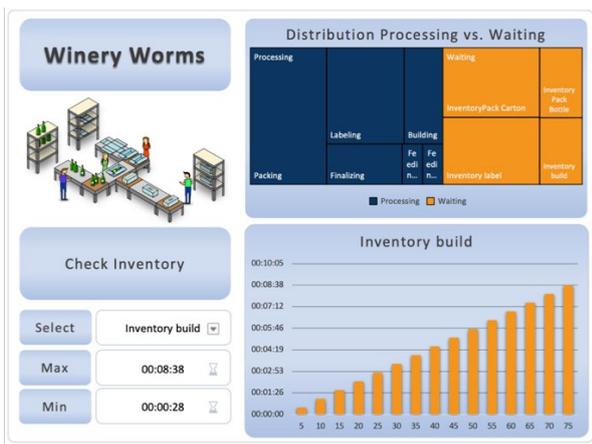


Figure 8: SSM Dashboard

The upper graph in the dashboard of figure 8 shows the distribution of waiting and processing times at the Winery in Worms. If a dashboard user wants to follow the origin of the large, orange tile of the inventory build label, this can happen elegantly over the selection menu beside *Select*. The graphics and times adapt accordingly.

CONCLUSION AND OUTLOOK

The current paper demonstrates an approach how to solve a problem commonly known from modeling and simulation, especially if Petri nets or other graphical concepts are used: The models tend to have a high model complexity which makes them hard to prepare, difficult to read, and error-prone.

Nonetheless, Petri nets and especially CPM can be applied successfully to problems in production and logistics as has been shown by Simon et al. (2021b), but there still is the necessity to overcome the model complexity problem.

The SSM approach that has been introduced here would be one possible solution if SSM could be derived from CPM automatically. The development of CPM patterns to model technical components of such systems and to provide SSM transformations for these patterns could possibly solve this task. One problem, however, would persist: the time complexity.

A first answer to this problem is ETM, as has also been demonstrated by Simon et al. (2021b). ETM are high-level Petri nets, too. The marking of a place represents the system state of a specific component at a specific moment in time, but the distinct moments considered for the different places of the net may vary from place to place. Necessarily, time information must be included in each token which indicates the moment a specific state was reached. From this stage, the next possible (local) state change can be calculated including the moment this change occurs. Especially if state changes occur rarely in the time scale chosen for the simulation, ETM tend to a lower time complexity when compared to CPM and SSM. If observing the system during a simulation run is not important, this is a practical solution.

Future work on new Petri net modeling techniques for timed systems will be the development of folding techniques for ETM. These Folded Event Systems (FES) are currently in their conceptual phase and are anticipated to have time and model complexities compared to CPM, ETM, and SSM as depicted in figures 9 and 10.

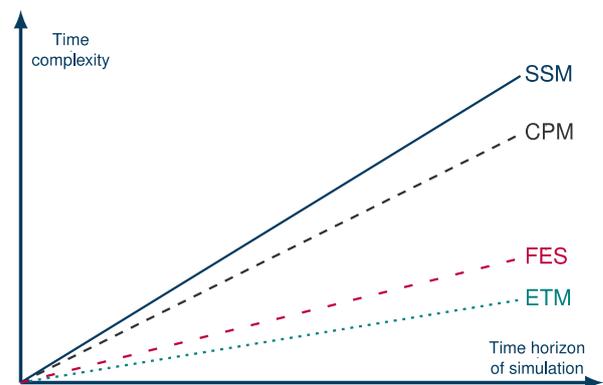


Figure 9: Anticipated Time Complexity Trend for FES

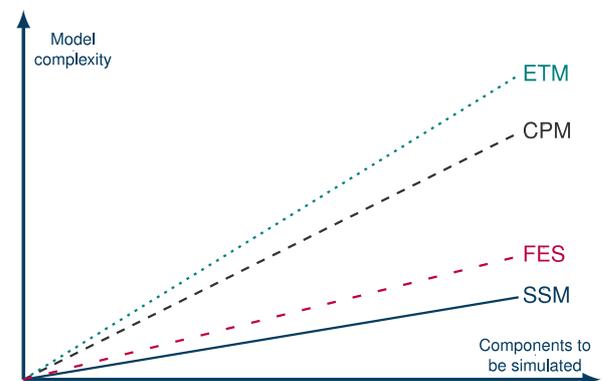


Figure 10: Anticipated Model Complexity Trend for FES

The previous consideration on the development of dashboards as based on the place *protocol* gives a hint on how the future research is conducted. With the aid of this place, it is possible to identify all moments in time when exactly one single clock pulse occurs in the SSM. If the clock pulses can be substituted by event pulses, the simulation speed can be increased dramatically.

REFERENCES

- Arona, S. and B. Barak. 2009. *Computational Complexity: A Modern Approach*. Cambridge University Press, UK.
- Batty, M. and P. M. Torrens. 2001. "Modeling complexity: The limits to Prediction", *UCL Working Papers Series* 36.
- Christensen, S. and L. Petrucci. 2000. "Modular Analysis of Petri nets". *The Computer Journal* 43, 3, 224-242.
- Fehling, R. 1992. *Hierarchische Petrinetze* (German, transl. Hierarchical Petri nets). Kovac, Hamburg, Germany.
- Garcia-Valles, F. and J. M. Colom, 1991. "Implicit places in net systems". *Proceedings of the 8th International Workshop on Petri Nets and Performance Models*, 104-113.
- Genrich, H. J. and K. Lautenbach. 1981. "System Modeling with High-Level Petri Nets". *Theoretical Computer Science* 13, 1, 109-135.
- Keller, W. 2002. "Clustering for Petri nets". *Theoretical Computer Science* 308, 145-197.
- van der Lans, R. 2012. *Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses*. Morgan Kaufmann Series on Business Intelligence, Elsevier, Amsterdam, the Netherlands.
- Popovics, G. and L. Monostori. 2016. "An approach to determine simulation model complexity". *Procedia CIRP* 52, 257-261.
- Recalde, L.; M. Silva; J. Ezpeleta; and E. Teruel. 2004. "Petri nets and manufacturing systems: An examples-driven tour". *Lectures on Concurrency and Petri Nets: Advances in Petri Nets*, 742-788. Springer, Berlin, Germany.
- Reisig, W. 2013. *Understanding Petri Nets*. Springer, Wiesbaden, Germany.
- Righinni, G. 1993. "Modular Petri nets for simulation of flexible production systems". *International Journal of Production Research* 31, 10, 2463-2477.
- Simon, C. 2008. *Negotiation Processes - The Semantic Process Language and Applications*. Shaker, Düren, Germany.
- Simon, C.; S. Haag; and L. Zakfeld. 2021a. "Research Agenda for Process Simulation Dashboards". In: *ECMS 2021: 35th International ECMS Conference on Modeling and Simulation*, 243-249.
- Simon, C.; S. Haag; and L. Zakfeld. 2021b. "Simulation of Push- and Pull-Processes in Logistics: Usage, Limitations, and Result Presentation of Clock Pulse and Event Triggered Models". *International Journal on Advances in Software* 14, 1&2, 88-106.

AUTHOR BIOGRAPHIES



CARLO SIMON studied Informatics and Information Systems at the University of Koblenz-Landau. For his PhD, he applied process thinking to automation technology in the chemical industry. For his state doctorate, he considered electronic negotiations from a process perspective. Since 2007, he is a Professor for Information Systems, first at the Provdavis School of Technology and Management Frankfurt and since 2015 at the Hochschule Worms. His e-mail address is: simon@hs-worms.de.



STEFAN HAAG holds degrees in Business Administration and Engineering as well as Economics with his main interests being related to modeling and simulation in graphs. After working at the Fraunhofer Institute for Systems and Innovation Research ISI Karlsruhe for several years, he is now a Research Fellow at the Hochschule Worms. His e-mail address is: haag@hs-worms.de.



LARA ZAKFELD graduated in International Logistics Management (B.A.) after completing an apprenticeship as a management assistant in freight forwarding and logistics services. She is currently pursuing a master's degree in Entrepreneurship and works as a Research Assistant at the Hochschule Worms. Her e-mail address is: zakfeld@hs-worms.de.

MICROBIAL GROWTH OF *LACTOBACILLUS DELBRUECKII* SSP. *BULGARICUS* B1 IN A COMPLEX NUTRIENT MEDIUM (MRS-BROTH)

Georgi Kostov*, Rositsa Denkova-Kostova**, Vesela Shopska*, Bogdan Goranov***, Zapryana Denkova***

*Department of Wine and Beer ** Department of Biochemistry and Molecular Biology, *** Department of Microbiology

University of Food Technologies, 4002, 26 Maritza Blvd., Plovdiv, Bulgaria

E-mail: george_kostov2@abv.bg; rositsa_denkova@mail.bg; vesi_nevelinova@abv.bg; dr.eng.bgoranov@gmail.com; zdenkova@abv.bg

KEYWORDS

Probiotics, growth kinetics, modeling, optimization, complex nutrient medium, MRS-broth.

ABSTRACT

The microbial growth of the probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1, cultivated in a complex nutrient medium (MRS-broth), was studied in the present work. The complex nutrient medium provides not only the carbon source necessary for the growth of biomass, but also all the additional sources of nitrogen, phosphorus and other components that the biomass needs for its growth. The use of non-structural mathematical dependences determines the optimal conditions (substrate concentration) for the accumulation of biomass or lactic acid, depending on the needs of the specific production.

INTRODUCTION

In recent years, there has been increased interest in the use of lactic acid bacteria of the species *Lactobacillus*, *Enterococcus*, *Pediococcus*, *Streptococcus*, *Lactococcus* and *Leuconostoc* for the development of probiotic and synbiotic preparations (Gibson, 2004). In order for a strain of these species to be classified as probiotic, it must meet a number of requirements, one of which is to allow the conduction of industrial cultivation (Saarela et al., 2002; Kostov et al., 2021).

To evaluate this property it is necessary to apply microbial kinetics, which can be used to assess parameters such as: specific growth rate (maximum and current value), specific rate of product accumulation (maximum and current value), different types of constants. saturation, inhibition, etc.). The combination of these parameters makes it possible to assess the growth of microorganism biomass (in particular lactic acid bacteria biomass), to determine the optimal growth conditions that can ensure the accumulation of biomass and/or metabolic products (Bouguettoucha et al., 2011). In their classical work Baily and Ollis, 1986, proposed different types of models to describe the microbial growth kinetics. These dependencies are based on the S-shaped nature of microbial growth and are divided into four main groups: non-structural non-segregated models; non-structural segregated models; structural non-

segregated models and structural segregated models (Bailey and Ollis, 1986). Nonstructural models view the growth of the microbial population as a whole. When applied, it is assumed that the microbial population grows in the conditions of unlimited food sources, unlimited space and lack of factors related to the vital activity of microorganisms. These models follow from the so-called equation of exponential growth (1) and the well-known Monod dependence (2):

$$\frac{dX}{d\tau} = \mu X \quad (1)$$

$$\mu = \mu_m \frac{S}{K_s + S} \quad (2)$$

where: μ_m - maximum specific growth rate, h^{-1} ; X - biomass concentration, g/dm^3 ; S - substrate concentration, g/dm^3 ; K_s - saturation constant, g/dm^3 .

A number of non-structural models have been developed based on the Monod equation and they have been usually named after the researcher who proposed them. Such examples are the Tiessier model, the Andrews and Noack model, the Hinshelwood model, the Aiba model, the Ghose and Tyagi model and others, that try to solve various aspects of microbial growth (substrate inhibition; product inhibition; product and substrate inhibition, etc.) (Bailey and Ollis, 1986; Bouguettoucha et al., 2011; Kostov, 2015; Muloiwa et al, 2020). In the present paper we are going to consider other examples as well (see Materials and methods).

Non-structural models describe only the amount of biomass and/or the amount of metabolites accumulated. Thus, they do not reflect the qualitative characteristics of the cell population and the changes that occur in it during cultivation. These changes can only be described by structural models. They are based on the material balance equations, but in their construction it is necessary to select the key changes taking place in the population. In this model type one works with the concentration of the corresponding variable in a volume unit of biophase, taking into account the cell density, the rate of component formation, the cell mass and more. These models are usually quite complex and include a large number of variables that do not always have a clear and precise biological meaning (Bailey and Ollis, 1986; Kostov, 2015; Shopska et al., 2019).

One of the most well known species of lactic acid bacteria is *Lactobacillus delbrueckii* ssp. *bulgaricus*.

Representatives of this species are included as starter cultures for the production of various types of food, as well as for the production of probiotic preparations (Arena et al., 2015; Maisto et al., 2021; Ivanov et al., 2021).

The ability to accumulate large amounts of biomass in the cultivation of lactic acid bacteria is very important for the production of probiotics. Complex nutrient media are usually used for the cultivation process. One of the most frequently used media for the cultivation of lactic acid media is MRS-broth medium (de Man, Rogosa and Sharpe). MRS-broth medium has been developed primarily for the cultivation of lactobacilli from various sources with the intention of producing a defined medium as a substitute for tomato juice agar. It is used for the cultivation of the whole group of lactic acid bacteria. The medium shows good productivity for nearly all lactic acid bacteria, but the original version is not selective. It was made selective for lactic acid bacteria by lowering the pH to 5.7 and the addition of 0.14% sorbic acid. Some strains from dairy sources show reduced growth rates in MRS. MRS agar is composed of tryptic digest of casein, beef extract, yeast extract, glucose, sorbitan monooleate, di-potassium hydrogen orthophosphate, magnesium sulfate, manganese (II) sulfate, ammonium citrate, sodium acetate, agar, and distilled or deionized water (Corry et al., 2003).

The main metabolite of lactic acid fermentation is lactic acid. It is known that its increasing concentration during fermentation has an inhibitory effect on the growth of the microbial population. The sensitivity to the accumulating lactic acid is strain-specific (Bouguettoucha et al., 2011; Gordeev et al., 2017).

The aim of the present work was to study the growth characteristics of the probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* when cultivated in a complex nutrient medium such as MRS-broth. The strain has demonstrated a number of probiotic characteristics and had been isolated from homemade yogurt (Goranov et al, 2015; Teneva et al., 2015). As already commented, in some cases, strains isolated from dairy products show reduced growth in MRS-broth medium. Six non-structural mathematical models (see Materials and methods) based on the Monod equation were used to model the microbial growth. The obtained data were used to determine the optimal concentrations of the complex food source in order to improve the accumulation of biomass or lactic acid.

MATERIALS AND METHODS

Microorganisms and cultivation conditions

The study was conducted with *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 isolated from home made yogurt (Goranov et al., 2015; Teneva et al., 2015).

The strain was cultivated in MRS-broth, produced by Merck, with the following qualitative and quantitative composition (g/dm³): peptone from casein - 10.0; meat

extract - 8.0; yeast extract - 4.0; D(+)-glucose - 20.0; dipotassium hydrogen phosphate - 2.0; Tween[®] 80 - 1.0; di-ammonium hydrogen citrate - 2.0; sodium acetate - 5.0; magnesium sulfate - 0.2; manganese sulfate - 0.05.

The cultivation was performed in a bioreactor with mechanical stirring, shown in Fig.1. The apparatus has a geometric volume of 2 dm³ and a working volume of 1.5 dm³ and is equipped with a Sartorius A2 control device, which includes all the measuring instruments for the fermentation process: temperature, pH, dissolved oxygen, etc. The fermentation process was carried out at a stirring speed of 150 rpm at 37±1°C.

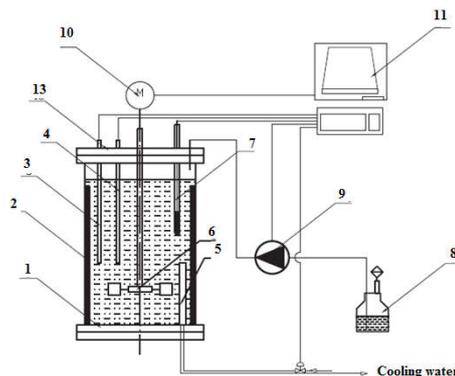


Figure 1: Laboratory Bioreactor

1 - vessel with a geometric volume of 2 dm³; 2 - baffles; 3 - temperature electrode (thermometer); 4 - cooling/heating device (water jacket); 5 - an additional cooling/heating device; 6 - turbine stirrer; 7 - pH/Eh electrode; 8 - fermentation medium/inoculum/pH adjustment medium; 9 - peristaltic pump; 10 - stirrer drive; 11 - Sartorius A2 control device;

Methods of analysis and nutrient medium for analysis

- Determination of titratable acidity (ISO/TS 11869:2012);
- Determination of number of viable lactobacilli cells (ISO 7889:2005).
- Nutrient media (ISO 7889:2005)
 - MRS-broth;
 - MRS-agar;
 - Saline solution.

Modeling of microbial growth and identification of parameters in kinetic models

The following system of differential equations was used to model the kinetics of microbial growth:

$$\begin{cases} \frac{dX}{d\tau} = \mu(\tau) X(\tau) \\ \frac{dP}{d\tau} = q(\tau) X(\tau) \\ \frac{dS}{d\tau} = -\frac{1}{Y_{X/S}} \frac{dX}{d\tau} - \frac{1}{Y_{P/S}} \frac{dP}{d\tau} \end{cases} \quad (3)$$

where: X – biomass concentration, g/dm³; P – lactic acid concentration, g/dm³; S – substrate concentration, g/dm³; Y_{P/S}, Y_{X/S} – yield coefficients; μ – specific growth rate, h⁻¹; q – specific lactic acid accumulation rate, g/(g.h).

The following dependences were used to model the biomass specific growth rate and the specific rate of lactic acid accumulation (Bailey and Ollis, 1986; Bouguettoucha et al., 2011; Kostov, 2015; Muloiwa et al, 2020):

- Monod model

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S} \\ q &= q_{p\max} \frac{S}{K_{SP} + S}\end{aligned}\quad (4)$$

- Haldane model

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \\ q &= q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}}\end{aligned}\quad (5)$$

- Aiba model

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S} \exp(-K_{PX}P) \\ q &= q_{p\max} \frac{S}{K_{SP} + S} \exp(-K_{PP}P)\end{aligned}\quad (6)$$

- Haldane-Aiba model

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \exp(-K_{PX}P) \\ q &= q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \exp(-K_{PP}P)\end{aligned}\quad (7)$$

- Haldane model for product inhibition

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \left(1 + \frac{P}{K_{PX}}\right) \\ q &= q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \left(1 + \frac{P}{K_{PP}}\right)\end{aligned}\quad (8)$$

- Haldane-Jerusalimski model

$$\begin{aligned}\mu &= \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \left(\frac{K_{PX}}{P + K_{PX}}\right) \\ q &= q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \left(\frac{K_{PP}}{P + K_{PP}}\right)\end{aligned}\quad (9)$$

where: μ_{\max} – maximum specific growth rate, h^{-1} ; $q_{p\max}$ – maximum specific rate of lactic acid formation, h^{-1} ; K_{SX} и K_{SP} – Monod constants for saturation of biomass and product by substrate, g/dm^3 ; K_{Xi} и K_{Pi} – substrate inhibition constants for biomass and product, g/dm^3 ; K_{PX} и K_{PP} – product inhibition constants for biomass and product, g/dm^3 .

The parameters in the kinetic equations are calculated by solving the system of differential equations using the Runge-Kuta method of the 4th row, by minimizing the sum of the squares of the difference between the experimental and model data. The software used was Microsoft Excel 2013 (Choi et al., 2014).

RESULTS AND DISCUSSION

The dynamics of the studied fermentation process in the complex nutrient medium MRS-broth is presented in Fig. 2.

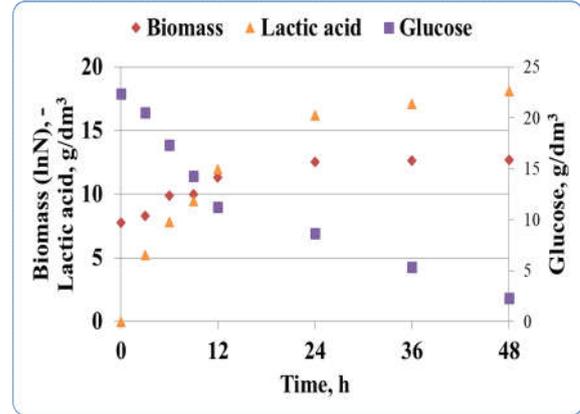


Figure 2: Dynamics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth in Bioreactor with Stirring

The data show that the fermentation process developed according to the trends for this process type. The duration of the lag phase was about 3 hours, after which the culture entered the exponential growth phase. The exponential growth phase lasted about 24 hours, and at the end of this phase a high number of viable cells was reached - 12.57 logarithmic units. In the next 24 hours, the culture was in the stationary phase, and it retained the high number of viable cells. In this phase, the substrate continued to be utilized at a high rate and lactic acid was constantly accumulating. At the end of the process, the lactic acid concentration reached 18.09 g/dm^3 and the substrate concentration decreased to 2.3 g/dm^3 . The unutilized substrate was due to the influence of the product and the substrate inhibition processes, which should be taken into account. This means that in order to optimize the process, opportunities for the complete utilization of the substrate should be sought, which is achieved by optimizing the concentration of the substrate. It is known that the process of cultivation of lactic acid bacteria is substrate and product dependent (inhibited) (Bouguettoucha et al., 2011; Kostov, 2015). The first two models we are going to discuss at are the classic Monod model and the Haldane model. The data for the kinetic parameters and the errors of the models are shown in Table 1, and the convergence of the models to the experimental data is given in Fig. 3 and Fig. 4.

Table 1: Kinetic Constants in the Different Models Used to Describe the Microbial Growth Kinetics of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1

	Monod	Haldane	Aiba	Haldane-Aiba	Haldane for product inhibition	Haldane-Jerusalimski
μ_{max} , h ⁻¹	0.072	0.038	0.576	0.049	0.113	0.011
K_{SX} , g/dm ³	43.05	32.51	213.78	47.59	65.69	17.55
K_{Xi} , g/dm ³	-	85.45	-	156.67	163.23	283.82
q_{pmax} , h ⁻¹	0.065	0.3982	0.266	0.063	0.256	0.065
K_{SP} , g/dm ³	11.13	132.36	29.71	0.005	29.12	41.39
K_{SPi} , g/dm ³	-	441.11	-	180.72	96.60	169.39
K_{PX} , g/dm ³	-	-	0.16	0.249	22.38	70.09
K_{PP} , g/dm ³	-	-	0.11	0.138	18.53	3.36
$1/Y_{x/s}$	0.3416	0.3055	0.6181	0.3318	0.3190	0.6680
$1/Y_{p/s}$	12	0.9820	2.2487	1.9694	0.0400	0.0400
$Y_{x/s}$	2.9277	3.2733	1.6176	3.0139	3.1348	1.4970
$Y_{p/s}$	0.0833	1.0183	0.4021	0.5078	25	25
R^2 (biomass)	0.7895	0.8473	0.8860	0.8712	0.9378	0.871
Error (biomass)	0.84	0.69	0.33	0.27	0.38	0.48
R^2 (product)	0.9507	0.9720	0.9763	0.983	0.9814	0.8895
Error (product)	2.63	2.42	1.42	1.38	1.42	1.80
R^2 (substrate)	0.9194	0.9361	0.9946	0.9907	0.9943	0.9697
Error (substrate)	2.75	2.25	1.27	1.23	1.27	1.12
S_{opt} (biomass), g/dm ³	-	52.71	-	86.35	103.55	70.58
S_{opt} (product), g/dm ³	-	241.63	-	0.96	53.04	82.37

The data from Fig. 3 and Fig. 4, as well as those in Table 1, show that the Monod model and the Haldane model agree very well with the experimental data.

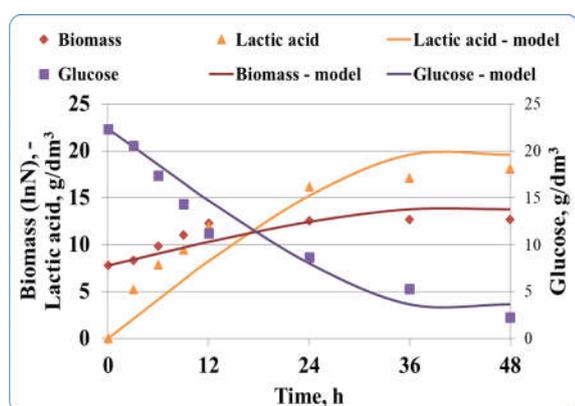


Figure 3: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Monod Model

The correlation coefficients range from 0.7895 to 0.9720. The values of the calculated errors vary in the range of 0.69 to 2.25. The Monod model gives almost twice the maximum specific growth rate (0.072 h⁻¹) compared to the Haldane model - 0.038 h⁻¹, due to the fact that the Haldane model takes into account the substrate inhibition of the lactic acid fermentation process. In both models there is an increased saturation constant of the substrate - 43.05 g/dm³ and 32.51 g/dm³, respectively, (in this case the substrate is equated to the concentration of the carbon source - glucose 20 g/dm³), which confirms the observation that glucose is the substrate limiting the fermentation process. The two

models also give different values with respect to the maximum specific lactic acid accumulation rate. The fact that the Haldane model gives about 6 times higher rate of acid formation (0.3982 h⁻¹) than the Monod model (0.065 h⁻¹) is very interesting. This also leads to significant differences in the saturation constants by product - 11.13 g/dm³ for the Monod model and 132.36 g/dm³ for the Haldane model. This difference shows that according to the Haldane model the rate of acid formation depends to a greater extent on the concentration of the limiting substrate.

It is interesting to determine theoretically at what substrate concentrations the culture will undergo substrate inhibition. Information on this is given by the substrate inhibition constants for biomass and product in the Haldane model - K_{Xi} and K_{SPi} , respectively. From the data presented in Table 1 it can be seen that the inhibitory effect of the substrate on cell proliferation and growth will begin to be observed at $K_{Xi} = 85.45$ g/dm³ and $K_{SPi} = 441.11$ g/dm³, which is 8.545% and 44.11% substrate (glucose) in the nutrient medium, respectively. K_{Xi} is close to the experimental results of various authors who found that at concentrations of the substrate (glucose) in the nutrient medium higher than 10%, its inhibitory effect on the specific growth rate of lactic acid bacteria becomes noticeable.

However, the K_{SPi} calculated by the Haldane model has an abnormally high value, which deviates greatly from the K_{Xi} . In this case, the value of K_{SPi} is real from mathematical point of view, but not from biological point of view, because such high glucose concentration in the medium will not allow the growth and accumulation of high numbers of viable cells that actively produce lactic acid. It should be noted that the

inhibition may be due not only to the carbon source, but also to some of the complex components in the nutrient medium. However, this is difficult to account for with non-structural models that are usually used to describe the fermentation process.

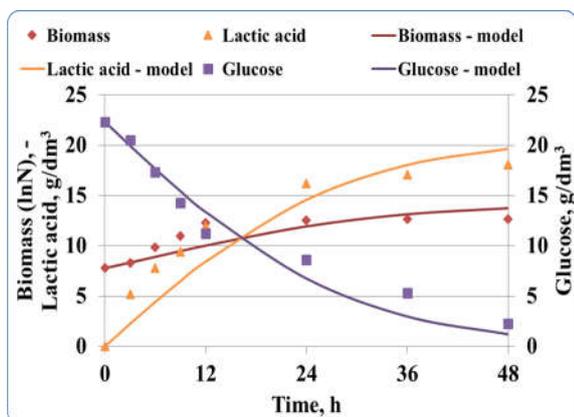


Figure 4: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane Model

It is also interesting to determine the metabolic (trophic) coefficients - $1/Y_{X/S}$ and $1/Y_{P/S}$, showing the consumption of substrate for biomass growth and for product synthesis. Since they are the reciprocal values of the respective economic coefficients, the latter can also be determined. From the data presented in Table 1 it is evident that both models show higher substrate consumption for biomass formation – the trophic coefficients being 0.3416 and 0.3055, respectively, while the economic coefficients being 2.9277 and 3.2733, respectively, and a smaller part of the substrate goes to lactic acid synthesis ($1/Y_{P/S}$ - 12 and 0.9820, respectively, and $Y_{P/S}$ - 0.083 and 1.0183, respectively). The Haldane mathematical model has an advantage over the Monod model - one can determine theoretically what the optimal substrate concentration will be at the optimal (maximum) specific growth rate:

$$\mu = \mu_{\max}^{opt} \Rightarrow S_{opt} = \sqrt{K_{SX} K_{SXi}} \quad (10)$$

Similarly, the optimal substrate concentration at which the rate of lactic acid synthesis will be optimal (maximum value) can be determined:

$$q_p = q_{p\max}^{opt} \Rightarrow S_{opt} = \sqrt{K_{SP} K_{SPi}} \quad (11)$$

Then, according to the Haldane model, S_{opt} for the growth and reproduction of the strain will be 52.71 g/dm³, which is 5.271% glucose in the nutrient medium. This value is also close to experimentally determined values by other authors (Bouguettoucha et al., 2011). Here, too, it should be noted that the concept of substrate should be considered as a balanced set of components that are necessary for cell growth. In this case, the result obtained is very close to the total amount of components in the used complex nutrient medium. For the acid formation rate S_{opt} is 241.63 g/dm³.

According to these data obtained from the Haldane model, it can be concluded that for *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 the substrate concentration should be in the range from 52.71 g/dm³ to 241.63 g/dm³, and if the substrate concentration is above 241.63 g/dm³ there will be complete inhibition of both the growth of the strain and its biosynthetic ability. Lactic acid fermentation is also a product-inhibited process (Bouguettoucha et al., 2011;), which is why the growth kinetics and lactic acid biosynthesis of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 with the Aiba model have been modeled. The results are shown in Fig.5, and the parameters of the model are presented in Table. 1.

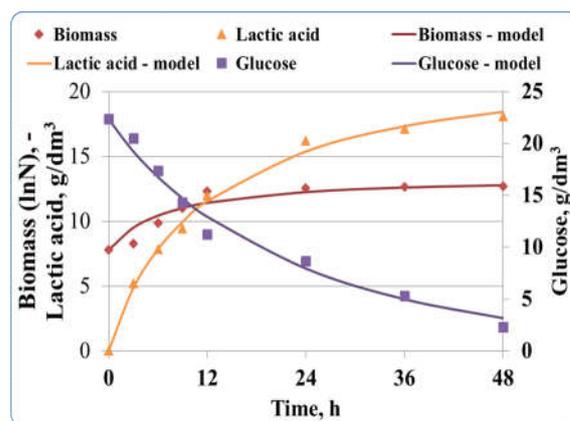


Figure 5: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Aiba Model

The Aiba model is characterized by high correlation values ranging from 0.8860 to 0.9946 and low identification errors. It gives relatively high rates of the specific growth rate - 0.576 h⁻¹ and 0.266 h⁻¹, but also confirms product inhibition. This is evidenced by the relatively close values of the constants K_{PX} and K_{SP} - 0.16 g/dm³ and 0.11 g/dm³. This in turn confirms that the cultivation process of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 must be carried out with neutralization of the lactic acid produced in order to achieve high maximum growth rate of the culture and high concentration of active cells. Low values of the product inhibition constants can also be considered as an indirect indicator of the sensitivity (resistance) of the strain to the acidic pH of the stomach, which means that this strain is good to be used in encapsulated form, as a probiotic strain. The presence of product inhibition increases the saturation constant value for biomass (213.78 g/dm³). The Aiba model again shows that most of the substrate is used for biomass formation. The values of the trophic coefficients - $1/Y_{X/S}=0.6182$ and $1/Y_{P/S}=2.2487$, and therefore the values of the economic coefficients $Y_{X/S}=1.6176$ and $Y_{P/S}=0.4021$ serve as a proof of this conclusion.

The data described so far show that lactic acid fermentation is both a substrate- and a product-inhibited process, which should be taken into account in its

modeling. Combined models such as the Haldane-Aiba model (equation 7), the Haldane model for product inhibition (equation 8) and the Haldane-Jerusalimski model (equation 9) can be used for this purpose. The results of the modeling of the cultivation process of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 are reflected in Fig. 6 to Fig. 8, as well as in Table 1.

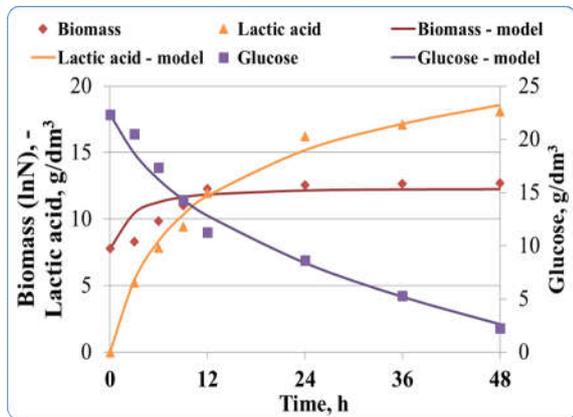


Figure 6: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane-Aiba Model

The data in Table. 1 show that all three models, including product and substrate inhibition, describe experimental data with high accuracy. The correlation coefficients vary in the range of 0.87 to 0.9943, the errors - in the range of 0.27 to 1.8 for the individual indicators.

The data in Table 1 show that the Haldane-Aiba model (Equation 7) and the Haldane-Jerusalimski model (Equation 9) predict lower biomass specific growth rate of 0.049 h^{-1} and 0.011 h^{-1} , respectively. The Haldane model for product inhibition (Equation 8) predicts growth rate of 0.113 h^{-1} . This difference is due to the approach used by the three models to describe the processes of product and substrate inhibition. The Haldane-Jerusalimski model determines lower value of the saturation constant - 17.55 g/dm^3 compared to the Haldane-Aiba model - 47.59 g/dm^3 and the Haldane model for product inhibition - 65.69 g/dm^3 . Despite this difference, the results are within the range expected in the cultivation of lactic acid bacteria. The data in Table 1 show that the three models predict growth inhibition by the substrate at concentrations above 15.67%, i.e. well above the current value of glucose in the MRS-broth medium. The models show that complete inhibition of growth by the substrate will occur only at glucose concentrations in the medium above 28.38%, and such values are not typical for nutrient media designed for the cultivation of lactic acid bacteria.

Data on the specific rate of lactic acid formation are of great interest in the enlisted models. The Haldane-Aiba model and the Haldane-Jerusalimski model give quite low values of q_{pmax} - 0.063 h^{-1} - 0.065 h^{-1} , which means that the MRS-broth medium is designed to allow enhanced synthesis of biomass at the expense of lactic

acid production. The Haldane model for product inhibition predicts more intense process of lactic acid biosynthesis, as evidenced by the significantly higher maximum specific rate of acid formation (Table 1).

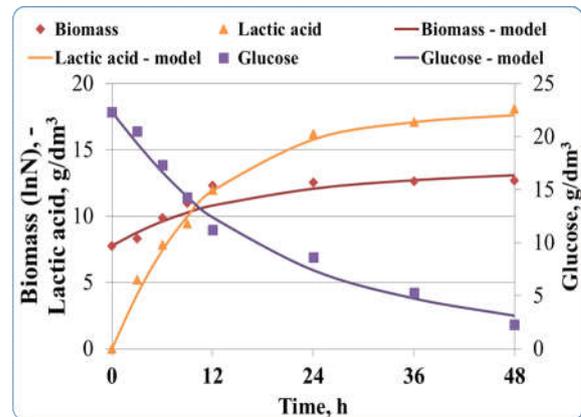


Figure 7: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane model for Product Inhibition

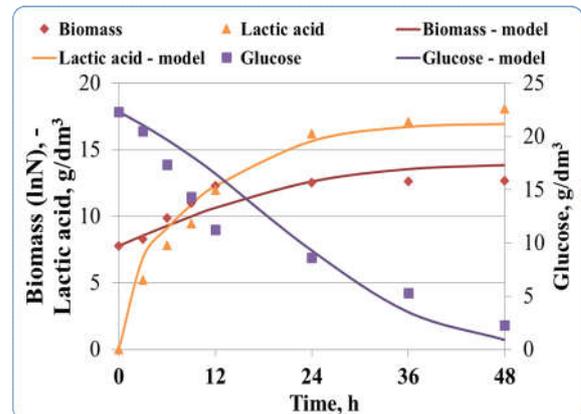


Figure 8: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane-Jerusalimski Model

However, the three models give different degrees of influence of lactic acid formation on cell growth. The Haldane-Aiba model shows relatively strong product inhibition due to low K_{PX} and K_{PP} values of 0.249 g/dm^3 and 0.138 g/dm^3 , respectively. The Haldane model for product inhibition gives higher K_{PX} and K_{PP} values - 22.38 g/dm^3 and 18.53 g/dm^3 , which according to this model means that the process is not product inhibited so strongly. The value of the product inhibition constant for the biomass given by the Haldane-Jerusalimski model is abnormally high. From a mathematical point of view, this value is correct and brings the values calculated by the model closer to the experimental ones, but its biological meaning is doubtful, as this is too high a concentration of lactic acid at which the specific growth rate would be half the maximum value. For K_{PP} this model gives a biologically realistic value - 3.36 g/dm^3 . The three models, including product and substrate inhibition, also allow the determination of the optimal

substrate concentrations to provide maximum (optimal) specific growth rate or acid formation (Table 1). The data presented in the table show that the Haldane-Aiba model and the Haldane-Jerusalimski model give close values of the optimal glucose concentration - 86.35 g/dm³ and 70.58 g/dm³, respectively. Once again, one must recall that the models determine the optimal concentration of the balanced set of components, rather than just the concentration of the carbon source in the medium. The Haldane model for product inhibition rather sets the substrate concentration limit (103.55 g/dm³) at which inhibition of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 growth will begin.

Similar conclusions can be made for S_{opt} for lactic acid synthesis. This time, however, the Haldane model for product inhibition and the Haldane-Jerusalimski model predict substrate concentrations that are characteristic of lactic acid bacteria. Only the Haldane-Aiba model showed an abnormally low value for S_{opt} - 0.96 g/dm³.

The results obtained (Fig. 2 to Fig. 8 and Table 1) do not support the statement cited in the introduction that the MRS-broth medium may not be suitable for the cultivation of strains originating from dairy products. In the cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth data show that the strain grew at relatively high growth rates and lower rates of lactic acid accumulation. The effects of substrate and product inhibition are normal for this type of fermentation, which allows relatively higher concentrations of viable cells to accumulate at the end of fermentation.

CONCLUSION

An important requirement for selection of strains to be included in the production of probiotic foods and preparations is the ability of the selected strains to be cultivated in industrial conditions and to accumulate high concentrations of viable cells. In the present work, the cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1, isolated from home-made yoghurt, cultivated in a complex culture medium (MRS-broth) was studied. Complex media provide a balanced set of components - carbon, nitrogen and phosphorus source, micro- and macroelements. These media are usually designed to provide optimal growth, but in some cases are unsuitable for certain strains. The obtained results show that the selected probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 grew at relatively high specific growth rates and accumulated moderate amounts of lactic acid, determined by moderate specific acid formation rates. The data show that the strain may be sensitive to lactic acid, which is why pH adjustment and neutralization of lactic acid accumulated in the medium can be applied in industrial cultivation. This will ensure complete absorption of the substrate by the cells and the accumulation of maximum cell numbers in the medium.

The data obtained show that, although of dairy origin, *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 can be cultivated in the complex nutrient medium MRS-broth.

REFERENCES

- Arena, M.P., G. Caggianiello, P. Russo, M. Albenzio, S. Massa, D. Fiocco, V. Capozzi, G. Spano. 2015. "Functional Starters for Functional Yogurt" *Foods*, 4 (1), 15-33. <https://doi.org/10.3390/foods4010015>
- Bailey, J. E. and Ollis, D. F. (1986). *Biochemical Engineering Fundamentals*, 2nd edition. McGraw-Hill.
- Bouguettoucha, A., B. Balanec and A. Amrane. 2011. "Unstructured Models for Lactic Acid Fermentation: A Review." *Food Technol. Biotechnol.*, 49 (1), 3–12.
- Choi, M., M. Saeed Al-Zahrani, and S. Y. Lee. 2014. "Kinetic model-based feed-forward controlled fed-batch fermentation of *Lactobacillus rhamnosus* for the production of lactic acid from Arabic date juice." *Bioprocess Biosyst Eng.*, 37, 1007–1015. <https://doi.org/10.1007/s00449-013-1071-7>
- Corry, J., G. Curtis, R., Baird. 2003. "Handbook of culture media for food microbiology", (Chapter de Man, Rogosa and Sharpe (MRS) agar), *Progress in Industrial Microbiology*, 37, 511-513. [https://doi.org/10.1016/S0079-6352\(03\)80066-8](https://doi.org/10.1016/S0079-6352(03)80066-8)
- Gibson, G. R. 2004. "From probiotics to prebiotics and a healthy digestive system". *J. Food Science*, 69 (5), M141- M143. <https://doi.org/10.1111/j.1365-2621.2004.tb10724.x>
- Goranov, B., V. Shopska, R. Denkova, G. Kostov, Georgi. 2015. "Kinetics of batch fermentation in the cultivation of a probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1" *Acta Universitatis Cibiniensis. Series E: Food Technology*, 19 (1), 61-72. <https://doi.org/10.1515/auft-2015-0006>
- Gordeev, L., A. Koznov, A. Skichko, and Y. Gordeeva. 2017. "Unstructured mathematical models of the lactic acid biosynthesis kinetics: A Review." *Theoretical Foundations of Chemical Engineering*, 51 (2), 175-190.
- ISO/TS 11869:2012. Fermented milks — Determination of titratable acidity — Potentiometric method
- ISO 7889:2005. Yogurt — Enumeration of characteristic microorganisms — Colony-count technique at 37 degrees C
- Ivanov, I., K. Petrov, V. Lozanov, I. Hristov, Zh. Wu, Zh. Liu, P. Petrova. 2021. "Bioactive compounds produced by the accompanying microflora in Bulgarian yoghurt" *Processes* 9 (1), 114. <https://doi.org/10.3390/pr9010114>
- Kostov, G. 2015. "Intensification of fermentation processes by immobilized biocatalysis". DSc Thesis, University of Food Technologies, Plovdiv. p. 307. (in Bulgarian).
- Kostov, G., R. Denkova-Kostova, V. Shopska, B. Goranov, Z. Denkova. 2021. „Comparative evaluation of *Lactobacillus plantarum* strains

through microbial growth kinetics”, In *ECMS 2021 Proceedings* Kh. Al-Begain, M. Iacono, L. Campanile, A. Bargiela (Eds.), European Council for Modeling and Simulation.
<https://doi.org/10.7148/2021-0165>

- Maisto, M., G. Annunziata, E. Schiano, V. Piccolo, F. Iannuzzo, R. Santangelo, R. Ciampaglia, G. C. Tenore, E. Novellino, P. Grieco. 2021. "Potential Functional Snacks: Date Fruit Bars Supplemented by Different Species of *Lactobacillus* spp." *Foods*, 10 (8), 1760. <https://doi.org/10.3390/foods10081760>
- Muloiwa, M., S. Nyende-Byakika, M. Dinka. 2020. "Comparison of unstructured kinetic bacterial growth models". *South African Journal of Chemical Engineering*, 33, 141-150.
<https://doi.org/10.1016/j.sajce.2020.07.006>
- Saarela, M., L. Zahteenmaki, R. Crittenden, S. Salminen, and T. Mattila-Sandholm. 2002. "Gut bacteria and health foods – the European perspective." *Int. J. Food Microbiol.* 78, 99-117.
- Shopska, V., R. Denkova, V. Lyubenova, G. Kostov. 2019. "Kinetic Characteristics of Alcohol Fermentation in Brewing: State of art and control of the fermentation process". In *Fermented Beverages*; Grumezescu, A.M., Holban, A.M., Eds.; Woodhead Publishing: Cambridge, UK, 2019; pp. 529–575.
<https://doi.org/10.1016/B978-0-12-815271-3.00013-0>
- Teneva, D., B. Goranov, R. Denkova, Z. Denkova. 2015. "Antimicrobial activity of *Lactobacillus delbrueckii* ssp. *bulgaricus* strains against *Candida albicans* NBIMCC 74". Scientific works of the University of Ruse, 54, series 10.4, 20-25.

ACKNOWLEDGEMENTS

This work were supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Healthy Foods for a Strong Bio-Economy and Quality of Life" approved by DCM № 577/17.08.2018 and by the project "Strengthening the research excellence and innovation capacity of University of Food Technologies - Plovdiv, through the sustainable development of tailor-made food systems with programmable properties", part of the European Scientific Networks National Programme funded by the Ministry of Education and Science of the Republic of Bulgaria (agreement № Д01-288/07.10.2020).

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is a full professor at the Department of Wine and Beer Technology at the University of Food Technologies, Plovdiv. He received his MSc degree in Mechanical Engineering in 2007, a PhD degree in Mechanical Engineering in the Food and Flavor Industry (Technological Equipment in the Biotechnology Industry) in 2007 at the University of Food Technologies, Plovdiv, and holds a DSc degree in

Intensification of Fermentation Processes with Immobilized Biocatalysts from 2015. His research interests are in the area of bioreactor construction, biotechnology, microbial population investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, and beer production.

VESELA SHOPSKA is an associated professor at the Department of Wine and Beer Technology at the University of Food Technologies, Plovdiv. She received her MSc degree in Wine-making and Brewing Technology in 2006 at the 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.

ROSITSA DENKOVA-KOSTOVA is an associated professor at the Department of Biochemistry and Molecular Biology at the University of Food Technologies, Plovdiv. She received her MSc degree in Industrial Biotechnologies in 2011 and a PhD degree in Biotechnology (Technology of Biologically Active Substances) in 2014. Her research interests are in the area of isolation, biochemical and molecular-genetic identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is a chief assistant professor at the department of Microbiology at the University of Food Technologies, Plovdiv. He received his PhD in 2015 from the 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 the Food Industry". His research interests are in the area of bioreactor construction, biotechnology, microbial population investigation and modeling, hydrodynamics and mass transfer problems, and fermentation kinetics.

ZAPRYANA DENKOVA is a full professor at the department of Microbiology at the 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.

SYNERGY BETWEEN SHUTTLES AND STACKER CRANES IN DYNAMIC HYBRID PALLET WAREHOUSES: CONTROL STRATEGIES AND PERFORMANCE EVALUATION

Giulia Siciliano, Yue Yu and Johannes Fottner
Chair of Materials Handling, Material Flow, Logistics
Technical University of Munich
Boltzmannstraße 15, Garching bei München, Germany
E-mail: giulia.siciliano@tum.de

KEYWORDS

Dynamic Hybrid Pallet Warehouse, Shuttle, Stacker Crane, Discrete Event Simulation, Control Strategies

ABSTRACT

This article considers two dynamic hybrid pallet warehouses obtained hybridizing a shuttle-based warehouse with stacker cranes. We begin by describing their design and characteristics. Afterwards, we explain the control algorithms that were developed for them. Next, we illustrate the modalities of the discrete event simulation study we ran to investigate their performance. In conclusion, we discuss the results in terms of throughput of the simulation study to individuate the field of application for the two layouts of dynamic hybrid pallet warehouses in comparison to stacker crane-based and shuttle-based warehouses.

INTRODUCTION

A dynamic hybrid pallet warehouse (DHPW) is a new kind of storage and retrieval system that has a shuttle tier on the base connected to the overlying storage layers through satellite stacker cranes. This arrangement allows a combination of the advantages of shuttle-based and stacker-crane-based warehouses (Eder et al., 2019) (Siciliano et al., 2020).

In recent years, another warehouse was investigated that contemplates the simultaneous use of shuttles and a stacker crane. This warehouse is denoted as autonomous shuttles and stacker crane (AS/SC) warehousing system. Its shuttles move orthogonally to the stacker crane's aisle and therefore can only use the Last In First Out (LIFO)

policy. In addition, so far only one stacker crane per aisle has been implemented. (Wang et al., 2020)

On the contrary, the shuttles of a DHPW can move in both directions of the plane and up to three stacker cranes per aisle have been coordinated and investigated in (Siciliano et al., 2022). To increase the throughput of a DHPW, specific order assignment strategies (Siciliano and Fottner, 2021) and specific stacker cranes' coordination policies (Siciliano et al., 2022) should be applied that take into consideration the complex nature of the connection between shuttle tier and multiple stacker cranes in a single aisle. To investigate the nature of the connection between shuttles and stacker cranes in more detail, and to find further applications for DHPWs, this article examines two additional warehouse arrangements, which we define as layout 2 and layout 3. We call the original DHPW with the channel storage above the shuttle base layout 1. In the following section, we describe the characteristics of layout 2 and layout 3 compared to layout 1.

Systems under consideration

Layout 2 and layout 3 have shuttle tiers on not only the base but also on the levels (Malik 2014), see Fig.1.

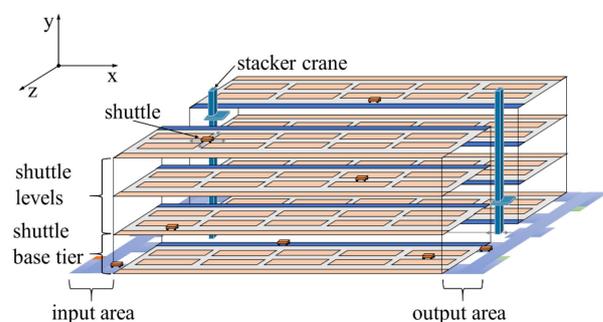


Figure 1: Structure of both Layouts 2 and 3

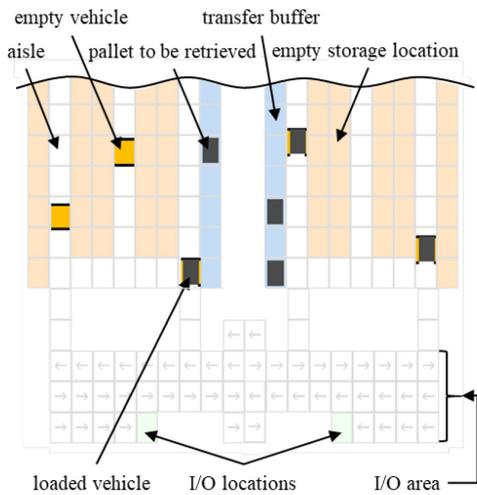


Figure 2: Screenshot of the Base Tier Model for both Layouts 2 and 3

Fork lift stacker cranes serve the transfer buffers of base tier and of the levels. Each shuttle remains in its zone i.e. left or right side of the aisle on a certain level in the warehouse. In layout 2, the shuttles cannot leave their level, while in layout 3 the shuttles can be transported by the stacker cranes between levels. The elements that make up the base tier of layout 2 and 3 are shown in Fig.2. The levels contain the same elements as the base tier, except for the fact that they lack input/output (I/O) areas. On one hand, having shuttle tiers on every level increases the investment and operational costs compared to layout 1. On the other hand, it enables better access to stored products compared to channel storage. Therefore, layout 1 can be seen as the result of the hybridization of a stacker crane-based warehouse through shuttles, while layout 2 and layout 3 are the hybridization of a shuttle-based warehouse through stacker cranes. Compared to a conventional shuttle-based warehouse with lifts, the stacker cranes' aisles in layout 2 and layout 3 offer a much more efficient means of material exchange between the base and upper levels. In fact, the transfer buffers on the base and on all levels along the whole length of the aisle provide many more exchange locations than the conventional few I/O locations of lifts. Thus, layouts 2 and 3 can achieve a higher throughput than conventional shuttle-based systems. In the following section we propose control strategies for layout 2 and layout 3.

CONCEPT DEVELOPMENT

To explain the control algorithms that were developed, we have to consider layout 2 and layout 3 separately. We implemented the algorithms in the cases of retrieval, storage and double cycles. A double cycle is the alternation of retrieval and storage orders for the shuttles. The same is true for the stacker cranes in the aisle. Therefore, retrieval and storage control strategies can be derived from the strategy for double cycles. We only discuss double cycles for the sake of brevity.

Layout 2

We first consider layout 2. The control strategy we developed for this in the case of double cycles is described in Fig. 3 for the shuttles on the base, in Fig. 4 for the shuttles on the levels, and in Fig. 5 for the stacker cranes. Abbreviations “ C_nS ” and “ C_nE ” indicate respectively start and end of connection n between shuttles and stacker cranes. The challenge compared to layout 1 is to connect and coordinate the stacker cranes with the shuttles on not only the base, but also on the different levels. For sake of completeness, we illustrate the connections between shuttles and stacker cranes for the execution of a double cycle. First, a shuttle on base executes a storage order by bringing a pallet from the I location to an available location of the transfer buffer. The shuttle then creates a storage order for the stacker crane to transport that pallet from the transfer buffer on the base to the transfer buffer of the target level, where it will be stored. The creation of such an order represents the start of one of four connection points between the control system of shuttles and that of stacker cranes. The end of connection is represented in the logic of the stacker crane by the examination of the availability status of the stacker crane. In case a stacker crane is available the storage order is executed and the pallet is delivered to the transfer buffer of the target level. At this point, the stacker crane creates a storage order for the shuttles on that level. This constitutes the start of another connection between shuttles and stacker cranes.

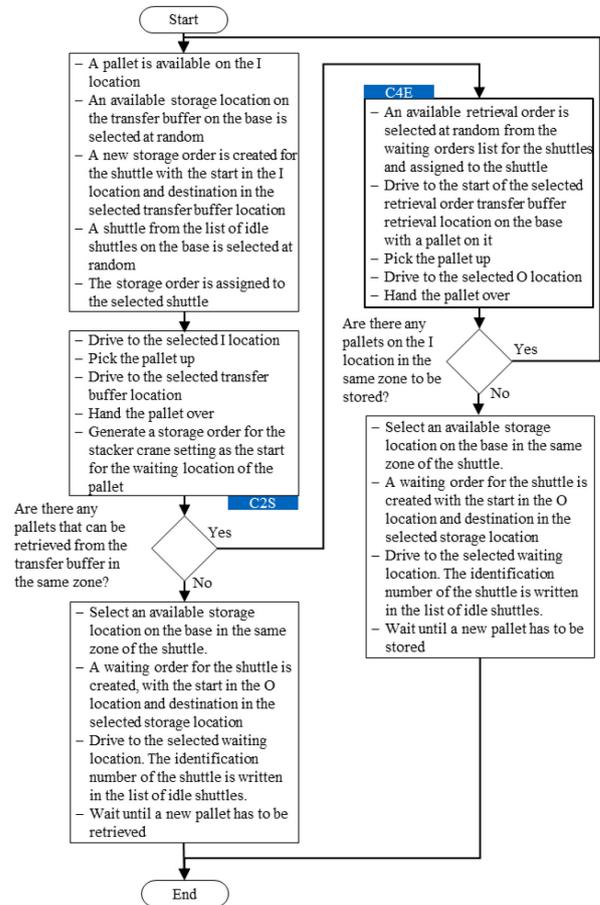


Figure 3: Control Logic – Layout 2, Double Cycles, Shuttles on Base

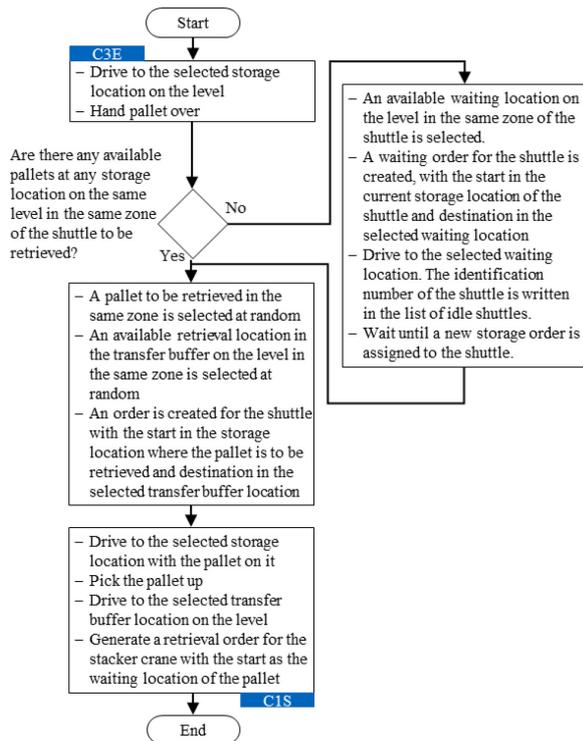


Figure 4: Control Logic – Layout 2, Double Cycles, Shuttles on Level

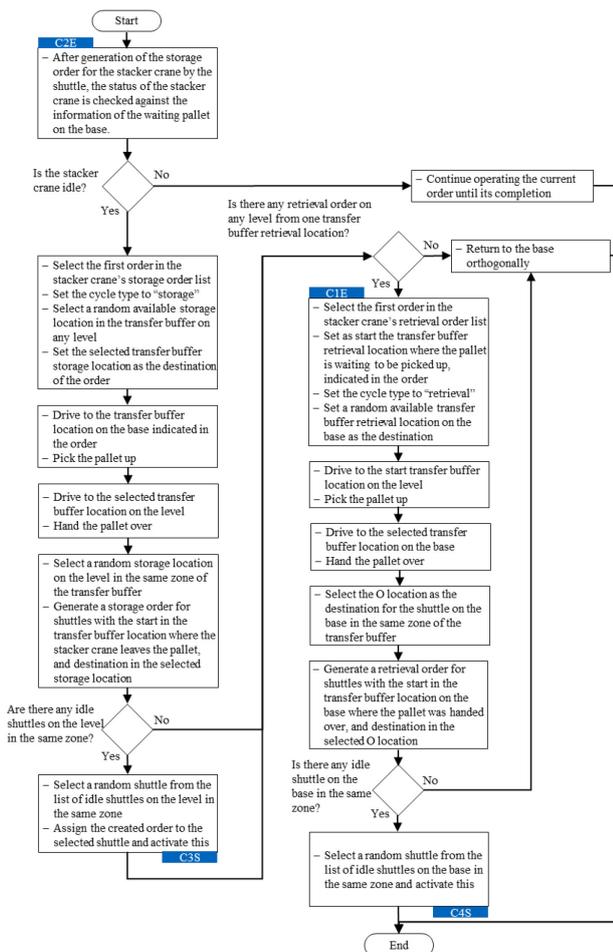


Figure 5: Control Logic – Layout 2, Double Cycles, Stacker Cranes

The end of the connection is in the logic of shuttles on level and is represented by the shuttle starting its route to pick the pallet on the transfer buffer and bring it to its final storage place. In the meantime, a shuttle on a level executes a retrieval order by moving a target pallet from its storage location to the transfer buffer. The shuttle generates then a retrieval order for the stacker crane. This generation is another connection between shuttles and stacker cranes. The end of connection is constituted by the stacker crane examining if there are retrieval orders to be executed. Next, the stacker crane performs the retrieval order by transporting the pallet to an available location of the transfer buffer on base. At this moment, another connection between stacker cranes and shuttles starts when the stacker crane generates a retrieval order for the shuttles on base. The end of connection is represented by the shuttle on base starting its route to execute the retrieval order. The shuttle transports the pallet from the transfer buffer to the O location and the double cycle is completed.

Layout 3

We now examine layout 3. The control algorithms we generated for the double cycles process in layout 3 is explained in Fig. 6 for the shuttles on the base, in Fig. 7 for the shuttles on the levels, and in Fig. 8. for the stacker cranes. The challenge, as opposed to layout 2, lies in the generation and correct assignment of transportation orders for the stacker crane to move shuttles between the different levels and the base, and of motion orders for the

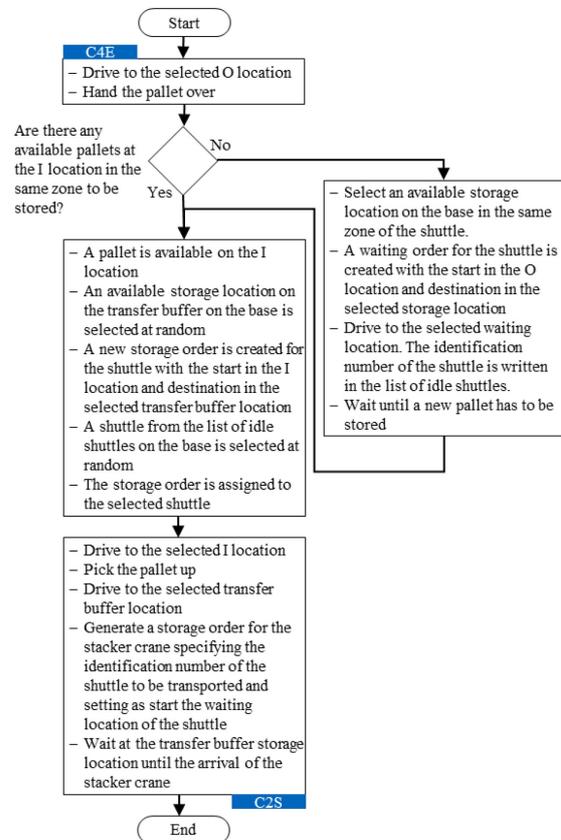


Figure 6: Control Logic – Layout 3, Double Cycles, Shuttles on Base

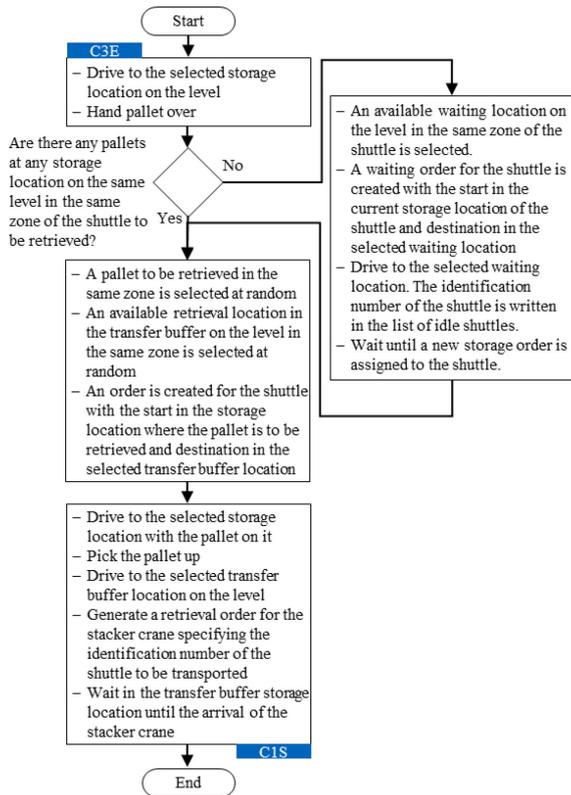


Figure 7: Control Logic – Layout 3, Double Cycles, Shuttles on Level

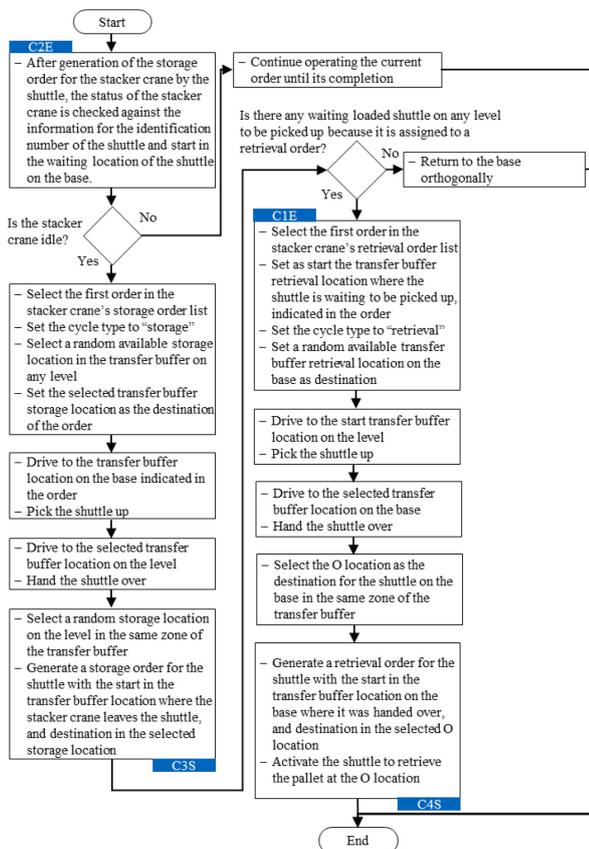


Figure 8: Control Logic – Layout 3, Double Cycles, Stacker Cranes

shuttles themselves. In fact, transportation orders should be created for not only loaded but also empty shuttles that have to be brought back from base to levels; motion orders should also be generated when an empty shuttle has to move, even if no pallet is to be picked or delivered. If the stacker cranes prove to be the bottleneck in the system, the shuttles without orders can wait directly on the transfer buffer. This saves travel time, when the stacker cranes are finally ready to exchange pallets, and energy, compared to having to drive to a waiting position in the storage locations as in layout 2.

SIMULATION STUDY

We implemented the model for layout 2 and layout 3 using the discrete event simulation environment Plant Simulation. To avoid deadlocks, the route of the shuttles on the different levels is based on the reservation of time windows, exactly like the shuttles on the base tier for layout 1 (Siciliano et al., 2020). This concept was initially introduced by (Kim and Tanchoco, 1991) and then further developed for shuttle fleets by (Lienert and Fottner, 2017a). An extensive description of the routing algorithm used for the shuttles on the levels can be found in (Lienert and Fottner, 2017b) (Lienert et al. 2020).

Parameters

The system we consider for layout 2 and layout 3 has two stacker cranes in a single aisle. We define a section as the area of the base or of a level comprised of two cross aisles. The different lengths of the aisle under consideration are two (38 m), three (54 m), four (68 m), five (83m) or ten (159 m) sections. There are three shuttle tier levels above the base. Both sides of the base have I/O area for pallets entering and leaving the warehouse. Each I/O area has two I/O locations, as shown in Fig. 2. The arrangement of cross aisles and storage aisles on the base is the same as in layout 1, see (Siciliano and Fottner, 2021), except for the I/O area. In fact, we discovered through experiments that the I/O area proposed in (Siciliano et al., 2020) creates an asymmetry in the dynamics of the shuttles for the right side of the warehouse compared to the left side for layouts 2 and 3. We therefore modified this as shown in Fig. 2 to guarantee symmetry, in other words the same performance for the right and left side of the warehouse, which resulted in an increased throughput.

The parameters used for the stacker crane in Tab. 1 and for the shuttles in Tab. 2 are provided by a manufacturer. Each experiment lasts 24 hours. We verified the model by comparing the analytically obtained travel time of individual vehicles with the simulated values (Siciliano et al., 2020). We then validated the travel time of the stacker crane and shuttles by comparing them with the values measured on the real subsystems, calculating the test positions of shuttles by the method in (Siciliano et al., 2021).

In the evaluation, we compare the throughput of layout 2 and layout 3 with following systems:

- Layout 1 with three channel storage levels above the base.

- Stacker crane-based warehouses whose throughput values are provided by a manufacturer.
- Shuttle-based warehouses, which we simulated in Plant Simulation. To make this comparable with DHPWs, we used the same shuttle tiers as for layout 2 and layout 3. The system has a total of four lifts i.e. two for each side of the warehouse, these being located at one third and two thirds of the length of the aisle.

The abbreviations used for the different systems examined from Fig. 9 to Fig. 14 are explained in the list of abbreviations at the end of this article.

Table 1: Stacker Crane Parameters

Parameter	Value
Speed (loaded)	0.6 m/s
Speed (empty)	1.0 m/s
Acceleration (loaded)	0.3 m/s^2
Acceleration (empty)	0.6 m/s^2
Turning time	6.6 s
Handover time	10.0 s

Table 2: Shuttle Parameters

Parameter	Value
Travel speed x	4.0 m/s
Travel acceleration x	0.5 m/s^2
Lifting speed y	1.0 m/s
Lifting acceleration y	1.0 m/s^2
Time of pallet handover	6.0 s
Time for positioning before channel	1.0 s

Evaluation

We first studied the throughput of layout 2 by varying the length of the aisle. For both of the processes of retrieval (Fig. 9) and of double cycles (Fig. 10), reducing the length of the aisle reduces the travel distance for the shuttles, resulting in an increase in throughput. However, this increase is not particularly high, so we can conclude that the length of the warehouse does not have a great influence on the throughput. It is significant for the scalability of the system, that up to a total of 64 shuttles, the shuttles remain the bottleneck in terms of the performance of the system for retrieval and double cycles. This means that simply increasing the number of shuttles would result in a further increase in throughput.

We now consider the behaviour of layout 3 as the length of the aisle changes. For the process of retrieval in Fig. 11, as for that of double cycles in Fig. 12, the length of the aisle has less influence on the throughput than in layout 2. Moreover, layout 3 allows a higher throughput than in

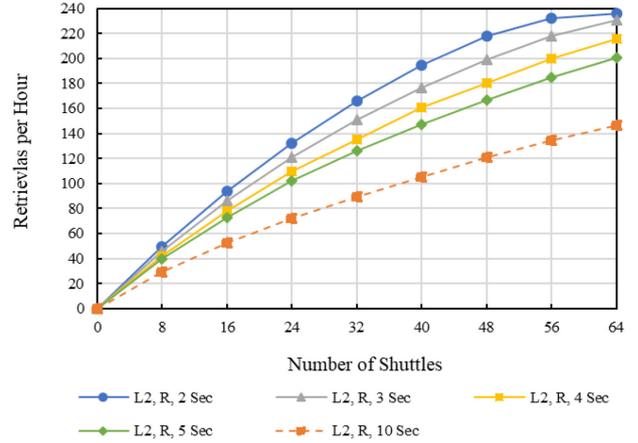


Figure 9: Retrieval Performance of Layout 2 Varying the Length of the Aisle from 2 Sections to 10 Sections

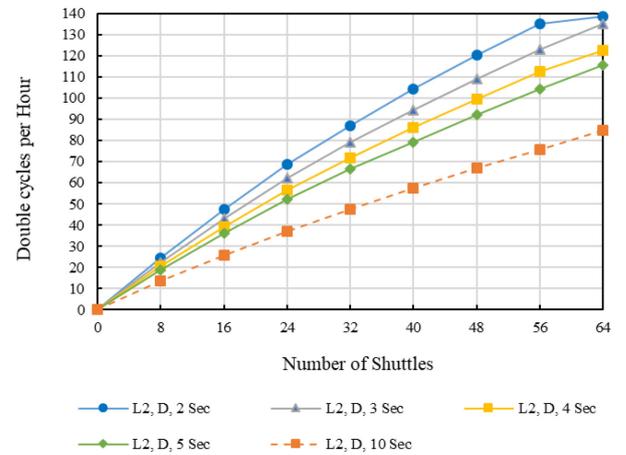


Figure 10: Double Cycles Performance of Layout 2 Varying the Length of the Aisle from 2 Sections to 10 Sections

layout 2 with a smaller number of shuttles. However, in the case of retrieval, layout 3 is limited by the bottleneck due to the stacker cranes, indicated by the plateaux in the curves, with a smaller number of shuttles than layout 2. Once the bottleneck of the stacker cranes is reached in layout 3, additional shuttles do not increase the throughput. Therefore, layout 3 is less scalable than layout 2. In Fig. 11 and Fig. 12 the results for 48 or more shuttles by two sections of layout 3 are not reported, because we do not recommend to use such a high number of shuttles in this case. The reason is that, when shuttles are able to change their levels, 48 or more shuttles are too many for the short layout of two sections and this causes congestions of shuttles near the transfer buffer of the base. As a consequence, throughput is reduced. This is a further demonstration of the lower scalability of layout 3 compared to layout 2. Not only do we compare layout 2 and layout 3 with each other, but also with other warehouses, as in Fig. 13 for retrieval and in Fig. 14 for double cycles. Layout 3 with four shuttles per level has a

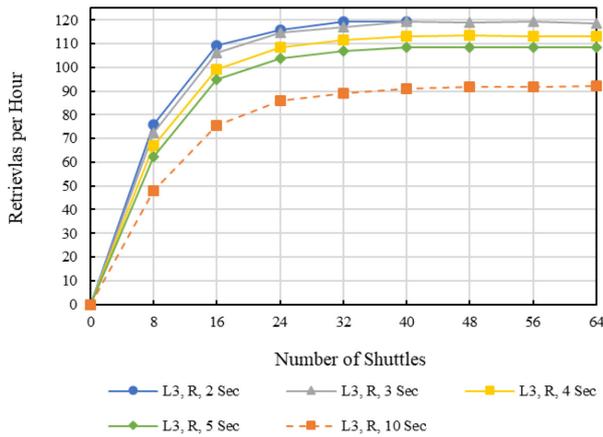


Figure 11: Retrieval Performance of Layout 3 Varying the Length of the Aisle from 2 Sections to 10 Sections

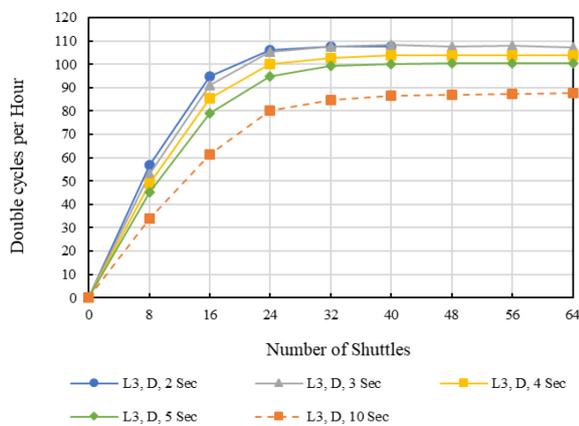


Figure 12: Double Cycles Performance of Layout 3 Varying the Length of the Aisle from 2 Sections to 10 Sections

throughput that is already higher than those of the other systems, in the case of both retrieval and double cycles.

By comparison, layout 2 needs up to 6 shuttles per level to provide a throughput that is not only higher than that of conventional stacker cranes but also of that of the shuttle-based warehouse with four lifts, which has comparable costs. Layout 1 achieves a higher throughput than that of conventional stacker crane-based warehouses, but one that is inferior to that of layout 2 and layout 3. The reason using stacker cranes, as in layouts 2 and 3, improves performance compared to using lifts is that the interface between lifts and shuttle tiers is made up of a reduced number of locations on the transfer buffer. Therefore, shuttles wait longer for a location or a pallet to become available than in the case of stacker cranes, which have locations on the transfer buffer all along the aisle.

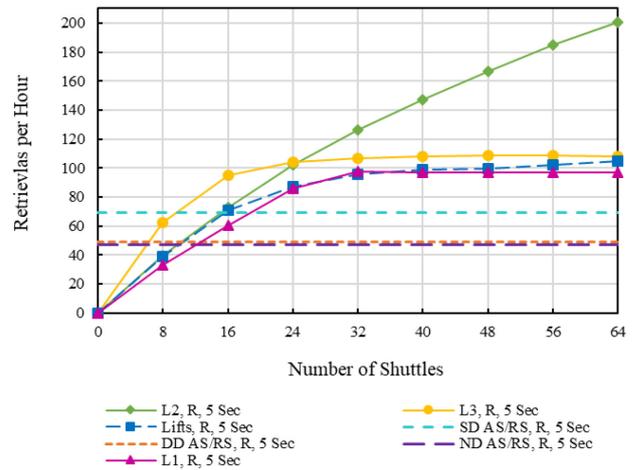


Figure 13: Comparison of Retrieval Performance between Different Warehouse Systems

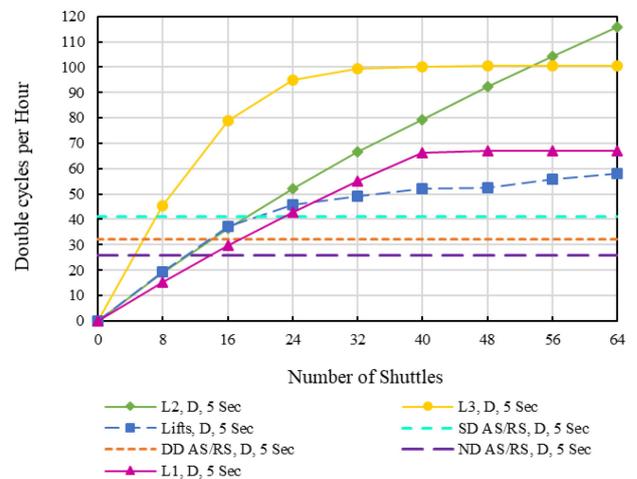


Figure 14: Comparison of Double Cycles Performance between Different Warehouse Systems

CONCLUSION AND OUTLOOK

In this article, we described the design and advantages of two warehouses, classifiable as DHPWs, which we call layout 2 and layout 3 respectively. We then proposed control strategies for each of these. Through a discrete event simulation study, we demonstrated that the length of the stacker cranes' aisle has no great influence on the throughput for either of them. With a small number of shuttles, layout 3 should be given preference over layout 2 because its poorer scalability is not yet dominant: moving shuttles to the levels where they are needed more urgently then overcompensates the additional orders for the stacker cranes. When using many shuttles, these additional orders lead to an earlier bottleneck, so that layout 2 is then preferable. All in all, the results of this paper are decision support in warehouse management for pallets insofar as they illustrate the performance benefits of substituting layout 1 to multi-depth stacker crane-based warehouses applications and of replacing the connection

to lifts in shuttle-based warehouses with a connection between shuttles and stacker cranes such as layouts 2 and 3. For future research, different coordination algorithms between shuttles and stacker cranes have to be investigated to further improve the throughput without having to increase the number of shuttles.

ACKNOWLEDGEMENTS

We are very grateful to Thomas Klopfenstein from the firm of Gebhardt Fördertechnik GmbH for providing us with the throughput values for conventional stacker crane-based warehouses.

Moreover, we would like to thank Joerg Eder, also from the firm of Gebhardt Fördertechnik GmbH, for the fruitful collaboration.

REFERENCES

- Eder, J., Klopfenstein, T., & Gebhardt, M., 2019. *Patent: Lagersystem zur Speicherung und Abgabe von Ladungsträgern*. DE102019211804, German Patent and Trade Mark Office (DPMA).
- Kim, C. W., & Tanchoco, J. M. A., 1991. Conflict-free shortest-time bi-directional AGV routing. *International Journal of Production Research* 29 (12): 2377-2391.
- Lienert, T., & Fottner, J., 2017. No more deadlocks—applying the time window routing method to shuttle systems. *Proceedings of the 31st European Conference on Modelling and Simulation (ECMS)*, 169–175.
- Lienert, T., & Fottner, J., 2017. Development of a generic simulation method for the time window routing of automated guided vehicles. *Logistics Journal: Proceedings*, Vol. 2017.
- Lienert, T., Wenzler, F., & Fottner, J., 2020. Simulation-based evaluation of reservation mechanisms for the time window routing method. *Proceedings of the 33rd European Conference on Modelling and Simulation (ECMS)*.
- Malik, O., 2014. *Patent Application Publication: Automated warehousing systems and method*. US20140086714A1, US Patent and Trademark Office (USPTO).
- Siciliano, G., Lienert, T., Fottner, J., 2020. Design, Simulation and Performance of a Highly-Dynamic, Hybrid Pallet Storage and Retrieval System. *Proceedings of the 19th International Conference on Modeling and Applied Simulation (MAS)*.
- Siciliano, G. & Fottner, J., 2021. Concept development and evaluation of order assignment strategies in a highly dynamic, hybrid pallet storage and retrieval system. *Proceedings of the 11th International Conference on Simulation and Modeling Methodologies (SIMULTECH)*, ISBN 978-989-758-528-9, ISSN 2184-2841, pp. 360-368.
- Siciliano, G., Durek-Linn, A., Fottner, J., 2022. Development and Evaluation of Configurations and Control Strategies to Coordinate Several Stacker Cranes on a Single Aisle for a New Dynamic Hybrid Pallet Warehouse. In: Shi X., Bohács G., Ma Y., Gong D., Shang X. (eds) *LISS 2021. Lecture Notes in Operations Research*. Springer, Singapore. https://doi.org/10.1007/978-981-16-8656-6_54
- Siciliano, G., Schuster, C. U., Fottner, J., 2021. Analytical method to determine the test positions for validation of a two-dimensional shuttle system model. *Proceedings of the 20th International Conference on Modeling & Applied Simulation (MAS)* 2021), pp. 21-28. <https://doi.org/10.46354/i3m.2021.mas.003>
- Wang, Y., Man, R., Zhao, X., Liu, H., 2020. Modeling of parallel movement for deep-lane unit load autonomous shuttle and stacker crane warehousing systems. *Processes* 8(1)

LIST OF ABBREVIATIONS

- L1 = Layout 1; L2 = Layout 2; L3 = Layout 3
- R = Retrieval process; D = Double cycles process
- Sec = Sections
- SD AS/RS = Single-deep storage stacker crane with telescopic forks
- DD AS/RS = Double-deep storage stacker crane with telescopic forks with relocations
- ND AS/RS = Nine-deep storage stacker crane with satellite without relocations

AUTHOR BIOGRAPHIES

GIULIA SICILIANO is a research associate at the Chair of Materials Handling, Material Flow and Logistics in the School of Engineering and Design at Technische Universität München. She holds a M.Sc. in Mechanical Engineering from Università degli Studi di Roma “Tor Vergata”. Her research interests lie in the design, control and simulation of automated warehouses, and in the development of artificial intelligence models for optimization of logistical processes.

YUE YU is a former master student of the Chair of Materials Handling, Material Flow and Logistics in the School of Engineering and Design at Technische Universität München and now process simulation engineer at Dräxlmaier Group. She holds a M.Sc. in Development, Production and Management. Her research interests include the design and simulation of automated warehouses.

JOHANNES FOTTNER is Professor and Head at the Chair of Materials Handling, Material Flow, Logistics in the School of Engineering and Design at Technische Universität München.

CALIBRATION MODEL FOR PERCEPTUAL COMPENSATION OF DEFECTIVE PIXELS OF SELF-EMITTING DISPLAY

Olga A. Basova

Moscow Institute of Physics and Technology
Institutskiy per. 9, Dolgoprudny 141701, Russia
Institute for Information Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia;
E-mail: basova.oa@phystech.edu

Anton S. Grigoryev

Institute for Information Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: me@ansgri.com

Dmitry P. Nikolaev

Institute for Information Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia;
LLC Smart Engines Service
Prospect 60-Letiya Oktyabrya 9, Moscow 117312, Russia
E-mail: dimonstr@iitp.ru

KEYWORDS

Displays, non-uniformity compensation, defective pixel compensation, display calibration, image enhancement, spatial filtering, spatial resolution, human visual system model, S-CIELAB.

ABSTRACT

In this paper, we study compensation of defective subpixels with insufficient maximum brightness. The aim of the compensation is to minimize the perceived image non-uniformity. Compensation of the displayed image non-uniformity is based on minimizing the perceived distance between the target (ideally displayed) and the simulated image displayed by the calibrated screen. In this work, we compare the efficiency of compensation depending on color coordinates we calculate color difference in. We investigated the behavior of compensations based on two different uniform color coordinates: CIELAB and Oklab. We examine the efficiency of the compensation on natural scenery images. It was found that Oklab shows better performance than CIELAB in terms of uniformity of perceived compensated image. However, taking into account the spatial properties of the human visual system using S-CIELAB preprocessing almost eliminates the difference between the color coordinates.

INTRODUCTION

Modern displays incorporate a large number of individual elements forming an image. Imperfect manufacturing techniques yield variation in the characteristics of the elements composing a display, which results in different luminance of these elements when receiving the same input signal. Furthermore, display elements age at different rates, which also leads to the variation of the characteristics among them (Arnold and Cok 2006; Harris 2007). If the input signal does not take into account these differences, then the uniform areas of the input image appear non-uniform on a screen (display non-uniformity problem), which significantly reduces the overall quality of the displayed image.

The problem of compensation of such distortions can be formulated for different types of displays: self-emitting (self-luminous elements with individually determined luminance, e.g. OLED or LED-array displays), transmissive (with optical filters as elements passing the light from the uniform source, e.g. liquid-crystal displays (LCD)), reflective (where the reflection coefficient of the outer light source is controlled, e.g. electronic paper and reflective LCDs), and transfective (which can function as transmissive or reflective, including a backlight dependent on ambient light). Each of these display types is characterized by a different type of distortion (Harris 2007; Uttwani et al. 2012).

In this work, we consider self-emitting displays. The latter do not suffer from non-uniform backlight luminance possible for LCDs, thus the main problem of non-uniformity in self-emitting displays is associated with the variation in the emission characteristics of individual pixels (see Figure 1).

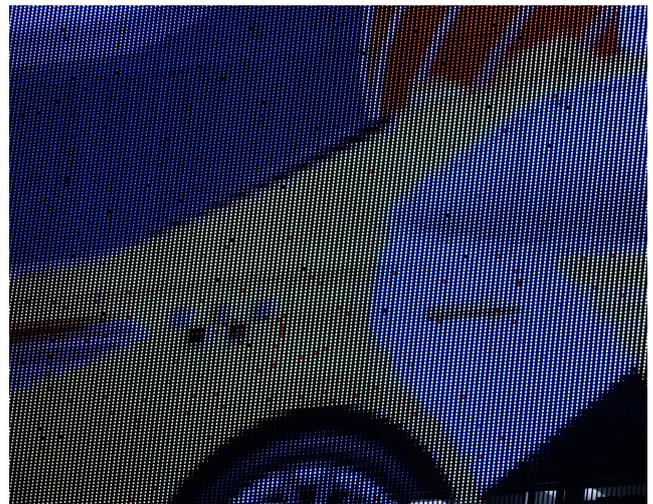


Figure 1: Defective Pixels on Large Format LED-array Display Can Be Easily Spotted and Significantly Reduce the Quality of the Displayed Image

The ideal display is a uniform display, i.e. a display with the same characteristics shared by all the

pixels. Equalization of properties of individual display elements involves a search for the compensating transformation of the input signals to a display so that the formed image and the image formed by an ideal display would be as close as possible.

The transformation should allow for efficient implementation, possibly in hardware, to avoid increasing the display latency.

In (Basova et al. 2020), we have suggested the display luminance non-uniformity model, which describes the pixel-wise variation as the element-wise multiplication by a 3D vector in the input signals to display matrix space. To calibrate a display within this model, we suggested the compensating transformation defined as the multiplication between a matrix 3×3 and the input signal to a pixel.

THE MODEL OF DISPLAY NON-UNIFORMITY

Let us consider a display of the following design: the size is H by W pixels, each pixel is composed of three self-luminous elements (subpixels) of different types: red, green and blue. The luminance of a subpixel, depending on the input signal, varies from zero to a certain maximum value determined by subpixel properties. The valid input signal values are limited to the range $[0, 1]$; all the values outside this interval are clipped to the interval edges (either 1 or 0). We denote the clipping operator to the interval $[0, 1]$ by *clip*.

Let us denote the spectral properties of the three subpixels of i -th pixel at maximum luminance as $e_1^i(\lambda), e_2^i(\lambda), e_3^i(\lambda)$ correspondingly, where $e_k^i(\lambda)$ is the spectral intensity function of wavelength λ . We call the latter the primaries. They form the color coordinate system defined by individual pixel characteristics, which we refer to as pixel coordinate system. The input signal to a subpixel S_k^i (where i is the index of the pixel and $k \in 1, 2, 3$ is the index of the subpixel) determines the ratio between subpixels' luminance and the maximum luminance. It can be defined as a point in pixel coordinate system. The i -th pixel emission spectrum is defined as

$$P^i(\lambda) = \sum_{k=1}^3 \text{clip}(S_k^i) e_k^i(\lambda).$$

In an ideal display, the three primaries of all pixels are identical. Let denote the ideal display primaries as $e_1(\lambda), e_2(\lambda), e_3(\lambda)$.

For the purposes of this work, we assume that all display subpixels can be described by one model of subpixel defects. In this paper, we consider probability model, which describe pixel defects. The model is based on the assumption that all display subpixels are independent. We assume that different subpixels of the same type differ only in peak luminance. Each subpixel can be either ideal or defective with probability p . If the subpixel is defective, its primary can be defined as

$$\tilde{e}_k^i(\lambda) = e_k(\lambda) \cdot d_k^i,$$

where d_k^i is the defectiveness coefficient sampled from a uniform distribution on the interval $[0,1]$ determining the factor by which subpixel's maximum luminance is lower than that of an ideal subpixel.

Therefore, the pixel emission spectrum can be written as

$$P^i(\lambda) = \sum_{k=1}^3 \text{clip}(S_k^i) d_k^i e_k(\lambda).$$

Thus, in the described model, the primaries of all subpixels of the same type are proportional (with the only exception for subpixels with maximum luminance of zero and non-defined primaries).

THE GENERAL APPROACH TO CALIBRATION

To compensate the display non-uniformity, we will search for such a signal transformation of the input signal to the display matrix, so that the image formed by a non-uniform display would be as close as possible to the image formed by an ideal (uniform) display.

There are two different approaches to compensation: compensation of image with defective pixels and display calibration based on gamut optimization.

For the first time, the idea of defective pixel compensation was introduced in the works (Kimpe et al. 2004, 2006). Authors propose an algorithm of an LCD display calibration. The algorithm uses pixels in the neighborhood of the defective pixel to provide overall image improvement. The masking neighboring pixels' input signal is modified in such a way as to increase the perceived uniformity of the area with the defective pixel. Perception of the displayed image is estimated using point spread function (PSF) which models the human visual system (HVS). Compensation is achieved by changing the input signal of the masking pixels. Later, in 2012 and 2015, the PSF-based compensation algorithm was patented (Kimpe 2012; Verstraete and Kimpe 2015). After a while, in the paper (Messing and Kerofsky 2006) a similar method for defective pixels was proposed. It was based on the contrast sensitivity function (CSF) of the HVS. In (Stellbrink 2007), an image compensation algorithm based on another HVS model was proposed which takes into account the masking effects of the visual perception.

It looks like the methods from aforementioned papers (Kimpe et al. 2004, 2006; Messing and Kerofsky 2006; Stellbrink 2007) are designed for image processing and not for estimation of a display compensation parameters. In other words, to estimate the parameters of defective pixel compensation one needs to apply the algorithm for every image. In contrast, we will estimate the calibration parameters of the display only once. Another difference from our problem formulation is that their distortion model does not allow to control the brightness of defective pixels, and do not seem to consider the case of clustered defects, working only with isolated defects.

In the paper (McFadden and Ward 2015), McFadden and Ward suggest a PSF-based algorithm for compensation of a grid distortion caused by gaps between

individual tiles of a display. The proposed algorithm reduces the apparent visibility of seams between individual tiles.

The idea of neighbor-based compensation of the defective pixels continues to be relevant to this day: for instance, in 2019 the patent on this topic was published (Jepsen et al. 2019). It describes the compensation of completely defective pixels on an LCD display. The brightness of surrounding pixels is increased to compensate the lost brightness output from non-functioning pixels. The patent describes three types of a spatial distribution function for apportioning additional brightness to surrounding pixels: (i) additional brightness is divided equally, (ii) allocating larger portions to closer surrounding pixel, (iii) diagonal pixels are received a larger amount of the additional brightness, since the human eye is more sensitive to horizontal or vertical lines. Also, the total brightness error associated with a given non-functioning pixel must be recomputed for each image frame.

All the algorithms mentioned above are based on various models of the HVS and work as an additional stage in image processing pipeline to adapt the input image to characteristics of the display. Another approach to correct for the defective pixels is by calibrating the display, and solutions implementing it are compatible with in-display implementation due to their independence of specific input signal.

One well-studied problem that is similar to calibration of displays with defective pixels is tiled display calibration. In such systems, each display tile has own white point chromaticity and maximum brightness. For example, the methods described in papers (Stone 2001; Bern and Eppstein 2003) allow for the complete restoration of the non-uniform display to a uniform state by the elimination of extremely bright and/or saturated colors, which could not be displayed by some separate parts of a tiled display. In other words, the algorithms reduce the gamut of the most bright and/or saturated sub-displays. The latter algorithm could be applied for calibration of self-emitting displays.

Another group of scientists have developed a system for calibration of OLED displays. They have published several patents describing a method for color and ageing compensation of an emission display (Chaji et al. 2017; Chaji 2019; Nathan et al. 2020). The algorithm described in the most recent patent (Nathan et al. 2020) compensates degraded pixels by supplying their respective driving circuits with greater voltages. The display data is scaled by a compression factor of less than one to reserve some voltage levels for compensating degraded pixels. In other words, the algorithms reduce the gamut of non-defective pixels.

Some display types, for instance, micro-LED, are difficult to produce with high uniformity, therefore the calibration of such displays is often implemented in software (Mao et al. 2017; Kim et al. 2020) rather than in hardware. The paper (Mao et al. 2017) proposes a calibration of an LED display (shown images have block artifacts typical for micro-LED-like displays, although

this is not specified in the article). The paper presents a calibration method based on the brightness correction coefficient map. The outputs of the algorithm are calibration coefficients for each channel, but their calculation does not take into account the HVS properties.

One of the most up-to-date works on self-emitting display calibration, particularly for micro-LED displays, is a paper from Samsung Research (Kim et al. 2020). It proposes two algorithms for calibration of the uniformity of micro-LED display. The first algorithm calculates a set of input to an output look-up tables (LUT). The second algorithm is based on 4D transform correction. It estimates the 4x4 matrix which works as an input to output converter at the on-device step. The authors note that the advantage of the 4D transform-based method compared to the older methods is that it can be fused with any other calibration algorithm which can be converted to a LUT.

All works mentioned above can be divided into two categories: (1) papers that describe a calibration of a display, but do not take into account the HVS properties, (2) methods that consider the HVS properties, but do not calculate a single display calibration, only enhancing individual displayed images. In our work, we try to combine these two approaches: we calculate a single set of compensation parameters for any images displayed on the screen, taking into account the HVS properties.

This work develops the approach proposed in (Basova et al. 2020). The suggested approach is based on the following. If the colors of the pixels within a neighborhood of a defective pixel are slightly corrected in a way that ensures that the average color of this area is closer to the desired one, then due to the small size of this area the HVS would perceive this as a whole: the same way the area composed of non-defective pixels with original (uncorrected) input signal would have been perceived. Thus, the proposed approach optimizes the compensation parameters based on the HVS model response, instead of the minimization of the emission difference.

The general flow of the conversion of the input signal is shown in Figure 2.

An image to be displayed is usually represented in standard RGB (sRGB) or other standardized color coordinates defined in relation to the CIE XYZ color space of the standard observer, which simulates a primary response of the HVS model.

Then compensating transformations are applied to the image. This results in a corrected subpixel signal which takes into account the non-uniformity of the subpixels.

The displayed image is perceived by the HVS. The perception is modeled by the CSF of the latter (Wuerger et al. 2002). The correspondence between the perceived image and the image displayed by an uniform display is the goal of the calibration.

Since an ideal display cannot be designed, and during the compensation parameters optimization it is not feasible to repeatedly show pairs of images to a person for similarity assessment, we need to develop a compu-

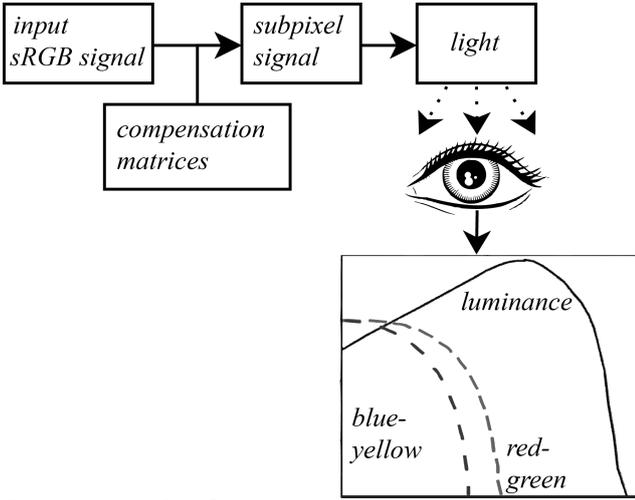


Figure 2: The Conversion of the Input Signal to a Display into a Perceived Image

tational model for such assessment, which should approximate the HVS perception of an image.

Thus, the compensation method development requires to choose the similarity assessment approach between two images: the ideal one and formed by a real display. To evaluate the similarity, we consider the averaged Euclidean distance in uniform color coordinates (CIELAB (Robertson et al. 1977) and Oklab (Ottosson 2020)) and two color image metrics: S-CIELAB (Zhang and Wandell 1996) and a hybrid model of Oklab and S-CIELAB.

SIMILARITY ASSESSMENT BETWEEN THE TWO IMAGES

CIE XYZ (hereafter XYZ) is a color space where spectral characteristics are determined by the spectral sensitivity functions of some average human eye, and its X, Y, and Z values are referred to as standard colorimetric color coordinates (Smith and Guild 1931). A simple Euclidean metric in this coordinate system does not approximate the perceived by a person difference between the colors. To eliminate this drawback, the coordinate system CIELAB (hereafter LAB) was derived from the XYZ color space (Robertson et al. 1977).

However, LAB coordinate system is not totally perceptually uniform. In particular, it is known that LAB and even advanced color difference formula CIEDE2000 (Luo et al. 2001) have drawbacks in large and very large color difference estimation (Abasi et al. 2020). This fact causes some problems in our task, because the large color differences are the most challenging in our task: the impact of defective pixels is the greatest when the pixels are significantly different from the surrounding ones (the background). In contrast, small color differences are not noticeable at all in some cases, and in other cases, it is easy to compensate it using the neighbour pixels.

Recently, the new uniform color coordinates system Oklab was proposed (Ottosson 2020). Based on it, coordinate systems for color picking were build (however, they have a non-Euclidean structure, so they are be-

yond the scope of this paper). That says that Oklab is potentially good for large color differences estimation.

We have compared the behaviour of LAB (ΔE_{ab}^*), CIEDE2000 (ΔE_{00}^*) and Oklab on three test sets of colors with large color difference. The results are shown in the Table 1. Each set contains three colors divided into two pairs (one color is the same between two pairs). The colors for last two sets was taken from (Abasi et al. 2020). It can be seen that the first pair of colors (a_3, a_2) is perceived closer to each other than the (a_1, a_2) pair, although LAB and CIEDE2000 shows the opposite situation. Additionally, it can be seen that the calculated color differences by LAB and CIEDE2000 for (b_1, b_2) is less than for (b_1, b_3). However, the difference between calculated color differences is small, while the perceived color differences between pair (b_1, b_3) is more significant than between (b_1, b_2). The same situation is for color pairs (c_3, c_2) and (c_1, c_2). In all described above cases, Oklab shows the better performance than LAB and CIEDE2000, so the usage of Oklab in defective display calibration task is promising.

Table 1: Comparison of the Color Difference Metrics Behaviour on Large Color Differences Examples

Color representation	LAB coordinates	ΔE_{ab}^*	ΔE_{00}^*	Oklab
	$a_1 = (44, -6.6, 28.3)$ $a_2 = (53.19, 0, 0)$	30.5	20.1	0.108
	$a_3 = (51.8, -10.27, 38)$ $a_2 = (53.19, 0, 0)$	39.4	21.2	0.095
	$b_1 = (60, -15, 6.5)$ $b_2 = (60, 11.5, -22.5)$	39.3	28.3	0.102
	$b_1 = (60, -15, 6.5)$ $b_3 = (71, 11.5, -22.5)$	40.8	29.7	0.146
	$c_1 = (60, -2.25, 10.5)$ $c_2 = (60, 21.5, -10.5)$	31.7	29.0	0.086
	$c_3 = (72, -2.25, 10.5)$ $c_2 = (60, 21.5, -10.5)$	33.9	30.6	0.128

Another attribute of the human visual system which is important for display calibration is that the human eye is less sensitive to chromatic and brightness differences in small details compared to larger ones. In other words, when the contrast of an initially resolvable image gradually decreases, the perception of uniformity is achieved before reaching the true zero contrast of the stimulus. Therefore, we are interested in the dependence of the minimum contrast which is still resolved (contrast sensitivity) on the spatial frequency of the stimulus. This dependence is described by contrast sensitivity functions (CSFs) (Bozhkova et al. 2019). CSFs

also model another spatial property of the visual system (however, it does not affect the compensation described in this paper): the lower sensitivity to low-frequency achromatic changes compared to higher-frequency ones. The shape of the CSF depends on the direction of the stimulus color contrast vector. Most studies consider the CSF along three directions which are assumed to be independent: the luminance axis, red-green, and blue-yellow directions. It is presumed that by knowing the sensitivity along these color directions, it is possible to predict the contrast sensitivity for any color pair.

To more realistically assess the difference between images defined in LAB color coordinates, the S-CIELAB metric was proposed in (Zhang and Wandell 1996). This metric takes into account the spatial properties of the human visual system (simulated via CSFs). S-CIELAB allows to approximate a perceived image difference for a given viewing distance (i.e. the distance between the observer’s eyes and the display) and resolution (dpi) of images.

S-CIELAB computation includes three steps:

1. spatial filtering of the images;
2. pixel-wise color difference metric computation for the compared images;
3. averaging of the difference map from the previous step into a single metric value.

The first step employs the CSF to simulate the dependence of human visual system sensitivity on the spatial frequency and chromaticity of the stimulus. The CSF is approximated using a weighted sum of Gaussian filters (each with its own weight w and blur parameter σ , the latter being scaled according to the viewing distance and image resolution). To apply the CSF more accurately, instead of the nonlinear LAB coordinates, the image is processed in linear opponent color coordinates along three basis directions: luminance, red-green, and blue-yellow. As the result of this spatial filtering, we obtain an image that contains only details visible at a given distance and resolution, with intensities modified according to spatial response of the HVS.

In the second step, for two images formed as described above and transformed back from opponent colors to LAB coordinates, the Euclidean metric (ΔE_{ab}^*) is computed pixel-wise to create a difference map, which is then averaged into a single value — the output of the S-CIELAB image difference metric.

The difference map calculation can be done in different color coordinates, such as those proposed in (Konovalenko et al. 2021; Ottosson 2020), or by different color difference metric, such as proposed in (Abasi et al. 2020). The spatial filtering method employed in S-CIELAB can be also utilized in different color coordinates, such as those proposed in (Konovalenko et al. 2021; Ottosson 2020). Furthermore, another model, N-CIELAB (Sai 2018), can be used to simulate the properties of the contrast sensitivity of the human visual system. This model was suggested for the compression optimization for JPEG and JPEG2000, as well as for the classification of images according to the level of the detail. Alternatively, the metric suggested in

(Frackiewicz and Palus 2017) for the assessment of the quantized color images can be used as the image quality metric.

In this work, we compare the calibrations obtained via the optimization of various color image metric: the mean Euclidean metric in the color coordinates LAB or Oklab, and image metric S-CIELAB or S-Oklab (spatial filtering from S-CIELAB and Oklab color metric calculation).

THE CONSIDERED CALIBRATION ALGORITHM

We described the algorithm for the calibration matrices calculation earlier in (Basova et al. 2020). In this article we describe the generalized version of the algorithm that estimate color image metric based on various models, i.e. the algorithm is in some sense parameterized by the color image metric.

The input of the algorithm is the primaries of the display pixels in the standard observer color space XYZ. These vectors can be determined, for example, by applying an input signal (1,0,0) in hardware RGB to all display pixels. Then for each pixel, the coordinates of its e_1 primary in XYZ can be obtained from the picture of the screen taken with the colorimetrically calibrated camera with sufficient resolution. Similarly, we can obtain coordinates of primaries e_2 and e_3 using (0, 1, 0) and (0, 0, 1) hardware RGB stimuli. The primaries of the display pixels thus represent the transition matrix B from individual pixel (hardware) space to the standard XYZ coordinates.

The output of the algorithm is an array of $H \times W$ (equal to the display shape in pixels) compensation matrices of size 3×3 . Let denote the array as C . Every single matrix of C is a calibration matrix for particular pixel of the display.

Then we will search for a such C , for which the following sum among all the pixels is minimized:

$$\sum_{S \in \{R, G, B, W\}} \nu_S \cdot D(J(S, C), I(S)) \rightarrow \min_C, \quad (1)$$

where I is the input image S rendered by the ideal (without defective pixels) display; $J(S, C)$ is the input image S corrected by the calibration matrices C and displayed on the calibrated screen with defective pixels; ν_S is the weight multiplier; and $D(., .)$ is the color difference between the two images of identical shape. The optimization is performed for uniform images S (all pixels of which are identical) of four colors: red (R), green (G), blue (B), and white (W). For greater increase in uniformity of achromatic areas, the white image was weighted with $\nu_S = 3$ and $\nu_S = 1$ for red, green, and blue images.

In this paper, we compare compensation algorithms where the color image metric D is calculated for various image representation: in color coordinates LAB or Oklab, or image representation by S-Oklab or S-CIELAB. We propose to explicitly consider in the minimization task (1) the parameter of color image metric. Let Ω

denote a function of image conversion into a specific color coordinates (LAB, OKLAB) or image representation (S-Oklab, S-CIELAB) depending on the variation of the calibration method under study. Then we propose to minimize the following functional:

$$\sum_{S \in \{R, G, B, W\}} \nu_j \cdot d(\Omega(J(S, C)), \Omega(I(S))) \rightarrow \min_C, \quad (2)$$

where $d(., .)$ is the average Euclidean distance between the two images of identical shape:

$$d(I_1, I_2) = \frac{1}{H} \frac{1}{W} \sum_{i=1}^H \sum_{j=1}^M \|I_1(i, j) - I_2(i, j)\|_2,$$

where i and j are pixel indexes and Ω has the following form:

$$\Omega(I) = \begin{cases} LAB(I), & \text{for LAB coordinates,} \\ OKLAB(I), & \text{for Oklab coordinates,} \\ SCIELAB(I, v), & \text{for S-CIELAB,} \\ SOKLAB(I, v), & \text{for S-Oklab,} \end{cases}$$

where $LAB(I)$ and $OKLAB(I)$ are transition functions from XYZ to LAB and Oklab color coordinates respectively, $SCIELAB(I, v)$ and $SOKLAB(I, v)$ are transition functions for viewing parameter v from XYZ to S-CIELAB or S-Oklab image representation respectively and have the following form:

$$SCIELAB(I, v) = LAB(M_O^{-1}[M_O I * f(v)]),$$

$$SOKLAB(I, v) = OKLAB(M_O^{-1}[M_O I * f(v)]),$$

where M_O is a transition matrix from XYZ to opponent color coordinates (Zhang and Wandell 1996), and $*$ is a channel-wise convolution with two-dimensional kernel f which for each channel has the following form:

$$f_j(v) = k_j \sum_i w_{i,j} e^{\frac{-(x^2+y^2)}{(\sigma_{i,j} v)^2}},$$

where j is an index of the opponent channel, values of $\sigma_{i,j}$ and $w_{i,j}$ are given in the article (Zhang and Wandell 1996); scale factor k_i is chosen so that f_i sums to 1; viewing parameter v is a multiplication of a display resolution (dpi) and a viewing distance in inches.

The corrected image displayed on the screen with defective elements is simulated as follows:

$$J(S, C) = B \cdot clip(C \cdot S),$$

where the clipping operator $clip$ ensures that the resulting pixel values are within the valid interval.

The input signal to the display $S(i, j)$ is specified in the linear color coordinates of an individual pixel (for each i, j S is a 3D column vector that corresponds to the input signal to this pixel).

The ideal display would render $S(i, j)$ as an image $I(S) = E \cdot S$, where $E = (e_1, e_2, e_3)$ are the non-defective pixel primaries.

To make compensation of brightness level possible, the maximum brightness of the gamut was reduced by 20%. The idea is that the impact of defective pixels is the greatest in the uniform image areas, where those pixels are significantly different from the surrounding ones, while on textured areas the defects are likely to be masked, so by performing optimization on uniform images only we improve the worst-case perceived uniformity.

In our current study, the optimization was implemented using the Adam algorithm (Kingma and Ba 2014). In further work the performance of other optimization methods, e.g. (Darkhovsky et al. 2014), should be explored for this problem.

EXPERIMENTS

The proposed algorithms were implemented in Python. The optimization was implemented using the Tensorflow library and was run on NVIDIA GPU (GeForce GTX 1080 Ti).

The following condition was set as the criterion of the optimization completion: if the average value of the error function over 300 iterations decreases by less than 10^{-4} . If the display characteristics do not meet the completion criterion, the maximum number of iterations was set to 90000. As the initial approximations of the compensation matrices, identity matrices were used.

In this work, to clearly present the simulation (so that even within the small area of an image or display different defective pixels could be observed), the pixel defect probability p was set to 0.005, which is greater than that officially specified by the manufacturer (LG 2021).

The calibrations obtained via the optimization of the four different color image metrics based on the following models were compared: LAB, Oklab, S-Oklab, and S-CIELAB for viewing distance 40 cm and 94 dpi. We studied the calibration results on real scenery images from the TID2013 dataset (Ponomarenko et al. 2013). In this work, we show several fragments of TID2013 images and its versions after the compensation. Figures 3, 4, and 5 show the images compensated by various calibration matrices.

The figure 3 shows that defective pixels are the most noticeable on uniform areas. For instance, defective pixels on a uniform background are clearly visible and cannot be fully compensated by the neighbours (see figures 3 (e) and 3 (f)). The figure 4 shows the perception of defective pixels on textured areas. For instance, we do not see any defective pixels on palm leaves, although there are some defective pixels, see the right palm in the figure 5 (there are a couple of purple pixels, which are compensated by green neighbours, see figures 5 (e) and 5 (f)). Defective pixels in the sky are visible in the images 4 (b-d), nevertheless the pixels are well compensated by calibration that takes into account spatial effects (see figures 4 (e) and 4 (f)). Thus, our initial hypothesis was confirmed: it is necessary to run optimization on uniform images, because the display

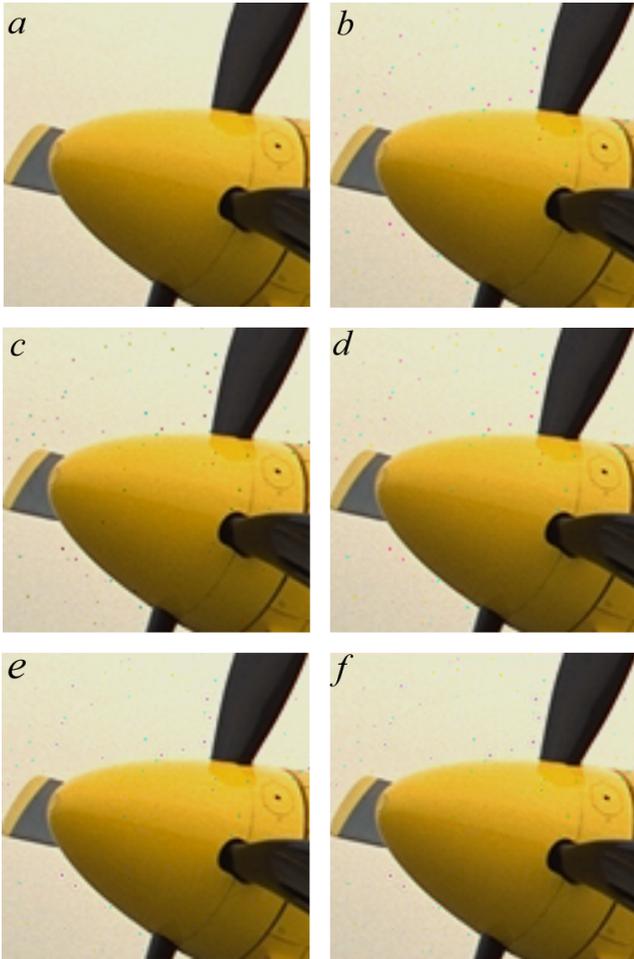


Figure 3: Illustration of changes in the structure of the observed defects: (a) input image, (b) uncalibrated image, (c-f) images after compensation obtained via color image metric based on LAB (c), Oklab (d), S-CIELAB (e), and S-Oklab (f)

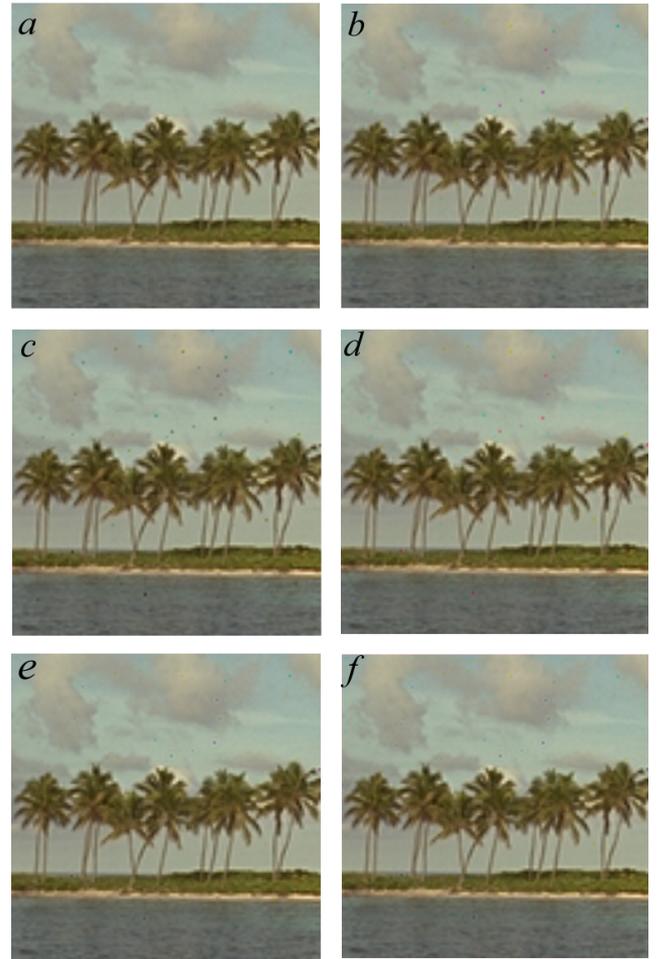


Figure 4: Illustration of changes in the structure of the observed defects: (a) input image, (b) uncalibrated image, (c-f) images after compensation obtained via color image metric based on LAB (c), Oklab (d), S-CIELAB (e), and S-Oklab (f)

artifacts are the most noticeable on them.

In addition to the above conclusion about the limits of applicability of the proposed algorithm, it can be seen that in all the figures (3, 4, and 5), LAB shows the worst performance: it makes some defective pixels even more noticeable than they are without calibration. In the case of calibration via Oklab color coordinates it can be seen that calibration improves the visual uniformity of the image and makes defective pixels slightly less (or at least the same) noticeable. However, the best perceived image uniformity among all compared color image metrics was achieved by using the models S-CIELAB or S-Oklab that take into account the spatial properties of the HVS. In addition, taking into account the spatial properties almost eliminates the difference between the color coordinates we estimate color differences in.

DISCUSSION

As a result of testing the calibration on natural scenery images, it was shown that defective pixels are mostly noticeable on uniform areas and almost invisible on textured ones. Thus, our initial hypothesis was

confirmed: it is necessary to optimize color difference on uniform images.

It was shown that if the spatial effects are not taken into account in the optimized functional, the results depend on the color difference formula (color space we calculate difference in) by which the color distance is optimized. Oklab led to better perceptual uniformity than LAB. In cases of large color differences (that is the most noticeable on a defective display), LAB tended to make perceptual differences even more visible. On the other hand, Oklab showed a good performance in display calibration tasks. The future work could be to compare other perceptually uniform color coordinate systems and color difference formulas.

Taking into account the spatial properties almost eliminates the difference between images obtained via different color coordinates we estimate color differences in. The difference between images is completely invisible from a given viewing distance for which the color difference was optimized. One of the reasons for this is that in both S-CIELAB and S-Oklab spatial filtering was performed in the same linear opponent color coordinates. The future work could be to study CSF

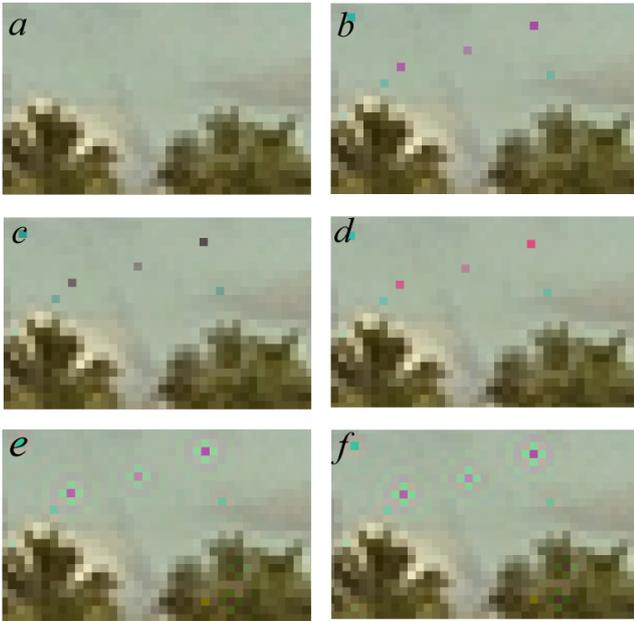


Figure 5: Illustration of changes in the structure of the observed defects: (a) input image, (b) uncalibrated image, (c–f) images after compensation obtained via color image metric based on LAB (c), Oklab (d), S-CIELAB (e), and S-Oklab (f)

simulation in other linear coordinates.

In the case of large differences between defective pixel and a uniform background, the proposed algorithm of display calibration is not able to compensate the defective pixel. Stand alone defective pixels with or without its compensation neighbours are detected by directionally selective neurons. It is known that in the brightness-spatial domain, the HVS is not linear, which is not modeled by the basic CSF and called the masking effect (Nadenau 2000). The term masking effect expresses that the perception of a signal (in our case, defective pixel and its compensation surround) is somehow inhibited by the masking signal (in our case, input image). We can see the masking effect in textured areas: deviations of the defective pixel’s color and its compensation neighbours are less noticeable on textured background. The future work could be to develop the new calibration model that generate a texture on the whole image that compensates defective pixels, for example, an chess-like pattern. Thus, the compensating pixels would be integrated into the overall texture of the image. Stand alone defective pixels and its compensation neighbours would be less noticeable on the entire textured image. This will increase the uniformity of displayed images at the cost of some display resolution reduction.

CONCLUSION

In this work, we consider the defective pixels compensation problem for a display to minimize the perceived non-uniformity of an image. The S-CIELAB image difference metric was used to evaluate the perceived non-uniformity. This method takes into account both the resolution of the human visual system and the

reduced sensitivity to the absolute values of the achromatic component of the image. We propose a reasonable compromise between uniformity of the displayed image, its maximum brightness, and computational efficiency of the algorithm.

The proposed calibration method can be used for factory calibration of displays to compensate for manufacturing variations. A more interesting, if complicated, use case would be compensation of aging effects of urban advertising displays. That would require some means of feedback, like a camera that periodically captures the displayed image for calibration.

We studied the efficiency of the display calibration on natural scenery images. It was shown that defective pixels on real scenery images are very noticeable on uniform background and they are almost invisible on textured background. Thus, our initial hypothesis was confirmed: it is necessary to run optimization on uniform images, because the artifacts of the display are the most noticeable on them. Moreover, we compared the efficiency of compensation depending on color coordinates we calculate color difference in. It was found that Oklab shows better performance than CIELAB in terms of compensation accuracy, however, taking into account the spatial properties of the human visual system using S-CIELAB almost eliminates the difference between the color spaces.

ACKNOWLEDGEMENTS

This work was supported by Russian Science Foundation (Project No. 20-61-47089).

REFERENCES

- Abasi, S.; M. Amani Tehran; and M.D. Fairchild. 2020. “Distance metrics for very large color differences.” *Color Research and Application*, 45(2):208–223.
- Arnold, A.D. and R.S. Cok. 2006. “Oled display with aging compensation.” US Patent 6,995,519.
- Basova, O.A.; A.S. Grigoryev; A.V. Savchik; D.S. Sidorchuk; and Nikolaev D.P. 2020. “On optimal visualization of images on photoemission displays with significant dispersion of efficiency of individual elements.” *Sensornyye sistemy [Sensory systems]*, 34(1):25–31. In Russian.
- Bern, M. and D. Eppstein. 2003. “Optimized color gamuts for tiled displays.” In *Proceedings of the nineteenth annual symposium on Computational geometry*, 274–281.
- Bozhkova, V.P.; O.A. Basova; and D.P. Nikolaev. 2019. “Mathematical models of spatial color perception.” *Information processes*, 19(2):187–199. In Russian.
- Chaji, G. 2019. “Compensation for color variations in emissive devices.” US Patent 10,181,282.
- Chaji, G.; J.M. Dionne; Y. Azizi; J. Jaffari; A. Hormati; T. Liu; and S. Alexander. 2017. “System and methods for aging compensation in amoled displays.” US Patent 9,786,209.
- Darkhovsky, B.S.; A.Y. Popkov; and Y.S. Popkov. 2014. “Method of monte carlo batch iteration to solving by global optimization problems.” *Informatsionnye Tekhnologii i Vychislitel’nye Sistemy [Information technology and computing systems]*, (3):39–52. In Russian.
- Frackiewicz, M. and H. Palus. 2017. “New image quality metric used for the assessment of color quantization algorithms.” In *Ninth International Conference on Machine Vision (ICMV 2016)*, volume 10341, 278–282.

- Harris, S. 2007. "Color and luminance uniformity correction for led video screens." signindustry.com/led/articles/2007-10-15-SH-PulseWidthModulationPWMCorrectionOfLEDDisplays.php3.
- Jepsen, M. L.; N. C. Loomis; B. Bastani; C. Vieri; C. Bralley; and S.B. Abercrombie. 2019. "Masking non-functioning pixels in a display." US Patent 10,354,577.
- Kim, K.; T. Lim; C. Kim; S. Park; C. Park; and C. Keum. 2020. "High-precision color uniformity based on 4d transformation for micro-led." In *Light-Emitting Devices, Materials, and Applications XXIV*, volume 11302, page 113021U.
- Kimpe, T. 2012. "Display assemblies and computer programs and methods for defect compensation." US Patent 8,164,598.
- Kimpe, T.; S. Coulier; and G. Van Hoey. 2006. "Human vision-based algorithm to hide defective pixels in lcds." In *Human Vision and Electronic Imaging XI*, volume 6057, page 60570N.
- Kimpe, T.; A. Xthona; and P. Matthijs. 2004. "Spatial noise and non-uniformities in medical lcd displays: Solution and performance results."
- Kingma, D.P. and J. Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.
- Konovalevko, I.A.; A.A. Smagina; D.P. Nikolaev; and P.P. Nikolaev. 2021. "Prolab: A perceptually uniform projective color coordinate system." *IEEE Access*, 9:133023–133042.
- LG. 2021. "Acceptable number of defective pixels lcd, oled modules of tvs and displays." www.lg.com/ru/support/product-help/CT20206007-1347276421471. In Russian.
- Luo, M.R.; G. Cui; and B. Rigg. 2001. "The development of the cie 2000 colour-difference formula: Ciede2000." *Color Research and Application*, 26(5):340–350.
- Mao, X.-Y.; R.-G. Wang; H.-B. Cheng; J. Miao; Y. Chen; and H. Cao. 2017. "Calibration of abnormal brightness area on the led display." In *ITM Web of Conferences*, volume 11, page 02001.
- McFadden, S.B. and P.A.S. Ward. 2015. "Improving image quality of tiled displays." In *International Conference Image Analysis and Recognition*, 22–29.
- Messing, D.S. and L.J. Kerofsky. 2006. "Using optimal rendering to visually mask defective subpixels." In *Human Vision and Electronic Imaging XI*, volume 6057, 236–247.
- Nadenau, M. 2000. *Integration of human color vision models into high quality image compression*. Ph.D. thesis, Citeseer.
- Nathan, A.; G. Chaji; S. Alexander; P. Servati; R.I. Huang; and C. Church. 2020. "Method and system for programming, calibrating and/or compensating, and driving an led display." US Patent 10,699,624.
- Ottosson, B. 2020. "A perceptual color space for image processing." <https://bottosson.github.io/posts/oklab/>.
- Ponomarenko, N. et al. 2013. "Color image database tid2013: Peculiarities and preliminary results." In *European workshop on visual information processing (EU-VIP)*, 106–111.
- Robertson, A.R. et al. 1977. "Cie recommendations on uniform color spaces, color-difference equations, and metric color terms." *Color Research and Application*, 2(5–6).
- Sai, S.V. 2018. "Metric of fine structures distortions of compressed images." *Computer Optics*, 42(5):829–837.
- Smith, T. and J. Guild. 1931. "The cie colorimetric standards and their use." *Transactions of the Optical Society*, 33(3):73–134.
- Stellbrink, J. 2007. "Comparison of vision-based algorithms for hiding defective sub-pixels." In *Image Quality and System Performance IV*, volume 6494, page 64940Q.
- Stone, M.C. 2001. "Color and brightness appearance issues in tiled displays." *IEEE Computer Graphics and Applications*, 21(5):58–66.
- Uttwani, P.K. et al. 2012. "Detection of physical defects in full color passive-matrix oled display by image driving techniques." *Journal of Display Technology*, 8(3):154–161.
- Verstraete, G. and T. Kimpe. 2015. "Optical correction for high uniformity panel lights." US Patent 9,070,316.
- Wuerger, S.M.; A.B. Watson; and A. Ahumada. 2002. "Towards a spatio-chromatic standard observer for detection." In *Human Vision and Electronic Imaging VII*, volume 4662, 159–172.
- Zhang, X. and B.A. Wandell. 1996. "A spatial extension of cielab for digital color-image reproduction." 27:731–734.

AUTHOR BIOGRAPHIES

OLGA A. BASOVA was born in Rostov-on-Don, Russia. She studied computer science, obtained her master's degree in 2019 at Moscow Institute of Physics and Technology. She has been developing image processing frameworks with the Vision Systems Lab at the Institute for Information Transmission Problems since



2018. Her research interests are image processing and enhancement methods, color appearance models. Her e-mail address is basova.aa@phystech.edu.

ANTON S. GRIGORYEV was born in Petropavlovsk-Kamchatskiy, Russia. Having graduated from Moscow Institute of Physics and Technology, he has been developing industrial computer vision systems with the Vision Systems Lab at the Institute for Information Transmission Problems since 2010. His research interests are image processing and enhancement methods, autonomous robotics and software architecture. His e-mail address is me@ansgri.com.



DMITRY P. NIKOLAEV was born in Moscow, USSR. He studied physics and computer science, obtained his master's degree in 2000 and his Ph.D. degree in 2004, all at 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 the computationally effective algorithms. His e-mail address is dmonstr@iitp.ru.



How to run a world record? A Reinforcement Learning approach

*Sajad Shahsavari, Eero Immonen
Computational Engineering and Analysis (COMEA)
Turku University of Applied Sciences
20520 Turku, Finland
Email: *sajad.shahsavari@turkuamk.fi

Masoomeh Karami, Hashem Haghbayan,
and Juha Plosila
Department of Computing
University of Turku (UTU)
20500 Turku, Finland

KEYWORDS

Optimal Control, Optimal Pacing Profile, Reinforcement Learning, Competitive Running Race

ABSTRACT

Finding the optimal distribution of exerted effort by an athlete in competitive sports has been widely investigated in the fields of sport science, applied mathematics and optimal control. In this article, we propose a reinforcement learning-based solution to the optimal control problem in the running race application. Well-known mathematical model of Keller is used for numerically simulating the dynamics in runner's energy storage and motion. A feed-forward neural network is employed as the probabilistic controller model in continuous action space which transforms the current state (position, velocity and available energy) of the runner to the predicted optimal propulsive force that the runner should apply in the next time step. A logarithmic barrier reward function is designed to evaluate performance of simulated races as a continuous smooth function of runner's position and time. The neural network parameters, then, are identified by maximizing the expected reward using on-policy actor-critic policy-gradient RL algorithm. We trained the controller model for three race lengths: 400, 1500 and 10000 meters and found the force and velocity profiles that produce a near-optimal solution for the runner's problem. Results conform with Keller's theoretical findings with relative percent error of 0.59% and are comparable to real world records with relative percent error of 2.38%, while the same error for Keller's findings is 2.82%.

I INTRODUCTION

I-A Background and motivation

A competitive runner attempts to optimize their *pacing* to minimize the time spent in covering a given distance. This optimal pacing (or velocity) profile is always a trade-off between moving faster and saving energy during the race, and is specific to the individual's physique characteristics. The optimal pacing problem has been widely studied in sports science (e.g. Abbiss and Laursen (2008); Casado, Hanley, Jiménez-Reyes, and Renfree (2021); Jones, Vanhatalo, Burnley, Morton, and Poole (2010)) as well as mathematical modeling (e.g. Alvarez-Ramirez (2002)) and optimal control theory (e.g. Keller (1974); Reardon (2013); Woodside (1991)). Through technological advances in high-performance computing as well as in wearable devices, which today are capable of predicting the running power on-line, the problem has also been addressed by computational optimization (e.g. Aftalion (2017); Maroński and Rogowski (2011)).

Finding the optimal running power profile that yields minimum race time is carried out by solving a dynamical optimization problem, subject to constraints from human metabolism and race conditions. Although theoretical solutions for this optimal control problem are proposed in prior research, they fail on more complicated, higher order mathematical models and constraints. On the other hand, numerical approaches, such as nonlinear programming in the so-called direct methods, suffer from *local convergence*, i.e., tending to converge to solutions close to the supplied first guesses, thus, requiring a good initial guess for the solution. This article addresses these issues by using probabilistic controller model trained by reinforcement learning (RL).

Optimizing the pacing strategy for a given race is a well-known open problem (Aftalion & Trélat, 2021). This problem is non-trivial because of: (1) the complexity of the human body metabolism (processes of aerobic and anaerobic energy production/consumption), (2) individual variations in the human physique, (3) mathematical modeling of human metabolism considering individual variations, and (4) resolving the corresponding optimal control problem based on the developed mathematical model.

The study of mathematical models of human body metabolism tries to specify the differential equations that describe the energy (or oxygen) variation in one's body based on the recovery rate and amount of applied work. Typically these models include a number of physiological parameters that characterize the athlete's body and need to be identified, individually, based on the experimental data on athlete's performance. Among these parameters, oxygen uptake rate ($\dot{V}O_2$, rate of oxygen recovery or energy production during the exercise) and critical power (CP, highest possible long-lasting rate of energy consumption) are the most significant ones. Beside running races, mathematical modeling for kinetic and metabolic variability has been used in a wide variety of other competitive sports such as road cycling (Wolf, 2019), swimming (McGibbon, Pyne, Shephard, & Thompson, 2018), horse racing (Mercier & Aftalion, 2020), wheelchair athletics (Cooper, 1990), etc.

The challenge of resolving optimal control profile is not restricted in the sport sciences. Generally, optimal control of dynamical systems governed by differential equations, as a fundamental problem in mathematical optimization, has numerous applications in scientific and engineering research. Such practical applications, for example, include space shuttle reentry trajectory optimization (Rahimi, Dev Kumar, & Alighanbari, 2013), unemployment minimization in economics and management (Kamien & Schwartz, 2012), robotic resource manage-

ment, task allocation (Elamvazhuthi & Berman, 2015) and etc., all of which are ultimately reduced to the problem of finding a sequence of decisions (control variables) over a period of time which optimizes an objective function. Analytical solutions such as Linear-Quadratic Regulator (LQR) can solve optimal control problem for linear systems with quadratic cost function, but usually systems are nonlinear and state-constrained. Numerical solutions for such nonlinear problems can be categorized into two classes: (1) indirect methods: apply first-order necessary optimality conditions based on Pontryagin's Maximum Principle to turn the optimal control problem into a multi-point boundary value problem, which can be solved by an appropriate solver, and (2) direct methods: discretize the problem on time, approximate control input, state, or both with parametric functions and iteratively identify the best parameter set which optimizes the objective function (Böhme & Frank, 2017).

In the current article, we propose a solution based on reinforcement learning to the optimal control problem in running races. In essence, the approach is to use a statistical parameterized control model driven towards the optimal solution by experiencing in the simulated race, with no *a priori* assumptions, whilst sufficiently generic to be applied on a variety types of dynamical models. Moreover, unlike the widely used direct methods in optimal control, here we optimize the parameterized *probabilistic policy function*, enabling it to explore the control trajectory space and iteratively improve its performance. We utilize the well-known mathematical model of Keller (1973) representing the dynamics of motion and human body's energy conservation in the running race application. We used theoretical optimal solution for this model provided by Keller (1974) to validate the solution of our proposed method, since physiological constants of a world-champion runner are also identified in Keller (1973), enabling us to reproduce the exact same dynamical model. It is interesting that while Keller assumes a 3-stage run (initial, middle, end), here this 3-part profile is not assumed *a priori*, but it is part of the solution.

I-B Contributions and key limitations

Overall, this paper presents a machine learning solution that is able to learn the optimal pacing profiles predicted by Keller's theory of competitive running. More specifically, the main contributions of the present work include:

- A neural network-based probabilistic policy model transforming the state of the system into optimal control action in continuous-space.
- A reinforcement learning-based parameter optimization procedure to iteratively improve the policy model's performance.
- Numerical simulation of the dynamical system that is essentially an implementation of Keller's mathematical model (differential equations governing runner's force, energy and velocity dynamics) temporally discretized by the Runge-Kutta approximation method (Tan & Chen, 2012).

- Validation discussion on the proposed method by comparing the predictions of our model with the theoretical results provided by Keller (1973).

It is important to note that the study of convergence of the learner and sensitivity of the approximation (time-discretization of the numerical model due to change in the step size) is performed qualitatively and not studied thoroughly in the present research. Another limitation of the current work is that the proposed RL procedure requires many number of simulated experiences in the environment. Nonetheless, the optimized model by simulated experience could potentially be used as a reasonable initial point of an on-line real-world system, thus reducing the number of required training data samples. Besides, even though the proposed analytical techniques in the literature tries to consider several characteristics/features of the runners' physical characteristics and the environment at the same time, the complex nature of modeling an individual runner makes the solution not to be globally suitable for all the runners. Different physical characteristics of the runners and different environments wherein the runners act provide a wide range of features that affect the constant parameters used in the models and/or might even change the mathematical modeling. Another fact is that the mental and psychological features that might significantly affect the modeling are mainly the result of runner and its environment interaction. In this paper, instead of a bottom-up structural/formal solution for the runner's optimal energy consumption, we propose a *high-level behavioural* model that can be trained and refined over time and can consider and accept many features all in one model. Moreover, the proposed computational approach herein is easily portable to different mathematical models, sports, terrains and even in different applications altogether, like battery electric vehicles.

I-C Relation to previous work

This work is directly linked to Keller (1973). This model is a pioneering mathematical model relying on Newton's law of motion and an energy conservation model assuming that the $\dot{V}O_2$ is constant during the race (the model is fully described in detail in Subsection II-A). Several modifications to this fundamental work have been introduced by considering: the effect of fatigue (Woodside, 1991), wind resistance and altitude (Quinn, 2004), variation in oxygen uptake rate (Behncke, 1997), and etc. To incorporate aerobic and anaerobic energy, a conceptual hydraulic model has been developed (Morton, 2006) comprised of two energy tanks: storage of aerobic energy which has infinite capacity but limited consumption rate, and storage of anaerobic energy which has unlimited consumption rate but is limited in capacity. Aftalion and Bonnans (2014) has utilized the dynamics of this hydraulic model and solved the optimal running problem for 1500-meter race by Runge-Kutta discretization scheme and nonlinear programming solver, all implemented in the Bocop toolbox (Bonnans, Martinon, & Grélard, 2012).

Reinforcement learning has been used for optimal control of dynamical systems in different applications (see for example J. Duan et al. (2019) for optimal charge/discharge control of hybrid AC-DC microgrids or Liu, Xie, and Modiano (2019) on computer network

queueing systems), but to the authors' knowledge it has not been used in optimal control of competitive sports.

I-D Organization of the article

Next sections of this paper is structured as follows: In section II, we formally define the runner's problem and details of the proposed reinforcement learning-based solution. Afterward, validation experiments are described and predictions of the control input for three track lengths are presented in section III. Finally, we conclude the paper in section IV and discuss future directions.

II PROPOSED METHOD

II-A Formal case study definition

Keller (1973) considered a mathematical model to incorporate the applied force dynamics and runner's energy conservation. The dynamical model is described shortly here for clarification.

The time T to run the track with length D is related to the velocity profile $v(t)$ by:

$$D = \int_0^T v(t) dt \quad (1)$$

Governing differential equation of runner's force balance is defined as:

$$\frac{dv(t)}{dt} + \frac{v(t)}{\tau} = f(t) \quad (2)$$

with $v(t)$ being the instantaneous runner's velocity, $f(t)$ the total propulsive force per unit mass applied by the runner and τ the damping/friction constant coefficient. Note that propulsive force $f(t)$ is under control of the runner through which the velocity is manipulated. The initial condition is $v(0) = 0$, i.e., initially the runner is at rest, and the upper bound limit for propulsive force is

$$f(t) \leq F_{max}, \forall t \in [0, T] \quad (3)$$

Dynamics of the runner's energy storage is described by differential equation:

$$\frac{dE(t)}{dt} = \sigma - f(t) \cdot v(t) \quad (4)$$

in which $E(t)$ is the quantity of available muscular energy per unit mass at given time, σ is the rate of energy recovery per unit mass at which energy is supplied to the body by the respiratory and circulatory systems (in excess of the non-running metabolism). Note that initially the runner has a certain amount of available energy in their muscles, i.e., $E(0) = E_0$, and energy can never be negative, i.e.,

$$E(t) \geq 0, \forall t \in [0, T] \quad (5)$$

The optimal control problem is to find $f(t)$, $v(t)$ and $E(t)$ such that (2)-(5) and initial conditions are satisfied and T defined by (1) is minimized. The four physiological constants τ , F_{max} , σ and E_0 and the distance D are given.

Keller (1974) attained analytical solution to this problem by using calculus of variation with optimization of constrained differential equations. He determined the

optimal velocity profile under three assumptions: (1) the race should be divided into exactly three distinct phases, (2) the runner should start with maximum force in the first phase ($f(t) = F_{max}$, for $t \in [0, t_1]$), and (3) the runner should deplete the energy source before the ending point and finish the race with zero energy in the third phase ($E(t) = 0$, for $t \in [t_2, T]$). With these assumptions, he carried out the velocity profile and found that the velocity in the middle phase (with no pre-assumption on it) is constant. He argued that these assumptions arise from the fact that in the optimal solution, variables with inequality constraints should appear at their extreme bounds (and become equality).

II-B Reinforcement learning formulation

Our proposed algorithm for finding optimal control inputs of the runner is based on *policy gradient reinforcement learning* method thereby we use (1) a fully connected feed-forward neural network as the parameterized probabilistic controller model, and (2) a numerical simulation of the dynamical model, that is described in Subsection II-A, to simulate the runner's environment and provide the reward signal. The neural network controller model is used as the parametric policy function to transform current state of the runner into predicted optimal propulsive force for the next time step. This predicted action value is then used in the numerical simulation to apply the propulsive force, update the runner's state accordingly and compute the reward value of the new state. Then the reward value will be used eventually to update the neural network parameters using gradient-based stochastic optimization, see Figure 1 for conceptual illustration.

The RL algorithm, in general, relies on trial and error, while enforcing actions in the *good trials* by rewarding them. In our consideration of the runner's problem, for example, positions close to the finish line are highly rewarded, while any constraint violation will terminate the run with considerably smaller reward value. Utilized policy gradient method is a family of RL algorithms that parameterize the policy function directly and optimize its parameters by gradient ascent on the performance objective (cumulative reward, also known as return). This optimization is usually performed on-policy, meaning that each update only uses data collected while acting according to the most recent version of the policy. The key idea underlying policy gradient methods in RL is to push up the probabilities of actions that lead to higher return, and push down the probabilities of actions that lead to lower return, until the model reaches to the optimal policy.

In this section, we describe the controller model that predicts the propulsive force for a given state, the simulated runner and the RL optimization procedure to identify the controller model parameters.

1) Neural network control model

The neural network controller model takes runner's current state (namely instantaneous position, velocity and available energy) as input and through several parametric hidden layers generates the corresponding predicted mean value of the normal distribution on runner's propulsive

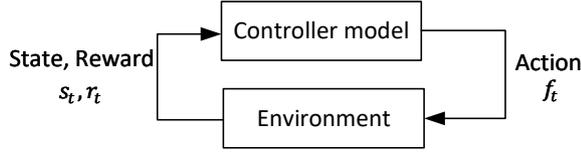


Fig. 1: Modules of reinforcement learning framework: (1) *Controller model*: reads the state and reward from current time of the environment and computes the next action f_t . Also, reward signal r_t is used to train the model, (2) *Environment*: reads the action input to advance execution of the dynamical model one step forward.

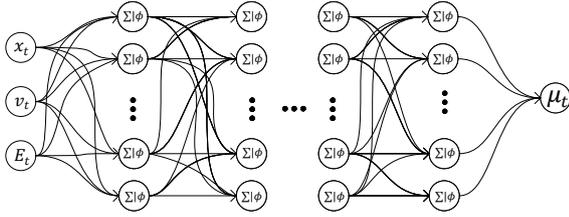


Fig. 2: Structure of a sample neural network as the predictor of the mean of the normal distribution on propulsive force, given runner's state.

force (as the next control input). Suppose that the runner's state in time t is $s_t = (x_t, v_t, E_t)$. If the corresponding output of the neural network for this input state is μ_t , then the next propulsive force is sampled from the probabilistic policy distribution of $\pi_t(f_t|s_t)$ as

$$f_t \sim \mathcal{N}(\mu_t, \sigma^2)$$

where σ^2 is assumed to be constant for all t .

Trainable parameters of the neural network are the set of weight matrices W_i and bias vectors B_i for each layer. Forward path of the neural network consists of computing the values of the stacked hidden layers by

$$z_i = W_i \cdot h_{i-1} + B_i$$

$$h_i = \tanh(z_i)$$

where $\tanh(\cdot)$ is used as the activation function. Finally, the only output of the network is weighted average of the last hidden layer values. Figure 2 shows the structure of the neural network control model.

2) Environment

In RL context, a simulated (or real) environment is needed to perform predicted action of the controller model and compute reward value of the selected action. We implemented a discretized version of dynamics of Keller's mathematical model (Equations (2) and (4)). This numerical simulation reads the instantaneous propulsive force and accordingly performs one step of update in the runner's state. We employed the Runge-Kutta method (RK4) to numerically approximate new values for the state variables (energy and velocity) based on their governing differential equations and previous values. RK4 is a forth-order iterative approximation method and uses a

weighted average of four slope values (evaluation of the derivative function) at the beginning, middle and end of the time interval Δt to approximate the next value of the variable. The accumulative error in RK4 is in order of $\mathcal{O}(\Delta t^4)$ compared to first-order Euler method (Butcher, 2016) with accumulative error in order of $\mathcal{O}(\Delta t)$. Thus, RK4 will lower the approximation error caused by time discretization and reduce the sensitivity of simulation to the time step size.

The *step* method of the runner's simulation will check the success/failure conditions after each update (to check if race is finished or failed). Particularly, successful races are recognized if $x_t = D$ (or greater than), and failures are recognized if one of the followings happens: (1) predicted propulsive force exceeds its maximum possible value ($f_t > F_{max}$), (2) runner runs out of energy ($E_t < 0$), or (3) irrationally, runner moves backward ($v_t < 0$).

In addition, in the RL framework a reward signal is needed for the model optimizer to form the objective function in the training process. Therefore, while step the runner forward, simulation model will compute a reward value for every executed time step. The lower the race time, the more the reward value should be. We use the following instantaneous reward function for each state-action (step):

$$r_t(s_t, f_t) = \begin{cases} \frac{1}{t} \cdot \frac{D}{D-x_t} & \text{if } x_t < D \text{ and not failed,} \\ \frac{1}{t} \cdot \frac{1}{\epsilon} & \text{if } x_t = D, \\ 0 & \text{if failed.} \end{cases} \quad (6)$$

where $s_t = (x_t, v_t, E_t)$ is runner's current state, f_t is predicted action control, *failed* is true if any of the three failure constraints mentioned above is violated and ϵ is a small value. This reward function exponentially increases the reward, as the runner's position approaches the finish line and integrates the race time with inverse relation.

Given a *trajectory* (sequence of states and actions in one run) as $\delta = (s_0, f_0, s_1, f_1, \dots, s_T)$, the cumulative reward (return) is defined as

$$R(\delta) = \sum_{t=0}^T r_t(s_t, f_t)$$

Now, if we consider the return function up to some intermediate time t as $R(\delta, t)$, then it will be proportional to the logarithmic barrier function of

$$R(\delta, t) \sim -\frac{1}{t} \cdot \log\left(1 - \frac{x_t}{D}\right)$$

which helps the optimization convergence (unlike discontinuous reward functions).

Figure 3 demonstrates functions r_t and $R(\delta, t)$ over time for a successful sample trajectory of a 400-meter track. Final peak of instantaneous reward is occurred when the simulation reaches to the finish line. Also, high values in the beginning is due to the term $1/t$ and beneficial in the sense that it leads the optimization toward forward movements at the start of the race. Steep portions in the cumulative reward graph will speed up the

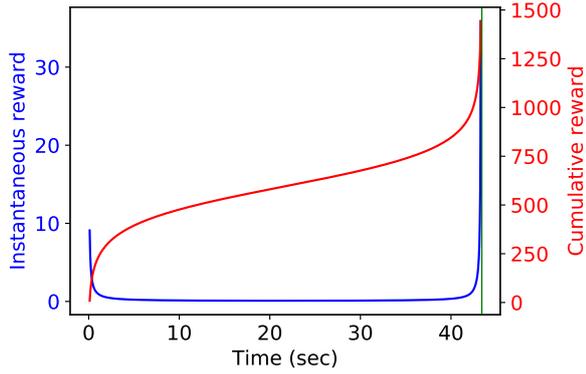


Fig. 3: Graph of (1) instantaneous reward over time (blue curve) which continuously and rapidly increases as the runner’s position x_t approaches the finish line, and (2) cumulative reward over time (red curve) which is the cumulative sum of instantaneous reward values up to time t .

stochastic optimization, since each training step is based on the gradient of expected return (details in the following Subsection).

3) Optimization

The weight parameters of the controller model neural network are initialized by Kaiming uniform initialization method (He, Zhang, Ren, & Sun, 2015): $W_i \sim \mathcal{N}(0, 2/n_i)$, where n_i is the number of inputs in layer i . These model parameters need to be identified. We consider the standard policy-gradient model-free actor-critic reinforcement learning algorithm (Degris, Pilarski, & Sutton, 2012; Y. Duan, Chen, Houthoof, Schulman, & Abbeel, 2016). This algorithm maximizes the expected return function

$$J(\pi) = \mathbb{E}_{\delta \sim \pi} [R(\delta)] \quad (7)$$

incrementally, to identify these parameters (denote all learnable parameters by θ) of the parametric probabilistic policy $\pi_\theta(f|s)$ by applying the gradient of (7) as:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\pi_\theta)|_{\theta_k} \quad (8)$$

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\delta \sim \pi_\theta} \left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(f_t|s_t) A^{\pi_\theta}(s_t, f_t) \right] \quad (9)$$

where $A^{\pi_\theta}(s_t, f_t)$ is the advantage function. The advantage function measures whether or not the action is better or worse than the policy’s default behavior (Schulman, Moritz, Levine, Jordan, & Abbeel, 2015) and is defined as:

$$A^{\pi_\theta}(s_t, f_t) = Q^{\pi_\theta}(s_t, f_t) - V^{\pi_\theta}(s_t) \quad (10)$$

with $Q^{\pi_\theta}(s_t, f_t)$ being the expected return when starting at state s_t , taking action f_t and then following the policy π_θ from there (on-policy action-value function) and $V^{\pi_\theta}(s_t)$ being the expected return when starting in state s_t and following the policy π_θ (Value function). These

two functions are defined, formally, as:

$$Q^\pi(s, f) = \mathbb{E}_{\delta \sim \pi} [R(\delta)|s_0 = s, f_0 = f] \quad (11)$$

$$V^\pi(s) = \mathbb{E}_{\delta \sim \pi} [R(\delta)|s_0 = s] \quad (12)$$

In practice, the expected values are estimated with a sample mean. If we collect a set of trajectories $\mathcal{D} = \{\delta_i\}_{i=1, \dots, N}$ (along with their return values) where each is obtained by employing policy π_θ in running the environment, the policy gradient is estimated with

$$\hat{g} = \frac{1}{N} \sum_{\delta \in \mathcal{D}} \sum_{t=0}^T \nabla_\theta \log \pi_\theta(f_t|s_t) \hat{A}(s_t, f_t) \quad (13)$$

The advantage function is also estimated by training a separate neural network to predict the value function $V^\pi(s)$. Training this model is performed along with training the policy by minimizing the mean squared residual error between estimated value for $V^\pi(s)$ and computed return value in training data.

Essentially, the training procedure consists of repetitively performing these operations: (1) collect training data \mathcal{D} by using current policy’s parameters θ_k and running the environment, (2) compute estimated value function $\hat{V}(s)$ for training data, (3) compute estimated advantage function $\hat{A}(s, f)$ for training data, (4) compute estimated gradient by Equation 13, (5) update the policy’s parameters (and also value function estimation model’s) by applying the gradient ascent (descent) step in Equation 8, and repeat the procedure from (1). For more detail about the RL training procedure see Achiam (2018); Mnih et al. (2016).

III VALIDATION EXPERIMENTS

III-A Setup

We trained separate neural network controller models for three distances 400, 1500 and 10000 meters. The fitted physiological parameters $\tau = 0.892s$, $F_{max} = 12.2m/s^2$, $\sigma = 41.54J/kg \cdot s$ and $E_0 = 2405.8J/kg$ are used from Keller (1973) to simulate a replication of the same dynamical model. Therefore, results in our model and Keller’s theoretical results are comparable. A one-layer neural network with 32 neurons in its hidden layer is used (with total of 161 learnable parameters) for both control input model and value function estimators. The model is trained for 40000 steps, in each step with a batch of 5000 training data samples (collected by running the environment with current policy). Adam optimizer (Kingma & Ba, 2014) has been employed for the stochastic optimization of neural network parameters. A fixed step size is used to for the numerical simulation of the dynamical model (0.1 seconds for 400-meter, 1 second for 1500- and 10000-meter track). The value for the normal distribution variance σ^2 has been set to 0.36.

III-B Results and discussion

Figure 4 shows the progress of training procedure for 400-meter-long track over training steps. Top curve shows the trajectory return value averaged in a batch over training steps, middle curve shows trajectory lengths

averaged in a batch, and the bottom curve shows the value of State-Value function averaged in a batch. The average trajectory length (middle curve) approaches the optimal value (43.9 seconds for the 400-m case) more rapidly at the beginning of the training (0-5K training steps) and then optimizer tries to improve the time, while not violating any constraint. In the bottom curve, it can be seen that the controller model steers the simulated runner into states with higher values that achieve higher rewards. Each training step consists of a single update in the parameters of policy model based on the computed loss over a batch of 5000 data samples. Similar training procedure is performed for the 1500 and 10000 meters. After completion of training steps, all intermediate models have been tested for the best performance and the best race time has been found among them. The predicted mean value for the propulsive force has been used as the control action and no sampling is performed in the test phase.

Table I shows the results of best race times found by the RL-based method along with the theoretical results of Keller and their percent errors with the actual world record times. Average percent error between our RL records and world records is 2.38% (while the error of Keller’s findings is 2.82%). The percent error between our RL records and Keller’s records is 0.59%.

Figure 5 shows the propulsive force profile, dynamics of runner’s position, velocity and available energy in race tracks with length 400, 1500 and 10000 meters (plotted in 5a, 5b and 5c respectively). Keller’s solution is also plotted for comparison. The force and velocity profiles for 1500 and 10000 meters are analogous to theoretical optimal solution which shows the proposed RL method is able to find the sub-optimal solution of the dynamical system. However, interestingly the force profile for the 400-meter track (top plot in Figure 5a) is different in shape with the theoretical solution, even though both result in close race times. The shape difference is due to the fact that RL method is capable of finding more *complex* profiles, compared to Keller’s solution that has a global view point of the solution with *a priori* assumptions which the race should be divided into exactly three phases. In other words, there are only two parameters in Keller’s solution that specify the solution completely: the duration of the first and last phase (t_1 and t_2 in Keller (1973)), compared to 161 parameters in our RL-based controller model. Another potential reason for the different profile shapes is higher resolution of numerical simulation (smaller time step value) for 400 meter that enables fine-grained control for decreasing the velocity smoothly in the ending part of the race. Note for example the ending part (in the time interval from about 38 to 44 seconds) of energy dynamics (bottom plot in Figure 5a) that is laid close to, but not crossing zero energy. In this situation, there is risk of constraint violation when the controller model advances with a relatively large time step. This risk increases especially with the *probabilistic sampling* in the force distribution that can generate a sample propulsive force value far from the mean, therefore violating a system constraint, even with an acceptable mean value lower than maximum force.

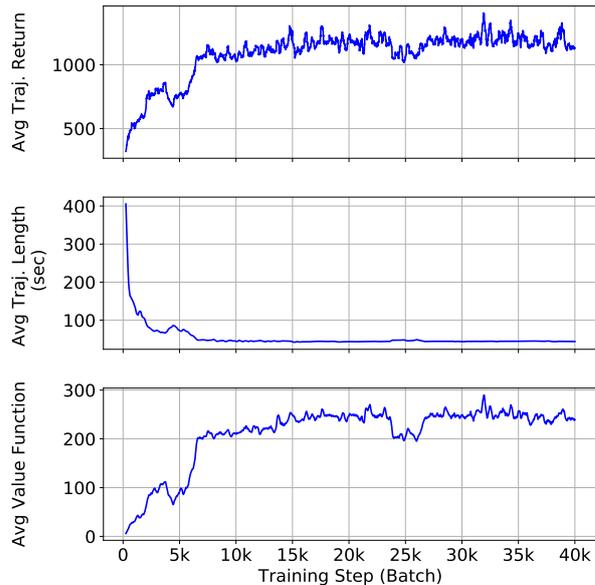


Fig. 4: Learning curves for 400-m track (return, length and evaluation of value function of states, all averaged over batch of training data).

IV CONCLUSIONS

In this article, we described a reinforcement-based solution to find the optimal propulsive force profile in the running race application. The mathematical model of Keller (1973) has been employed to model the dynamics of motion and available energy in the runner’s body. A numerical simulation model has been implemented for this mathematical model using Runge-Kutta approximation method, and along with a reward calculation with a designed logarithmic barrier function, has been used as the environment in the RL framework. Then, the parameters of the neural network-based probabilistic controller model have been trained by policy-gradient model-free actor-critic RL algorithm. Obtained optimal solution conforms with the theoretical results, proving the concept of the proposed method. Unlike theoretical analysis of the problem, the proposed method does not require imposing any *a priori* assumptions regarding the characteristics of the solution, enabling the method to (1) be applied on any dynamical system, and (2) find more complicated solutions in the extended search space.

Despite capability of solving the optimal control problem, the proposed RL-based method has some limitations. First, the training procedure requires to run many trials (experiences) in the simulated environment fitted to the athlete’s body. This limits capability of the method in the on-line application from the scratch. However, an off-line trained model with simulated data can initialize the on-line adaptation with real-world sensory data from the runner. Another limitation of the method is that it does not guarantee the optimality of the solution which is a direct consequence of stochastic optimization used to maximize the total reward.

Future research work on the topic can focus on exploiting more realistic mathematical model for the energy

TABLE I: Race time results for 1972 world records, Keller’s solution and our trained controller model (along with their percent errors). The first two data are obtained from Keller (1973).

Track Length	World Record (sec)	Keller’s Theory Record (sec)	Our RL Record (sec)	Keller’s Theory Error (%)	Our RL Error (%)
400 meters	44.5	43.27	43.9	-2.76	-1.34
1500 meters	213.1	219.44	219.9	2.97	3.19
10000 meters	1659.4	1614.1	1616.0	-2.72	-2.61

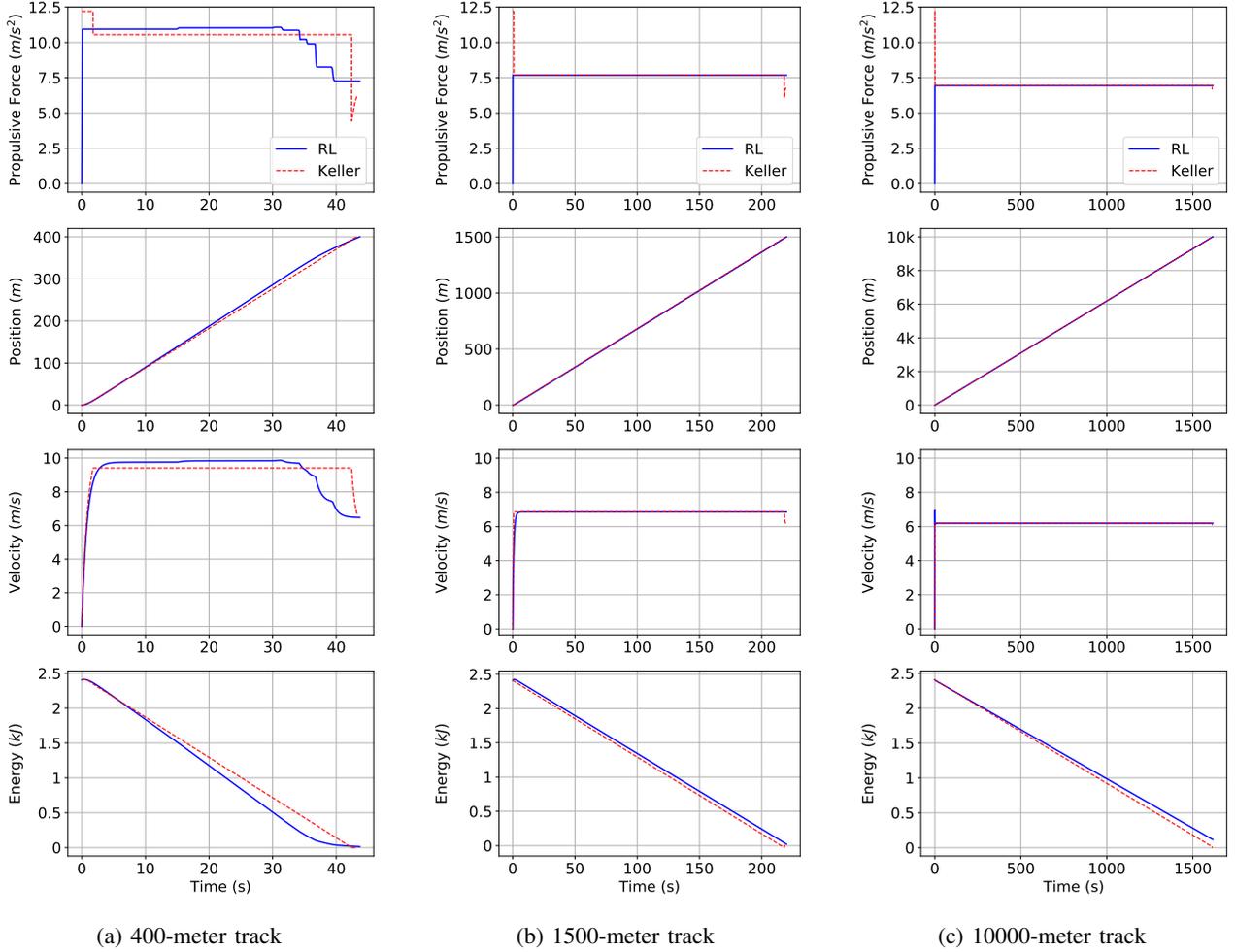


Fig. 5: Propulsive force profile (predicted with our RL method in blue and Keller’s in dashed red) and resulting position, velocity and energy dynamics of the runner.

and motion dynamics (with potentially variable oxygen uptake rate and critical power during the race). Another interesting direction is to incorporate variable variance in the controller model and study the exploration behaviour in the search space due to probabilistic sampling. Here, the neural network predicts the time- and state-dependant variance σ_t^2 of the normal distribution along with its mean, regulating the exploration range during the run and potentially reduce the risk of inequality constraints violation, especially at the end of the track. Sensitivity analysis on the time step size may also be investigated in the future.

SOURCE CODE

The source code of the framework is available on: <https://github.com/COMEA-TUAS/rl-optimal-control-keller>

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from Academy of Finland (ADAFI project).

REFERENCES

Abbiss, C. R., & Laursen, P. B. (2008). Describing and understanding pacing strategies during athletic competition. *Sports Medicine*, 38(3), 239–252.

- Achiam, J. (2018). Spinning up in deep reinforcement learning. *URL* <https://spinningup.openai.com>.
- Aftalion, A. (2017). How to run 100 meters. *SIAM Journal on Applied Mathematics*, 77(4), 1320–1334.
- Aftalion, A., & Bonnans, J. F. (2014). Optimization of running strategies based on anaerobic energy and variations of velocity. *SIAM Journal on Applied Mathematics*, 74(5), 1615–1636.
- Aftalion, A., & Trélat, E. (2021). Pace and motor control optimization for a runner. *Journal of Mathematical Biology*, 83(1), 1–21.
- Alvarez-Ramirez, J. (2002). An improved peronnet-thibault mathematical model of human running performance. *European journal of applied physiology*, 86(6), 517–525.
- Behncke, H. (1997). Optimization models for the force and energy in competitive running. *Journal of mathematical biology*, 35(4), 375–390.
- Böhme, T. J., & Frank, B. (2017). Indirect methods for optimal control. In *Hybrid systems, optimal control and hybrid vehicles: Theory, methods and applications* (pp. 215–231). Cham: Springer International Publishing. doi: 10.1007/978-3-319-51317-1_7
- Bonnans, F., Martinon, P., & Grélard, V. (2012). *Bocop-a collection of examples* (Unpublished doctoral dissertation). Inria.
- Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.
- Casado, A., Hanley, B., Jiménez-Reyes, P., & Renfree, A. (2021). Pacing profiles and tactical behaviors of elite runners. *Journal of Sport and Health Science*, 10(5), 537–549.
- Cooper, R. A. (1990). A force/energy optimization model for wheelchair athletics. *IEEE transactions on systems, man, and cybernetics*, 20(2), 444–449.
- Degrís, T., Pilarski, P. M., & Sutton, R. S. (2012). Model-free reinforcement learning with continuous action in practice. In *2012 american control conference (acc)* (pp. 2177–2182).
- Duan, J., Yi, Z., Shi, D., Lin, C., Lu, X., & Wang, Z. (2019). Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid ac–dc microgrids. *IEEE Transactions on Industrial Informatics*, 15(9), 5355–5364.
- Duan, Y., Chen, X., Houthoofd, R., Schulman, J., & Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning* (pp. 1329–1338).
- Elamvazhuthi, K., & Berman, S. (2015). Optimal control of stochastic coverage strategies for robotic swarms. In *2015 IEEE international conference on robotics and automation (icra)* (pp. 1822–1829).
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).
- Jones, A. M., Vanhatalo, A., Burnley, M., Morton, R. H., & Poole, D. C. (2010). Critical power: implications for determination of v_{O2max} and exercise tolerance. *Medicine & Science in Sports & Exercise*, 42(10), 1876–1890.
- Kamien, M. I., & Schwartz, N. L. (2012). *Dynamic optimization: the calculus of variations and optimal control in economics and management*. courier corporation.
- Keller, J. B. (1973). A theory of competitive running. *Physics today*, 26(9), 42–47.
- Keller, J. B. (1974). Optimal velocity in a race. *The American Mathematical Monthly*, 81(5), 474–480.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Liu, B., Xie, Q., & Modiano, E. (2019). Reinforcement learning for optimal control of queueing systems. In *2019 57th annual allerton conference on communication, control, and computing (allerton)* (pp. 663–670).
- Maroński, R., & Rogowski, K. (2011). Minimum-time running: a numerical approach. *Acta of Bioengineering and Biomechanics/Wroclaw University of Technology*, 13(2), 83–86.
- McGibbon, K. E., Pyne, D., Shephard, M., & Thompson, K. (2018). Pacing in swimming: A systematic review. *Sports Medicine*, 48(7), 1621–1633.
- Mercier, Q., & Aftalion, A. (2020). Optimal speed in thoroughbred horse racing. *Plos one*, 15(12), e0235024.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937).
- Morton, R. H. (2006). The critical power and related whole-body bioenergetic models. *European journal of applied physiology*, 96(4), 339–354.
- Quinn, M. (2004). The effects of wind and altitude in the 400-m sprint. *Journal of sports sciences*, 22(11–12), 1073–1081.
- Rahimi, A., Dev Kumar, K., & Alighanbari, H. (2013). Particle swarm optimization applied to spacecraft reentry trajectory. *Journal of Guidance, Control, and Dynamics*, 36(1), 307–310.
- Reardon, J. (2013). Optimal pacing for running 400-and 800-m track races. *American Journal of Physics*, 81(6), 428–435.
- Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- Tan, D., & Chen, Z. (2012). On a general formula of fourth order runge-kutta method. *Journal of Mathematical Science & Mathematics Education*, 7(2), 1–10.
- Wolf, S. (2019). *Applications of optimal control to road cycling* (Unpublished doctoral dissertation).
- Woodside, W. (1991). The optimal strategy for running a race (a mathematical model for world records from 50 m to 275 km). *Mathematical and computer modelling*, 15(10), 1–12.

AUTHOR BIOGRAPHIES

SAJAD SHAHSAVARI works as Researcher at Turku University of Applied Sciences, and is a PhD candidate at University of Turku, Department of Computing, Finland.

EERO IMMONEN is an Adjunct Professor at Department of Mathematics at University of Turku, Finland, and works as Principal Lecturer at Turku University of Applied Sciences, Finland.

MASOOMEH KARAMI is a PhD candidate in Autonomous Systems Lab at University of Turku, Department of Computing, Finland.

HASHEM HAGHBAYAN is a post-doctoral researcher at University of Turku, Department of Computing, Finland.

JUHA PLOSILA is a Professor in the field of Autonomous Systems and Robotics at the University of Turku, Department of Computing, Finland.

ELONGATED BOUNDARIES DETECTOR PARAMETERS OPTIMISATION BASED ON GENERATION OF SYNTHETIC DATA FROM AERIAL IMAGERY

Ekaterina Panfilova
Evocargo LLC, Moscow, Russia
V. A. Trapeznikov Institute of
Control Sciences, RAS,
Profsoyusnaya street, 65, Moscow, 117997, Russia
E-mail: mipt.epanfilova@gmail.com

Anton Grigoryev
Evocargo LLC, Moscow, Russia
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: me@ansgri.com

Vladimir Burmistrov
Evocargo LLC, Moscow, Russia
E-mail: burbiksvy@gmail.com

KEYWORDS

Boundaries detector, synthetic dataset, road markings detection, Optuna toolkit

ABSTRACT

The detector of elongated boundaries in the image, such as road marking lines, rails, etc., is an important component of the visual system of a highly automated vehicle (HAV). It is used by HAV to solve self-localization problems or maintain the position inside the lane. Testing and optimisation of computer vision algorithms include preparing of the datasets that are usually labelled manually. Synthetic data of various kinds can reduce the complexity of algorithms development. In this paper we describe an approach to generation of synthetic data for testing elongated boundaries detectors that works with road images obtained from the cameras mounted on the HAV and converted to bird's eye view. This approach is based on cutting of raster aerial imagery and corresponding vector markup of target objects in the aerial imagery in various ways.

There is a class of elongated boundaries detectors, the first stage of which is the background suppression on the image and thus highlighting of the target lines. For them, we propose a method for creating a dataset consisting of images with a suppressed background, specifically, images of white elongated lines on a black background. The lines' shape will be similar to the target lines of the detector. With such a dataset you can tune the parameters, which affect stages of the algorithm, following the suppression of the background in the image.

In this paper we also consider elongated boundaries detector. Its parameters fix the geometric model of the target lines. Automatic optimisation of the quality of

such a detector is possible using the Optuna toolkit, but it requires a large dataset. In this paper, we propose an approach to optimisation of the detector on a synthetic dataset. The effectiveness of this approach is confirmed by testing on real data.

INTRODUCTION

The detector of elongated boundaries in the images of the road is an important component of the visual system of a highly automated vehicle (HAV). The elongated boundaries in the images of the road can be, for example, road markings lanes, traffic-way borders, tram tracks, etc. (see fig. 1).

The information about the boundaries location in the image are necessary to solve the problems such as map-relative localization of the HAV (Ziegler et al. 2014; Shipitko and Grigoryev 2018), maintaining the position of the HAV inside the lane (Wang et al. 2005) or lane departure warnings (Jung and Kelber 2004; Hsiao et al. 2008).

There are various approaches to the detecting elongated boundaries in the world. Approaches based on machine learning methods: such as structural forest (Xiao et al. 2016), convolutional neural networks (Teplyakov et al. 2021), recurrent neural networks (Zheng et al. 2020), combinations of different types of neural networks (Dong et al. 2021; Zhang et al. 2021), neural networks which analyze not only input image but also use engineering features (Erlygin and Teplyakov 2021). Approaches based on classical image processing methods: for example, image background suppression in order to highlight target lines and then approximation of the highlighted lines by polylines using the Hough transform (Jang et al. 2014) or the window Hough transform (Panfilova and Kunina 2020). The Canny edge detector (Canny 1986) or other edge

detectors can also be used to highlight target lines (Tropin et al. 2019; Mousavi et al. 2019), in combination of probabilistic Hough transform (Hou et al. 2016; Li et al. 2016) or the least squares method (Yoo et al. 2013), applied for polyline approximation.

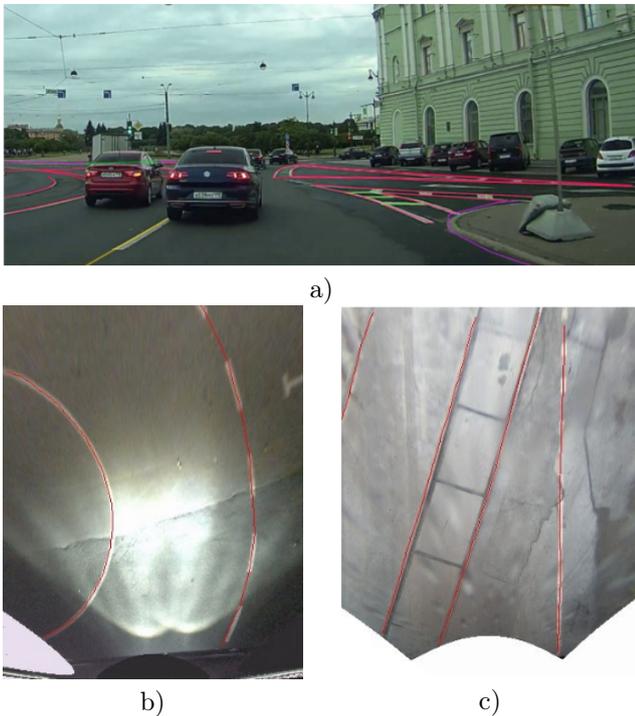


Fig. 1: Examples of elongated boundaries in the images of the roads: a) colour highlights solid and dash-lines, traffic-way borders and safety islands, b) bird's eye view image, colour highlight solid and dash lines, c) bird's eye view image, colour highlights tram tracks and dash-lines

A dataset, usually being marked manually, is required in order to tune the parameters and evaluate the quality of the detector, like of any other computer vision algorithm. There are datasets of roads' images obtained by a camera mounted on a vehicle: TuSimple (<https://github.com/TuSimple/tusimple-benchmark>), CalTech (<http://www.mohamedaly.info/research/lane-detection>), CuLane (Pan et al. 2018), but they are not suitable for the detectors which input data is bird's eye view image. They do not contain parameters of cameras that can be used to convert images to bird's eye view. Moreover, when HAV is launched at a certain territory, it is necessary to tune detector parameters based on data as close as possible to real one, that is, images of the area on which the HAV will move.

In this article we will present method of generating synthetic datasets consisting of bird's eye view images of the road. For the detectors, first step of which is background suppression, we propose a dataset of drawn white road marking lines on a black background. To show the practical applicability the elongated boundaries detector (Panfilova and Kunina 2020) will be optimised on the synthetic data. Than the quality of optimised parameters will be checked on a real data.

The rest of the paper is structured as follows: section

II presents the algorithms of generation of synthetic dataset of two types, section III shows the practical applicability of synthetic datasets and section IV summarises the results.

SYNTHETIC DATASETS GENERATION

In this section we will present the algorithms for creating synthetic datasets of two types: aerial imagery synthetic dataset (see fig. 3) and drawn road markings synthetic dataset (see fig. 4). These datasets will be based on a raster aerial imagery (see fig. 2a) and its vector markup of target objects (see fig. 2b).

Aerial imagery synthetic dataset

This type of a dataset consists of bird's eye view images of the road with a user-specified image resolution. The line length as well as the total length of all lines in the image are limited from below. Such a dataset can be used to train and test elongated boundaries detectors that work with bird's eye view images.

Further, we will formulate the precise statement of the problem and present its solution.

Problem statement

Let we have

- Raster aerial imagery I_{map} size of w_{map} by h_{map}
- Its markup of detector target lines (such as road markings lanes, trails, etc.) – $VectMarkup_{map}$
- Image resolution – number pixels per metre is ppm_{map}

It is required to create a dataset (see fig. 3) such as:

- Bird's eye view images of the roads
- Image size is w by h
- Result image resolution is ppm_{im}
- The pixel length of each line is greater than the threshold – thr_{line}
- The total pixel length of all lines in the image is greater than the threshold – thr_{lines}

Algorithm of aerial imagery synthetic dataset generation

The main idea of the algorithm is to copy regions on the aerial imagery in different positions and for each position determine and save markup lines which appears in the region. The region and the obtained markup must be scaled to satisfy the requirements of result images in the dataset. To obtain different positions of the region, the starting point on the aerial imagery is randomly selected, and then the region is shifted horizontally, vertically, and rotated around its centre with fixed deltas. You can see the pseudocode of this algorithm below (Algorithm 1).

The result of the algorithm 1

To obtain the aerial imagery synthetic dataset, we got an aerial imagery of a closed area “Kalibr”, Moscow, Russia (a certain territory used for autonomous vehicle testing) size of $w_{map} = 6475$ by $h_{map} = 7098$

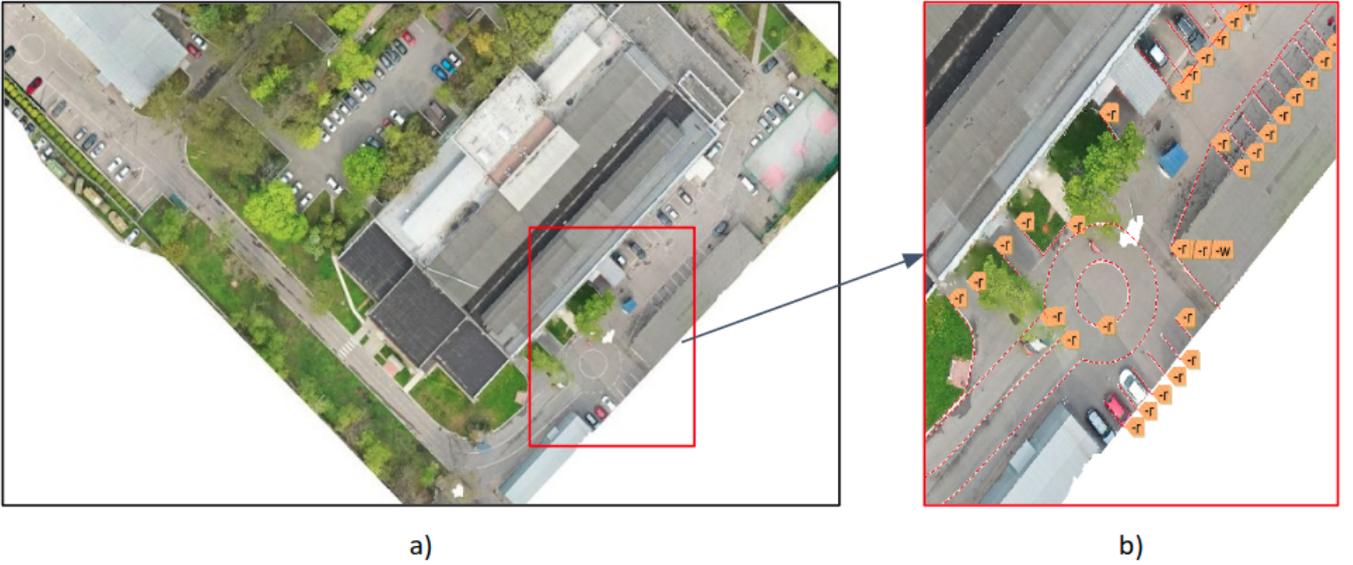


Fig. 2: a) Fragment of aerial imagery. b) Markup of target lines in the aerial imagery fragment.

Algorithm 1 Generation of aerial imagery synthetic dataset

Input: $I_{map}, VectMarkup_{map}, ppm_{map}, w, h, ppm_{im}, sh_x, sh_y, \Delta\alpha, thr_{line}, thr_{lines}, output_path$

Output: Set of bird's eye view images size of w by h at a resolution ppm_{im} and corresponded them files of target lines markup in the $output_path$ directory

```

1:  $scale\_f \leftarrow \frac{ppm_{map}}{ppm_{im}}$ 
2:  $W_w \leftarrow w * scale\_f; W_h \leftarrow h * scale\_f$ 
3:  $x \leftarrow random(0; W_w); y \leftarrow random(0; W_y)$ 
4: ScaleParams( $sh_x, sh_y, thr_{line}, thr_{lines}, scale\_f$ )
5: while  $x < w_{map} - 2 * W_w$  do
6:   while  $y < h_{map} - 2 * W_h$  do
7:      $I_{wind} \leftarrow$  region on  $I_{map}$  with top left coordinate  $(x, y)$  size of  $2W_w \times 2W_h$ 
8:      $VectMarkup_{wind} \leftarrow$  objects in  $VectMarkup_{map}$ , which appear in  $I_{wind}$  and which length  $> thr_{line}$ 
9:     for  $i \leq \frac{360^\circ}{\Delta\alpha}$  do
10:       $\alpha = i * \Delta\alpha$ 
11:       $I_{turn} \leftarrow Turn(\alpha, x + W_w, y + W_h, I_{wind})$  - rotation around point  $(x + W_w, y + W_h)$  by  $\alpha$ 
12:       $VectMarkup_{turn} \leftarrow Turn(\alpha, x + W_w, y + W_h, VectMarkup_{wind})$ 
13:       $I_{res} \leftarrow$  region on  $I_{turn}$  with top left coordinate  $(x + \frac{W_w}{2}, y + \frac{W_h}{2})$  size of  $W_w \times W_h$ 
14:       $VectMarkup_{res} \leftarrow$  objects in  $VectMarkup_{turn}$ , which appear in  $I_{res}$ 
15:       $sum\_length \leftarrow \mathbf{SumLen}(VectMarkup_{res})$ 
16:      if  $sum\_length < thr_{lines}$  then
17:        continue
18:      end if
19:      Scale( $I_{res}, VectMarkup_{res}, \frac{1}{scale\_f}$ )
20:      Save( $I_{res}, VectMarkup_{res}, output\_path$ )
21:    end for
22:     $y = y + sh_y$ 
23:  end while
24:   $x = x + sh_x$ 
25: end while

```

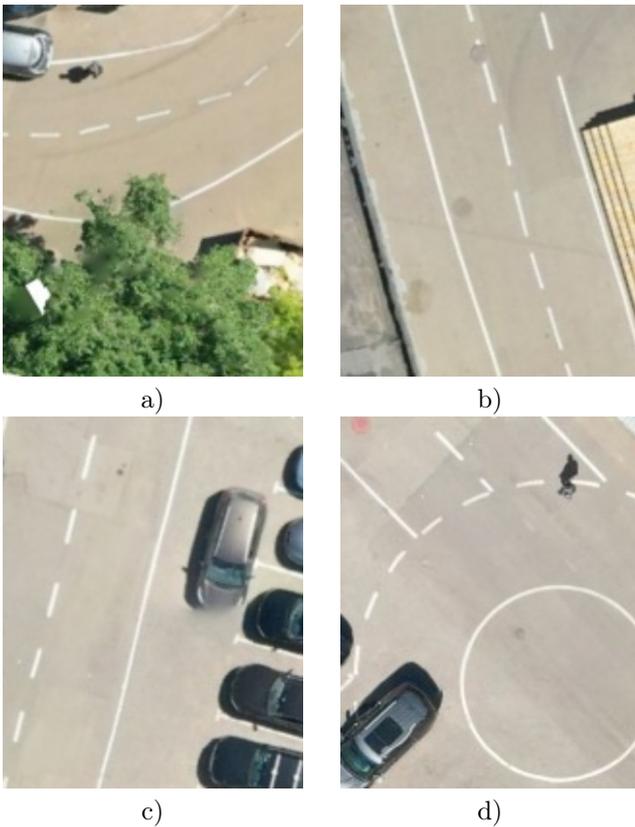


Fig. 3: Examples of images from aerial imagery dataset

pixels with image resolution $ppm_{map} = 14.5785$. Required image parameters in the dataset are $w = 320$, $h = 400$, $ppm_{im} = 60$. The given algorithm parameters are $sh_x = 160$, $sh_y = 200$, $\Delta\alpha = 60^\circ$, $thr_{line} = 30$ pixels, $thr_{lines} = 120$ pixels. As a result we got dataset consisted of 5337 images. You can see the images examples of the resulting dataset in fig. 3, the full dataset is available at <https://zenodo.org/record/6054552>.

Drawn road markings synthetic dataset

This type of a dataset consists of images like white elongated lines in the road markings shape on the black background. As in previous dataset type the image resolution set by user. The line length and the total line lengths are limited from below. We can also set line thickness and the degree of image blur. Such a dataset can be used only for the road lines detectors, first step of which is background suppression. So the user can optimise the parameters that correspond the detector's steps after background suppression.

Further we will formulate the precise statement of the problem and present its solution.

Problem statement

Let us have:

- Vector marking of the detector target lines on the aerial imagery size of w_{map} by h_{map} — **VectMarkup_{map}**. We will call this entity a **vector map of an aerial imagery**.
- Aerial imagery resolution - number of pixels per meter is **ppm_{map}**

It is required to create a dataset such as:

- The images of white lines on a black background (see fig. 4). The lines are geometrically shaped as the target lines of the detector.
- Line thickness - **thk_{line}**
- Image size is **w** by **h**
- Image resolution is **ppm_{im}**
- Measure of lines blurring by Gaussian kernel size of **kernel_w** by **kernel_h** set by sigma equal to **sigma_x = sigma_y = sigma**

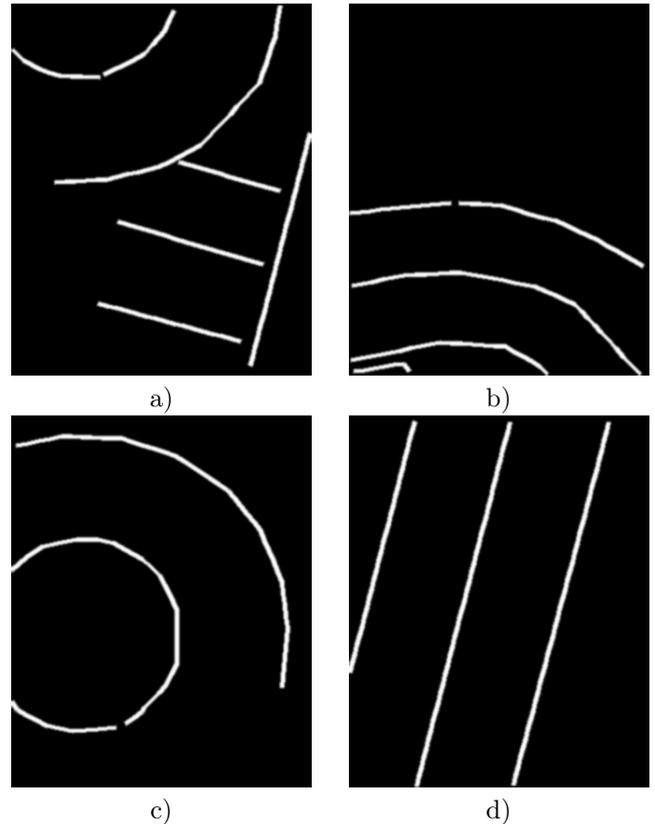


Fig. 4: Examples of images from the drawn road markings synthetic dataset

Algorithm of drawn road markings synthetic dataset generation

The main idea of the algorithm is to consider different positions of the region on the vector map of the aerial imagery. All objects of the vector map that appear into the fixed region are drawn on a black image of a given size with a certain line thickness and then blurred with a Gaussian window. To obtain different window positions, the starting point on the map is randomly selected, and then the region is shifted horizontally, vertically, and rotated around region centre by fixed deltas. You can see the pseudocode of this algorithm below (Algorithm 2).

The result of the algorithm 2

To obtain synthetic dataset we took the same as for the first algorithm aerial imagery of closed territory "Kalibr". So parameters of the aerial imagery are

Algorithm 2 Generation of drawn road markings synthetic dataset

Input: $VectMarkup_{map}$, ppm_{map} , w , h , ppm_{im} , sh_x , sh_y , $\Delta\alpha$, thr_{line} , thr_{lines} , thk_{line} , $kernel_w$, $kernel_h$, $sigma$, $output_path$

Output: Set of images of white lines on a black background and corresponded them files of target lines markup in the $output_path$ directory

```
1:  $scale\_f = \frac{ppm_{map}}{ppm_{im}}$ 
2:  $W_w \leftarrow w * scale\_f$ ;  $W_h \leftarrow h * scale\_f$ 
3:  $x \leftarrow random(0; W_w)$ ;  $y \leftarrow random(0; W_y)$ 
4: ScaleParams( $sh_x$   $sh_y$ ,  $thr_{line}$ ,  $thr_{lines}$ ,  $scale\_f$ )
5: while  $x < w_{map} - 2 * W_w$  do
6:   while  $y < h_{map} - 2 * W_h$  do
7:      $I_{wind} \leftarrow$  region with top left coordinate  $(x, y)$  size of  $2W_w \times 2W_h$ 
8:      $VectMarkup_{wind} \leftarrow$  objects in  $VectMarkup_{map}$ , which appear in  $I_{wind}$  and which length  $> thr_{line}$ 
9:     for  $i \leq \frac{360^\circ}{\Delta\alpha}$  do
10:       $\alpha = i * \Delta\alpha$ 
11:       $VectMarkup_{turn} \leftarrow Turn(\alpha, x + W_w, y + W_h, VectMarkup_{wind})$  - rotation around  $(x + W_w, y + W_h)$ 
12:       $I_{turn} \leftarrow$  region with top left coordinate  $(x + \frac{W_w}{2}, y + \frac{W_h}{2})$  size of  $W_w \times W_h$ 
13:       $VectMarkup_{res} \leftarrow$  objects in  $VectMarkup_{turn}$ , which appear in  $I_{turn}$ 
14:       $sum\_length \leftarrow SumLen(VectMarkup_{res})$ 
15:      if  $sum\_length < thr_{lines}$  then
16:        continue
17:      end if
18:      Scale( $VectMarkup_{res}$ ,  $\frac{1}{scale\_f}$ )
19:       $I_{res} \leftarrow$  3-channel RGB image size of  $w$  by  $h$  filled with  $(0, 0, 0)$  pixels – black colour
20:       $clr \leftarrow (255, 255, 255)$  – set white colour
21:      Draw( $I_{res}$ ,  $VectMarkup_{res}$ ,  $thk_{line}$ ,  $clr$ )
22:      GaussianBlur( $I_{res}$ ,  $kernel_w$ ,  $kernel_h$ ,  $sigma$ )
23:      Save( $I_{res}$ ,  $VectMarkup_{res}$ ,  $output\_path$ )
24:    end for
25:     $y = y + sh_y$ 
26:  end while
27:   $x = x + sh_x$ 
28: end while
```

the same. The required properties of the image are $w = 320$, $h = 400$, $ppm_{im} = 60$. The given algorithm parameters are $sh_x = 320$, $sh_y = 400$, $\Delta\alpha = 120$, $thk_{line} = 5$, $kernel_x = kernel_y = 7$, $sigma = 1$, $thr_{line} = 30$ pixels, $thr_{lines} = 120$ pixels. As a result we got a dataset consisted of 661 images. The images example you can find in the figure 4. The full dataset is available at <https://zenodo.org/record/6054552>.

EXPERIMENTAL RESULTS

Elongated boundaries detector

In order to test the practical applicability of the synthetic dataset, we considered the elongated boundaries detector from the article (Panfilova and Kunina 2020). This detector accepts as input a bird’s eye view image.

Its first stage is the suppression of the background of the image and thus the highlighting of the target lines. The second stage is the approximation of the highlighted lines by polylines and filtration them by length. The quality of this algorithm was tested on the open dataset (see Fig. 3 (a)) presented in the proceedings (Panfilova et al. 2021). The dataset consists of road images from the the closed territory “Kalibr”, Moscow, Russia, files with target lines markup and the parameters of the camera on which the dataset was collected. Knowing the camera’s parameters we can convert images to bird’s eye view (see fig. 5). The algorithm parameters were tuned by an expert. The found line in the image was considered TP if its segments were close enough to the target line and coincided with the target

line direction (see proceedings (Panfilova et al. 2021) for a more detailed description of the quality metrics). Precision of the algorithm is 0.43, recall - 0.73, F-score - 0.54.

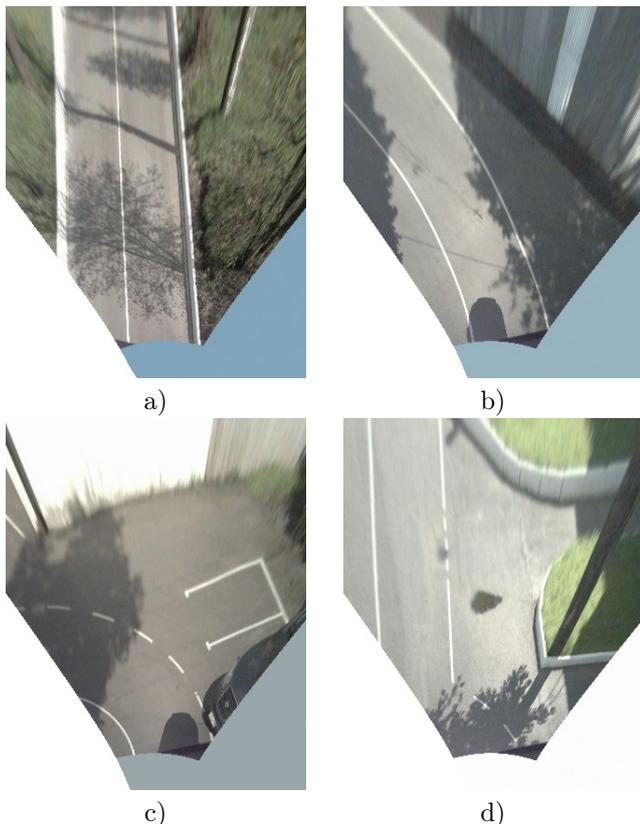


Fig. 5: Examples of images from the open dataset (Panfilova et al. 2021) converted to bird's eye view

Optuna optimisation using synthetic dataset.

In the second section we presented two synthetic datasets: aerial imagery (see fig. 3) and drawn road markings (see fig. 4). In order to optimise the detector parameters and evaluate the detector we merged these datasets together and randomly split merged data into train and evaluation sub-samplings. Train sub-sampling consists of 1300 images: 1000 images of aerial imagery dataset and 300 images of drawn markings line dataset. Test sub-sampling consists of 4998 images: 4337 images of aerial imagery dataset and 661 images of drawn road markings dataset.

To optimise the detector parameters we used Optuna toolkit (Akiba et al. 2019). It allows to optimise model based algorithms like the detector (Panfilova and Kunina 2020). The target value was F-score. We ran the optimisation of the detector on the train subsampling of the synthetic dataset 6 times to get the F-score distribution with randomly chosen starting parameters and with best parameters after optimisation. The distributions were calculated for the train and test sub-samplings of synthetic dataset and for the open dataset of real data obtained from the same territory as aerial imagery for the synthetic dataset (Panfilova et al. 2021). As a result we got the quality of the detector

Increase of F-score value during detector parameters optimisation on the synthetic dataset

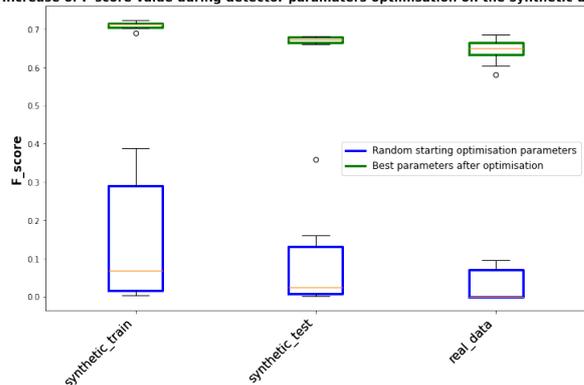


Fig. 6: Boxplot of F-score distribution

on the real data with optimised parameters on average 0.54 higher than the quality of one with randomly chosen starting parameters for optimisation process. You can see the visualisation of the obtained results in the fig. 6.

CONCLUSIONS

In this paper we proposed the approach to generation of synthetic datasets of two types: aerial imagery dataset (see fig. 3) and drawn road markings on a black background dataset (see fig. 4). The main advantage of this method is that you only need aerial imagery and you have to markup the target lines on it just once. The aerial imagery dataset intended to be used for train and test elongated boundaries detectors which input image is a bird's eye view image of the road. The drawn markings dataset was proposed for training the detectors first stage of which is background suppression and whose input image is also a bird's eye view image of the road. In this case you will be able to optimise not all detector's parameters but those that correspond the stages followed by the background suppression. The generated synthetic datasets from the aerial imagery of the "Kalibr", Moscow, Russia are available at <https://zenodo.org/record/6054552>.

Moreover, we showed the practical applicability of the synthetic datasets. Considering the elongated boundaries detector from the proceedings (Panfilova and Kunina 2020), we optimised it on the train part of synthetic dataset and show the quality gain on a real data. Thus, using synthetic data can reduce time of data collecting and it markup and gives acceptable detector quality on a real data.

Further improvement of the synthetic data generation may be focused on simulating different weather conditions, like puddles, shadows, wet paved road and etc.

REFERENCES

- Akiba, Takuya; Shotaro Sano; Toshihiko Yanase; Takeru Ohta; and Masanori Koyama. 2019. "Optuna: A next-generation hyperparameter optimization framework." In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

- Canny, John. 1986. "A computational approach to edge detection." *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.
- Dong, Yongqi; Sandeep Patil; Bart van Arem; and Haneen Farah. 2021. "A hybrid spatial-temporal deep learning architecture for lane detection." *arXiv preprint arXiv:2110.04079*.
- Erlygin, L. A. and L. M. Teplyakov. 2021. "Improvement of a line segment detector based on a neural network by adding engineering features." *Sensory systems*, 35(1):50–54. DOI: 10.31857/S0235009221010042.
- Hou, Changzheng; Jin Hou; and Chaochao Yu. 2016. "An efficient lane markings detection and tracking method based on vanishing point constraints." In *2016 35th Chinese Control Conference (CCC)*, 6999–7004.
- Hsiao, Pei-Yung; Chun-Wei Yeh; Shih-Shinh Huang; and Li-Chen Fu. 2008. "A portable vision-based real-time lane departure warning system: day and night." *IEEE Transactions on Vehicular Technology*, 58(4):2089–2094.
- Jang, Ho-Jin; Seung-Hae Baek; and Soon-Yong Park. 2014. "Lane marking detection in various lighting conditions using robust feature extraction."
- Jung, Claudio Rosito and Christian Roberto Kelber. 2004. "A lane departure warning system based on a linear-parabolic lane model." In *IEEE Intelligent Vehicles Symposium, 2004*, 891–895.
- Li, Yadi; Ligu Chen; Haibo Huang; Xiangpeng Li; Wenkui Xu; Liang Zheng; and Jiaqi Huang. 2016. "Night-time lane markings recognition based on canny detection and hough transform." In *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 411–415.
- Mousavi, Seyed Muhammad Hossein; Vyacheslav Lyashenko; and Surya Prasath. 2019. "Analysis of a robust edge detection system in different color spaces using color and depth images." *Компьютерная оптика*, 43(4):632–646.
- Pan, Xingang; Jianping Shi; Ping Luo; Xiaogang Wang; and Xiaoou Tang. 2018. "Spatial as deep: Spatial cnn for traffic scene understanding." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Panfilova, E. I. and I. A. Kunina. 2020. "Using window hough transform for detecting elongated boundaries in an image." *Sensory systems*, 34(4):340–353. DOI: 10.31857/S0235009220030075.
- Panfilova, Ekaterina; Oleg S Shipitko; and Irina Kunina. 2021. "Fast hough transform-based road markings detection for autonomous vehicle." In *Thirteenth International Conference on Machine Vision*, volume 11605, page 116052B.
- Shipitko, Oleg and Anton S Grigoryev. 2018. "Ground vehicle localization with particle filter based on simulated road marking image." In *ECMS*, 341–347.
- Teplyakov, Lev; Kirill Kaymakov; Evgeny Shvets; and Dmitry Nikolaev. 2021. "Line detection via a lightweight cnn with a hough layer." In *Thirteenth International Conference on Machine Vision*, volume 11605, page 116051B.
- Tropin, D. V.; Y. A. Shemyakina; I. A. Konovalenko; and I. A. Faradjev. 2019. "Localization of planar objects on the images with complex structure of projective distortion." *Информационные процессы*, 19(2):208–229.
- Wang, Jin; Stefan Schroedl; Klaus Mezger; Roland Ortloff; Armin Joos; and Thomas Passegger. 2005. "Lane keeping based on location technology." *IEEE Transactions on Intelligent Transportation Systems*, 6(3):351–356.
- Xiao, Liang; Chuanxiang Li; Dawei Zhao; Tongtong Chen; and Bin Dai. 2016. "Road marking detection based on structured learning." In *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, 2047–2051.
- Yoo, Hunjae; Ukil Yang; and Kwanghoon Sohn. 2013. "Gradient-enhancing conversion for illumination-robust lane detection." *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1083–1094.
- Zhang, Ce; Yu Han; Dan Wang; Wei Qiao; and Yier Lin. 2021. "A network that balances accuracy and efficiency for lane detection." *Mobile Information Systems*, 2021.
- Zheng, Tu; Hao Fang; Yi Zhang; Wenjian Tang; Zheng Yang; Haifeng Liu; and Deng Cai. 2020. "Resa: Recurrent feature-shift aggregator for lane detection." *arXiv preprint arXiv:2008.13719*, 5(7).
- Ziegler, Julius; Henning Lategahn; Markus Schreiber; Christoph G Keller; Carsten Knöppel; Jochen Hipp; Martin Haueis; and Christoph Stiller. 2014. "Video based localization for berth." In *2014 IEEE intelligent vehicles symposium proceedings*, 1231–1238.

AUTHOR BIOGRAPHIES

EKATERINA PANFILOVA was born in Moscow, Russia. She studied applied physics and mathematics, and obtained her Master degree in 2019 in Moscow Institute of Physics and Technology. Currently she is a Phd. student in V. A. Trapeznikov Institute of Control Science, RAS.



Since 2018, she has been working as a junior researcher at the Vision Systems Lab of the Institute for Information Transmission Problems and since 2021 - at Evocargo LLC as a software engineer. Her research activities focus on the areas of computer vision and image processing.

Her email address is mipt.epanfilova@gmail.com.

VLADIMIR BURMISTROV was born in Kiev, Ukraine. Currently he is pursuing a bachelor's degree in computer science and computer technology at the Higher School of Economics. Since 2020, he has been working in Evocargo LLC as a software engineer. His research interests are focused on robotics and data



pipelines architecture.

His e-mail address is burbiksvy@gmail.com.

ANTON GRIGORYEV was born in Petropavlovsk-Kamchatskiy, Russia. Having graduated from Moscow Institute of Physics and Technology, he has been developing industrial computer vision systems with the Vision Systems Lab at the Institute for Information Transmission Problems since 2010. Currently he is also working at Evocargo LLC as a leading software engineer. His research interests are image processing and enhancement methods, autonomous robotics and software architecture.



His e-mail address is me@ansgri.com.

A Model for Predicting the Amount of Photosynthetically Available Radiation from BGC-ARGO Float Observations in the Water Column

Frederic Stahl¹, Lars Nolle^{1,2}, Ahlem Jemai³, Oliver Zielinski^{1,3}

¹German Research Center for Artificial Intelligence (DFKI), Marine Perception
Oldenburg, Germany

Email: {Frederic_theodor.stahl | Lars.Nolle, Oliver.Zielinski}@dfki.de

²Jade University of Applied Science, Department of Engineering Sciences
Wilhelmshaven, Germany

Email: Lars.Nolle@jade-hs.de

³Carl von Ossietzky University of Oldenburg, Institute for Chemistry and Biology of the Marine Environment
Oldenburg, Germany

Email: {Ahlem.Jemai | Oliver.Zielinski}@uol.de

KEYWORDS

Machine Learning, BGC-Argo Floats, Underwater light field, PAR, Downwelling Irradiance

ABSTRACT

Modern oceanography uses, amongst other platforms, automated diving devices, which are drifting with the ocean current whilst continuously collecting vertical profiles of environmental parameters. One of the important parameters is photosynthetically available radiation (PAR). It was studied in this work whether the PAR values can be reconstructed by combinations of measurements from the remaining onboard sensors with specific wavelength. If a reconstruction of PAR is possible, this would allow allocating the sensor with a further specific wavelength instead of PAR. Having available more spectral information could for example enable natural scientists to better distinguish phytoplankton or UV radiation. Therefore, data from three different expeditions from different regions of the world were used to model PAR using multiple linear regression and regression trees (RT). Multiple linear regression achieved an R^2 value of 0.970 for the combined dataset and RT achieved an R^2 value of 0.960. Hence, the models are accurate enough to predict the PAR parameter without the need for a dedicated PAR sensor. Thus the PAR sensor reading could be replaced with measurements of an additional wave length.

INTRODUCTION

Modern operational oceanography uses a plethora of different autonomous platforms [1]. Among them, the nearly 4000 Argo floats [2], automated diving devices, drifting with the ocean current and collect continuous vertical profiles from a depth ~ 2000 m, evolved to be a core component. With Argo float data being transmitted via the Iridium or Argos satellite systems, data is publicly and freely available via two global data assembly centers

(GDAC) typically within 24 hours (see Argo website <https://argo.ucsd.edu>).

While Argo started with a three sensor setup aiming at physical oceanographic information, there has been a significant increase in bio-optical instrumentation on Argo, leading to the biogeochemical Argo (short BGC-Argo) initiative [3]. Together with this increase in sensors, accompanied by the data management and quality control processes, demand for machine learning has been on the rise [3,4].

In this context, the BGC-Argo community suggested to re-configure the Ocean Color Radiometer (OCR) to dismiss the fourth channel, originally designed to record PAR measurement, since this could potentially be reconstructed from the three available distinct channels, measuring wavelengths at 380 nm, 412 nm, and 490 nm. In this study, a machine learning approach is provided, that models the entire wavelength ranges of PAR from the three wavelengths. This enables including a further specific wavelength and thus increase the flexibility of the device [1].

RADIOMETRIC PROFILING FLOAT OBSERVATIONS

The underwater light field is one of the six essential variables measured by so-called BGC-Argo Floats [6]. Featuring the multispectral technology, the OCR-504 from SATLANTIC Inc./Sea-Bird Scientific, USA [7] is used to routinely measure the radiometric observation at four channels. Three channels 380 nm, 412 nm and 490 nm were selected as they are related to the main variations in underwater optical properties. The fourth channel is dedicated to measure PAR. Figure 1 shows the Argo APEX Float WMO7900562, deployment on 27th of September 2019 in the western Mediterranean, with attached sensors, including the radiometer (left) and the radiometer OCR-504 (right).



Figure 1: Argo APEX platform with attached sensors, including

The PAR parameter is commonly used to disclose the overall light available for the primary production in natural waters and allows for the integration of downward irradiance between 400 nm and 700 nm. Recently, Jemai et al. [4] provided a review of radiometric measurements on Argo floats. They, as well as Organelli et al. [8], emphasized the need for more spectral information, from multi- to hyperspectral instrumentation. This platform provided the data that was used for the modelling as described below.

The data used in this study is publicly available at <ftp://ftp.ifremer.fr/ifremer/argo/dac/coriolis>. The dataset represents the German contribution within the BGC-Argo program. The data was acquired by four floats deployed at different sites, one (WMO 7900585) in the North Atlantic, one (WMO7900562) in the Mediterranean Sea, and two (WMO7900579 and WMO7900580) in the Baltic Sea. Radiometric observations were collected during the ascent phases every two or five days in the upper layer, and sampling was carried out at 2 dbar vertical resolution for all floats.

PRELIMINARY ANALYSIS AND PROCESSING OF THE DATA

Data from three different expeditions from different regions were used. From the Mediterranean Sea one dataset with 13,068 data instances was used; from the Baltic Sea two datasets were used, one with 1,373 data instances and the other with 1,274 data instances and Atlantic Ocean with 4,403 data instances. Some data instances contained a very small amount of missing values, these data instances were removed. Missing value

are caused by malfunction of the float. In total, 20,079 instances were available after deletion of missing values.

In order to establish the correlation between the different sensors, a scatter matrix with all float datasets concatenated was plotted as can be seen in Figure 2. It can be seen that there is generally a good correlation between all sensors, except for P (pressure). What can also be seen is that for P below 100 dbar (equivalent to an approximate depth of 100 m), all sensors produce low values. The reason for this is that light at this depth is fully absorbed by the water. Since this is the case for all sensors, it has no effect on the correlation between all optical sensors. The fact that P does not correlate with the other sensors has subsequently also been confirmed with the Institute for Chemistry and Biology of the Marine Environment. Therefore, it was decided to exclude P from modelling.

MODELLING

Two different methods for modelling were selected, Multiple Linear Regression [9] and Regression Trees [10]. The reason for choosing these two techniques is that they produce predictive models for continuous target variables. All data instances were used for the modelling process with input variables downwelling irradiance at 380 nm, 412 nm, 490 nm and target variable PAR. 30% of the data instances were randomly selected without replacement to be included into a test set and the remaining instances were used to fit the models.

In total, ten models were produced, two for each float location, i.e. one Regression Tree and one Multiple linear regression equation, and two models for all float data combined.

Models based on Multiple Linear Regression

For the modelling standard Multiple Linear Regression [9] was used without forcing an intercept. In this paper the Scikit-learn implementation (<https://scikit-learn.org/stable/>) of Multiple linear regression was used. The results are presented in Figures 3-7. The figures plot the true PAR values versus the predicted PAR values.

Figure 3 shows the result of Multiple Linear Regression using the combined dataset comprising all float locations.

The R^2 value was 0.97. The resulting model can be found in Equation 1.

$$\begin{aligned} \text{PAR} = & 2582.06 * \text{Ed}_{380} \\ & - 1715.67 * \text{Ed}_{412} \\ & + 1023.94 * \text{Ed}_{490} - 1.57 \end{aligned} \quad (1)$$

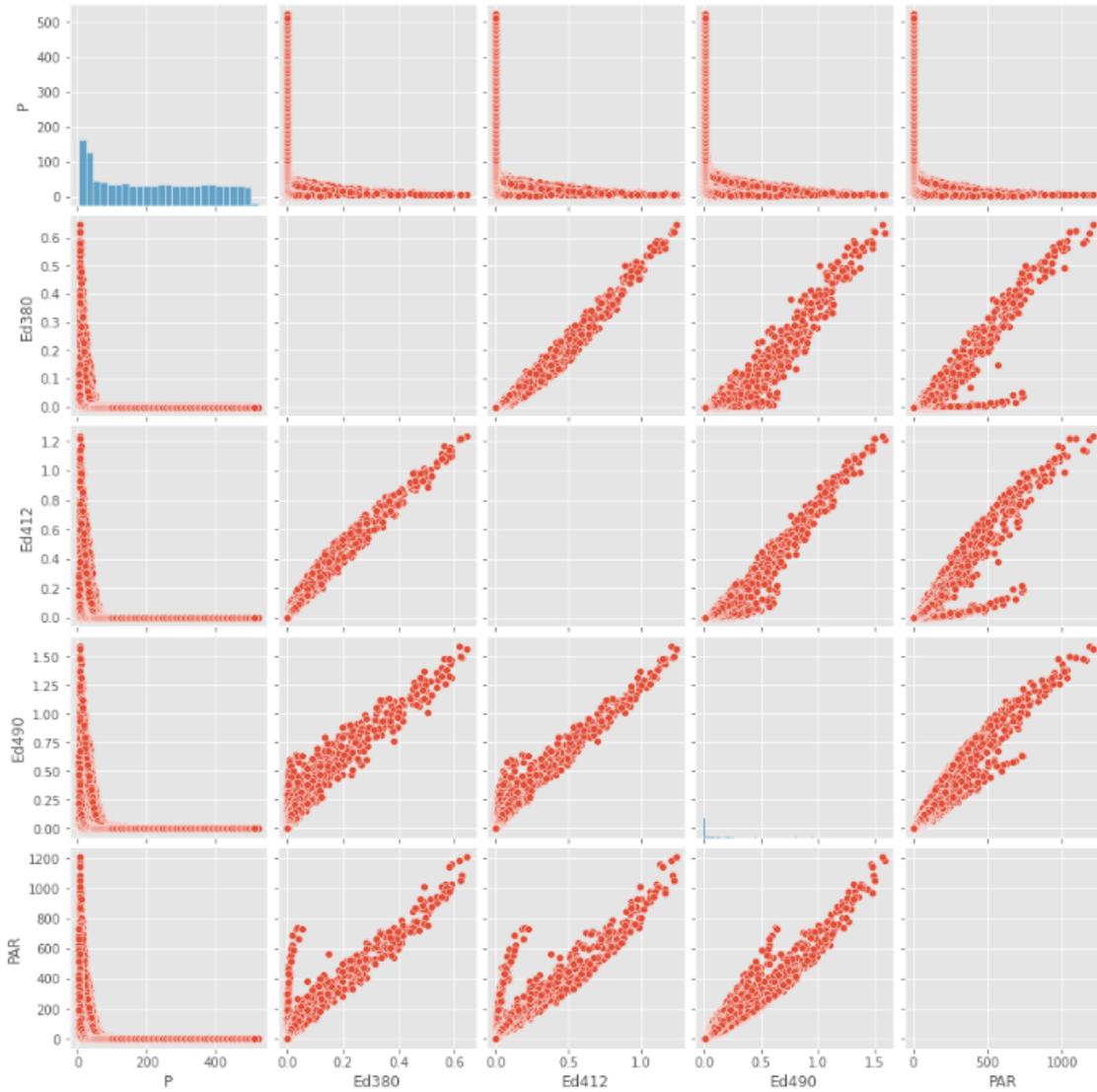


Figure 2: Dependency of sensors of the float

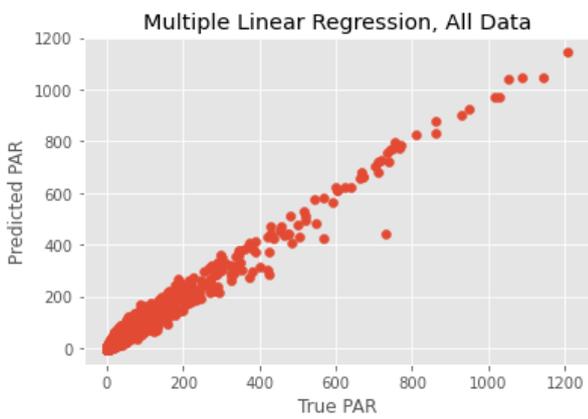


Figure 3: Multiple linear regression on all data

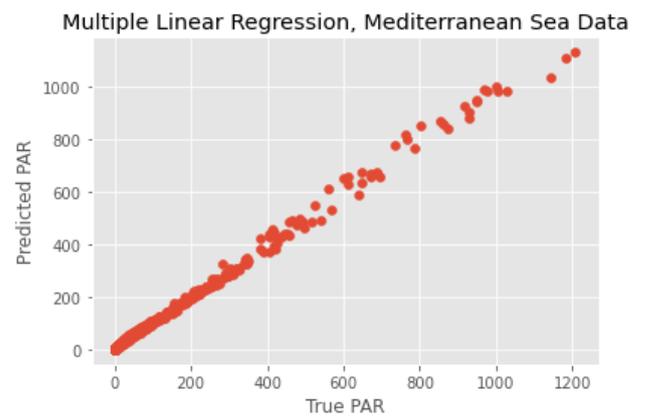


Figure 4: Multiple linear regression on Mediterranean Sea data

Figure 4 shows the result of multiple linear regression using the dataset comprising data for the Mediterranean Sea float location.

The R^2 value was 0.997. The resulting model can be found in Equation 2.

$$PAR = 1744.62 \cdot Ed380 - 726.90 \cdot Ed412 \quad (2)$$

$$+578.50*Ed490-1.14$$

Figure 5 shows the result of multiple linear regression using the dataset comprising data for the Baltic Sea float location 1.

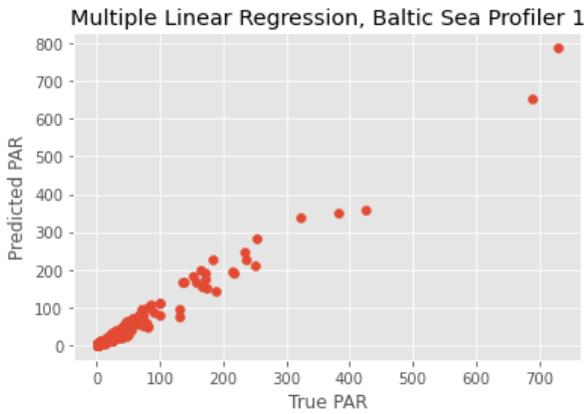


Figure 5: Multiple linear regression on Baltic Sea float 1

The R^2 value was 0.991. The resulting model can be found in Equation 3.

$$\begin{aligned} \text{PAR} = & 14321.34 * Ed380 \\ & -2350.74 * Ed412 \\ & +1168.52 * Ed490 + 1.88 \end{aligned} \quad (3)$$

Figure 6 shows the result of multiple linear regression using the dataset comprising data for the Baltic Sea float location 2.

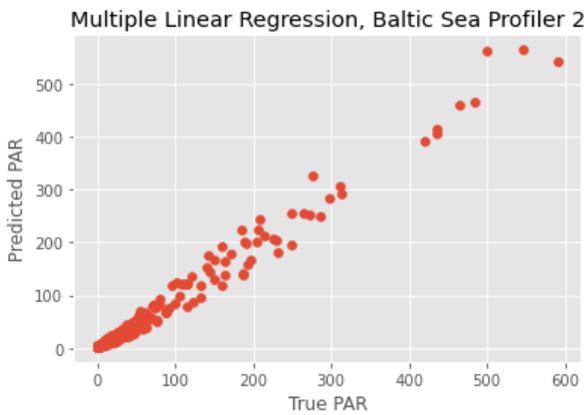


Figure 6: Multiple linear regression on the Baltic Sea float 2

The R^2 value was 0.983. The resulting model can be found in Equation 4.

$$\begin{aligned} \text{PAR} = & 3644.03 * Ed380 \\ & -200.34 * Ed412 \\ & +966.75 * Ed490 + 1.44 \end{aligned} \quad (4)$$

Figure 7 shows the result of multiple linear regression using the dataset comprising data for the Atlantic Ocean float location.

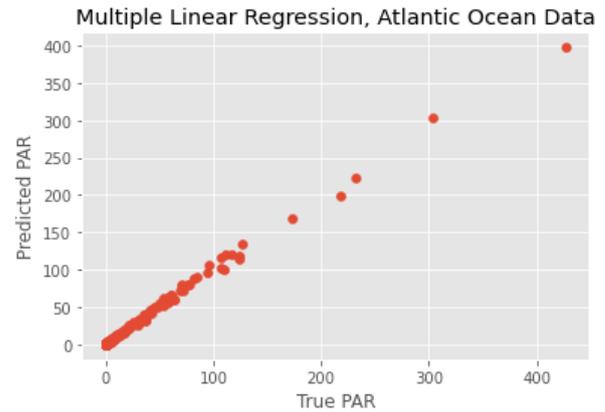


Figure 7: Multiple linear regression on Atlantic Ocean Data

The R^2 value was 0.996. The resulting model can be found in Equation 5.

$$\begin{aligned} \text{PAR} = & 805.10 * Ed380 \\ & +203.80 * Ed412 \\ & +494.75 * Ed490 - 0.38 \end{aligned} \quad (5)$$

Models based on Regression Trees

A regression tree algorithm generating a binary tree was used in this research. The central task was to find a split that leads to an optimal separation of data [10]. In this paper the Scikit-learn implementation of regression tree was used, which makes use of the Gini Importance [11] to choose an attribute to split on. Figures 9 to 13 plot the predicted PAR values versus the groundtruth. When compared with the results for linear regression, it can be seen that groups of the plotted data points are aligned horizontally. This is because the regression tree predicts value ranges rather than individual values. Figure 8 shows the resulting tree for the combined dataset. Due to the complexity of the tree, the parameters of the subsequent tree based models are omitted.

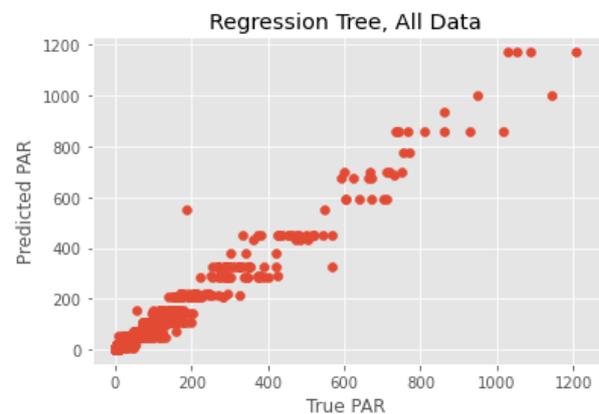


Figure 9: Regression tree on all data

```

|--- Ed490 <= 0.39
| |--- Ed490 <= 0.10
| | |--- Ed490 <= 0.02
| | | |--- Ed490 <= 0.01
| | | | |--- Ed412 <= 0.00
| | | | | |--- value: [0.08]
| | | | |--- Ed412 > 0.00
| | | | | |--- value: [2.52]
| | | |--- Ed490 > 0.01
| | | | |--- Ed380 <= 0.00
| | | | | |--- value: [12.49]
| | | | |--- Ed380 > 0.00
| | | | | |--- value: [5.54]
| | |--- Ed490 > 0.02
| | | |--- Ed380 <= 0.00
| | | | |--- Ed490 <= 0.04
| | | | | |--- value: [32.61]
| | | | |--- Ed490 > 0.04
| | | | | |--- value: [67.39]
| | |--- Ed380 > 0.00
| | | |--- Ed490 <= 0.06
| | | | |--- value: [17.13]
| | | |--- Ed490 > 0.06
| | | | |--- value: [41.63]
| |--- Ed490 > 0.10
| | |--- Ed490 <= 0.20
| | | |--- Ed380 <= 0.00
| | | | |--- Ed490 <= 0.12
| | | | | |--- value: [102.96]
| | | | |--- Ed490 > 0.12
| | | | | |--- value: [161.96]
| | |--- Ed380 > 0.00
| | | |--- Ed490 <= 0.15
| | | | |--- value: [59.81]
| | | |--- Ed490 > 0.15
| | | | |--- value: [94.03]
| |--- Ed490 > 0.20
| | |--- Ed380 <= 0.01
| | | |--- Ed490 <= 0.28
| | | | |--- value: [231.32]
| | | |--- Ed490 > 0.28
| | | | |--- value: [335.38]
| | |--- Ed380 > 0.01
| | | |--- Ed380 <= 0.05
| | | | |--- value: [123.24]
| | | |--- Ed380 > 0.05
| | | | |--- value: [184.44]
|--- Ed490 > 0.39
| |--- Ed380 <= 0.28
| | |--- Ed380 <= 0.18
| | | |--- Ed380 <= 0.03
| | | | |--- Ed490 <= 0.50
| | | | | |--- value: [450.25]
| | | | |--- Ed490 > 0.50
| | | | | |--- value: [539.81]
| | | |--- Ed380 > 0.03
| | | | |--- Ed490 <= 0.55
| | | | | |--- value: [255.22]
| | | | |--- Ed490 > 0.55
| | | | | |--- value: [339.21]
| | |--- Ed380 > 0.18
| | | |--- Ed490 <= 0.62
| | | | |--- Ed490 <= 0.52
| | | | | |--- value: [344.61]
| | | | |--- Ed490 > 0.52
| | | | | |--- value: [394.69]
| | | |--- Ed490 > 0.62
| | | | |--- Ed490 <= 0.82
| | | | | |--- value: [459.72]
| | | | |--- Ed490 > 0.82
| | | | | |--- value: [508.59]
| |--- Ed380 > 0.28
| | |--- Ed490 <= 1.16
| | | |--- Ed380 <= 0.38
| | | | |--- Ed490 <= 1.07
| | | | | |--- value: [605.62]
| | | | |--- Ed490 > 1.07
| | | | | |--- value: [689.01]
| | | |--- Ed380 > 0.38
| | | | |--- Ed380 <= 0.41
| | | | | |--- value: [688.43]
| | | | |--- Ed380 > 0.41
| | | | | |--- value: [741.71]
| | |--- Ed490 > 1.16
| | | |--- Ed490 <= 1.32
| | | | |--- Ed380 <= 0.47
| | | | | |--- value: [802.03]
| | | | |--- Ed380 > 0.47
| | | | | |--- value: [888.95]
| | | |--- Ed490 > 1.32
| | | | |--- Ed380 <= 0.59
| | | | | |--- value: [1021.58]
| | | | |--- Ed380 > 0.59
| | | | | |--- value: [1126.55]

```

Figure 8: Regression tree structure induced on the combined dataset.

Figure 9 shows the result of the regression tree using the combined dataset comprising all float locations. The fit of the regression tree model on all data combined resulted in $R^2 = 0.960$. Figure 10 shows the result of the regression tree using the Mediterranean dataset. The fit of the regression tree model on the Mediterranean Sea data resulted in $R^2 = 0.989$. Figure 11 shows the result of the regression tree using the Baltic Sea dataset float location 1.

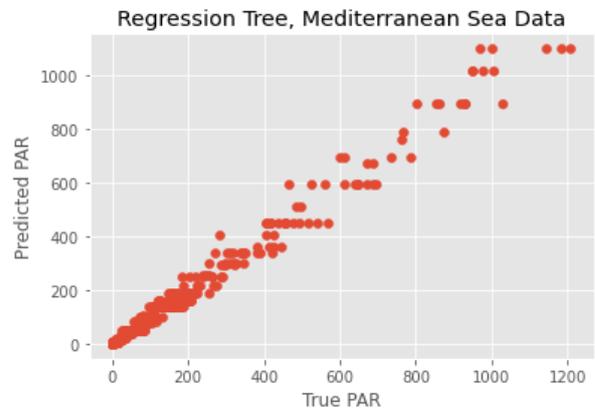


Figure 10: Regression tree on Mediterranean data

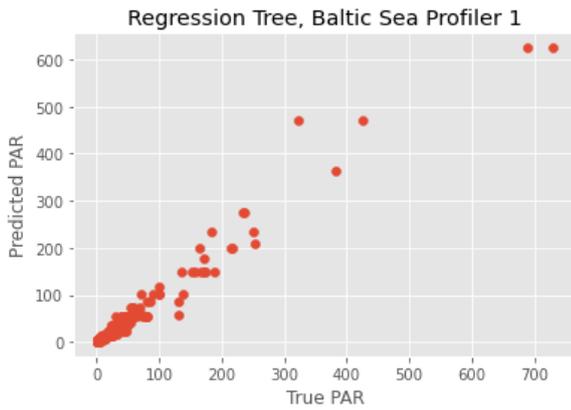


Figure 11: Regression tree on Baltic Sea float 1 data

The fit of the regression tree model on the Baltic Sea location 1 data resulted in $R^2 = 0.973$.

Figure 12 shows the result of the regression tree using the Baltic Sea dataset float location 2.

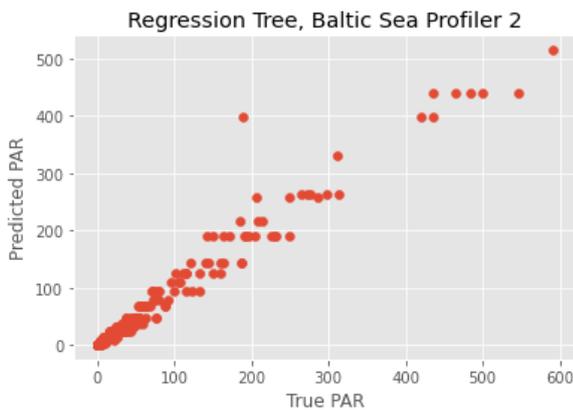


Figure 12: Regression tree on Baltic Sea float 2 data

The fit of the regression tree model on the Baltic Sea location 2 data resulted in $R^2 = 0.963$.

Figure 13 shows the result of the regression tree using the Atlantic Ocean dataset.

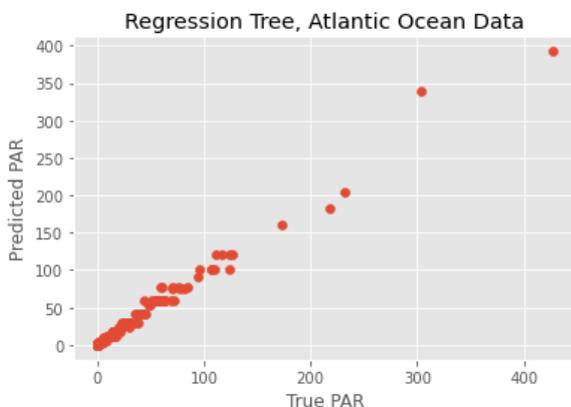


Figure 13: Regression tree on Atlantic Ocean data

The fit of the regression tree model on the Atlantic Ocean data resulted in $R^2 = 0.988$.

RESULTS AND DISCUSSION

In Table 1 the R^2 values for the different models are compared. As it can be observed, the R^2 values for the Multiple linear regression models are marginally better than those for the regression tree. It is assumed that this is caused by inherent discretization of predicted results at the leaf nodes of the regression tree.

Table 1: R^2 values for different models using Multiple linear regression (MLR) and Regression Tree (RT)

R ² values		
Dataset	MLR	RT
Combined	0.970	0.960
Mediterranean Sea	0.997	0.989
Baltic Sea Float 1	0.981	0.973
Baltic Sea Float 2	0.983	0.963
Atlantic Ocean	0.996	0.988

Furthermore, it can be seen that for both Multiple linear regression and regression trees based models the R^2 values for the combined datasets are slightly worse than models tailored for individual locations. This can be explained by influences of environmental parameters, for example salinity, which are different at the different location. These parameters were not available as input parameters for the models. However, the models are accurate enough to predict the PAR parameter without the need for a dedicated PAR sensor. Thus, PAR can be replaced by a specific wavelength enabling recording of more spectral information.

The next steps in this research is to use non-linear machine learning methods in order to increase the accuracy further.

REFERENCES

1. Roemmich D., Alford M.H., Claustre H., Johnson K., King B., Moum J., Oke P., Owens W.B., Pouliquen S., Purkey S., Scanderbeg M., Suga T., Wijffels S., Zilberman N., Bakker D., Baringer M., Belbeoch M., Bittig H.C., Boss E., Calil P., Carse F., Carval T., Chai F., Conchubhair D.Ó., d'Ortenzio F., Dall'Olmo G., Desbruyeres D., Fennel K., Fer I., Ferrari R., Forget G., Freeland H., Fujiki T., Gehlen M., Greenan B., Hallberg R., Hibiya T., Hosoda S., Jayne S., Jochum M., Johnson G.C., Kang K., Kolodziejczyk N., Körtzinger A., Le Traon P.-Y., Lenn Y.-D., Maze G., Mork K.A., Morris T., Nagai T., Nash J., Naveira Garabato A., Olsen A., Pattabhi R.R., Prakash S., Riser S., Schmechtig C., Schmid C., Shroyer E., Sterl A., Sutton P., Talley L., Tanhua T., Thierry V., Thomalla S., Toole J., Troisi A., Trull T.W., Turton J., Velez-Belchi P.J., Walczowski W., Wang H., Wanninkhof R., Waterhouse A.F.,

- Waterman S., Watson A., Wilson C., Wong A.P.S., Xu J., Yasuda I. (2019) On the Future of Argo: A Global, Full-Depth, Multi-Disciplinary Array, *Frontiers in Marine Science*, Vol. 6, Article 439.
2. Sloyan, B. M., Roughan, M., Hill, K. (2018) Global Ocean Observing System, *New Frontiers in Operational Oceanography*, 75-89.
 3. Claustre H., Bernard S., Berthon J, Bishop J., Boss E., Coatanoan C., D'Ortenzio F., Johnson K., Lotliker A., Ulloa O. (2011) Bio-Optical Sensors on Argo Floats, In: Claustre, H. (ed.) *Reports and Monographs of the International Ocean-Colour Coordinating Group*, Dartmouth, Canada, p. 1-89, JRC67902.
 4. Jiang Y., Gou Y., Zhang T., Wang K., Hu C. (2017) A machine learning approach to argo data analysis in a thermocline, *Sensors*, Vol. 17, No. 10, Article 2225.
 5. Jemai A., Wollschläger J., Voß D., Zielinski O. (2021) Radiometry on Argo Floats: From the Multispectral State-of-the-Art on the Step to Hyperspectral Technology, *Frontiers in Marine Science*, Vol. 8, Article 676537.
 6. Claustre H., Johnson K. S., Takeshita Y. (2020) Observing the global ocean with biogeochemical-Argo. *Annual Review of Marine Science*, 12, 23–48.
 7. SATLANTIC (2013) Operation manual for the OCR-504, SATLANTIC Operation Manual SAT-DN-00034, Rev. G, p 66.
 8. Organelli E., Leymarie E., Zielinski O., Uitz J., D'Ortenzio F., Claustre H. (2021) Hyperspectral radiometry on Biogeochemical-Argo floats: A bright perspective for phytoplankton diversity, in: *Frontiers in Ocean Observing: Documenting Ecosystems, Understanding Environmental Changes, Forecasting Hazards*, Kappel E.S., Juniper S.K., Seeyave S., Smith E., Visbeck M. (eds), *A Supplement to Oceanography*, Vol. 34, No. 4.
 9. Freedman D.A. (2009) *Statistical models: theory and practice*, Cambridge University Press.
 10. Breiman L., Friedman J.H., Olshen R.A., Stone, C.J. (2017) *Classification and regression trees*, Routledge.
 11. Nembrini S., König I.R., Wright M.N. (2018) The revival of the Gini importance?, *Bioinformatics*, Vol. 34, Issue 21, pp 3711-3718.

AUTHOR BIOGRAPHIES

FREDERIC STAHL is Senior Researcher at the German Research Center for Artificial Intelligence (DFKI). He has been working in the field of Data Mining for more than 15 years. His particular research interests are in (i) developing scalable algorithms for building adaptive models for real-time streaming data and (ii) developing scalable parallel Data Mining algorithms and

workflows for Big Data applications. In previous appointments Frederic worked as Associate Professor at the University of Reading, UK, as Lecturer at Bournemouth University, UK and as Senior Research Associate at the University of Portsmouth, UK. He obtained his PhD in 2010 from the University of Portsmouth, UK and has published over 65 articles in peer-reviewed conferences and journals.

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. He is also affiliated with the Marine Perception research group of the German Research Centre for Artificial Intelligence (DFKI). His main research interests are AI and computational optimisation methods for real-world scientific and engineering applications.

AHLEM JEMAI is a PhD candidate at *Carl von Ossietzky University of Oldenburg*, Germany. She is working on the “Spectral Argo-N” and the “Deep Argo 2025” projects that are utilising the Argo technology. The projects are focused on the assessment of hyperspectral light conditions in oceanic and coastal waters through ocean color remote sensing and bio-optical models. She is also a co-worker in the project “Meteor Fjord Flux” on Expedition Meteor 179.

OLIVER ZIELINSKI is head of the research group “Marine Sensor Systems” at the Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University of Oldenburg. Since 2019 he is also heading the research department Marine Perception at the German Research Center for Artificial Intelligence (DFKI). After receiving his Ph.D. degree in Physics in 1999 from University of Oldenburg, he moved to industry where he became scientific director and CEO of “Optimare Group,” an international supplier of marine sensor systems. In 2005, he was appointed Professor at the University of Applied Science in Bremerhaven, Germany. In 2007, he became Director of the Institute for Marine Resources (IMARE). He returned to the Carl von Ossietzky University of Oldenburg in 2011. His area of research covers marine optics and marine physics, with a special focus on coastal systems, smart sensors, and operational observatories involving different user groups and stakeholders.

Taking randomness for granted: the complexities of applying random number streams in simulation modelling

Maximilian Selmair

Tesla Manufacturing Brandenburg SE
15537 Grünheide (Mark), Germany
mselmair@tesla.com

Abstract—Uncertainty, as a constant companion of our world, is one major reason why simulation modelling takes precedence over static calculations to achieve accurate predictions. Computational random number generators are able to algorithmically determine values on the basis of random distributions, which utilise seed values to calculate streams of random numbers. This deterministic approach to replicating seemingly non-deterministic numbers ensures stochastic models to be reproducible at any time – one of the major requirements of simulation models. However, there are some pitfalls in the application of random number streams in modelling and simulation, which may even mislead experienced developers. In addition to a general introduction of the history of random number generators, this article shares empirical considerations and means by which the utilisation of random number streams can be improved to deliver valid and reliable results.

Keywords—Randomness; Modelling and Simulation; Seed Values; Random Number Generators

I. INTRODUCTION AND THE HISTORICAL PERSPECTIVE

The digital replication of any system or process involving randomness requires a method to generate or obtain numbers that are both random and reproducible. Practical examples for random occurrences in the field of production and logistics are service times, interarrival times and maintenance occurrences. This article focuses on the history of random number generators as well as the potential drawbacks of utilising deterministic random number generators for simulation modelling. The first paragraph introduces how random values can be generated efficiently from a predetermined probability distribution in order to provide data sets for simulation modelling.

The modest beginnings of generating random numbers date back to over a century ago. (Hull and Dobell, 1962; Dudewicz and Dalal, 1971; Morgan, 1984). The earliest samples were not generated by computer, but literally carried out by hand: flipping coins, throwing dice, dealing out game cards or the lottery draws. Even today, most lotteries are still operated in this manner to avoid fraud allegations. As early as the 20th century, gamblers were joined by statisticians in their quest to explore random numbers and mechanised devices were built to generate random numbers more efficiently. Particular examples from the late 1930s are e.g. Kendall and Smith (1938) – the use of a spinning disk to select values from a

turntable containing a hundred thousand random digits. Only two years later, electric circuits based on randomly pulsating vacuum tubes were used to deliver random digits at much higher rates of up to 50 per second. Such a device, referred to as a random number machine, was used by the British General Post Office to pick the winners of the Premium Savings Bond lottery (Thomson, 1959). Electronic devices were also used by the Rand Corporation (2001) to produce a sequence of a million random numbers. Some past examples of different approaches were picking numbers randomly out of phone books and using digits in an expansion of π , such as 0.1415926535, 0.8979323846, 0.2643383279, etc. (Tu and Fischbach, 2005). Another notable example of retrieving random numbers was proposed by Yoshizawa et al. (1999), who described a physics-based random number generator that relied on the radioactive decay of the nuclide Americium-241.

As computers and, later on, simulation modelling became more relevant, computational random number generators began to gain popularity. A first attempt in this direction was the use of the previously mentioned Rand Corporation's table in a computer's memory (Rand Corporation, 2001). This solution depended on substantial memory requirements and a vast amount of time to retrieve new values. Research in the 1940s and 1950s developed towards more deterministic strategies of generating random numbers. These values were sequential, which means that each new number was determined by one or sometimes several of its predecessors according to a fixed algorithm. The first known deterministic generator was proposed by von Neumann (1951). The well-known mid-square method is demonstrated in the subsequent example.

A researcher may begin with a four-digit positive integer $Z_0 = 1234$ and square it to obtain an integer with at least eight digits. The central four digits of this eight-digit number constitutes the next four-digit number, Z_1 . By placing a decimal point on the left of Z_1 , the first random number between 0 and 1 (U_1) is yielded. Following this procedure, an unlimited sequence of deterministic "random" values can be created. Table I lists the first few examples.

At first sight, the mid-square method seems to provide a seemingly suitable set of random numbers. However, a main disadvantage of this method is the strong tendency of the

i	Z_i	U_i	Z_i^2
0	1234	—	01522756
1	5227	0.5227	27321529
2	3215	0.3215	10336225
3	3362	0.3362	11303044
4	3030	0.3030	09180900
5	1809	0.1809	03272481

Table I: Sample calculation for a set of random numbers with the mid-square method based on value 1234

generated values to converge to zero.

A more fundamental objection to the mid-square method is that it does not yield "random" values at all, if one considers random values to be unpredictable in nature. That is, if we know one number, we have to acknowledge that it determines the succeeding numbers as the method stipulates. In more recent times, the first number in a sequence or stream of random numbers is referred to as *seed value*. With a given Z_0 , the whole sequence of Z_i 's and U_i 's is determined. This characteristic applies to all deterministic generators. These deterministic random generators are often faulted to be generating pseudo-random numbers. In the author's opinion, pseudo-random is a misnomer, indeed it is an oft-quoted remark by von Neumann (1951), who declared that "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin. For, as has been pointed out several times, there is no such thing as a random number – there are only methods to produce random numbers, and a strict arithmetic procedure of course is not such a method... We are here dealing with mere "cooking recipes" for making digits..." (von Neumann, 1951).

Although this quote dates back to over 70 years ago, it does not seem to have lost its relevance today. Furthermore, it is rarely stated that von Neumann proceeds in the same paragraph that these "recipes" "probably can not be justified, but should merely be judged by their results. Some statistical study of the digits generated by a given recipe should be made, but exhaustive tests are impractical. If the digits work well on one problem, they seem usually to be successful with others of the same type" (von Neumann, 1951). Lehmer (1951) offered a more practice-oriented definition: "A random sequence is a vague notion embodying the idea of a sequence in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests traditional with statisticians and depending somewhat on the use to which the sequence is to be put" (Lehmer, 1951).

Since Lehmer's proposal, further formal definitions and empirical considerations about randomness were formulated in the subsequent decades:

- "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin. For, as has been pointed out several times, there is no such thing as a random number – there are only methods to produce random numbers, and a strict arithmetic procedure of course is not such a method" von Neumann (1951).

- "[A] sequence is random if it has every property that is shared by all infinite sequences of independent samples of random variables from the uniform distribution" Franklin (1963).
- "[...] random numbers should not be generated with a method chosen at random. Some theory should be used" Knuth (1968).
- "The generation of random numbers is too important to be left to chance" Coveyou (1969).
- "[In statistics] you have the fact that the concepts are not very clean. The idea of probability, of randomness, is not a clean mathematical idea. You cannot produce random numbers mathematically. They can only be produced by things like tossing dice or spinning a roulette wheel. With a formula, any formula, the number you get would be predictable and therefore not random. So as a statistician you have to rely on some conception of a world where things happen in some way at random, a conception which mathematicians don't have" LeCam (1988).
- "Sequences of random numbers also inevitably display certain regularities. [...] The trouble is, just as no real die, coin, or roulette wheel is ever likely to be perfectly fair, no numerical recipe produces truly random numbers. The mere existence of a formula suggests some sort of predictability or pattern" Peterson (1998).
- "The practical definitions of randomness – a sequence is random by virtue of how many and which statistical tests it satisfies and a sequence is random by virtue of the length of the algorithm necessary to describe it [...]" Bennett (1999).

The author agrees with most researchers that deterministic generators, if appropriately designed, can replicate numbers which would have been generated by independent draws from the $U(0,1)$ distribution and would also pass a series of statistical tests. In summary, in the domain of simulation modelling, deterministic random number generators are used to replicate stochastic occurrences and measure their influence on a system. These generators allow simulation modellers to achieve reproducible scenarios with an unlimited set of seemingly non-determined values. In simulation modelling, a replication is defined as having the same set of parameters, but different stochastic influences. Each replication is based on a predetermined *seed value* (e. g. 1, 2, 3, ...), defined as a specific reproducible stream of random numbers. The purpose of this article is to address the complexities of simulation modelling with random number streams, which are associated with the risk of generating skewed, unreliable and invalid results.

II. MODELLING PITFALLS

In simulation modelling, the term *iteration* refers to different combinations of parameter values. A single iteration is composed of a series of *replications*, their number determined by a sensitivity analysis, which differ in terms of their internal seed of randomness. Models that do not contain any internal stochastic are referred to as *deterministic*; however, these lie outside the focal point of this article. In *stochastic* modelling,

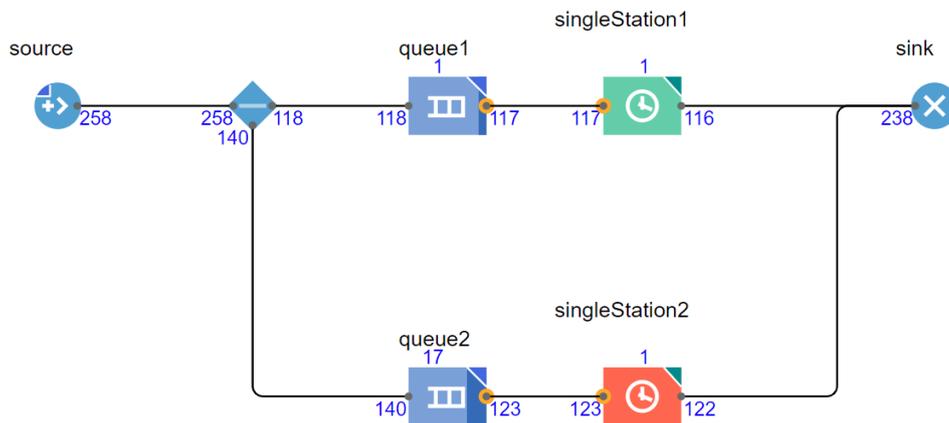


Figure 1: Examined process flow modeled with *AnyLogic*

a set of replications is computed in order to ascertain the influence of internal randomness on the final results. This methodology is referred to as the *Monte Carlo Experiment* (Shonkwiler and Mendivil, 2009).

In order to be able to compare two iterations with each other, it is vital that both are subject to the the same random stochastic behaviour. These can be e.g. machine maintenance occurrences, interarrival times of supply or interfering influences. If such random influences differ between iterations, the results, for instance differences in a Key Performance Indicator (KPI) of interest, may well be caused by the actual internal randomness. As such, the risk of attributing a result to a certain combination of parameters is considerable. Additionally, the adjustment of parameters (e. g. a different rule set or policy) may not correlate with any improvement of a KPI. The following conditions have been identified as potential causes of deviation in random behaviour in two or more iterations with the same seed of randomness:

- 1) When decisions in a model are stipulated by parameters, and a particular decision causes consumers of randomness, that is, every level of every variable, to require random numbers in a different sequence, e. g. job shop machine scheduling models
- 2) When parameters affect the number of consumers of randomness, e. g. Automated Guided Vehicle (AGV) fleet with an initially randomly distributed battery level
- 3) When parameters manipulate the model's initialisation sequence, e. g. data-driven models vs. applied random distributions

The following sections are intended to illustrate how these conditions can lead to different stochastic behaviour, even if the same seed of randomness is utilised. An example is provided to illustrate each item.

The first example presented is a basic process flow consisting of a source, two queues with subsequent servers and a sink, see Figure 1. The model's parameters were set as follows:

- Source, interarrival time:

30 seconds, exponentially distributed

- Single stations 1 & 2
process time: 1.5 - 2.5 minutes, uniformly distributed
mean time between failures: 60 hours, exp. distributed
mean time to repair: 15 hours, based on the Erlang distribution with $n = 2$
- There are two policies that decide whether an agent, for instance a customer, chooses the first or the second server. Policy 1 is a random-only decision with a 50:50 chance. With Policy 2, arriving agents will always choose the shorter queue. If both queues are of the same length, a random 50:50 chance is calculated.

If Policy 1 were to yield a higher throughput than the "smarter" Policy 2, it is deemed self-evident that the results were skewed by the difference in stochastic behaviour, caused by any one or combination of the above mentioned conditions. To examine this scenario, a Monte Carlo simulation trial run was performed on both iterations (Policy 1 and 2), each with 10^5 replications. The number of replications was determined by the results of a sensitivity analysis carried out at an earlier point. A different seed value was allocated to each replication, which provides a stream of random numbers for the consumers source, policy and both servers. Each simulated iteration covers a duration of 24 actual hours.

The expected results were that the policy which considers the length of the queues generated a greater throughput than expected. In order to ascertain that this is the case in every replication, a further analysis showed that in 14.8% of replications, Policy 1 (random queue) lead to a higher throughput than Policy 2 (shortest queue). In order to investigate this further, the number of server maintenance occurrences was analysed and the findings showed that, in some cases, two iterations with the same random seed and stream were correlated with different numbers of maintenance occurrences per server. This seems to indicate that we are, in fact, comparing not only two different policies, but also two different number consumption patterns. Even though these differing patterns

are only obvious in the 14.8% of the cases, it is suggested that this methodological shortcoming affects all cases. It has a certain influence on all the other scenarios, but in these cases the impact is less substantial and not noticeable when one only regards the number of maintenance occurrences. However, an assessment of the down-time duration is likely to detect stochastic differences in all replications. Focusing on those 14.8% of cases where the throughput of Policy 1 (random queue) is higher, it is notable that the number of maintenance occurrences is lower than when applying Policy 2 with the same seed value.

Why does the randomness differ between two iterations with the same random seed? As mentioned above, the mid-square method generates random numbers in such a manner that each random number is based on the preceding one. Therefore, the sequence of numbers is considered to be predetermined. In the process flow described above, there are several consumers of random numbers. Depending on the chosen policy, all the consumers of random numbers (source, policy, servers) retrieve a random number whenever the stochastic influence necessitates it. For instance, Policy 1 (random queue) consumes a random number every time an agent needs to decide whether it joins Queue 1 or 2. In contrast, Policy 2 (shortest queue) only consumes a random number when both queues are of the same length, which only rarely occurs. Here, the sequence of random number consumers differs substantially and this leads to a different pattern of maintenance occurrences and process times at all subsequent processes.

The question that arises from these deliberations is how can researchers achieve the same stochastic influence for two different parameter settings, in this case iterations with the same random seed value. The solution appears to lie within the grasp of the researcher, who can separate all random occurrences within the model that are actually independent of each other. In the presented scenario, this refers to the interarrival time of the source, the random decision of the applied policy, the process times of both servers and their maintenance occurrences. More specifically, each random number consumer needs to retrieve a new random number from its own stream. If this is the case, the sequence of consumption will no longer influence the generated random numbers for the other consumers.

Keeping random streams separate from each other is also proposed to solve the issue addressed in condition 2) where parameters affect the consumed random numbers during the initialisation phase of a simulation model. A logistics scenario considering the number of AGVs that are involved in a specific material flow system lends itself to illustrate the case in point. If the AGV initially retrieves a random number to establish its initial battery level, the sequence is directly influenced by the number of AGVs. Consequently, this leads to very different random behaviour from the very beginning of the simulation. This unwanted deviation can be prevented by either allocating one specific stream of random numbers for each AGV or one stream that only provides the random initial battery levels.

Point 3 directly relates to the topic of model initialisation. Here, the issue arises when a model is capable of replicating

both, evaluating actual historical data or applied random distributions. Both usually result in a different sequence in the initialisation phase and therefore in a different sequence of random numbers.

The conditions described apply intrinsically to simulation modelling in general, regardless of the software used. Depending on the software utilised for modelling, the described pitfalls must be tackled differently. *FlexSim*, for example, offers a total of 100 random streams by default. They can be allocated by only using digits from 1 to 100 as the respective parameters in any distribution function. Beyond that, *FlexSim* allows the user to create a new stream by using the command `getstream(current)`, where `current` refers to the individual consumer of random numbers whom a single stream is dedicated to commencing as of the first retrieval. In contrast, there is an unlimited number of seed values available in *AnyLogic*, which can be used to initialise an unlimited number of Java objects by using, for instance, `Random r = new Random(1);` for a seed value of 1. These objects can be used as a parameter in any distribution function instead of the default random generator of the model. *PlantSimulation*, on the other hand, designates an individual stream for each consumer automatically. Despite the advanced settings and suggestions that *PlantSimulation* offers, the developer requires a firm understanding of the design of their simulation environment and the requirements to maintain an empirically sound simulation when utilising random numbers. In summary, while not explicitly naming all common simulation tools, they all offer functions to maintain control over random streams.

III. CONCLUSION

Based on this author's previous reviews of the pertinent literature in this domain, it is suggested that few simulation studies tend to the exact assessment of random occurrences from one replication to the another. Modellers frequently rely on the software's default random generator, as its main purpose is to simply provide random numbers. Yet, depending on the design of the simulation study, default settings may not suffice when empirically reliable results require a more thorough approach.

In this article, a brief historical review of random number generators was provided and some substantial pitfalls in their application to simulation modelling were highlighted. In this particular context, it may appear as if some modellers do not separate their random streams as their design may require it. This oversight can lead to considerable differences in the behaviour of random number consumers as well as the results of the simulation, even if the same initial seed value is used. As such, assuming the same stochastic behaviour may lead to imprecise results and their comparison does not lead to reliable conclusions. However, for many cases this difference of the model's stochastics may go unnoticed or perhaps be attributed to the parameter settings. Finally, a number of suggestions were made to remedy these procedural shortcomings of replicating randomness.

REFERENCES

- Bennett, D. (1999), *Randomness*, Harvard University Press, ISBN: 978-0674107465.
- Coveyou, R.R. (1969), 'Random number generation is too important to be left to chance', *Applied Probability and Monte Carlo Methods and modern aspects of dynamics*, pp. 70–111.
- Dudewicz, E.J. and Dalal, S.R. (1971), 'Allocation of observations in ranking and selection with unequal variances', *Optimizing Methods in Statistics*, Elsevier, pp. 471–474, ISBN: 9780126045505.
- Franklin, J.N. (1963), *Mathematics of Computation* (8), pp. 28–59.
- Hull, T.E. and Dobell, A.R. (1962), 'Random Number Generators', *SIAM Review* 4 (3), pp. 230–254, ISSN: 0036-1445.
- Kendall, M.G. and Smith, B.B. (1938), 'Randomness and Random Sampling Numbers', *Journal of the Royal Statistical Society* 101 (1), p. 147, ISSN: 09528385.
- Knuth, D.E. (1968), *The Art of Computer Programming*, Stanford University: Addison-Wesley, ISBN: 0-201-89684-2.
- LeCam, L. (1988), 'Interview', *More Mathematical People* (8), p. 174.
- Lehmer, D.H. (1951), 'Mathematical methods in large-scale computing units', *Annu. Comput. Lab. Harvard University* 26, pp. 141–146.
- Morgan, B.J.T. (1984), *Elements of simulation*, London and New York: Chapman and Hall, 351 pp., ISBN: 978-0412245909.
- Peterson, I. (1998), *The Jungles of Randomness: A Mathematical Safari*, Penguin Books Ltd, ISBN: 978-0140271720.
- Rand Corporation (2001), *A million random digits with 100,000 normal deviates*, Santa Monica, Calif.: Rand Corporation.
- Shonkwiler, R.W. and Mendivil, F. (2009), *Explorations in Monte Carlo methods*, Undergraduate texts in mathematics, Dordrecht and Heidelberg: Springer, ISBN: 978-0-387-87837-9.
- Thomson, W.E. (1959), 'Ernie - A Mathematical and Statistical Analysis', *Journal of the Royal Statistical Society. Series A (General)* 122 (3), p. 301, ISSN: 00359238.
- Tu, S.-J. and Fischbach, E. (2005), 'A Study on the Randomness of the Digits of π ', *International Journal of Modern Physics C* (16), pp. 281–294.
- Von Neumann, J. (1951), 'Various Techniques Used in Connection with Random Digits', *Monte Carlo Method*, ed. by A.S. Householder, G.E. Forsythe and H.H. Germond, vol. 12, National Bureau of Standards Applied Mathematics Series, Washington, DC: US Government Printing Office, pp. 36–38.
- Yoshizawa, Y., Kimura, H., Inoue, H., Fujita, K., Toyama, M. and Miyatake, O. (1999), 'Physical random numbers generated by radioactivity', *Journal of the Japanese Society of Computational Statistics* 12 (1), pp. 67–81, ISSN: 0915-2350.

Finite - Discrete - Element Simulation

FEM STUDY ON THE STRENGTH INCREASING EFFECT OF NITRIDED SPUR GEARS

Jakab Molnár*
Péter T. Zwierczyk
Attila Csobán

Department of Machine and Product Design
Faculty of Mechanical Engineering
Budapest University of Technology and Economics
1111, Műegyetem rkp. 3, Budapest, Hungary
E-mail: molnar.jakab@gt3.bme.hu
*Corresponding author

KEYWORDS

finite element method, nitriding, DANTE heat treatment software, spur gears, surface endurance limit

ABSTRACT

In this research, the strength increasing effect of nitriding was investigated on a small series of spur gears with a module of 1 mm. The nitriding process was simulated in ANSYS FEM software with DANTE heat treatment external add-on. Only a simplified transient heat treatment model was created to study the basics and the effects of nitriding on spur gears. As a result, the nitrogen distributions, the hardness of the gear tooth flanks, the residual stresses and deformations were calculated with DANTE. Approximate surface endurance limit could be calculated analytically for the analysed spur gears according to the local endurance limit formula of Kloos and Velten. As the surface endurance limit value increased because of the nitriding process, the calculated allowable torque of the driving gear also increased. Even though the higher allowable torque and the initial residual stress increased the contact stress of the gear pairs, the calculated contact stress remains acceptable.

INTRODUCTION

Heat treatment is a controlled, predefined process of temperature variation to provide the required material properties (strength, toughness, etc.) by deliberately changing the microstructure, the initial stress state, and the mechanical, physical, or chemical properties of the finished part, without modifying its geometry (Csizmazia 2003). Properly modifying the microstructure and the material's properties can increase strength and fatigue resistance, leading to longer service life. Since it's not always required to perform heat treatment overall to the whole part, depending on the application of loads, heat treatments will be carried out only on critical surfaces of the particular section the part (Cserjésné et al. 2015). Nitriding is one of the most commonly used surface hardening processes (local heat treatment) for low-alloy steel gears, where only on the critical surface layers (gear flanks) will nitrogen be diffused to create a hardened, wear-resistant surface. Since nitriding modifies the

material properties of the gear working surface (gear flanks), it's essential to analyse how nitriding affects the working conditions of the meshing gears.

The fast and continuous development of finite element software and modules allows us to analyse more and more specific engineering subjects. A general engineering problem requires several different types of calculations (structural, thermal, electromagnetic, etc.) to be carried out. The results of different types of calculations have an impact on each other, that's why it's necessary to establish a coupling between each calculation (the input of stress calculations could be a thermal calculation). Coupled FEM analysis provides an excellent tool to study complex engineering problems, so several factors that were simplified or neglected previously can be reconsidered.

In today's state-of-the-art finite element method, there is limited ability to validate realistic heat treatment parameters in FEM gear simulations. The influence of real tooth surface hardness on the tooth surface contact stress during gear mesh is significant and is essential for building an accurate finite element gear analysis. The exact nature of tooth surface hardness (which depends on given gear parameters) can be investigated using coupled finite element analysis. The hardness profile created during nitriding can be used to analytically calculate the surface endurance limit for a specified nitrided surface case layer. In order to accurately determine the contact stress during the meshing of the gears, it is necessary to know the exact tooth surface endurance stress values that will be developed during the heat treatment.

This research studied the gear nitriding capabilities of the US-developed DANTE finite element-based heat treatment software. The main objective was to calculate the actual surface hardness and endurance limit for predefined spur gears with FEM. In light of the findings, the allowable torque of the driving gear could be calculated, and a new study was created for analysing the effect of residual stress on the contact stress of the meshing spur gears.

METHOD

DANTE is a standalone finite element-based heat treatment software, but it's also available as a software add-on for modern finite element software. DANTE version 5-1b with ACT2-7 was used with ANSYS 2020 R2 environment during the analyses. A small series of gear pairs was studied during the research, with the following gear parameters.

Table 1. The main geometrical parameters of the analysed gear pairs

z1	i	z2	d _{w1}	d _{w2}	a _w	b
[-]	[-]	[-]	[mm]	[mm]	[mm]	[mm]
17	1	17	17	17	17	10.2
17	4	68	17	68	42.5	10.2
17	6	102	17	102	59.5	10.2
30	1	30	30	30	30	18
30	4	120	30	120	75	18
30	6	180	30	180	105	18
40	1	40	40	40	40	24
40	4	160	40	160	100	24
40	6	240	40	240	140	24

Only spur gears were analysed with the module of 1 mm. The gear pairs have zero-backlash and addendum modifications. The CAD models of the analysed spur gear pairs were exported from KISSsoft design software. In DANTE, the nitriding process of one gear was investigated at a time, so for each gear, a complete set of results was available. Because of the high computational time of the analyses, only a 0.1 mm slice of one gear tooth could be analysed per gear pair.

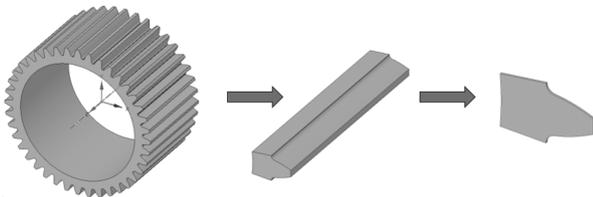


Figure 1. Simplified gear geometry for the nitriding analysis

The geometry preparation process for 2D contact stress analysis was the same as it was discussed in depth in a previous paper (Molnár et al. 2021). The applied material for the analysed gears was 42CrMo4 (AISI 4140). During the nitriding analysis, the material was directly selected from the DANTE material database. Since the contact stress analysis was performed in a standalone ANSYS study (independent from DANTE), the material properties were imported as follows:

Table 2. Material properties used for 42CrMo4 in analytical calculations and FEM contact analysis

Material property	Value of
Base fatigue limit [MPa]	785
Ultimate Strength [MPa]	950
Young's Modulus [MPa]	210 000
Poisson's ratio [-]	0.3

The residual stress results were imported as initial stress load from DANTE nitriding results for the contact analysis.

Finite Element Model

First, the nitriding process was studied with DANTE, and after that, in light of the results, a contact analysis was created to study the effect of nitriding on contact stress.

The DANTE-based nitriding process is built up from a series of coupled simulations, so the results of each sub-simulation significantly impact the results of the other simulations, making it crucial to accurately define the simulation parameters. The nitridation simulation can be divided into the following sub-simulation steps: 0. Geometry and FEM model preparation I. Nitriding model setup II. Thermal model setup III. Stress model setup. The order of the sub-simulations is not interchangeable, they are strongly dependent on each other. The relationship between each sub-simulation can be seen in the following figure.

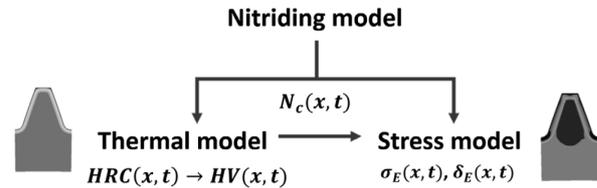


Figure 2. Process of nitriding in DANTE

The nitriding and thermal models require transient thermal simulations, while the stress model requires static mechanical simulation. In the nitriding model, the nitrogen distribution (history) will be calculated, and this result is used as an input for both the thermal and stress models. The hardness distribution and the thermal load for the stress model will be calculated in the thermal model. As a final step, the thermal history will be imported as an external load to the stress model, so can the residual stress and deformation be calculated. After the nitriding analysis, a 2D contact stress analysis was created to study the effect of residual stress and increased torque on the contact stress of the spur gears.

Preparation, the initial model

The initial step of a coupled simulation is to create the base model containing the prepared gear geometry, the FEM mesh, and base analysis settings. With the help of the initial simulation model, it can be ensured that the calculation is carried out with the same settings and geometry for each sub-simulation. As mentioned above, only a 0.1 mm-thick slice of one gear teeth could be analysed for each gear. The initial model provided the finite element mesh used in every sub-simulation for each study. According to the specifications of DANTE, only first-order linear elements can be used. The global element size was chosen as 0.1 mm equal to the gear slice thickness. During the nitriding simulation, it's crucial to properly register the amount of nitrogen diffusion, the thermal and phase transformation and stress gradients close to the surface, so a very fine mesh is required close

to the gear flanks. The created nitrided layer is divided into a few-micrometre-thick white layer and a larger diffusion zone, the finite element mesh. Since the extension of the diffusion layer is larger than the white layer, it's not required to use very fine mesh in the entire nitrided depth region. The growth rate was defined to gradually increase the element size of the mesh in the diffusion region, the maximum thickness of the refinement region was equal to the total expected case depth. The nitriding case depth can be approximate according to the research of Gustav Niemann (G. Niemann et al. 1965), in the case of spur gears with a module of 1 mm, the maximum nitriding case depth is near 0.2 mm. Since the FEM calculated case depth may differ from the recommended value, the refinement region depth was set to 0.4 mm. This study used 0.001 mm (1 μm) element size near the surface, and coarser elements were used away from the fine surface. An example of the used FE mesh structure can be seen in Figure 3.

Table 3. Main parameters of the FEM mesh

Property	Value
Global element size [mm]	0.1
Element size at the gear flank [mm]	0.001
Element size near case depth [mm]	0.05
Refinement depth [mm]	0.5
Number of nodes [-]	5998
Number of elements[-]	2902

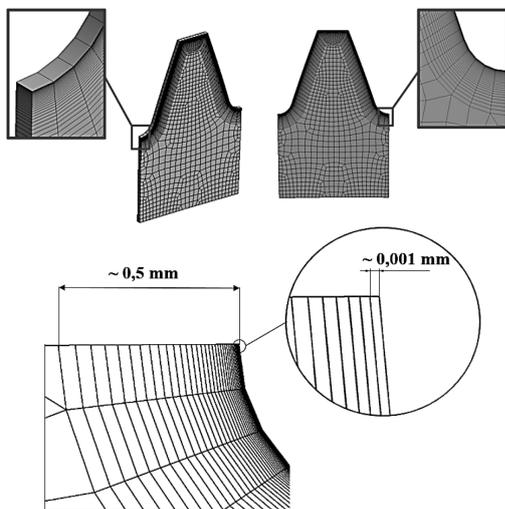


Figure 3. FE mesh of the gear slices ($z=240$)

Nitriding Model

In a DANTE nitridation model, a transient thermal analysis modelled the diffusion of nitrogen. This model only considers effective nitriding without pre-heating or cooling. Single-stage nitriding was defined with a constant temperature of 525 $^{\circ}\text{C}$ (Cserjésné et al. 2015). The effective nitriding time (processing steps) was uniformly predicted to be 28 880 s (8 hours) based on the research of Gustav Niemann (G. Niemann et al. 1965)

and our test runs and results for the case depth thickness. Multiple nitriding time was investigated, the results of the test runs will be detailed in the next chapter. The film coefficient was set to $0.001 \frac{W}{\text{mm}^2 \cdot ^{\circ}\text{C}}$. The other DANTE specific nitriding settings were left as default values. For thermal boundary conditions, the gear teeth flank surface was selected as a convection surface. The output of the nitriding model was the nitrogen distribution (history) in the gear body, which was used as an input for both the thermal and stress models.

Thermal Model

Based on the effective nitriding history, the thermal model retrieves the hardness distribution (profile) after the full heat-treatment process. In the thermal model (transient thermal analysis), not only the single-stage effective nitriding is considered, but also pre-heating and cooling during the heat treatment process, so the initial heat load for the stress model can be calculated. The complete nitriding process (pre-heating, nitriding, cooling) has to be considered in separate simulation steps with different time values. Since the effective nitriding process has already been investigated in the nitriding model, and the resulting nitrogen distribution data were imported, the duration of the effective nitriding was taken to be symbolically 1 s (DANTE recommendation). The duration of pre-heating and cooling was taken to be equal to 1800 s (0.5 hours), based on the test runs. In this case the initial temperature for the transient thermal simulation was selected to normal 20 $^{\circ}\text{C}$ room temperature. The other DANTE specific nitriding settings were left as default values. Based on the calculation, it was possible to retrieve the hardness distribution after heat treatment, and also to calculate the allowable torque for the driving gear. The output of the thermal model was the temperature distribution (history), which serves as the initial heat load input parameter for the stress model.

Stress Model

In the stress model (static mechanical stimulation), both the nitrogen distribution and the temperature history were required from the previous sub-simulations to determine the residual stress and deformation after the nitriding process. The calculated temperature history was imported and considered as an initial external load in the stress simulation. No other load was present during this simulation. The initial temperature for the static mechanical stimulation was selected to normal 20 $^{\circ}\text{C}$ room temperature. The other DANTE specific nitriding settings were the same as in the thermal model.

The static mechanical simulation required that a sufficient number of degrees of freedom of the geometry be constrained to run the static calculation successfully. Because of the cyclic symmetry of the gear teeth, frictionless contact was applied on the side slice surfaces of the gear body. Another boundary condition was the fixation of the bottom side of the gear body's face surface so as to take into account the support effect of the missing

part of the gear body. The displacement of the top face surface was not fixed in the normal direction (axial displacement) because thermal expansion of the gear body is not restricted in that direction. The boundary conditions of the gear slice are shown in the following figure.

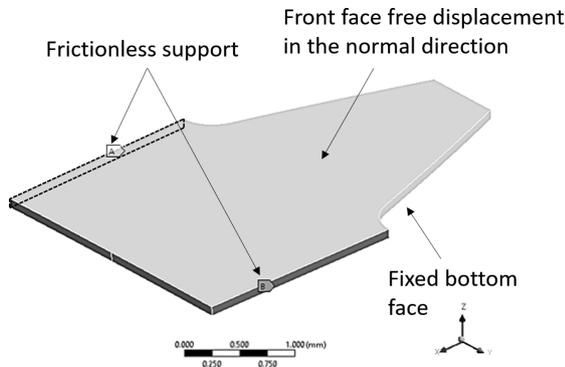


Figure 4. Boundary conditions for stress model

The output of the stress model was the residual stress and residual deformation of the gear body. The contact analysis model used the residual stress tensor as an input stress state.

Contact Analysis Model

In the contact analysis model, the effect of the residual stress with increased driving gear torque was analysed on the surface contact stress. A static mechanical model was created with linear elastic mechanical properties. In the present study, the analysed spur gears had a gear width - operating pitch diameter ratio ($\frac{b}{d_w}$) of 0.6 (disc-type components), that's why 2D plane stress was considered. The corresponding elements of the residual stress tensor were imported to the 2D analysis as an initial stress state. The surface endurance limit and the applied driving gear torque were calculated from the thermal model's surface hardness results. The driven gear was fixed in each case, and the driving gear was loaded with the calculated torque. The model settings, boundary conditions and FEM mesh were the same as it was detailed (see above) in our previous paper (Molnár et al. 2021). The boundary conditions of the simulation can be seen in Figure 5.

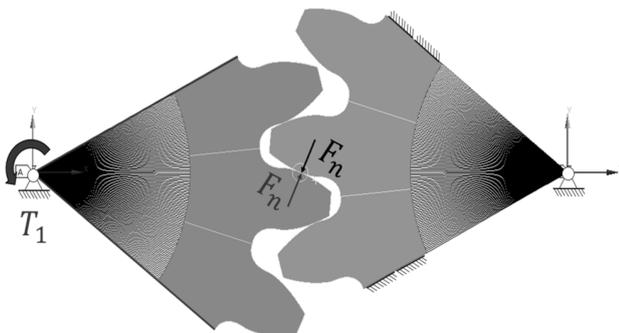


Figure 5. Boundary conditions for contact analysis

RESULTS

In this chapter, the calculated parameters will be discussed. The result of the nitriding model was the nitrogen distribution (history) of the effective nitriding process, which was an input for both the thermal and stress analyses. In the thermal analysis, the hardness distribution (profile) was calculated. An example can be seen for the hardness distribution after the heat treatment process in the following figure.

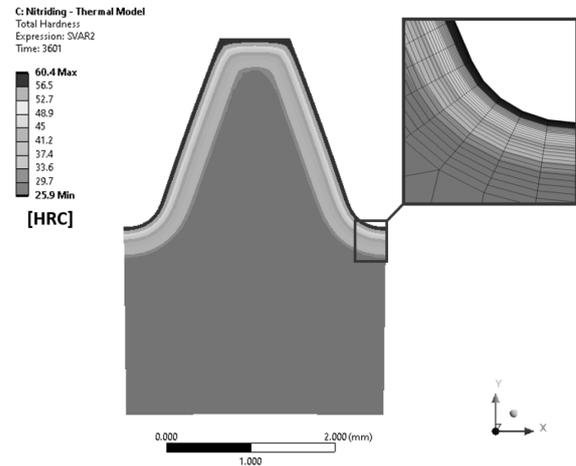


Figure 6. Visualization of hardness distribution after nitriding ($z=240$)

The hardness values were calculated as Rockwell's (HRC) hardness values. Figure 6. confirms that the simulated hardness distribution is reasonable for the thermal model. A uniform hardness distribution was achieved, the profile of which corresponds to the curvature of the gear tooth surface. The hardness value was maximal at the gear tooth surface, and it decreased continuously far from the gear tooth surface, there was no abrupt change in the hardness values. The resulting distribution pattern is in accordance with the hardness profiles found in the literature (M. A. Terres et al. 2017). The average surface hardness was 60 HRC for the analysed spur gears, the difference was negligible. The nitriding time and case depth were determined based on different hardness profiles. The hardness distribution and case depth for the different test durations are shown in the following figure.

Development of tooth flank hardness

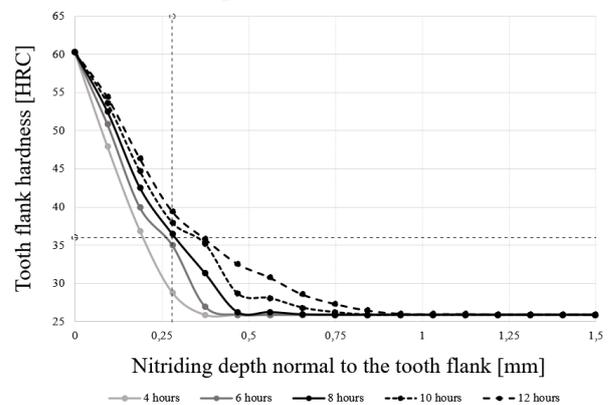


Figure 7. Development of tooth flank hardness for different nitriding time ($z=17$)

The nitriding time is the duration needed for the required case depth to form below the tooth surface. The required case depth can also be approximated from the core hardness. Approximately the effective case depth is given at a value of 10 HRC from the core hardness. As it was mentioned above, in the case of spur gears with a module of 1 mm, the maximum nitriding case is near 0.2 mm, according to Gustav Niemann (G. Niemann et al. 1965). The nitriding time was tested with 2-hour increments in the time range of 4, 6, 8, 10, 12 hours respectively. As shown in Figure 7, the nitriding time of 8 hours meets both the depth and hardness conditions for the required case depth, the optimal nitriding time is 8 hours.

The residual stress and deformation can be calculated based on the previously calculated nitriding and thermal (load) history in the stress model. The calculation generated a residual stress tensor, which contains the residual stress state after the heat treatment process. This tensor can be used as an input initial stress state in further analyses. An example can be seen in the next two pictures for the distribution and development of the residual stress tensor's Normal-Z component.

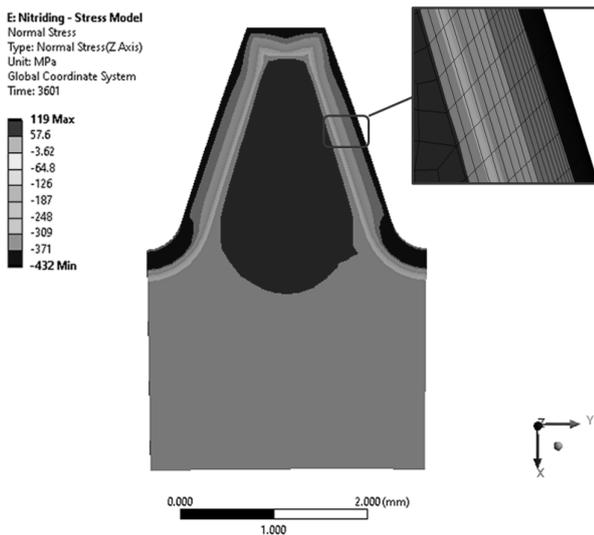


Figure 8. Normal-Z component of the residual stress tensor (z=240)

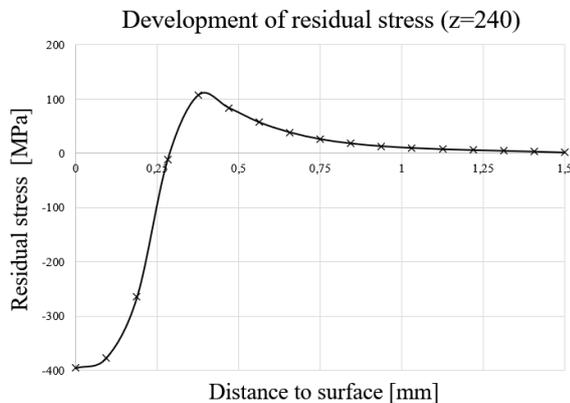


Figure 9. Development of the residual stress tensor's Normal-Z component (z=240)

Figure 8-9. shows that residual compressive stress is generated near the gear flank surface, and tensile stress is generated in the core below the gear tooth surface. The residual compressive stress follows the curvature of the gear tooth flank, and its maximum value is located at the close subsurface layers of the gear tooth surface, where nitrogen concentration and hardness were maximal. The residual stress was uniformly distributed along with the elements, there was no sudden change in the stress value. The average surface residual compressive stress was 400 MPa for the analysed spur gears, the difference was negligible. The characteristics and the values of the residual stress are in good correspondence with X-ray diffraction measures in the literature (M. A. Terres et al. 2019). Because of the compressive characteristic of the residual stress, it can be predicted that the residual compressive stress present in the nitrided case will prevent crack propagation, however, further studies would be necessary to prove this hypothesis.

An example of the residual deformation after the nitriding process can be seen in the following figure.

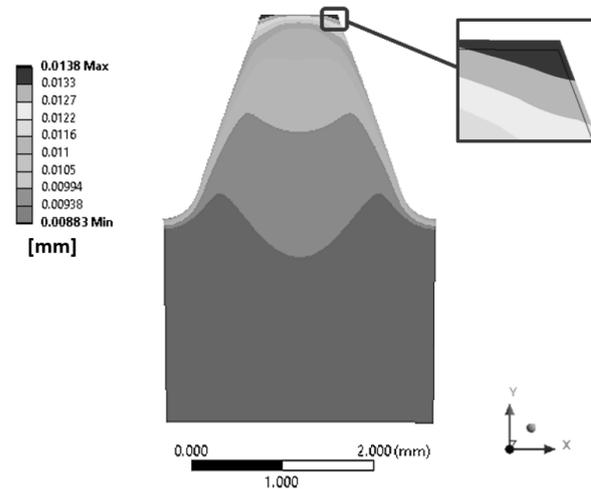


Figure 10. Distribution of the residual deformation after nitriding process (z=240)

As shown in Figure 10, the residual deformation was negligible, but the curvature of the gear flank influenced it. The maximal deformation value was 0.014 mm for spur gear with gear teeth of 240. These results confirm the hypothesis that deformations after nitriding are negligible (Cserjésné et al. 2015). In the case of pre-heat treatment gear modifications, it is recommended to consider the deformation value before prescribing the modification parameters.

Based on the average tooth surface hardness and residual compressive stress values, the local fatigue limit of the tooth surface could be calculated. The local surface endurance limit can be estimated according to the Kloos and Velten (K.H. Kloos et al. 1984) formula:

$$\sigma_{Hlim} \approx \sigma_{Wo} \left(1 - \frac{\sigma_m + \sigma_{res}}{R_m} \right) \left(1 + \sqrt{\frac{1600}{HV^2} \cdot \chi^*} \right) \quad (1)$$

Where σ_{Hlim} is the local surface endurance limit, σ_{W0} is the base fatigue limit, σ_m is the mean applied stress, σ_{res} is the residual stress after heat treatment, R_m is ultimate tensile strength, HV is the average hardness of the tooth flank, χ is the applied relative stress gradient factor (reciprocal of the gear width). In this study, fully reversed loading was considered (zero mean stress), and only the hardness and residual stress parameters were used from the results of DANTE nitriding, the base endurance limit and ultimate strength were present as initial parameters before nitriding, they were not recalculated. The hardness value in Rockwell was converted to Vickers hardness. The difference of the calculated endurance limits for each gear was negligible, therefore a global surface endurance limit was considered. The calculated global surface endurance limit was multiplied by reducing factors from DIN 3990-2. The calculated value of the endurance limit for contact stress is 1065 MPa (for spur gear with a module of 1 mm). The result shows good correspondence with the values found in the literature (DIN 3990-5 1987).

After calculating the global surface endurance limit of the analysed spur gears, the allowable torque could be analytically calculated for the driving gears, according to Gy. Erney (Gy. Erney 1983):

$$T_1 \leq \frac{a_w^3 \cdot \sigma_{Hlim}^2 \cdot \sin(\alpha) \cdot \cos(\alpha) \cdot \frac{b}{d_w} \cdot u}{(u + 1)^4 \cdot K_A \cdot C_B \cdot E \cdot 87,5} \quad (2)$$

Where: T_1 is the calculated torque of the driving gear, a_w is the centre distance, α is the pressure angle (in this study 20°), $\frac{b}{d_w}$ is the gear width – working pitch diameter factor, u is the gear teeth ratio, K_A is the service factor, C_B is a stress factor, E is Young's modulus of the analysed gear material. The calculated torques were used as the load of the driving gear in the contact analysis, the calculated torques can be found in Table 4.

Table 4. Calculated driving gear torques

z_1 [-]	i [-]	z_2 [-]	d_{w1} [mm]	d_{w2} [mm]	T_1 [Nmm]
17	1	17	17	17	2000
17	4	68	17	68	3100
17	6	102	17	102	3400
30	1	30	30	30	10700
30	4	120	30	120	17000
30	6	180	30	180	18200
40	1	40	40	40	25200
40	4	160	40	160	40300
40	6	240	40	240	43200

In the 2D contact analysis, both the calculated torque and the corresponding elements of the residual stress tensor were taken into account. The residual stress created an initial stress state for the analysis, and the increased torque condition assumed the nitrided state of the spur gear with linear elastic properties. Since the contact stress generally is a compressive stress type, the minimum principal stress was analysed during the studies.

The following figure shows an example of the resulting contact stress distribution.

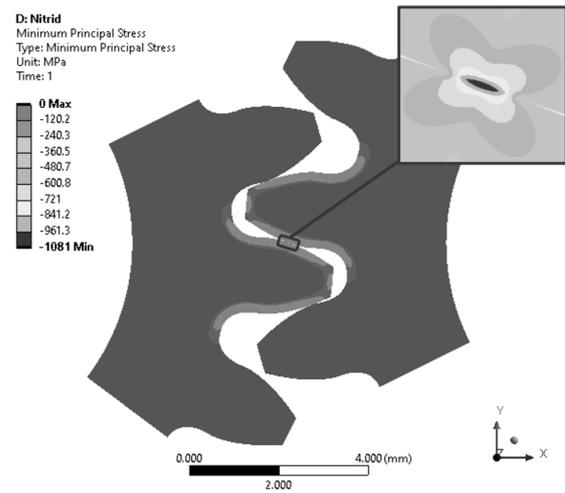


Figure 12. Contact stress distribution with imported residual stress ($z=17$)

In this model, numerically calculated contact stress values (without considering the residual stress) provide sufficiently accurate results with the analytically calculated Hertzian contact stress values, as it was demonstrated in our previous paper (Molnár et al. 2021). Considering the residual stress, the calculation gave the value of contact stress as the sum of the theoretical contact stress and the residual compressive stress, because the gear teeth were already in a pre-stressed initial compressive state before the engagement. The maximal principal stress was achieved. The maximum contact stress value was achieved during the analysis of the gear pairs with the smallest number of gear teeth ($z=17$), the peak value of the contact stress was 1080 MPa, as it is shown in Figure 12. The figure shows that the contact distribution is slightly distorted in the contact point, but its character is still Hertzian. It can be established that the calculated maximum contact stress (1080 MPa) reaches the calculated endurance limit (1065 MPa). The results show that although the higher allowable torque and the initial residual stress increased the contact stress of the gear pairs, the calculated contact stress remains acceptable.

Summary

Increased pinion torque can be calculated for nitrided spur gears, based on the increased hardness and tooth surface fatigue limit. The optimum nitriding case depth provides excellent operating properties, deeper case was not necessary for this study. It turned out that the number of teeth has no effect on the nitriding case depth or the nitriding time, but the module had a significant influence on both nitriding time and nitriding depth. After the nitriding process, residual stress and deformation were present, which influenced the operating condition of the spur gears. The residual deformation was negational. In the case of pre-heat treatment tip relief modification, the effect of residual deformation should be considered when specifying the tip relief parameters. The residual

compressive stress present in the nitrided case will prevent crack propagation, however, further studies would be necessary to prove this hypothesis. The results show that although the higher allowable torque and the initial residual stress increased the contact stress of the gear pairs, the calculated contact stress remains acceptable. According to my experiment with the heat treatment module of DANTE, I can establish that the software provides an excellent tool for simulating heat treatment of spur gears.

DISCUSSION

The presented study was limited only to cylindrical spur with zero modification and backlash. Neglections and simplifications were made in the construction process of the studies. Because of the high computational time of the heat treatment analyses, only a 0.1 mm slice of one gear tooth could be studied per gear pair in the nitriding analyses. In the future, we aim to compare our results from the FEM heat treatment model with laboratory test results.

ACKNOWLEDGMENTS

Hereby, the authors would like to thank DANTE Solutions Inc. and Econ Engineering Ltd. for making this study possible.

The research reported in this paper and carried out at BME has been supported by the NRDI Fund (TKP2020 NC, Grant No. BME-NCS) based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology

REFERENCES

- K.H. Kloos et al., 1984. „Berechnung der Dauerschwingfestigkeit von plasmanitrierten bauteilähnlichen Proben unter Berücksichtigung des Härte- und Eigenspannungsverlaufs“, *Konstruktion* 36 (5): 181
- F. Csizmazia, 2003. “Hőkezelés“, Széchenyi István Egyetem, Győr
- S. Á. Cserjésné et al., 2015. “Nitridálás – korszerű eljárások és vizsgálati módszerek”, Miskolci Egyetem, Miskolc
- J. Molnár et al., 2021. “Analysis Of Tip Relief Profiles For Involute Spur Gears”, *ECMS*, Volume 35, Issue 1
- G. Niemann et al., 1965. “Maschinenelemente”, Band 2, Springer, Verlag, 1965

- M. A. Terres et al., 2017. “Low Cycle Fatigue Behaviour of Nitrided Layer of 42CrMo4 Steel”, *International Journal of Materials Science and Applications*, 6(1): 18-27
- M. A. Terres et al., 2019. “Experimental and analytical study of residual stresses relaxation in nitrided 42CrMo4 parts”, *Materialwissenschaft und Werkstofftechnik*. 50: 844-855.
- DIN 3990-5:1987-12, Tragfähigkeitsberechnung von Stirnrädern; Dauerfestigkeitswerte und Werkstoffqualitäten.
- Gy. Erney, 1983., “Fogaskerekek.”, Műszaki Könyvkiadó, Budapest

AUTHOR BIOGRAPHIES



JAKAB MOLNÁR is a technical assistant at Budapest University of Technology and Economics, Department of Machine and Product Design, where he received his M.Sc. degree in 2022. His primary interest is in spur and helical gears, the development of gearboxes, and FEM. His e-mail address is: molnar.jakab@gt3.bme.hu and his webpage can be found at: <http://gt3.bme.hu>



PÉTER T. ZWIERCZYK is an assistant professor at Budapest University of Technology and Economics Department of Machine and Product Design, where he received his M.Sc. degree and then completed his Ph.D. in mechanical engineering. His main research field is the railway wheel-rail connection. He is a member of the finite element modelling (FEM) research group. His email address is: z.peter@gt3.bme.hu, and his webpage can be found at: <http://gt3.bme.hu>



ATTILA CSOBÁN is an assistant professor at Budapest University of Technology and Economics. Member of the Association of Hungarian Inventors since 2000. Member of the Entrepreneurship Council of the Hungarian Research Student Association since 2006. Member of the Hungarian Academy of Sciences (MTA) public body since 2012. Gold level member of the European Who is Who Association since 2013. Research field: gear drives, gearboxes, planetary gear drives, cycloidal drives. His email address is: csoban.attila@gt3.bme.hu, and his webpage can be found at: <http://gt3.bme.hu>

APPLICATION OF THE FINITE ELEMENT METHOD TO DETERMINE THE VELOCITY PROFILE IN AN OPEN CHANNEL

Daria Wotzka

Faculty of Electrical Engineering, Automatic Control and Informatics

Opole University of Technology

ul. Prószkowska 76

45-758 Opole, Poland

E-mail: d.wotzka@po.edu.pl

KEYWORDS

Finite element methods, laminar flow, turbulent flow, open channel

ABSTRACT

This paper contains the results of work on the simulation of laminar and turbulent flows using the finite element method. In particular, exemplary literature references are indicated, boundary and initial conditions are described, and numerical results are illustrated, including fluid velocity distributions and profiles in a cylindrical open channel structure.

INTRODUCTION

Urban and rural development, caused by a constant increase in the number of inhabitants, results in an increase in the amount of sewage which flows into the central sanitary sewer infrastructure. While the construction of new housing developments is associated with the simultaneous construction of sanitary sewers with an appropriately sized volume, in some inner-city areas it is not physically or economically feasible to extend existing sewers, usually open channels, which are currently operating often at the limit of their capacity.

Improving the monitoring system of sanitary networks is now an important industry to assess its hydraulic performance. The main problem faced by water and sewerage companies concerns the systematic collection of, among others, the volume and velocity of sewage flow in open sewers (Synowiecka *et al.*, 2014) and the thickness of sewer sludge in overflow collectors (Kalinowski, 2016). Lack of supervision over the proper operation of the sewage disposal system may cause leakage, resulting in a potential threat to life and health of the region's inhabitants and environmental contamination. The access to the current parameters of the system operation enables the estimation of the load in particular areas and the detection of undesirable phenomena, such as the occurrence of ponding and flaring in the canal or exceeding of maximum fills (Synowiecka *et al.*, 2014).

One of the areas of current scientific and research work is the construction of simulation models of

sanitary systems, which can be used to predict the behaviour of networks under different operating conditions. The values of sewage system operating parameters, which are recorded on a regular basis, constitute an important and necessary element of the process of calibration and validation of mathematical models.

The phenomenon of fluid flow in sewers has been described in detail to some extent in the literature. Equations and mathematical models characterizing the flow through channels of different shapes, such as trapezoid, rectangle or cylinder, have been indicated (Chow, 1988; Jobson and Froehlich, 1988; Sturm, 2001; Basu, 2019a, 2019b). However, sewerage systems, especially in large agglomerations are diverse, consisting of channels of different sizes and shapes with numerous crossing points. In Poland, many water and sewerage companies do not yet have metering systems for wastewater or rainwater infrastructure. Measurement of linear velocity of flow in successive network sections is still a technological challenge and an important element of development of existing sewage infrastructure. The problem in question arises from the necessity of installing the measuring device usually in very polluted canals, in a possibly non-invasive way, not causing any disturbances in the proper flow of waste water.

The main objective of the conducted scientific and research work is to develop a device for measuring the volume flow rate in the sewage system, which has an implementation potential (Wotzka, 2022). In particular, the device should meet the condition of possibly low invasiveness in the normal flow of sewage in an open sewer. An additional constraint is the condition of low production cost of the developed device. It has been proposed that ultrasonic acoustic signals be used to determine the velocity using a cross-correlation method. The cross-correlation method allows the calculation of the linear velocity of particles moving inside an open channel. To calculate the flow rate, it is necessary to know the area A through which the fluid flows. In order to determine the velocity profile in the different layers of a cylindrical open channel, a computer model was developed and a series of simulations were performed as part of the research

work. The model was made in COMSOL Multiphysics environment using the CFD module, which uses algorithms based on the finite element method (FEM) to approximate solutions of partial differential equations.

The scope of the conducted research included the analysis of the influence of the following model parameters on the obtained velocity distributions:

- fluid level $H=\{0.05; 0.09; 0.12\}$ m in the channel,
- channel length $L=\{1; 10\}$ m,
- object input velocity $v=\{0.2; 0.3\}$ m/s.

In the following sections of the paper the numerical method used and the simulation results obtained are presented.

NUMERICAL METHOD IN OPEN CHANNEL FLOW STUDIES

Numerical methods are now an important alternative to experimental studies due, among other things, to the high-powered computing computers available today and a number of software solutions that offer a relatively affordable application of the CFD method (Ramanathan, 2013; Lai and Kuowei Wu, 2019).

One of the areas of application of numerical methods is the study of changes in the geometry of river currents occurring during floods, particularly near bridges (Adhikary, Majumdar and Kostic, 2009) and studies of sediment transport modeling (Nations, no date; *Hydrology and Sediment Transport.*, 2010; Lai and Kuowei Wu, 2019). Another area of research is work related to hydraulic structures, which includes channel widening which provides a transition from a narrow to a relatively wide channel cross-section (Najmeddin, 2012). At the transition point, the flow tends to separate from the spreading sidewalls and forms turbulent vortices when the divergence angle exceeds a threshold value. This phenomenon can cause unwanted flow energy loss and sidewall erosion. The author of (Najmeddin, 2012) presented the results of a study on adjusting the lift in the vertical plane to eliminate flow separation. He used CFD modeling in this study, which allowed a systematic investigation of the effects of different divergence angles, lift heights and Froude number for subcritical flow. The modeling results were validated using analytical solutions under simplified conditions and available experimental data for a limited number of cases.

Another issue investigated is the occurrence of arcs in both natural and artificial open channels. Due to the change in pressure and centrifugal force values, as well as the interaction between these two forces, strong secondary flows are generated in the bends of open channels, which in turn cause a full three-dimensional complex fluid flow in the area where the channel bends. Many numerical studies have been conducted to model the characteristics of open channel bends. For example, the author of (Gholami *et al.*, 2015) used CFD methods to study the flow

depth and velocity field in acute angle bends, including considering two phases water and air.

On the other hand, in the paper (Gandhi, Verma and Abraham, 2010) the authors used CFD method to study the flow velocity profile in rectangular open channels for determining the optimal number and location of flow sensors.

BOUNDARY AND INITIAL CONDITIONS IN THE NUMERICAL MODEL

The test object is a cylinder filled with liquid up to a certain height H , flowing along the object with a constant average velocity v , determined at the surface of the inlet opening. In the considered modeling task using the FEM method, the Navier-Stokes equations are solved under laminar (1-2) and turbulent (2-5), incompressible fluid flow conditions. The dependent variables in the model are the field components of velocity u, v, w and pressure p . In the task, a constant temperature $T=293.15$ K and zero initial conditions: $u=(0,0,0)$ m/s, $p=0$ Pa were assumed.

$$\rho(\mathbf{u} \cdot \nabla)\mathbf{u} = \nabla \cdot [-p\mathbf{I} + \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)] + \mathbf{F}, \quad (1)$$

where: \mathbf{u} – velocity vector, ρ – density, p – pressure, μ – dynamic viscosity, \mathbf{F} – volumetric force vector, \mathbf{I} – intensity vector.

$$\rho\nabla \cdot (\mathbf{u}) = 0, \quad (2)$$

$$\rho(\mathbf{u} \cdot \nabla)\mathbf{u} = \nabla \cdot [-p\mathbf{I} + (\mu + \mu_T)(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)] + \mathbf{F}, \quad (3)$$

where: μ_T – dynamic viscosity of the turbulent model $k-\epsilon$.

$$\rho(\mathbf{u} \cdot \nabla)k = \nabla \cdot \left[\left(\mu + \frac{\mu_T}{\sigma_k} \right) \nabla k \right] + p_k - \rho\epsilon, \quad (4)$$

where: $k=5.202 \cdot 10^{-7}$ m²/s³, kinetic energy, $\epsilon=2.412 \cdot 10^{-9}$ m²/s³ dispersion coefficient, $\sigma_k=1$ turbulent model parameter, p_k - A component of the turbulent kinetic energy source.

$$\rho(\mathbf{u} \cdot \nabla)\epsilon = \nabla \cdot \left[\left(\mu + \frac{\mu_T}{\sigma_\epsilon} \right) \nabla \epsilon \right] + c_{\epsilon 1} \frac{\epsilon}{k} p_k - c_{\epsilon 2} \rho \frac{\epsilon^2}{k}, \quad (5)$$

where: $\sigma_\epsilon=1.3$, $c_{\epsilon 1}=1.44$, $c_{\epsilon 2}=1.92$ are the parameters of the turbulent model.

$$\mu_T = c_\mu \rho \frac{k^2}{\epsilon}, \quad (6)$$

$$p_k = \mu_T [\nabla\mathbf{u} : \nabla\mathbf{u} + (\nabla\mathbf{u})^T], \quad (7)$$

where: $c_\mu=0.09$ is a turbulent model parameter.

Fig. 1-3 depict a visualization of the object along with the boundary conditions B1-B4. The following boundary conditions were assumed in the task.

Boundary condition B1 - non-slip fixed wall
Laminar flow is calculated assuming $\mathbf{u} = 0$.
Turbulent flow is calculated with (8-11).

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad (8)$$

$$[(\mu + \mu_T)(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)]\mathbf{n} = \rho \frac{u_\tau}{u^+} \mathbf{u}_{\text{tang}}, \quad (9)$$

$$\mathbf{u}_{\text{tang}} = \mathbf{u} - (\mathbf{u} \cdot \mathbf{n})\mathbf{n}, \quad (10)$$

$$\nabla k \cdot \mathbf{n} = 0, \epsilon = \rho \frac{c_\mu k^2}{\kappa_\nu \delta_w^+ \mu}, \quad (11)$$

where: \mathbf{n} – normal vector, u_τ – tangential speed, u^+ – friction speed, $\kappa_\nu = 0.41$, $\delta_w^+ = 0.2$ are the parameters of the turbulent model.

The wall condition applies to fluid flow with stationary walls. No-slip is the default boundary condition for modeling solid walls. A no-slip wall is one in which the fluid velocity relative to the wall velocity is zero.

Boundary condition B2 - open edge with normal stress $f_0 = 0 \text{ N/m}^2$. Laminar flow is calculated with (12). Turbulent flow is calculated with (13-14).

$$[-p\mathbf{I} + \mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)]\mathbf{n} = -f_0\mathbf{n}, \quad (12)$$

$$[-p\mathbf{I} + (\mu + \mu_T)(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)]\mathbf{n} = f_0\mathbf{n}, \quad (13)$$

$$\begin{cases} \nabla k \cdot \mathbf{n} = 0, \nabla \epsilon \cdot \mathbf{n} = 0, \\ \text{if } \mathbf{u} \cdot \mathbf{n} \geq 0, \\ k = \frac{3}{2} (U_{\text{ref}} I_T)^2, \epsilon = c_\mu^{\frac{3}{4}} \frac{k^{\frac{3}{2}}}{L_T} \\ \text{if } \mathbf{u} \cdot \mathbf{n} < 0. \end{cases} \quad (14)$$

where: U_{ref} – reference speed, I_T – intensity, L_T – length are the scale parameters of the turbulent model.

The open edge condition describes the boundaries of the domain that are in contact with a large volume of fluid that can both flow into and out of the domain beyond the object region. The normal stress condition f_0 means that $f_0 \approx p$.

Boundary condition B3 - constant pressure with backflow suppression $p_0 = 0$, $\hat{p}_0 \leq p_0$. Laminar flow is calculated with (15). Turbulent flow is calculated with (16-17). This condition is applicable at boundaries for which there is outflow from the domain. In order to obtain a proper numerical solution, it is recommended to consider the inlet conditions when determining the outlet condition. For example, if velocity is specified at the inlet, pressure can be specified at the outlet and vice versa.

$$[-p\mathbf{I} + \mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)]\mathbf{n} = -\hat{p}_0\mathbf{n}, \quad (15)$$

$$[-p\mathbf{I} + (\mu + \mu_T)(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)]\mathbf{n} = -\hat{p}_0\mathbf{n}, \quad (16)$$

$$\nabla k \cdot \mathbf{n} = 0, \nabla \epsilon \cdot \mathbf{n} = 0, \quad (17)$$

Specifying the velocity vector at both the inlet and the outlet can cause convergence problems. Selecting the appropriate outlet conditions for the Navier-Stokes equations is a non-trivial task. This option determines the normal stress, which in most cases is approximately equal to the pressure. The tangential component of the stress equals zero. If the reference pressure p_{ref} is 0, then the pressure p_0 at the domain boundary is the absolute pressure. Otherwise p_0 is the relative pressure at the boundary.

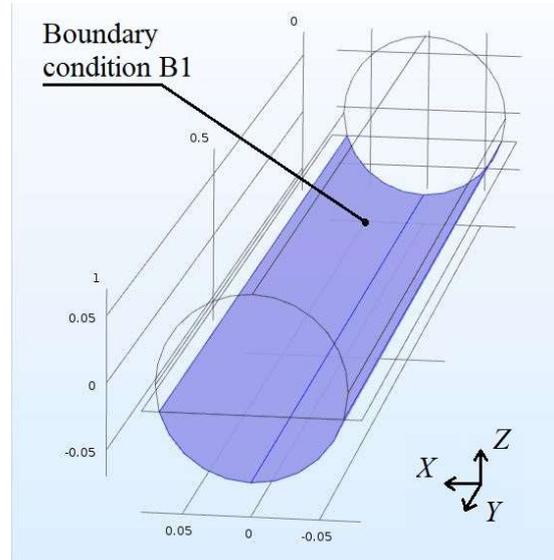


Fig. 1 Visualization of the boundary condition B1.

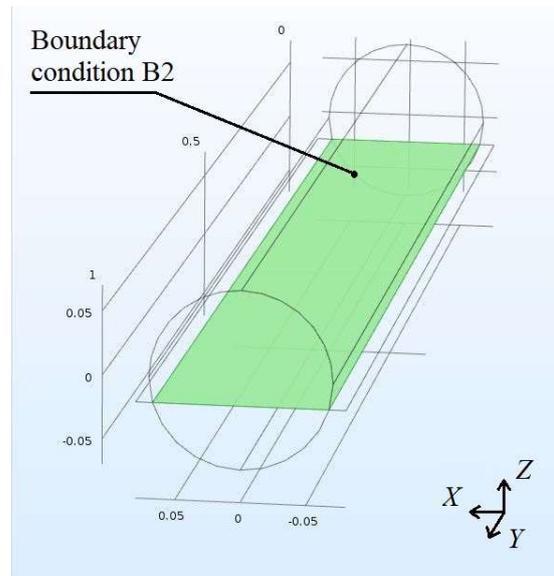


Fig. 2 Visualization of the boundary condition B2.

Boundary condition B4 - normal inlet speed $U_0=0.2$ m/s, $\mathbf{u} = -U_0 \mathbf{n}$, where \mathbf{n} is the boundary normal directed out of the area, and, a U_0 is the normal inflow velocity, $U_{\text{ref}} = U_0$, $k = (U_{\text{ref}} I_T)^2$, $\epsilon = \frac{k^{3/2}}{L_T}$.

This condition is applicable at boundaries for which there is a net flow to the interior of the domain. In addition, the following other parameters are specified in the task:

- reference temperature $T_{\text{ref}} = 293.15$ K,
- reference pressure level $p_{\text{ref}} = 1$ atm,
- absolute pressure $p_{\text{bw}} = p$ [Pa] + p_{ref} ,
- speed value U [m/s] calculated as $U = \sqrt{(u)^2 + (v)^2 + (w)^2}$, where u, v, w are the dependent variables computed by the numerical method in the domain under consideration.

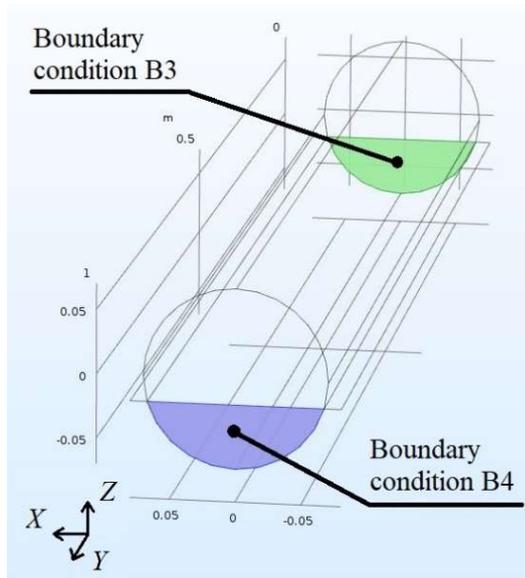


Fig. 3 Visualization of boundary conditions B3 and B4.

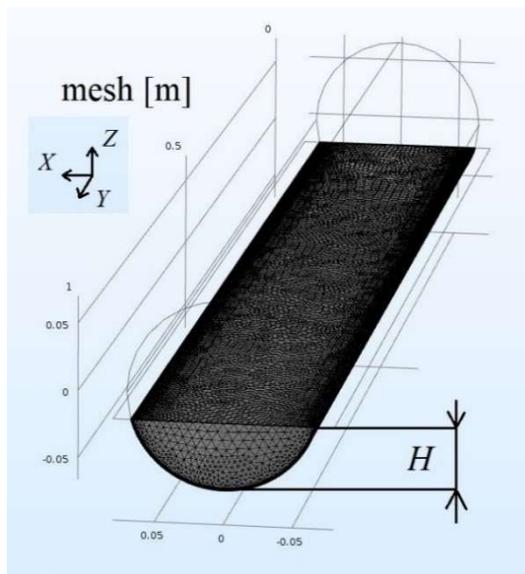


Fig. 4 Visualization of the discretization mesh used in the numerical model.

Table 1 lists the values of the discretization mesh parameters used in the numerical model. A visualization of the grid elements is shown in Fig. 4.

Table. 1 Parameters used to discretize the object into finite elements.

Description of grid statistics	Value
Minimum element quality	120.6 mm
Average component quality	690.4 mm
Maximum element size	7.9 mm
Minimum element size	1.5 mm
Maximum element growth rate	1.13

The simulation study used a Windows 10 computer with an Intel® Core™ i7-5960X CPU@3.00 GHz, with eight cores. A RAM size of 32 GB was used. The computation time of each simulation depended on the size of the object and ranged from a few minutes to several hours. The FEA algorithms used were Smoothed Aggregation AMG and PARDISO, which are implemented in the COMSOL Multiphysics environment.

RESULTS OF NUMERICAL CALCULATIONS

Fig. 5 illustrates the position of the lines parallel to the ground and with respect to the Y axis, for which the values of the fluid velocity inside the object are visualized in the following section. In particular, these are the positions with respect to the Z-axis: $Z=0$, $Z=-0.02$, $Z=-0.05$ m, which correspond to different heights for the example model, with liquid level in the channel $H=0.09$ m. Fig. 6 illustrates the position of the lines perpendicular to the bottom and relative to the Y axis. The following section visualizes the liquid velocity values calculated at the positions, which are labeled in Fig. 6 as: Location A - near the liquid outlet, Location B - center of the area, Location C - near the liquid inlet to the object. In Fig. 7-10, surface distributions of fluid velocity inside a channel of length $d=1$ m and diameter $\phi=0.15$ m are illustrated along with isotopes. Fig. 7, 9 and 10 show the results for laminar flow, Fig. 8 shows the results for turbulent flow. The flow velocity in each case was $v_{\text{wlot}}=0.2$ m/s. The liquid level for the results illustrated in Figs. 7 and 8 was $H=0.12$ m, in Fig. 9 the liquid level was $H=0.09$ m, in Fig. 10 the liquid level was $H=0.05$ m.

The lines forming the flow profile are visible in the individual images. Characteristic areas are shaped differently in the inlet area (location C), in the center (location B) as well as in the fluid outlet area from the domain (location A). Slight differences between laminar and turbulent flow can also be seen at the boundary with the wall.

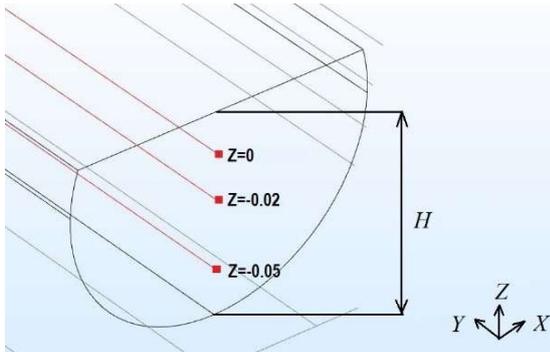


Fig. 5 Visualization of position of the lines on the Z axis parallel to the ground for which the calculation results are illustrated. Model assumes a liquid level of $H=0.09$ m.

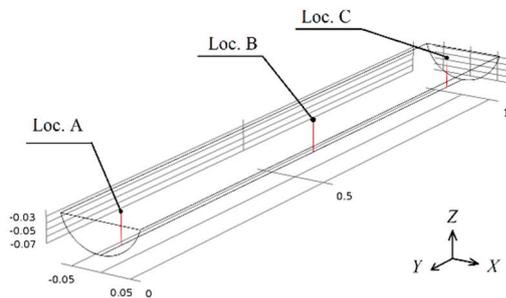


Fig. 6 Visualization of lines perpendicular to the ground of the lines on the Y-axis for which the calculation results were visualized. Liquid level of $H=0.05$ m was assumed.

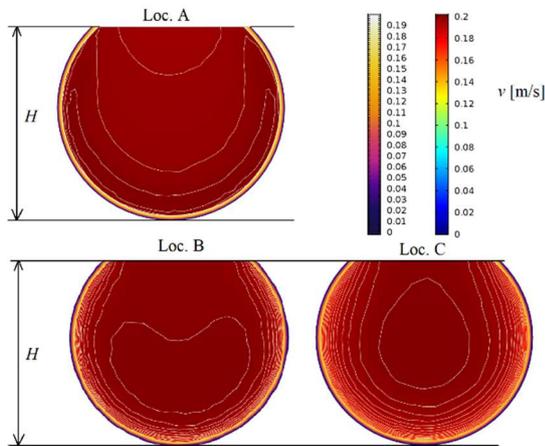


Fig. 7 Surface distributions of fluid velocities inside a channel with length $L=1$ m and diameter $\phi=0.15$ m, fluid height $H=0.12$ m and assumed laminar flow $v_{wlot}=0.2$ m/s, at locations A, B and C, as shown in Fig. 6.

Figs. 11-13 illustrate the fluid velocity distributions along a channel $\phi=0.15$ m in diameter filled with fluid of height $H=0.12$ m. Fig. 11 shows the results for a channel of length $L=10$ m, and Figs. 12 and 13, for a channel of length $L=1$ m. The presented curves illustrate the change in velocity at different positions relative to the fluid height H , denoted by Z in Fig. 5.

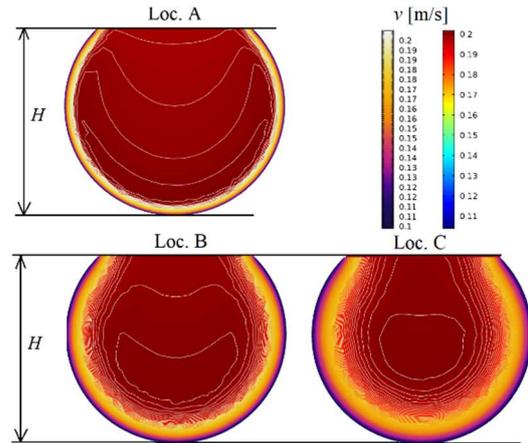


Fig. 8 Surface distributions of fluid velocities inside a channel with length $L=1$ m and diameter $\phi=0.15$ m, fluid height $H=0.12$ m and assumed turbulent flow $v_{wlot}=0.2$ m/s, at locations A, B and C, as shown in Fig. 6.

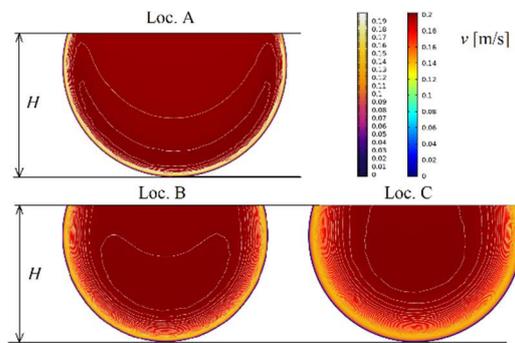


Fig. 9 Surface distributions of fluid velocities inside a channel with length $L=1$ m and diameter $\phi=0.15$ m, fluid height $H=0.09$ m and assumed laminar flow $v_{wlot}=0.2$ m/s, at locations A, B and C, as shown in Fig. 6.

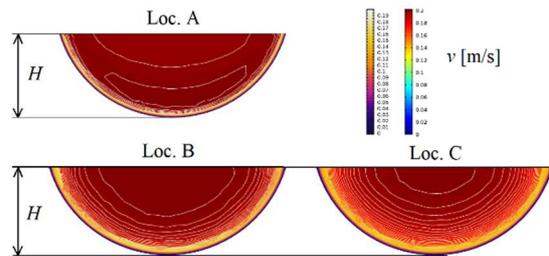


Fig. 10 Surface distributions of fluid velocities inside a channel with length $L=1$ m and diameter $\phi=0.15$ m, fluid height $H=0.05$ m and assumed laminar flow $v_{wlot}=0.2$ m/s, at locations A, B and C, as shown in Fig. 6.

The flow velocity at the model inlet was $v_{wlot}=0.2$ m/s each time. Figs. 11 and 12 show the results for laminar flow, Fig. 13 for turbulent flow. The Figures show decreases in the initial velocity, equal to $v_{wlot}=0.2$ m/s at the entrance, to 0.18 m/s at the exit for the object with $L=10$ m, to 0.195 m/s for the object with $L=1$ m and laminar flow, and to 0.193 m/s for turbulent flow.

Figs. 14-16 illustrate the velocity distributions of the fluid across the channel with a diameter $\phi=0.15$ m. The flow velocity at the model inlet was $v_{wlot}=0.2$ m/s each time. Figs. 14 and 15 show the results for a fluid-filled channel with height $H=0.12$ m, for laminar and turbulent flow, respectively. In Fig. 16, the results are illustrated for laminar flow in a liquid-filled object up to a height of $H=0.05$ m.

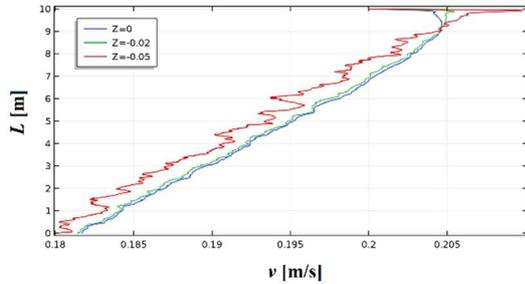


Fig. 11 Laminar flow velocity distribution along the channel of length $L=10$ m and diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to Z marked letter in Fig. 5.

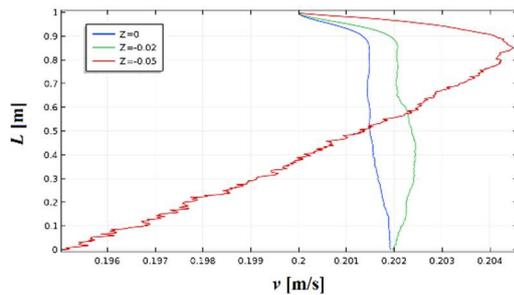


Fig. 12 Laminar flow velocity distribution along the channel with length $L=1$ m and diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to Z marked letter in Fig. 5.

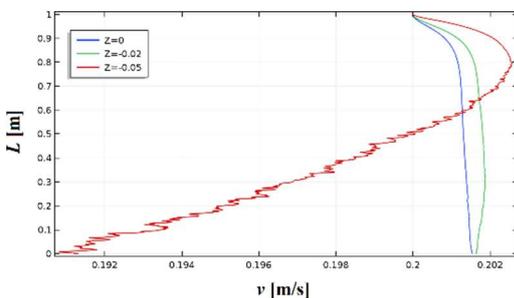


Fig. 13 Turbulent flow velocity distribution along the channel with length $L=1$ m and diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to Z marked letter in Fig. 5.

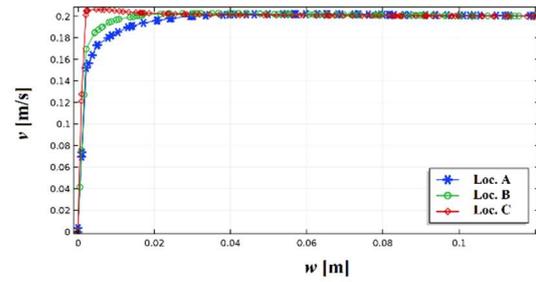


Fig. 14 Laminar flow velocity distribution across the channel of diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to A, B, C in Fig. 6.

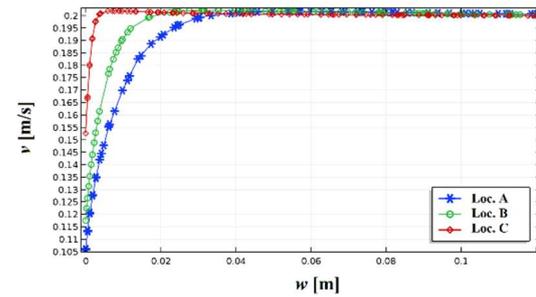


Fig. 15 Turbulent flow velocity distribution across the channel of diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to A, B, C in Fig. 6.

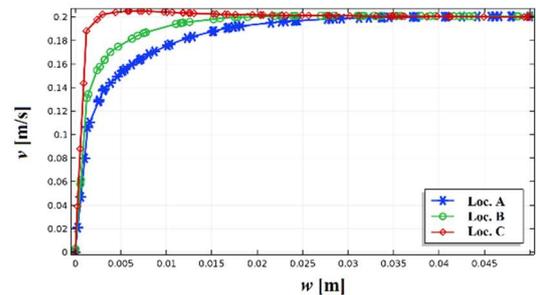


Fig. 16 Laminar flow velocity distribution across the channel of diameter $\phi=0.15$ m, water level $H=0.05$ m and assumed flow value $v_{wlot}=0.2$ m/s. The positions correspond to these marked with the A, B, C in Fig. 6.

Fig. 17 shows the velocity distribution along a channel with length $L=10$ m and diameter $\phi=0.15$ m, water level $H=0.12$ m and assumed flow rate $v_{wlot}=0.3$ m/s as a horizontal plot. The velocity line contours depict velocity gradients, irrelevant inside the channel and significant at the channel walls. Fig. 18 shows the velocity distribution along the channel with the same parameters as a slice plot.

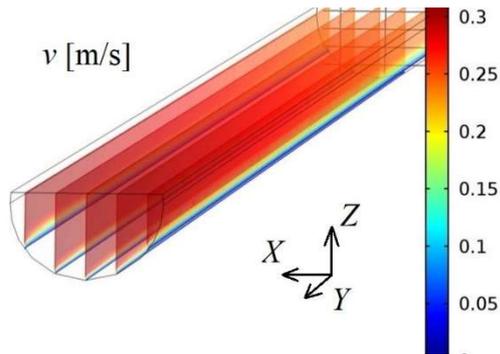


Fig. 18 Velocity distribution along the channel of length $L=10$ m, diameter $\phi=0.15$ m, water level $H=0.09$ m and assumed flow rate $v_{wlot}=0.3$ m/s as a slice plot.

CONCLUSIONS

This paper presents the results of the CFD theoretical analyses performed, which involved computer simulations using the FEM method.

As a result of the numerical calculations carried out in the simulation model, velocity profiles were determined in channels with different diameters and levels of their liquid filling and with different lengths. Moreover, the influence of the input velocity on the obtained velocity distributions in particular locations of the considered area was investigated.

Based on the analysis of the obtained relations, it was found that the individual velocity layers can be approximated by circles of increasing diameters and it was assumed that the area A can be approximated by the relation (18). With height h as the input variable, measured by the device under development, and radius r as the input variable, measured independently.

$$A = \frac{1}{2} r^2 \left(2 \cos^{-1} \left(\frac{r-h}{r} \right) - \sin \left(2 \cos^{-1} \left(\frac{r-h}{r} \right) \right) \right), \quad (18)$$

where: A – the area of a segment of a circle of radius r and height h .

The results of the discussed simulation studies confirm the theoretical assumptions made for the development of the measuring device, and equation (18) can be used to calculate the flow velocity in the measuring device mounted in an open channel.

REFERENCES

- Adhikary, B. D., Majumdar, P. and Kostic, M. (2009) 'CFD simulation of open channel flooding flows and scouring around bridge structures', in *Proceedings of the 6th WSEAS International Conference on Fluid Mechanics*, pp. 106–113.
- Basu, S. (2019a) 'Chapter V - Velocity and Force Type Flow Meter', in Basu, S. (ed.) *Plant Flow Measurement and Control Handbook*. Elsevier, pp. 395–539. doi: 10.1016/b978-0-12-812437-6.00005-6.
- Basu, S. (2019b) 'Open channel flow measurement.', in

Plant Flow Measurement and Control Handbook. Fluid, Solid, Slurry and Multiphase Flow. Chapter III, pp. 257–331. doi: 10.1016/B978-0-12-812437-6.00003-2.

- Chow, V. Te (1988) *Open Channel Hydraulics*. McGraw-Hill Book Company. doi: 10.1016/B978-0-7506-6857-6.X5000-0.
- Gandhi, B. K., Verma, H. K. and Abraham, B. (2010) 'Investigation of flow profile in open channels using CFD', in *IGHM*. IIT Roorkee, India, pp. 243–251.
- Gholami, A. *et al.* (2015) 'Simulation of open channel bend characteristics using computational fluid dynamics and artificial neural networks', *Engineering Applications of Computational Fluid Mechanics*, 9(1), pp. 355–369. doi: 10.1080/19942060.2015.1033808.
- Hydrology and Sediment Transport*. (2010). doi: 10.1007/698_2010_67.
- Jobson, H. E. and Froehlich, D. C. (1988) *Basic hydraulic principles of open-channel flow, U.S. GEOLOGICAL SURVEY, Report 88-707*.
- Kalinowski, M. (2016) 'Problemy Monitoringu Przepływu Ścieków I Miąższości Osadów W Przelazowych Kolektorach', *Journal of Civil Engineering, Environment and Architecture*, 63(22), pp. 149–164. doi: 10.7862/rb.2016.156.
- Lai, Y. G. and Kuowei Wu (2019) 'A three-dimensional flow and sediment transport model for free-surface open channel flows on unstructured flexible meshes', *Fluids*, 18(4), pp. 1–19. doi: 10.3390/fluids4010018.
- Najmeddin, S. (2012) *CFD Modelling of Turbulent Flow in Open-Channel Expansions*. Concordia University.
- Nations, F. and A. O. of the U. (no date) *Data Requirements for Sediment Transport Models of Rivers*.
- Ramanathan, V. (2013) *Applications of Computational Fluid Dynamics for River Simulation : State of the Practice*.
- Sturm, T. (2001) *Open hydraulics channel*. McGraw-Hill Book Company.
- Synowiecka, J. *et al.* (2014) 'Pomiary na czynnych sieciach kanalizacji deszczowej i ogólnospławnej', *Inżynieria Ekologiczna*, 39, pp. 187–197. doi: 10.12912/2081139X.62.
- Wotzka, D. (2022) *Koncepcja, wykonanie i badania urządzenia do pomiaru strumienia objętości ścieków w kanale otwartym*. SIM z. 568. OW Politechnika Opolska.

AUTHOR BIOGRAPHIE

DARIA WOTZKA received the M.Sc. degree in Computer Science from the Technische Universität Berlin, Germany and the Ph.D. degree in Electrical Engineering from the Opole University of Technology, Poland. She is a lecturer and research fellow at the Opole University of Technology, Poland. Her research interests include data mining, modeling and simulation applied in engineering and medicine.

DEVELOPMENT OF A 2D DISCRETE ELEMENT SOFTWARE WITH LABVIEW FOR CONTACT MODEL IMPROVEMENT AND EDUCATIONAL PURPOSES

László Pásthly

Department of Machine and Product Design, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
E-mail: pasthy.laszlo@edu.bme.hu

József Gráff

Department of Mechatronics, Optics and Mechanical Engineering Informatics, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
E-mail: graff@mogi.bme.hu

Kornél Tamás

Department of Machine and Product Design, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
E-mail: tamas.kornel@gt3.bme.hu

ABSTRACT

In this study a two-dimensional discrete element software has been developed in LabVIEW (National Instruments) environment. It has an easy-to-use, graphical user interface, and the graphical programming interface makes it easy to understand and modify program code. The software can be used for educational purposes, as the basics and practical application of the discrete element method can be easily illustrated, and it can also be used for research purposes, as new geometric and contact models can be implemented and tested in a relatively short time. The functionality of the program was illustrated by simulating a gravitational deposition of 300 randomly generated particles.

INTRODUCTION

The discrete element method (DEM) has been developed for modeling granular materials (Cundall and Strack 1979). The method has many applications, which include soil (Tamás et al. 2013), agricultural grain crops (Horváth et al. 2019), chemical, food and pharmaceutical materials, railway crushed stone (González 2018), and stone construction (Bagi 2007) modeling. More advanced models are also able to calculate heat conduction between particles (Li et al. 2019), electrostatic charge states (Hogue et al. 2008), and wear of geometries effected by contacted particles (Schramm et al. 2020).

The main steps of the calculation are the determination of the dimension (two or three-dimensional) of the problem, the definition of particle geometry (circle/sphere, particles built up from multiple circles/spheres, polyhedral particles), the definition of the geometry of the rigid bodies, the definition of the properties of the particles according to the selected contact model (density, normal and shear stiffness, sliding and rolling friction coefficient, damping factor, normal and shear strength, thermal conductivity, etc.), determination of the utilized time step, creation of the assembly of particles (with the use of gravitational deposition, isotropic compaction, growth in range, etc.) and running the simulation and finally evaluating the

results (retrieve the speed, volume flow, force, voltage, temperature, cohesion, porosity, etc. data and representation of data in diagrams). The main objective of the simulation is to model the movement of the particles in the assembly under the influence of different forces.

Basically, discrete element software can be divided into two groups: There are software for research use and software for industrial use. In general, the source code for software used by researchers for example PFC 2D, PFC 3D (Eychenne 2007), UDEC, 3DEC (Israelsson 1996), Yade (Šmilauer et al. 2010), LIGGGHTS (Berger et al. 2015), and ESyS-Particle (Weatherley 2009) are free to develop, but are less easy to use, because knowledge of several programming languages is required at the same time. In contrast, in industrial software (e.g. EDEM (Dun et al. 2016), and Rocky DEM (Fonte 2015)) the possibilities are more limited, however they are easier and faster to use thanks to their simplified, user-friendly programming languages, or more complex user interfaces.

However, there is currently no example of discrete element software which has a user-friendly interface, and at the same time its source code is open and easy to understand and modify thanks to the graphical programming interface of LabVIEW.

The aim of this study was to develop a two-dimensional discrete element software in LabVIEW environment which can be used both for educational purposes and for rapid, preliminary testing of new contact models. Accordingly, in addition to the basic steps of discrete element method are implemented in the software, our objective was to have an easy-to-use user interface and easy-to-understand and modify source code.

In the course of development of the discrete element software, the following simplifications have been implemented: The particles were modeled with a unit-height cylinder shape in the same size. A two-dimensional simulation was performed in which the cylindrical particles appeared as circles. The material parameters of the particles and the boundary wall were the same. The Hertz-Mindlin contact model (Yang et al. 2020) was used to calculate the contact forces.

In addition, the further aim was to demonstrate the operation of the software by simulating a gravitational deposition of an assembly of 300 randomly generated particles.

THEORETICAL BACKGROUND

One of the most common contact models is Hertz-Mindlin (Yang et al. 2020), which can be used to model non-cohesive, flexible materials (for example dry polymer granules). This model has been implemented in the software.

The relationship between two ball shaped particles in contact is shown on Figure 1.

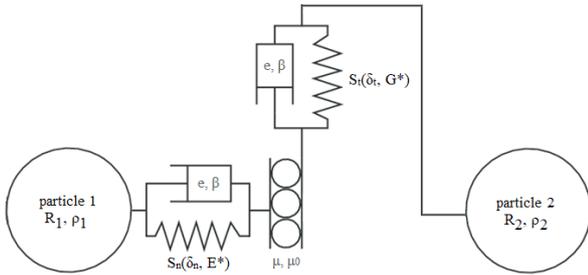


Figure 1: Hertz-Mindlin contact model, ρ_1, ρ_2 [kg/m³] – density of particles, R_1, R_2 [m] – radius of sphere or circle particles, E^* [Pa] – equivalent Young's modulus, G^* [Pa] – equivalent shear modulus, μ, μ_0 [-] – sliding and rolling friction coefficient, e [-] – coefficient of restitution, β [-] – damping factor, δ_n [m] – normal overlap of particles, δ_t [m] – tangential overlap of particles, S_n [N/m] – equivalent normal stiffness, S_t [N/m] – equivalent shear stiffness

The equivalent Young's modulus of two particles in contact (E^* [MPa]):

$$\frac{1}{E^*} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \quad (1)$$

Where E_1 [MPa] is the Young's modulus of particle 1, E_2 [MPa] is the Young's modulus of particle 2, ν_1 [-] is the Poisson's ratio of particle 1 and ν_2 [-] is the Poisson's ratio of particle 2.

The shear modulus (G [MPa]):

$$G = \frac{E}{2 \cdot (1 + \nu)} \quad (2)$$

Where E [MPa] is the Young's modulus and ν [-] is the Poisson's ratio of a particle.

The equivalent radius of two particles in contact (R^* [m]):

$$\frac{1}{R^*} = \frac{1}{R_1} + \frac{1}{R_2} \quad (3)$$

Where R_1 [m] is the radius of particle 1 and R_2 [m] is the radius of particle 2.

The mass of a particle, assuming cylindrical particles of unit height (m [kg]):

$$m = \rho \cdot R^2 \cdot \pi \quad (4)$$

Where ρ [kg/m³] is the density and R [m] is the radius of a particle.

The equivalent mass of two particles in contact (m^* [kg]):

$$m^* = \left(\frac{1}{m_1} + \frac{1}{m_2} \right)^{-1} \quad (5)$$

Where m_1 [kg] is the mass of particle 1 and m_2 [kg] is the mass of particle 2.

The contact forces can be calculated with the following equations.

The normal elastic force (F_{ne} [N]):

$$F_{ne} = \frac{4}{3} \cdot E^* \cdot \sqrt{R^*} \cdot \delta_n^{\frac{3}{2}} \quad (6)$$

Where E^* [Pa] is the equivalent Young's modulus, R^* [m] is the equivalent radius and δ_n [m] is the normal overlap of particles.

The normal damping force (F_{nd} [N]):

$$F_{nd} = -2 \cdot \sqrt{\frac{5}{6}} \cdot \beta \cdot \sqrt{S_n \cdot m^*} \cdot v_{nrel} \quad (7)$$

Where β [-] is the damping factor, S_n [N/m] is the equivalent normal stiffness, m^* [kg] is the equivalent mass and v_{nrel} [m/s] is the relative normal velocity of particles.

The equivalent normal stiffness (S_n [N/m]):

$$S_n = 2 \cdot E^* \cdot \sqrt{R^*} \cdot \delta_n \quad (8)$$

Where E^* [Pa] is the equivalent Young's modulus, R^* [m] is the equivalent radius and δ_n [m] is the normal overlap of particles.

The total normal force (F_n [N]):

$$F_n = F_{ne} + F_{nd} \quad (9)$$

Where F_{ne} [N] is the normal elastic force and F_{nd} [N] is the normal damping force.

The tangential elastic force (F_{te} [N]):

$$F_{te} = -S_t \cdot \delta_t \quad (10)$$

Where S_t [N/m] is the equivalent shear stiffness and δ_t [m] is the tangential overlap of particles.

The equivalent shear stiffness (S_t [N/m]) can be calculated similarly to the equivalent normal stiffness:

$$S_t = 8 \cdot G^* \cdot \sqrt{R^*} \cdot \delta_t \quad (11)$$

Where G^* [Pa] is the equivalent shear modulus, R^* [m] is the equivalent radius and δ_t [m] is the tangential overlap of particles.

The tangential damping force (F_{td} [N]):

$$F_{td} = -2 \cdot \sqrt{\frac{5}{6}} \cdot \beta \cdot \sqrt{S_t \cdot m^*} \cdot v_{trel} \quad (12)$$

Where β [-] is the damping factor, S_t [N/m] is the equivalent shear stiffness, m^* [kg] is the equivalent mass and v_{trel} [m/s] is the relative tangential velocity of particles.

Finally the total tangential force (F_t [N]), which is limited by friction:

$$F_t = \begin{cases} F_{te} + F_{td}, & \text{if } |F_{te} + F_{td}| < \mu_0 \\ F_n \cdot \mu, & \text{if } |F_{te} + F_{td}| \geq \mu_0 \end{cases} \quad (13)$$

Where F_{te} [N] is the tangential elastic force, F_{td} [N] is the tangential damping force, F_n [N] is the total normal

force, μ [-] is the sliding friction coefficient and μ_0 [-] is the rolling friction coefficient of particles.

Since the forces act parallel (normal direction) and perpendicular (tangential direction) to the straight section connecting the centers of the two contacted particles, it is necessary to determine the relative velocities in this local coordinate system, which makes it possible to calculate the contact forces. Thus, the velocities of the particles have to be transformed into the local coordinate system, and then the contact forces acting on the particles has to be transformed back into the global, horizontal (x), and vertical (y) coordinate system for summation. These operations required the use of scalar multiplications, which are presented below.

The relative velocity vector of i and j particles in contact in the normal-tangential coordinate system ($\mathbf{v}_{rel,i,j(n,t)}$ [m/s]):

$$\begin{aligned} \mathbf{v}_{rel,i,j(n,t)} &= \begin{pmatrix} v_{rel,i,j,n} \\ v_{rel,i,j,t} \end{pmatrix}_{(n,t)} \\ &= \begin{pmatrix} (\mathbf{v}_{i(x,y)} - \mathbf{v}_{j(x,y)}) \cdot \mathbf{r}_{ij(x,y)} \\ |(\mathbf{v}_{i(x,y)} - \mathbf{v}_{j(x,y)}) - (\mathbf{v}_{i(x,y)} - \mathbf{v}_{j(x,y)}) \cdot \mathbf{r}_{ij(x,y)} \cdot \mathbf{r}_{ij(x,y)}| \end{pmatrix}_{(n,t)} \end{aligned} \quad (14)$$

Where $v_{rel,i,j,n}$ [m/s] is the relative velocity in the normal direction, $v_{rel,i,j,t}$ [m/s] is the relative velocity in the tangential direction, $\mathbf{v}_{i(x,y)}$ [m/s] is the velocity vector of particle i in the global (x-y) coordinate system, $\mathbf{v}_{j(x,y)}$ [m/s] is the velocity vector of particle j in the global (x-y) coordinate system and $\mathbf{r}_{ij(x,y)}$ [m] is a unit-length vector pointing from the center of particle i to the center of particle j in the global (x-y) coordinate system.

The contact force vector acting on particle i from the contact with particle j in the global (x-y) coordinate system ($\mathbf{F}_{i,j(x,y)}$ [N]):

$$\mathbf{F}_{i,j(x,y)} = \begin{pmatrix} F_{i,j,x} \\ F_{i,j,y} \end{pmatrix}_{(x,y)} = \begin{pmatrix} F_{i,j,n} \cdot r_{ijx} + F_{i,j,t} \cdot r_{ijy} \\ F_{i,j,n} \cdot r_{ijy} + F_{i,j,t} \cdot r_{ijx} \end{pmatrix}_{(x,y)} \quad (15)$$

Where $F_{i,j,x}$ [N] is the x direction component of the contact force, $F_{i,j,y}$ [N] is the y direction component of the contact force, $F_{i,j,n}$ [N] is the normal direction component of the contact force and $F_{i,j,t}$ [N] is the tangential direction component of the contact force.

The moment acting on particle i from the contact with particle j ($M_{i,j}$ [Nm]) were also taken into account according to the following equation:

$$M_{i,j} = F_{i,j,t} \cdot R^* \quad (16)$$

Where $F_{i,j,t}$ [N] is the tangential direction component of the contact force and R^* [m] is the equivalent radius of the particles.

The acceleration and angular acceleration of particles are calculated by Newton's laws, therefore the acceleration and angular acceleration of a particle are proportional to the force and torque acting on it.

Acceleration vector (\mathbf{a} [m/s²]) and angular acceleration (α [rad/s²]) of a circular particle performing a planar motion:

$$\mathbf{a} = \frac{\sum \mathbf{F}}{m} \quad (17)$$

$$\alpha = \frac{\sum M}{\Theta} = \frac{\sum M}{\frac{m \cdot R^2}{2}} \quad (18)$$

Where $\sum \mathbf{F}$ [N] is the vector sum of the forces, $\sum M$ [Nm] is the sum of the moments, m [kg] is the mass, Θ [kgm²] is the moment of inertia and R [m] is the radius of particle.

The velocity vector, position vector, angular velocity, and angular position of the particles can be calculated from the state of motion in the old time step and the state of acceleration and angular acceleration calculated in the current time step. To do this, a second-order system of ordinary differential equations needs to be solved, for which several numerical methods are available. Our aim was the accuracy and fast calculation. However, these two conditions are difficult to meet at the same time because accuracy is at the expense of the speed of calculation and vice versa. Multi-step formulas require less calculation, although they are generally less accurate. However, in the simulation behavior of textiles (Gräff, J. et al. 2004) the simultaneously applied second-order predictor and corrector gave a surprisingly good result for the second-order system of nonlinear differential equations, so this method was used here as well.

The first-order Adams-Bashforth integral formula (Bashforth and Adams 1883) was used in the first time step to determine the velocity vector (\mathbf{v}_1 [m/s]) and angular velocity (ω_1 [rad/s]) of particles:

$$\mathbf{v}_1 = \mathbf{v}_0 + \mathbf{a}_0 \cdot \Delta t \quad (19)$$

$$\omega_1 = \omega_0 + \alpha_0 \cdot \Delta t \quad (20)$$

Where \mathbf{v}_0 [m/s] is the initial velocity vector, \mathbf{a}_0 [m/s²] is the initial acceleration vector, ω_0 [rad/s] is the initial angular velocity, α_0 [rad/s²] is the initial angular acceleration and Δt [s] is the length of a time step.

After the first time step, for the determination of velocity vector (\mathbf{v}_i [m/s]) and angular velocity (ω_i [rad/s]) of particles in the timestep i , the second-order Adams-Bashforth integral formula (Bashforth and Adams, 1883) was used:

$$\mathbf{v}_i = \mathbf{v}_{i-1} + \frac{3 \cdot \mathbf{a}_{i-1} - \mathbf{a}_{i-2}}{2} \cdot \Delta t \quad (21)$$

$$\omega_i = \omega_{i-1} + \frac{3 \cdot \alpha_{i-1} - \alpha_{i-2}}{2} \cdot \Delta t \quad (22)$$

Where \mathbf{v}_{i-1} [m/s] is the velocity vector in time step $i-1$, \mathbf{a}_{i-1} [m/s²] is the acceleration vector in time step $i-1$, \mathbf{a}_{i-2} [m/s²] is the acceleration vector in time step $i-2$, ω_{i-1} [rad/s] is the angular velocity in time step $i-1$, α_{i-1} [rad/s²] is the angular acceleration in time step $i-1$, α_{i-2} [rad/s²] is the angular acceleration in time step $i-2$ and Δt [s] is the length of a time step.

The second-order Adams-Moulton integral formula (Moulton 1926) was used in all time steps to determine the position vectors (\mathbf{r}_i [m]) and angular positions (φ_i [rad]) of particles, since the predictor already determined the velocities and angular velocities in the old time step.

The position vectors (\mathbf{r}_i [m]) and angular positions (φ_i [rad]) of particles in time step i :

$$\mathbf{r}_i = \mathbf{r}_{i-1} + \frac{\mathbf{v}_{i-1} + \mathbf{v}_i}{2} \cdot \Delta t \quad (23)$$

$$\varphi_i = \varphi_{i-1} + \frac{\omega_{i-1} + \omega_i}{2} \cdot \Delta t \quad (24)$$

Where \mathbf{r}_{i-1} [m] is the position vector in time step $i-1$, \mathbf{v}_i [m/s]: is the velocity vector in time step i , \mathbf{v}_{i-1} [m/s] is the velocity vector in time step $i-1$, φ_i [rad] is the angular position in time step i , φ_{i-1} [rad] is the angular position in time step $i-1$, ω_i [rad/s] is the angular velocity in time step i , ω_{i-1} [rad/s] is the angular velocity in time step $i-1$ and Δt [s] is the length of a time step.

Overall the contact forces are determined by the Hertz-Mindlin model, followed by the accelerations and angular accelerations by Newton's laws, and finally the velocity, angular velocity, position, and angular position of the particles by the Adams-Bashforth and Adams-Moulton numerical integral formulas.

RESULTS

The structure of the program is shown on Figure 2. The main parts are the graphical display, the detection of the contacts, the calculation of the overlaps and contact vectors, the calculation of the forces and torques, the calculation of the accelerations and the determination of the motion state. The program is able to plot the changes in potential, kinetic and total energy and the so-called unbalanced force ratio, which is the ratio of mean summary force on particles and mean force magnitude in contacts (Šmilauer et al., 2010), in real time, and it is also possible to save the data of the energy and unbalanced force ratio.

During initialization, the input data is determined for the calculation cycle. It is possible to place the particles randomly in the simulation space. An algorithm then generates the specified number of particles in random locations so that it does not come into contact with the particles placed so far. And it is also possible for the user to determine the position of the particles. Another important task of initialization is the conversion of units into the SI system (for example conversion of particle radius from mm to m).

The input parameters of the calculation cycle are the material parameters (density, Young's modulus, shear modulus, rolling friction coefficient, sliding friction

coefficient and damping factor), the geometrical parameter (particle radius), parameters of initial motion (initial position, velocity, acceleration, angular position, angular acceleration), the gravity field, the length of the time step, and the calculation time. Instead of specifying the time step, it is also possible to set the simulation to run until the STOP button is pressed.

During the graphical display, each particle is drawn in its position. In the case of contacts detection the program examines the contact for each pair of particles using the Pythagorean theorem, it calculates the distance between the particles and compares this value to twice the radius of the particles. If the distance between the particles is less than twice the radius, the two particles are in contact.

Also, the contact of the particles with the wall is detected if the coordinates of the particles are less than or greater than a certain value, the particles are in contact with the wall.

The overlaps between the particle pairs in contact and the particle-wall pairs in contact are then determined: the difference between twice the particle radius and the distance of particles gives the overlap in the normal direction of the particles. On a similar principle, the overlap of the particles in contact with a wall is calculated.

In addition, in the case of particle-to-particle connections, so-called contact vectors are defined, which point from the center of the particle in contact to the center of the other particle in contact, and are of unit length. This is necessary, because the contact forces between the particles are calculated in the direction of the contact vectors (the normal direction), and in the direction perpendicular to the contact vectors (tangential direction). Because of this, the velocity data also has to be transformed into the local coordinate system (with normal and tangential directions) of the particles in contact. After calculating the contact forces, the forces are transformed back into the global coordinate system (with directions x and y) using scalar multiplications.

The moments from the contacts are obtained as the product of the tangential forces and the radius of the particles.

The particle-to-particle and particle-to-wall contact forces and moments are calculated in a separate function and then the forces and torques acting on a single particle are summed in the global coordinate system.

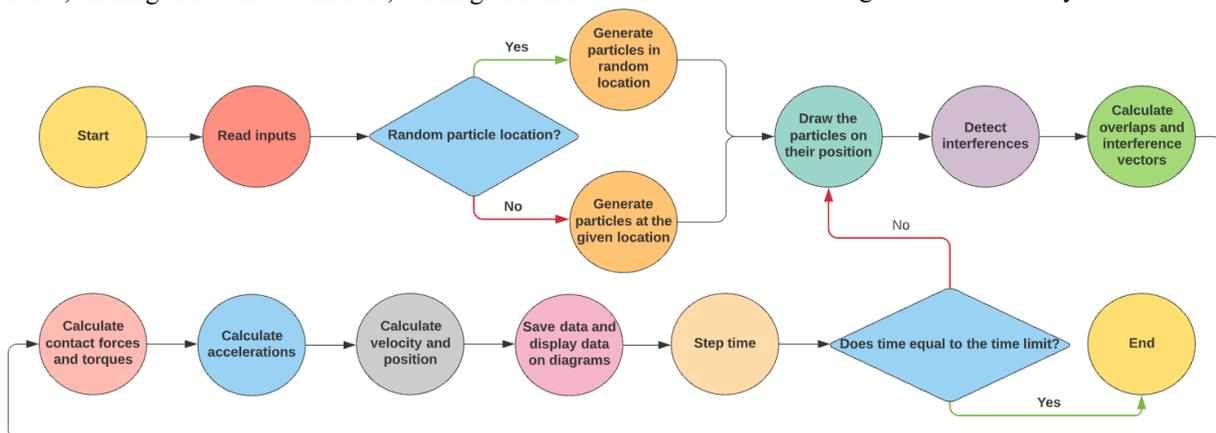


Figure 2: Structure of the developed discrete element program in LabVIEW

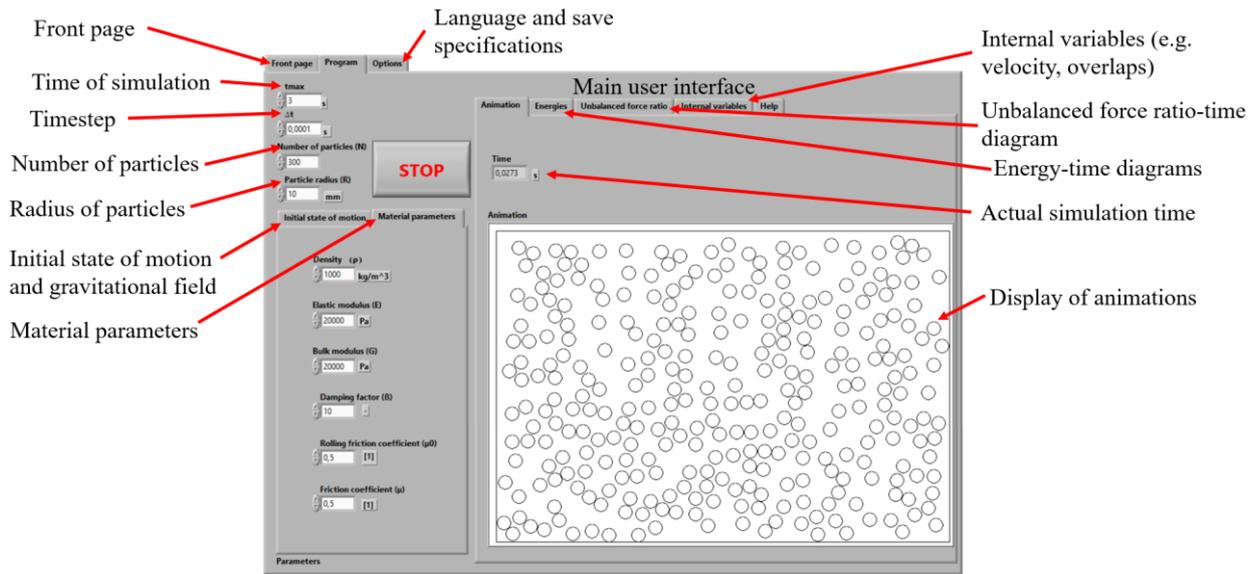


Figure 3: User interface of the developed “BMEDEM2D” discrete element software

In addition to the particle-to-particle and particle-wall contact forces and moments, the user-specified gravitational force acting uniformly on all particles is taken into account too when calculating the acceleration state.

The velocity and displacement state of the current time step are determined using the integral formulas described in equations (19) - (24).

Once the motion state has been determined, the simulation cycle starts again or the simulation ends if one of the stop conditions are met.

While running the program, the changes in the kinetic, potential and total energy, as well as the so-called unbalanced force ratio is automatically plotted. Its significance is in the case of gravitational deposition simulations, where the simulation is usually stopped when the unbalanced force ratio falls below a certain level.

After the simulation, it is possible to save the data displayed on the charts. When the “Save Data” option is turned on, the data is saved in a text file with headers and a user-specified file name. It is then possible to read and evaluate the data in spreadsheet software.

When designing the user interface (Figure 3), we aimed for a simple, easy-to-understand design as well as a structure similar to commercial software. In the upper left corner, it is possible to enter the simulation time, the length of time step, particle radius, and number of particles. Geometric parameters, gravity field, and material parameters can be entered in the tabs on the left. During the simulation, the movement of the particles can be observed on the animation tab on the right. Another tab shows the kinetic, potential and total energies, a third shows the unbalanced force ratio, and a fourth shows other internal variables (for example the current position of particles). The simulation can be stopped at any time by pressing the STOP button in the upper left corner.

The operation of the software is illustrated by a simulation of a gravitational deposition, during which the

particles left alone in the specified gravitational field gradually settle on top of each other.

Table 1 shows the parameters used in the simulation.

Table 1: Parameters of the gravitational deposition simulation

Name	Notation	Quantity	Unit
Density	ρ	1000	kg/m ³
Young's modulus	E	20000	Pa
Shear modulus	G	20000	Pa
Damping factor	β	10	-
Sliding friction coefficient	μ	0.5	-
Rolling friction coefficient	μ_0	0.5	-
Particle radius	R	0.01	m
Number of particles	N	300	-
Gravitational field (in vertical direction)	g	-9.81	m/s ²
Length of time step	Δt	$1 \cdot 10^{-4}$	s

Figure 4 shows the deposition the assembly of 300 particles with a random initial position in a vertical gravitational field. Figure 5 shows the energy-time diagrams and Figure 6 shows the unbalanced force ratio-time diagram of the assembly.

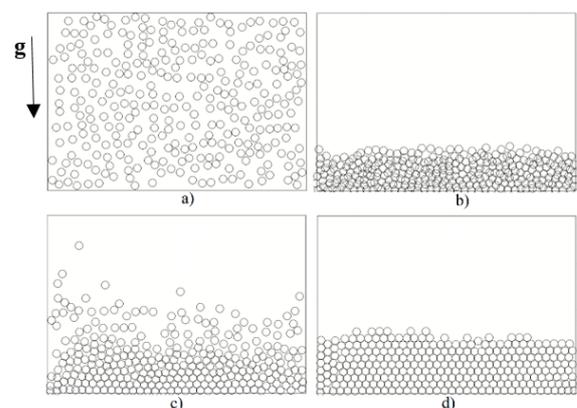


Figure 4: Deposition of 300 particles depends on the actual time step, a) 0 s, b) 0.23 s, c) 0.45 s, d) 3 s

The particles fall down relatively quickly (Figure 4 b), but then the elastic energy stored in the overlaps is so large that the upper particles bounce back from the top of the assembly, but no longer reach their original height (Figure 4 c) as their total energy is reduced due to damping. This process is repeated until the particles have enough energy to separate again after the collisions.

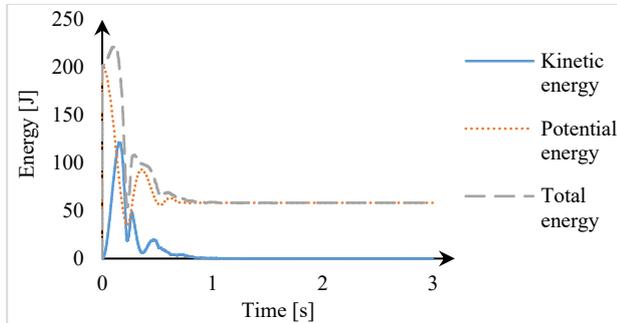


Figure 5: The resulted energy-time diagrams

At the beginning there is an increase in the kinetic energy and total energy of the assembly. Then the energies begin to decrease due to damped collisions with the wall. After a certain time (approximately after 1 s), the energies no longer change significantly. At this point, however, the assembly is not yet considered to be completely settled, as some displacements are still made by a few particles, as a result of which they may find themselves in a more energetically stable position.

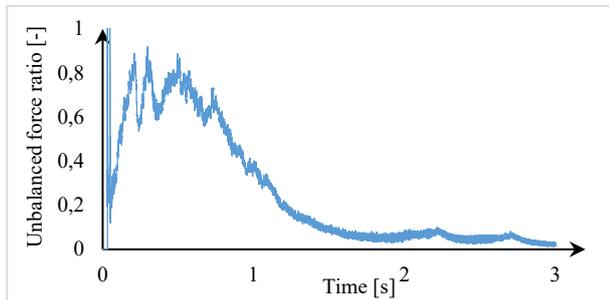


Figure 6: Unbalanced force ratio-time diagram

The unbalanced force ratio cannot be interpreted at the beginning of the simulation, which is due to the fact that neither particle is in contact with another, so the denominator of the quotient required for calculation of the ratio is zero. However, after this initial stage, the particles begin to collide with each other and the value of the unbalanced force ratio becomes interpretable (maximum value is 2.4). In the rest of the simulation, due to the damped collisions, the particles settle on top of each other, and in parallel, the unbalanced force ratio shows a decreasing trend. At the end of the simulation, i.e. at 3 s, its value is 0.013. Local maximum values of the unbalanced force ratio can also be observed when the particles are in a less energetically stable position. However, by overcoming these potential barriers, the unbalanced force ratio becomes always lower than before, so the particles get into an energetically more stable position.

In the future, it may also be worthwhile to reduce the time required for the gravitational deposition by changing the damping.

CONCLUSION

In this study, a user-friendly discrete element software capable of simulating granular materials has been developed in LabVIEW environment, which is able to calculate changes in the state of motion due to particle-to-particle and particle-wall contacts. The theoretical background of the computational method was described, as well as the operation of the software, then the operation of the software was demonstrated by simulating the gravitational settling of an assembly of 300 randomly placed particles. Realistic results were obtained during the presented simulation. Due to the damping in the contact model, the total energy of the particles continuously decreased and then fluctuated slightly around a minimum value. The unbalanced force ratio also showed a declining trend, i.e., the particles became more and more balanced over time.

Thanks to its easy-to-use interface, the software makes it easy to prepare and evaluate simulations, and due to the graphical programming interface it is also simple to modify parts of the program, allowing new contact models to be improved and tested quickly. The operation of the simulation was demonstrated through a practical example.

In the future, we plan to expand the software with other features. These include the presence of rigid, moving bodies arbitrarily located in the simulation space, as well as particles of different sizes within a simulation. In addition to the discrete element method, it is also planned to integrate the finite element method into the software, which would make it possible to model the deformation of solids.

ACKNOWLEDGEMENT

The research reported in this paper and carried out at BME has been supported by the NRDI Fund (TKP2020 NC, Grant No. BME-NCS) based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology.

The research has been partially supported by the NKFI OTKA K-138642 grant of the Hungarian Ministry for Innovation and Technology.

The publication of the work reported herein has been supported by ETDB at BME.

REFERENCES

- Bagi, K. 2007. "A diszkrét elemek módszere". *BME Department of Structural Mechanics, Budapest, 5-12.*
- Bashforth, F. and J. C. Adams 1883. "An Attempt to test the Theories of Capillary Action by comparing the theoretical and measured forms of drops of fluid. With an explanation of the method of integration employed in constructing the tables which give the theoretical forms of such drops" *Cambridge.*
- Berger, R.; C. Kloss; A. Kohlmeyer and S. PirkerBerger; 2015. "Hybrid parallelization of the LIGGGHTS open-source DEM code". *Powder technology*, vol. 278, pp. 234–247.

- Cundall, P. A. and O. D. L. Strack, 1979. "A discrete numerical model for granular assemblies", *Géotechnique*.
- Dun, G.; H. Chen; Y. Feng; J. Yang; A. Li and S. Zha. 2016. "Parameter optimization and test of key parts of fertilizer allocation device based on EDEM software". *Transactions of the Chinese Society of Agricultural Engineering*, vol. 32, no. 7, pp. 36–42.
- Eychenne, J. 2007. "Documentation de la plateforme PFC (version 1.0)". *Phonologie du Français Contemporain Phonologie du Français Contemporain*, p. 41.
- Fonte, C. B. 2015. "DEM-CFD coupling: mathematical modelling and case studies using ROCKY-DEM® and ANSYS Fluent®". in *Proceedings of the 11th International Conference on CFD in the Minerals and Process Industries, CSIRO, Melbourne, Australia*, pp. 7–9.
- González, J.; E. Oñate and F. Salazar. 2018. "Numerical analysis of railway ballast behaviour using the Discrete Element Method". *Monograph CIMNE*.
- Gräff, J. and J. Kuzmina. 2004. "Cloth simulation using mass and spring model". In *GÉPÉSZET 2004, Proceedings of the Fourth Conference on Mechanical Engineering*. Budapest, Magyarország: Budapest University of Technology and Economics 781 p. pp. 443-447. , 5 p.
- Hogue, M.D.; C.I. Calle; P.S. Weitzman and D.R. Curry. 2008. "Calculating the trajectories of triboelectrically charged particles using Discrete Element Modeling (DEM)". *Journal of Electrostatics*, vol. 66, no. 1, pp. 32–38.
- Horváth, D.; T. Poós and K. Tamás. 2019. "Modeling the movement of hulled millet in agitated drum dryer with discrete element method". *Computers and Electronics in Agriculture*, vol. 162, pp. 254–268.
- Israelsson, J. I. 1996. "Short descriptions of UDEC and 3DEC" *Developments in geotechnical engineering*, vol. 79, Elsevier, pp. 523–528.
- Li, Y.; H. Xu; C. Jing; J. Jiang and X. Hou. 2019. "A novel heat transfer model of biomass briquettes based on secondary development in EDEM", *Renewable Energy*, vol. 131, pp. 1247–1254.
- Moulton, F. R. 1926. "New methods in exterior ballistics" *University of Chicago Press*.
- Rackl M.; F. Top; C.P. Molhoek and D.L. Schott 2017. "Feeding system for wood chips: A DEM study to improve equipment performance", *Biomass and Bioenergy*, vol. 98, pp. 43–52. (Mar).
- Schramm, F.; Á. Kalácska; V. Pfeiffer; J. Sukumaran; P.De. Baets and L. Frerichs. 2020. "Modelling of abrasive material loss at soil tillage via scratch test with the discrete element method" *Journal of Terramechanics*, vol. 91, pp. 275–283.
- Šmilauer, V.; E. Catalano; B. Chareyre; S. Dorofeenko; J. Duriez; A. Gladky; J. Kozicki; C. Modenese, L. Scholtès; L. Sibille; J. Stránský and K. Thoeni. 2010. "Yade reference documentation" *Yade Documentation*, vol. 474, no. 1.
- Tamás K.; I.J. Jóri and A.M. Mouazen. 2013. "Modelling soil–sweep interaction with discrete element method", *Soil and Tillage Research*, vol. 134, pp. 223–231.
- Yang, W.; M. Wang; Z. Zhou; L. Li; G. Yang and R. Ding. 2020. "Research on the relationship between macroscopic and mesoscopic mechanical parameters of limestone based on Hertz Mindlin with bonding model" *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, vol. 6, no. 4, pp. 1–15.
- Weatherley, D. 2009. "ESyS-Particle v2. 0 user's guide".

AUTHOR BIBLIOGRAPHIES



LÁSZLÓ PÁSTHY is an MSc mechanical engineering student at the Budapest University of Technology and Economics, Hungary where he received his BSc degree. He specializes in machine design and is also involved in the research of coupled discrete element method (DEM) – finite element method (FEM) simulations. His e-mail address is: pasthy.laszlo@edu.bme.hu and his web-page can be found at https://gt3.bme.hu/oktatoi_oldal.php?lepes=4&oid=190.



JÓZSEF GRÄFF is a master teacher at Budapest University of Technology and Economics where he received his MSc degree as a mathematical engineer. His professional fields are numerical methods and simulations with multi steps integrators. His e-mail address is: graff@mogi.bme.hu and his web-page can be found at https://mogi.bme.hu/oktatoi_oldal.php?lepes=4&oid=63



KORNÉL TAMÁS is an assistant professor at Budapest University of Technology and Economics where he received his MSc degree and then completed his PhD degree. His professional field is the modelling of granular materials with the use of discrete element method (DEM). His e-mail address is: tamas.kornel@gt3.bme.hu and his web-page can be found at https://gt3.bme.hu/oktatoi_oldal.php?lepes=4&oid=162.

APPLICATION OF THE EXTENDED FINITE ELEMENT METHOD IN THE AIM OF EXAMINATION OF CRACK PROPAGATION IN RAILWAY RAILS

Dániel Bóbis*

Péter T. Zwierczyk

Tamás Máté

Department of Machine and Product Design

Faculty of Mechanical Engineering

Budapest University of Technology and Economics

Műegyetem rkp. 3., H-1111 Budapest, Hungary

E-mail: bobis.daniel.@gt3.bme.hu

*Corresponding author

KEYWORDS

Extended finite element method (X-FEM), Ansys Mechanical APDL, Stress intensity factor (SIF), Rolling contact fatigue (RCF), Head check (HC), Crack propagation, Railway.

ABSTRACT

In this research, a rolling contact fatigue (RCF) caused crack the so-called head check (HC) was studied by the extended finite element method (X-FEM). The formation and treatment of these cracks is a major problem in the railway industry worldwide. The study aimed to explore the capabilities of the X-FEM concerning examinations of crack propagation in rolling contact conditions. The paper introduces the X-FEM from a practical point of view, with an emphasis on the application of the method, rather than focusing on the exact values. Results showed that the method can be used adequately for such complex issues as examination of crack propagation. However, the possibilities of this technique are quite limited in practice yet due to technical reasons.

INTRODUCTION

In this paper, a finite element (FE) modeling technique is introduced, called extended finite element method (X-FEM). The aim of this study is to explore the possibilities of the above mentioned method in practice regarding the examination of crack propagation in the research area of railway wheel-rail connection. On both the wheel and the rail side micro-cracks can appear for different reasons that must be inspected and handled properly. Rolling contact fatigue (RCF) is one of the main reasons for crack initiations. The phenomenon is investigated widely but still has numerous unanswered questions. The overall purpose of the research is to gain a better understanding on the nature of these cracks.

Overview of the Examined Issue

This research focuses on the so-called head check (HC), which refers to a kind of multiple hairline cracks in the railhead. The formation of the HC is a typical

manifestation of the RCF-caused failures of the railway rail. They are located on the gauge corner parallel to each other at a typical slanted angle as Figure 1 depicts.

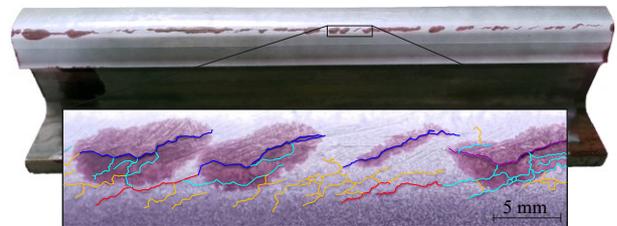


Figure 1: Typical HCs on a Part of a Rail. As a Result of a Former Liquid Penetrant Inspection, Cracks are Highlighted in Red. At the Bottom, Similar Cracks are Represented with the Same Color.

Csizmazia (Csizmazia and Horvát, 2014) summarized the initiation and propagation mechanisms of these cracks in addition to the methods used to eliminate them. In brief, HC usually appears in sharper curves on the outer rail where the wheel-rail connection is more adverse. Due to the plastic strain, significant hammer-hardening effect occurs on the upper layers of the rail. As the material hardens, its elasticity declines. When the highly concentrated loads exceed the resistance of the rail steel, micro-cracks initiate and start to grow towards the railhead. Regarding the stages of crack growth, fluid significantly influences the propagation process, since it can be forced to the crack by contact pressure, which produces hydraulic pressure inside the crack.

Spalling of the surface is usually caused by these micro-cracks, which can lead to the development of larger and more dangerous defects. Therefore, it is important to inspect and treat HCs properly. The treatment should be done at the early stages of the formation by removing the damaged layer of the rail, mainly via re-grinding processes. Detection systems are generally based on eddy current technology. The accuracy of these systems might be unreliable in some cases, hence the damaged layer is conservatively

overestimated. Since the elimination processes are highly expensive, the extent and frequency of the maintenance is a key issue from an economic point of view.

As there is a high level of uncertainty around the phenomenon, every study made on the subject can be valuable in order to understand the problem better and handle it more efficiently.

Methods

Linear Elastic Fracture Mechanics

All of the used methods are within the theory of linear elastic fracture mechanics (Anderson, 2005). It assumes that the material behavior is brittle and can be described by Hooke's law.

Cracks can be loaded by three different modes according to Figure 2. In mode I, the load is normal to the crack plane. Mode II corresponds to in-plane shear load loading and Mode III refers to out-of-plane shear.

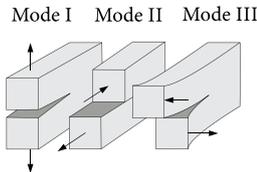


Figure 2: Crack Loading Modes (Anderson, 2005)

The stress field can be expressed via the K stress intensity factors (SIF) nearby the crack tip since it is a singular point. According to the crack loading modes, different SIFs can be distinguished. In this study, SIFs are calculated numerically.

Fatigue Crack Propagation

The typical fatigue crack growth behavior of metals is shown in Figure 3.

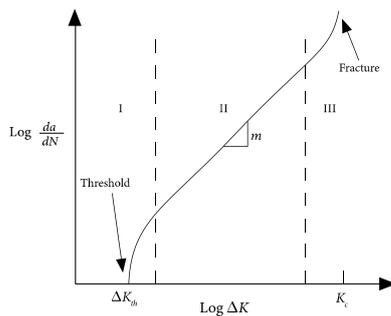


Figure 3: Fatigue Crack Grow (Anderson, 2005)

The range where this curve is linear (marked II) can be described with the commonly used formula called Paris' Law. According to the formula, the crack propagation rate is

$$\frac{da}{dN} = C\Delta K^m, \quad (1)$$

where C and m are material constants determined

experimentally (Anderson 2005). Regarding this specific study, ΔK is derived from the SIFs calculated previously.

Extended Finite Element Method

The finite element method (FEM) is a very common and frequently used numerical tool applied in many areas of engineering. Its great advantage is that it can also be used for more complex geometries. However, the calculated results are strongly dependent on the quality of the mesh. Thus, in cases where the solution contains non-smooth behavior like singularities or discontinuities in the displacement field, the classical finite element method cannot be applied with sufficient efficiency. A promising method for modeling different discontinuities is the extended finite element method. The theoretical background of the method is well explained in (Koei, 2015) and in (Zhuang, 2014).

The basis of the X-FEM approach is the extension of the displacement field by special enrichment functions to describe the effect of discontinuities, including cracks. In general, the displacement field can be written as

$$u^h(x) = \sum_I N_I(x) \cdot u_I(x) + \sum_J \Psi_J(x) \cdot q_J(x), \quad (2)$$

where $N(x)$ denotes the shape function of the standard finite element, and $u_I(x)$ is the standard degree of freedom (DOF). $\Psi_J(x)$ is the enrichment shape function chosen for the specific discontinuous problem. In the study, this term describes the current geometry of the crack as well as the effect of the crack on the displacement field. The newly added DOFs are denoted by the $q_J(x)$.

As a result of this kind of approach to the modeling of cracks, X-FEM has notable benefits. Cracks are independent from the FE mesh, thus neither remeshing is required during the analysis, nor drastic mesh refinement is necessary near the crack tip. Figure 4 shows the significance of it by comparing the classical and extended FE methods.

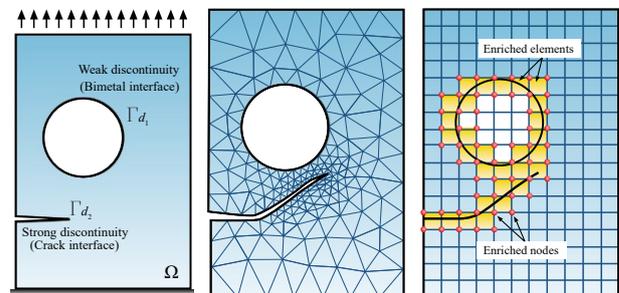


Figure 4: Modeling of Discontinuities with Different Techniques. The Picture in the Middle Shows an Adaptive (Classical) FE Mesh Which is Regenerated as the Crack Propagates. Significant Refinement is Noticeable Near the Crack tip. In the Right Picture, an X-FEM based Structural Mesh can be Seen with Far Fewer and Better Quality Elements (Koei, 2015)

Nevertheless, the use of the X-FEM modeling technique in practice is quite complex mainly due to the lack of a graphical user interface at least regarding the tools that were used in this study. Until now there is a limitation in both the available technical support and experience gained on the use of the method.

FINITE ELEMENT ANALYSIS

The aim of the research is to create a specific model by the application of the X-FEM, which is able to examine HC-type cracks in order to get a better understanding on the development of these cracks. Since the practical use of this method is quite complex, significant simplifications were taken during the research. There might be some inaccuracy in the input data applied. Consequently, the main emphasis is placed on the possibilities of the method, rather than focusing on the exact values of the results

Workflow and Software Environments

Regarding the X-FEM-based analysis, Ansys Mechanical APDL version 2020 R1 was used. Since the required functions are not yet available through the graphical user interface, the model was configured in a text file with Notepad++ and post-processed in Microsoft Excel. The workflow is shown in Figure 2. During the configuration, iteration steps were needed, because the enrichment area where the crack propagates is not known, as well as the optimal threshold value ($K_{eq,th}$) where the propagation starts.

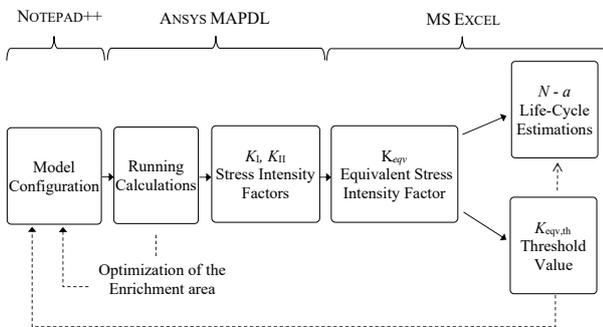


Figure 5: Software Products Used and the Workflow of the Analysis

Geometry

The examined geometry is two-dimensional in the longitudinal section of the rail with a predefined initial crack, corresponding to Figure 6 and Figure 7. The plane strain behavior of the model assumes that the extent of the crack is perpendicular to the section and relatively large. In contrast to this coarse approximation, head checks can be imagined as a countless number of spatial surfaces growing into the deeper layers of the real head at a slanted angle. In addition to this, they are typically located parallelly and close to each other. The connection between the surfaces of the initial crack is frictional. The frictional coefficient $\mu_{crack} = 0,15$ assumes dry conditions.

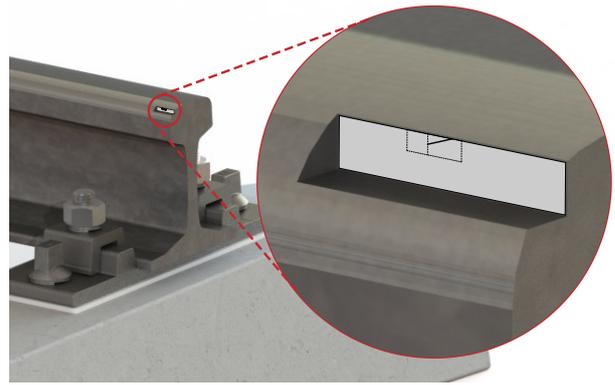


Figure 6: Interpretation of the Two-dimensional Geometry Examined

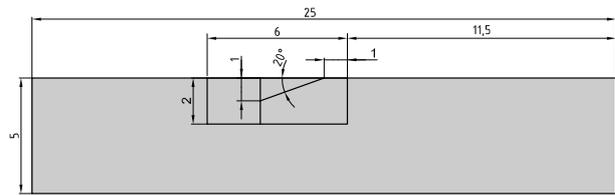


Figure 7: Dimensions of the Geometry Including a 1 mm Deep Initial Crack

Finite Element Mesh

The FE mesh consists of 0.012, 0.05, and 0.2 mm size 4-node quadrilateral-shaped elements corresponding to Figure 8. The large-scale mesh refinement was needed because loads were defined in time as a series of tiny steps, and in every load step 20 elements cracks in case the fracture criteria is satisfied. Special elements were used in order to describe the effect of the crack. Elements highlighted in yellow in Figure 4 have an enriched degree of freedom, thus they can crack. This yellow highlighted region is called the enrichment area, the extent of this region is critical. Red elements had already been cracked and were formed from the yellow elements. The green element has the crack tip which is the basis of the subsequent computations.

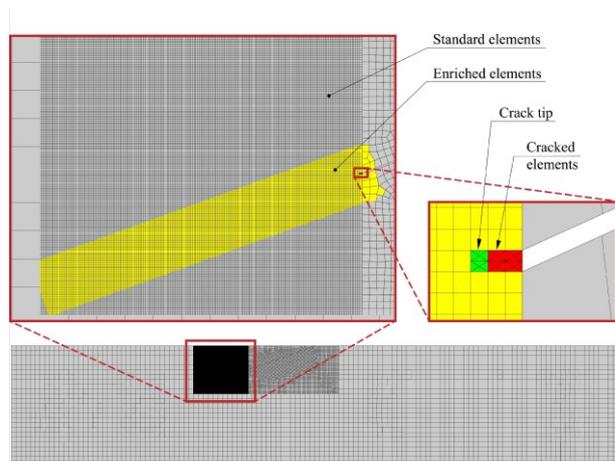


Figure 8: Finite Element Mesh with the Special Elements Describing the Crack

Loads and Boundary Conditions

The loading model assumes that the locomotive is accelerating and the frictional state between the wheel and the rail is at the limit of sticking. The contact stress distribution between the wheel of a passenger carriage and the rail was calculated by Zwierczyk in his Ph.D. dissertation (Zwierczyk, 2015). The stress distribution shown in Figure 9 was used during the analysis, assuming it does not differ significantly on the rail side.

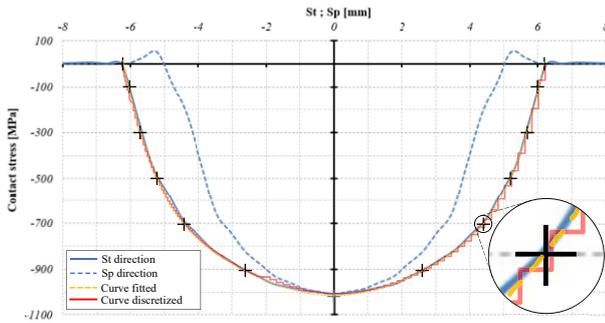


Figure 9: Contact Pressure Under the Wheel and the Discretization of it.
Adapted from (Zwierczyk, 2015)

A six-order polynomial curve was fitted to the marked points. Then, it was discretized by applying the right amount of pressure onto each FE element located on top of the examined geometry, using the equation of the curve.

The analysis is quasi-static, hence the movement of the loads is defined separately as a series of steps. The loads are moving from the left to the right in 25 load steps, each load step representing a position as Figure 10 shows. Above that resolution, results are not differing significantly, but the analysis would be computationally much more expensive. During the analysis, loads are passing through the geometry five times.

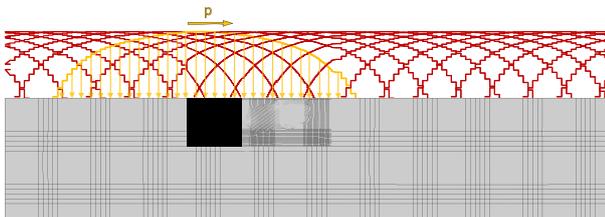


Figure 10: Representation of the Moving Loads. During the Analysis, Loads Highlighted Yellow are Changing Their Position According to the Red Curves.

Proportion to the contact pressure, frictional forces are applied as well. The direction of the frictional force is opposed to the direction of the movement. The frictional coefficient $\mu_{\text{wheel-rail}}=0,15$ assumes dry conditions.

Boundary conditions are shown in Figure 11.

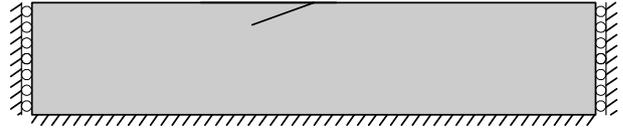


Figure 11: Boundary Conditions

At the bottom of the geometry, a fixed constraint is defined. On the sides, only the vertical DOFs are free. The latter might influence the behavior of the rail, but in this stage of the model, these constraints are acceptable. Wider geometry would require a lot more computational resources.

Material Properties

Material properties are tabulated in Table 1.

Table 1: Material Properties

Material Property	Value	Unit
Young's modulus	200 000	MPa
Poisson ratio	0.3	1
Paris constant (C)	$1.0 \cdot 10^{-8}$	1
Paris constant (m)	1.13	1

The values are correlating to that Zwierczyk used in his calculations (Zwierczyk, 2015). Paris constants were elected from a report by Aglan (Aglan, 2011), assuming that the rail steel is bainitic.

Further Analysis Settings

The duration of the load steps is 1 s. Since 25 load steps were defined over 5 cycles, the total time of the analysis is 125 s. For the sake of accuracy, each load step was divided into 20 equally spaced subsets. This results to be the crack propagation rate uniform over time, by splitting the same number of elements per step. Furthermore, it provides a satisfactory convergence of the nonlinear analysis.

The method calculates the SIFs in each step, and in case the predefined value is exceeded, then the splits of the elements result in a Δa crack increment. Life-cycle estimations are based on Equation (1), using the above-mentioned variables. Since the procedure gives a large number of ΔN cycles in each increment, the calculated lifetime is not directly correlating to the number of how many times the loads are passed through the geometry or to the size of the FE elements. The direction of the propagation is arbitrary and always turns into the critical angle which is recalculated in every moment.

RESULTS

The propagated crack in the fifth load cycle is shown in Figure 12. For the sake of illustration, it is shown in the position where $t=105$ s which is not the critical point.

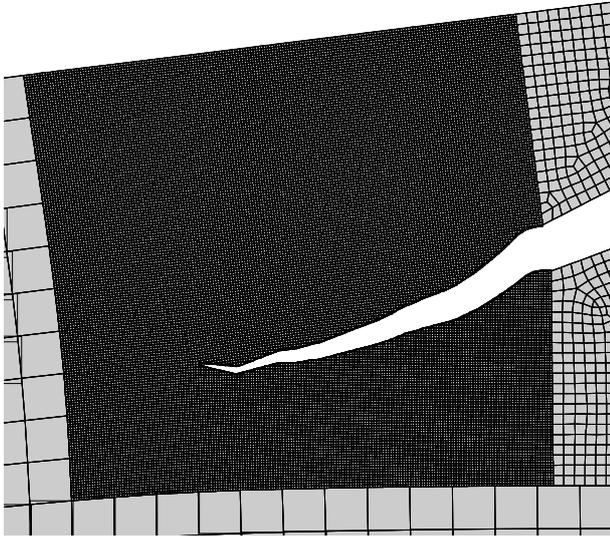


Figure 12: Crack Propagation in the Fifth Load Cycle.
Scale of Deformations is 50:1

The characteristic of the distribution of equivalent von Mises stress around the crack tip is shown in Figure 13. It is shown in the critical time instant of the fifth load cycle. The values are not representing the reality, indeed, since the crack tip is a singularity and the material behavior is linear elastic.

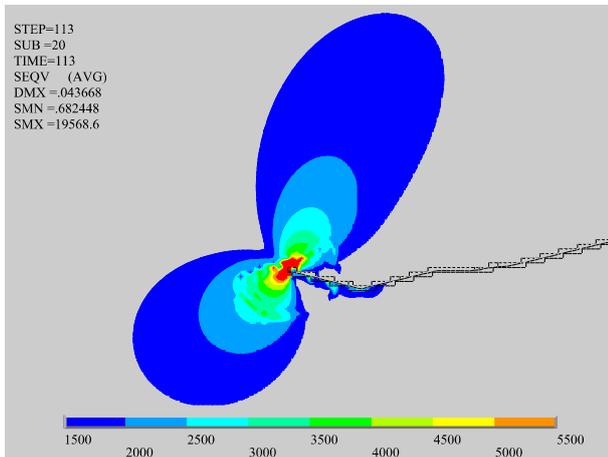


Figure 13: Equivalent von Mises Stress in MPa
Around the Crack Tip

The stress state of the crack tip can be described via the SIFs. Changes in SIFs during the simulation are shown in Figure 14.

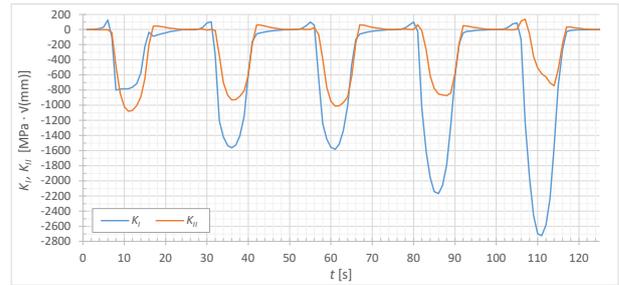


Figure 14: Stress Intensity Factors in Function of Time

It can be seen that the compressing Mode I crack loading is dominant, and the magnitude of it is increasing together with the crack length. As a consequence of it, the peak value of the equivalent stress intensity factor (K_{eqv}) is decreasing over time.

The equivalent stress intensity factor is calculated from SIFs. The positive values of the K_{eqv} are shown in Figure 15 in the function of the subsets which are proportional to time.

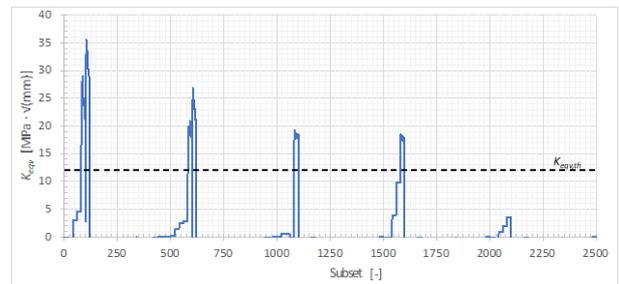


Figure 15: Equivalent Stress Intensity Factor in the
Function of Subsets. Jagged Line Shows the Threshold
Value Above Which the Crack Propagates.

After the fourth load cycle, the threshold value is not reached, therefore the crack stopped growing. This shows consistency with the reality, as the development of HC is driven by the trapped fluid (which was neglected in this study), by which hydro pressure tends to tear up the crack. In other words, the compression of the cracked surfaces is not that dangerous in the case of dry conditions from the crack propagation point of view.

The equivalent stress intensity factor is used directly in the lifetime estimations ($K_{eqv}=K$) based on Equation (1). The overall result of this is shown in Figure 16.

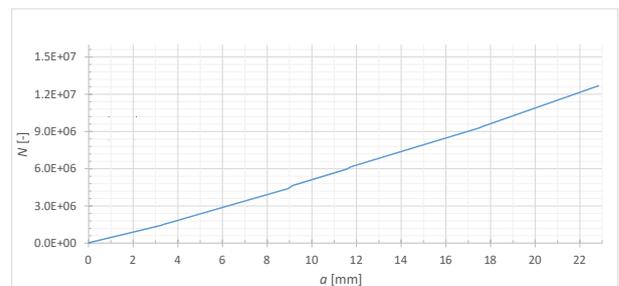


Figure 16: Life-Cycle Estimation Based on Paris' Law.
 N denotes the theoretical cycles, a is the crack length.

In Figure 17 a comparison of results of different examinations on HC can be seen. At the top, a non-destructive, liquid penetrant examination shows cracks on the rail surface. At the bottom, HC is shown in a cross-section of the rail under a microscope.

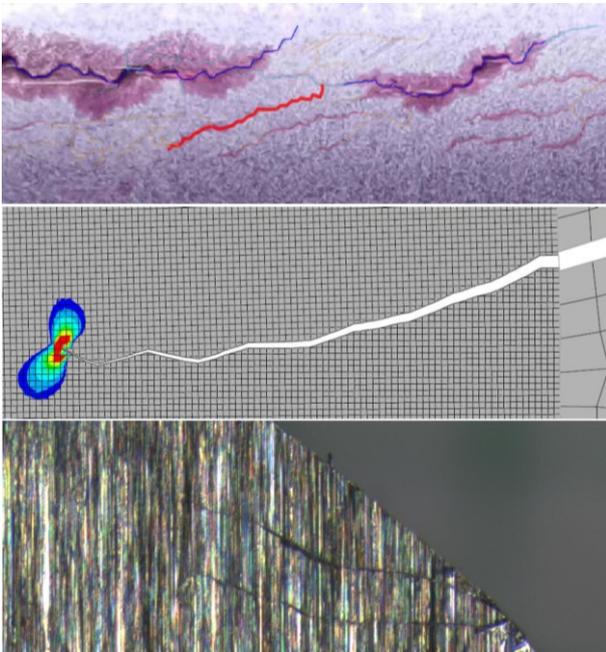


Figure 17: Comparison of the Laboratory and X-FEM Results

Even though these results are interpreted from different aspects and the numerical results are made with simplifications, the similarities are promising and prove that the phenomenon is worth to be examined from this kind of approach as well.

CONCLUSIONS

In conclusion, applying the extended finite element method in practice has met the expectations of the theoretical background. Based on the results of the above study this method has been proven to be adequate to examine such complex problems as the propagation of HC-type cracks. Using the extended finite element method arbitrary crack growth simulations can be implemented without the need of complicated meshing techniques and cracks can be examined in detail.

The software environment used (Ansys Mechanical APDL) is programmable, providing the possibility to define complex loading conditions and other specific parameters, which were essential in the study. The above environment provides the possibility to make fatigue-based calculations, so the method is able to predict lifetime as well.

However, it is important to mention that the method has numerous limitations in practice.

- Branching the crack is not possible, as such element-types are not supported.

- The contact of the cracked surfaces within an FE element is non-configurable, which might cause stability problems. By default, this is a frictionless, normal penalty.
- Loads and boundary conditions cannot be applied on cracked elements, which makes it difficult to take into account the effects of the trapped liquid.
- Regarding fatigue-based calculations, only Paris' law can be applied..
- Material model must be isotropic linear and elastic.

Furthermore, the functions required for this method are not available from the graphical user interface at the moment, the model has to be programmed by text, which makes the configurational process quite complex and time-consuming.

Results in this paper may not fully cover the real behavior of HCs, since the aim of the research is to gain experience on the X-FEM. Due to the above mentioned difficulties some significant simplifications were taken during the analysis, which may have an impact on the behavior of the HC:

- Cracks were examined in plane
- Only an extracted crack was examined, the possible effects on each other were neglected.
- The effect of the trapped and pressurized fluid was neglected

Besides above-mentioned simplifications, the validity of Paris' law might be also questionable, since the RCF phenomenon differs from the classical fatigue problems at some point. Regarding future research, results should be further validated

Theoretically, the method and tools allow the extension of the simulation to three-dimensional space. However, based on this study, such complex tasks as the examination of the HCs expanded in space can be managed in this way only taking into account considerable doubts. The study showed clearly how significant and important it is to support researchers and engineers with user-friendly software products.

Due to the fact that the presented model is coarse and significantly simplified, numerous recommendations can be identified for further development of the model. For instance, it would be important to model the trapped and pressurized fluid between the cracked surfaces. More cracks should be defined and examined simultaneously, as well as the input data should be determined more precisely, especially in case of the loads and material properties.

Since it is hard to find better numerical tools for this kind of problems, it is worth implementing the above-recommended developments.

The relevance of the FE analyses is that they could provide specific information for more accurate maintenance plans in the future.

REFERENCES

- Aglan H. A. 2011. "Fatigue Crack Growth and Fracture Behavior of Bainitic Rail Steels". Technical Report. Tuskegee University, Tuskegee. 57 p. Available: https://rosap.ntl.bts.gov/view/dot/23497/dot_23497_DS1.pdf
- Anderson, T. L. 2005. "Fracture Mechanics: Fundamentals and Applications". Boca Raton: CRC Press. 640 p. ISBN 978-1-4200-5821-5.
- Csizmazia F. and Horvát F. 2014. (In Hungarian) "A sínfej-hajszálrepedések műszaki és gazdasági alapú kezelése". *Sínek Világa*. 56 no. 5. pp. 13-21.
- Khoi A. R 2015. "Extended Finite Element Method: Theory and Applications". Pondicherry: SPi Publisher Services. 584 p. ISBN 978-1-118-45768-9.
- Zhuang Z. et. al. 2014. "Extended Finite Element Method". Waltham: Academic Press. 286 p. ISBN 9780124078567
- Zwierczyk P. T. 2015. "Thermal and stress analysis of a railway wheel-rail rolling-sliding contact". Ph.D. theses, Budapest University of Technology and Economics, Budapest.

AUTHOR BIOGRAPHIES



DÁNIEL BÓBIS is a final year M.Sc. student at the Budapest University of Technology and Economics, studying mechanical engineering. He started to investigate cracks in railway rails by participating at a Students' Scientific Conference in 2020 and is willing to continue the research. His email address is bobis.daniel@gt3.bme.hu.



PÉTER T. ZWIERCZYK is the deputy head of the Department of Machine and Product Design and assistant professor at Budapest University of Technology and Economics where he received his M.Sc. degree and then completed his Ph.D. in mechanical engineering. His main research field is the railway wheel-rail connection. He is member of the finite element modelling (FEM) research group. His e-mail address is z.peter@gt3.bme.hu.



TAMÁS MÁTÉ is a PhD student at the Budapest University of Technologies and Economics Department of Machine and Product Design where he received his M.Sc. degree in mechanical engineering. His research field is engaged to crack propagation in railway wheels and rails. His e-mail address is mate.tamas@gt3.bme.hu.

Modelling and Simulation of Cyber-Physical-Systems

MODELLING AGV OPERATION SIMULATION WITH LITHIUM BATTERIES IN MANUFACTURING

Ozan Yesilyurt
Marius Kurrle
Andreas Schlereth
Miriam Jäger

Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Nobelstraße 12, 70569 Stuttgart, Germany

Alexander Sauer

Fraunhofer Institute for Manufacturing Engineering and Automation IPA & Institute for Energy Efficiency in Production, EEP, University of Stuttgart
Nobelstraße 12, 70569 Stuttgart, Germany

E-Mail: ozan.yesilyurt@ipa.fraunhofer.de

KEYWORDS

Simulation modeling, manufacturing, automated guided vehicles

ABSTRACT

This paper describes the development of a production simulation model with automated guided vehicle (AGV) operation to prepare relevant production data validating an approach using AGV batteries as energy storage to reduce peak loads in a manufacturing company. First, the definition of AGV and the simulation modeling approach are introduced. Then, the systematic literature review methodology is described to explore relevant existing simulation models with AGV operation. With the help of this information, a simulation model is designed and developed. The last sections include the experiments performed with the simulation model, analysis, and the following results. The results show that the developed simulation can be used to generate data to evaluate the above-described approach in production.

INTRODUCTION

Industrial manufacturing faces significant challenges due to the increasing importance of sustainability and the rise of complexity in markets. First, unlike conventional transportation systems, battery-driven AGVs produce little emissions and have high energy efficiency. Still, reducing energy consumption is essential to reach a firm's own or external greenhouse gas reduction goals and, at the same time, to benefit economically from lower energy costs and higher energy security (Roesch et al. 2019). Compared to conventional vehicles, battery-driven AGVs need more time to charge. Consequently, energy consumption and charging time can be very volatile, which has to be considered in real-life use cases to manage a possible decrease in costs and emissions (Ma et al. 2021; Pfeilsticker et al. 2019; Zhu et al. 2018). Second, the steady increase in the complexity of markets poses a challenge on manufacturing companies, and their value creation networks increasingly facing new challenges (Bauernhansl et al. 2014). The manufacturing companies are forced to respond to such events with non-automated operations, high stock levels, and significant

lead times. As a result, decisions can be postponed, and remedial action can occur late. To solve these problems, simulations can forecast the planned changes in manufacturing because they acquire relevant results for practical implementations (VDI-Gesellschaft Produktion und Logistik 2014). With the new developments in the AGV field (Shihua Li et al. 2018), the relevance of using AGVs for companies is growing (Kunst 2018). To achieve the goal of optimal production despite the challenges of volatile power consumption and complex markets, simulation models can help companies to gather data for validating real-life use cases. This paper aims to generate realistic data for validating the approach of using AGV batteries as energy storage in production to minimize peak loads. Next, the problem statement and objective target of this paper are introduced.

PROBLEM STATEMENT AND OBJECTIVE TARGET

Manufacturing companies face a complex environment regarding current and future energy supply with their factories. The companies are charged for electricity based on two principles. One is for energy consumption and the other for peak demand. Besides the energy consumption contracts, the companies should agree with the electricity providers on contractual price models depending on the highest peak production load. The manufacturing companies experience these peak loads in their production and pay high amounts of money for generating them (Kurnik et al. 2017). Different concepts and applications of the energy storage systems such as stationary and electric vehicle (EV) batteries were studied and developed in manufacturing plants to minimize peak loads and enhance savings for the companies.

In this paper, only the electrical energy storage devices of the AGV are considered to achieve the same goals. To validate this approach, some data such as AGV and energy consumption data is needed from a company. After contacting the different companies, one manufacturing company agreed to cooperate. The cooperated company sent the energy consumption data for one year. However, the company does not possess any

AGV. Therefore, it is planned that a simulation should generate the required AGV data. The goal is to develop a simulation model according to the company's logistics processes that simulate a production line with AGV. The simulation model should generate availability, position, state of the charge status of the AGVs, and availability of the charging stations. In summary, the scope of this paper is to develop a simulation model to generate and analyze the data on the energy consumption of AGVs.

STATE OF THE ART

In the following, first, AGVs are defined. Then, the definition of a simulation and a recommended approach to develop a simulation are introduced. Last, the systematic literature review identifies the existing AGV simulation concepts.

Automated Guided Vehicles

AGVs are in-plant, ground-based material flow systems consisting of automatically controlled vehicles whose main task is to transport the material (VDI-Gesellschaft Produktion und Logistik 2005). The most crucial flexibility criteria for material handling technology include the ability to be integrated into an existing production environment, transport a wide variety of goods, and adapt to productivity fluctuations.

AGVs offer the highest degree of flexibility among all automated material handling technologies. Other advantages of AGVs include minimal infrastructure measures, use of existing paths, and the possibility of easy replacement with another vehicle or a conventional forklift. Concerning energy flexibility, AGVs can also serve as storage units that establish resilience by creating buffers (Roesch et al. 2019). This is particularly important for enabling companies to create energy flexibility, which is the base for trading in dynamic energy markets (Pfeilsticker et al. 2019). Because of the wide range of possible applications, there are almost no restrictions on the design of the AGV (Stegmüller & Zürn 2014; Ullrich & Albrecht 2019).

Simulation Modelling

This section introduces the definition of a simulation, a recommended approach to generate a simulation, and simulation modeling methods. A simulation represents a physical system and its related processes in a model. Its goal is to obtain transferable results for practical applications (VDI-Gesellschaft Produktion und Logistik 2014). The terms system and model are related to the term simulation. The system is a collection of components and their properties, which are connected by interdependencies (Hall and Fragen 1956). A model is an abstract image of a system (Eley 2012). If computers are used for the necessary calculations in the simulation, this is called a computer simulation. For this purpose, the model must be available in a mathematical-logical form and implemented in a computer program. These

computer programs are considered simulation tools (Eley 2012).

A simulation experiment is the reproduction of the behavior of a system with a model over a certain period of time (VDI-Gesellschaft Produktion und Logistik 2014). This particular period is called simulation time. On the other hand, the simulation time represents the time progressing in the existing system (Eley 2012). According to (VDI-Gesellschaft Produktion und Logistik 2014) the following approach is recommended to create a simulation:

1. formulation of problems,
2. test of simulation-worthiness,
3. formulation of targets,
4. data collection and data analysis,
5. modeling,
6. execution of simulation runs,
7. result analysis and
8. documentation.

The approach above is used to create a simulation model in this paper. The following section describes the systematic literature review to identify existing AGV simulation concepts.

RELATED WORK

The approach for finding related literature is based on the systematic approach according to (Jan Vom Brocke et al., 2009). It represents a patterned system for identifying and selecting relevant literature for the research field. (Jan Vom Brocke et al. 2009) identified five required steps for a literature review. In these steps, first, the research framework is established, second, the topic is conceptualized. Then, the literature review and literature analysis are conducted. Finally, a research agenda completes the research (Vom Brocke et al. 2009).

This literature search aims to find existing research approaches and implementations of AGV use in simulations. To conduct the literature search, search terms are created by using synonyms and closely related terms illustrated in Table 1.

Table 1: Search terms

Context Synonyms	Scope	Topic Synonyms
production	simulation	"AGV battery"
manufacturing	routing	
	modeling	

After that, Boolean operators are used to merge the search terms into a search string. Wild cards (*) are used to consider the plural of search terms and exclude forms of words in the literature search. The following search string is applied to find relevant literature in different databases.

("Production*" OR "Manufacturing*") AND ("simulation*" OR "routing*" OR "modeling*") AND ("AGV*" OR "AGV* battery*")

Five different databases are chosen to conduct the search string to find relevant papers. The databases' technical orientation and the search results' scientific relevance are considered in the selection of the databases. Figure 1 shows the selected databases and the methodology of the multi-stage filtering system of the search results.

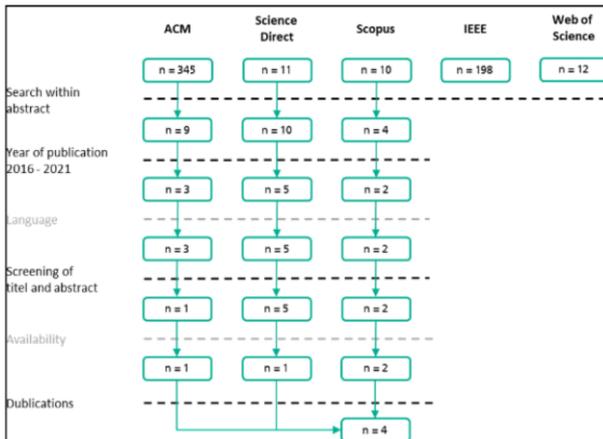


Figure 1: Results of the literature search with the multi-stage filtering system

Four relevant publications are found after conducting the literature search with the methodology of the multi-stage filtering system. In the next section, these publications are described in detail.

Existing AGV Simulation Concepts and Applications

The four different papers on AGV simulation concepts in production are introduced in this section.

The research paper by Ndiaye et al. 2016 introduces an AGV transportation system defined by a layout, several vehicles, several parking spaces, and a vehicle management policy. First, a simple formula determines the minimum required number of vehicles. Then, a discrete-event simulation model is used to evaluate different layouts and vehicle-dispatching policies. Initially, eight vehicles were required to meet the transportation demand, but with some optimizations, this number could be reduced to four vehicles. While these optimizations reduce the number of vehicles, they incur additional costs during the implementation phase. This means that savings achieved by reducing the number of vehicles are lost during the implementation phase in terms of software. The paper focuses on reducing the number of AGVs. However, generating the AGV and charging station data is required to verify the above-described approach.

Zhan et al. 2019 describe two-stage battery-charging strategies proposed for AGVs equipped with lithium-ion batteries to improve utilization. In stage 1, two routing decisions are developed. These are the nearest charging station (NCS) and the charging station with minimum delay (MDCS). In stage 2, the duration of each operation is reduced considering the charging characteristics of the lithium-ion battery. A real case is adopted to illustrate the applicability and effectiveness of the proposed approach. These methods help to improve manufacturing at a short-term capacity to meet the market demand. This paper shows that the loading strategy of MDCS performance is better than the loading strategy of NCS, in terms of AGV utilization and overall performance. This means that improving the utilization of AGV contributes to increasing the production of the system. The charging station with minimum delay methodology idea was taken from this paper and implemented in the new simulation. Nevertheless, the work of the researchers does not solve the problem statement described in the paper.

The research of Mousavi et al. 2017 used a fuzzy hybrid genetic algorithm (GA)-particle swarm optimization (PSO) algorithm with a comparison with three other algorithms (GA, PSO, and hybrid GA-PSO). Comparing the four algorithms results showed that the Fuzzy Hybrid-GA-PSO yields the lowest production time and AGV numbers. However, a difference was observed between the performance of Fuzzy-Hybrid-GA-PSO and Hybrid-GA-PSO. The only significant improvement over Hybrid-GA-PSO concerned the computation time. The AGV system simulation with Flexsim software proved the practicality of the developed model and the studied algorithms. The focus of this paper was to compare different optimization algorithms. Therefore, the results of this paper cannot be used for the described problem statement.

Mousavi et al. 2017 focused on multi-objective AGV scheduling in a flexible manufacturing system (FMS) using GA, PSO, and hybrid GA-PSO algorithms. A model for AGV task scheduling was developed. The comparison of the three algorithms shows that the hybrid GA-PSO provides the lowest production runtime and AGV numbers. It was found that after optimization, despite a slight increase in the total AGV running time (loaded and unloaded), reducing the idle time of the AGVs improved the operating efficiency of the AGVs. Consistent with the experimental results, FlexSim software has been used to prove the feasibility of the developed model and the suitability of the optimization algorithms for the scheduling problem. The developed model can be applied to any FMS. It can be applied to optimize the objectives separately or in a combinatorial way. Various algorithms are reviewed to enable and optimize the multi-scheduling of AGVs. This paper was out of scope because the problem statement could not be solved with the help of this paper.

The systematic literature review results show no existing AGV simulation to solve the above-described problem statement. Therefore, it is decided to develop a new AGV simulation to generate the required data that can be applied to a realistic situation. The next section describes the case study, including the concept of the simulated logistics process. After that, the simulation model is described.

CASE STUDY

Introduction of the company

The company for which the simulation was developed is a chemicals manufacturer in the synthetic leather and textile coating industry. The company's factory embraces 4000 m² and consists of a warehouse and two production lines. The company's warehouse is located between two production lines (with a diameter of 30 meters). Between 70-100 tons of products are transported per day. Two employees work three shifts per day in the company, and one employee transports a barrel with a forklift truck during one trip. With the help of this information, the assumptions of the concept are described in the next section.

Concept Description

The simulation model should be developed with the AGV operation. The following assumptions in the simulation model are made for the logistics processes of the company to realize a realistic simulation:

- The simulation duration is set to 1-year simulation time from 01.01.2021 to 31.12.2021, because, to calculate how many peak loads can be covered with AGV batteries, AGV data for one year should be generated.
- The transport processes of two products (chemicals) are considered.
- The interval of the production orders (min. in 3 minutes and max. in 27 minutes) is calculated for one year per day using the respective energy consumption data integrating it into the simulation.
- An AGV has an average speed of 1 m/s (approx. 4 km/h).
- The SoC limits of the AGV are predefined (20-90% for the lithium battery saving). If the lower SoC limit is exceeded, AGV should drive to the charging station, which has a minimum delay to charge the AGV battery.
- All AGVs have an initial value of 50% for the charge state of the battery.
- One AGV garage is located in the factory layout and consists of two AGVs. The employees have been replaced with AGVs.
- AGVs can transport the products to both production lines.
- Two charging stations are simulated so that AGV batteries are charged.

The data from Kuka KMP 1500 AGV (KUKA AG 2016) is used as the AGV data. It is shown in Table 2.

Table 2: Kuka AGV data

Feature	Value
maximum payload	1,500 kg
battery capacity	104 Ah (extended battery version)
charging current	52 A
charging voltage	96 V
battery energy density	9,984 Wh
charging time	2 hours (up to 100%)
driving consumption	13 A (min. 8 hours)

After the assumptions are determined, the simulation model is developed. The next chapter presents the model description.

MODEL DESCRIPTION

The simulator *Tecnomatix Plant Simulation* (SimPlan AG 2021) was chosen for its wide selection of objects used in production processes and pre-existing models using battery-driven AGVs. In addition, the simulation can be precisely adapted to the case study using custom-written *SimTalk* program language. The simulation model (see Figure 2) is based on Steffen Bangsow's training example which includes AGVs using tracks (Bangsow 2021).

At the beginning of the simulation, the AGVs launch in the garage below the transport routes. The number of spawned AGVs is defined through the variable *numAGV*, which, according to the concept, is set to two. The source of the products is connected to a buffer out that transfers the products to the AGVs. There are two products ("Part1", "Part2") in the simulated scenario that have the same characteristics but different destinations. According to the table *Workplan* that defines the routes of the products, one part must be transported to the first station on the right side, and, accordingly, the other part to the second station on the left side of the model. Therefore the AGVs operate on two separate paths which provide for collision avoidance. Besides the production sequence, *Workplan* defines the setup and processing times of the stations as well. These are set to avoid AGV queues before the source. If there would still occur, the following AGV would wait until the demanding one has received the product. This model is flexible and can be extended by adding more stations as long as the *Workplan* gets updated continuously.

The number of products generated by the source in each interval can be varied. The interval is set to one day in the case study. Thus, the daily production quantity can be

defined. This was implemented in the simulation by a generator that changes the time gap in which the source generates the products according to the table *ProductInterval*. The two endpoints of the transport routes are represented by two station objects in the simulation. In front of each station is a buffer that deals with the incoming products. In the simulation, the buffer serves the purpose of holding parts if components in the station cannot be processed in time. For the simulated case, a buffer is not crucial because capacities, setup, and processing times are calculated accordingly so that the products can be dispatched at the right time. However, since the simulation should be extendible in the future, the buffer was retained as a linking component. The buffer type is set to a queue. Consequently, products exit the buffer in the same order as they arrived (First-In-First-Out-principle).

All objects must be provided with the correct parameters to ensure that the correct production sequence can be created and automated later on. The simulation's most important object is the AGV itself. It is equipped with custom methods that define the logic of its operation. The most relevant method is *doJob*, which defines that an empty AGV should drive to the source to pick up a new part and afterward drive to the destination of the new job. The method also describes the procedure of battery charging. If the AGV's charge level goes below the defined battery reserve threshold, the vehicle checks the availability of the charging stations and drives to the next unoccupied station. This only occurs when the AGV is idle and therefore has no current job. After arrival, the AGV charges up to the specified capacity (90% of maximum capacity).

In *Plant Simulation*, properties of objects and process logic can be automated by so-called methods written in the programming language *SimTalk*. Methods can rely on tables and/or write new data in existing tables. The presented simulation consists of eight methods that ensure the correct execution of the aspired process. For this paper, the methods were adapted to the described conditions. In addition, the method *writeData* was created to save the data generated during the simulation in tables.

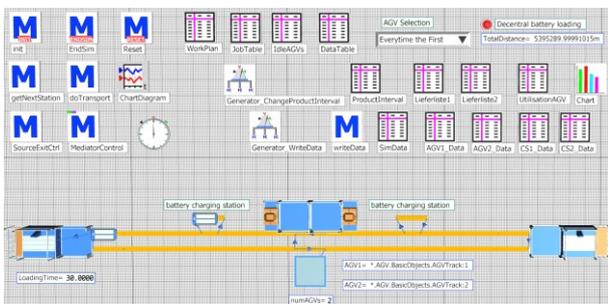


Figure 2: Production simulation model with AGV operation

RESULTS

The results of the developed simulation experiment are shown and interpreted in this section. The simulation time is set to one year for collecting data per minute so that the AGV batteries' approach can be validated. The simulation generated data for two AGVs and two charging stations. A total of 525,600 data was created per AGV and charging station in this simulation experiment. The simulated example charging station data is illustrated in Figure 3. They are current time, charging station ID, and charging station availability.

Identifier	CurrentTime	csiD	Measurement	csiavailability
0	2:01:00	1	ChargingStation	1
1	1:00:00.000	1	ChargingStation	1
2	2:00:00.000	1	ChargingStation	1
3	3:00:00.000	1	ChargingStation	1
4	4:00:00.000	1	ChargingStation	1
5	5:00:00.000	1	ChargingStation	1
6	6:00:00.000	1	ChargingStation	1
7	7:00:00.000	1	ChargingStation	1
8	8:00:00.000	1	ChargingStation	1
9	9:00:00.000	1	ChargingStation	1
10	10:00:00.000	1	ChargingStation	1

Figure 3: Charging station simulation data

The simulated example AGV data is shown in Figure 4. They are current time, AGV ID, AGV availability, state of charge of the AGV battery, AGV position on the X-axis, and the Y-axis, AGV velocity.

Identifier	CurrentTime	agvID	Measurement	agvavailability	soAGV	agvLatPosition	agvLongPosition	velochyAGV
0	2:01:00	1	AGV	0	50.00	13.50	0.00	1.00
1	1:00:00.000	1	AGV	0	49.88	0.80	2.00	1.00
2	2:00:00.000	1	AGV	0	49.82	0.80	2.00	1.00
3	3:00:00.000	1	AGV	0	49.76	0.80	2.00	1.00
4	4:00:00.000	1	AGV	0	49.69	0.80	2.00	1.00
5	5:00:00.000	1	AGV	0	49.63	0.80	2.00	1.00
6	6:00:00.000	1	AGV	0	49.56	0.80	2.00	1.00
7	7:00:00.000	1	AGV	0	49.50	0.80	2.00	1.00
8	8:00:00.000	1	AGV	0	49.44	0.80	2.00	1.00
9	9:00:00.000	1	AGV	2	49.28	24.20	2.00	1.00
10	10:00:00.000	1	AGV	0	49.20	27.20	2.00	1.00

Figure 4: AGV simulation data

After the simulation data was collected, the following AGV results were obtained and shown in Table 3. Both AGVs work approximately only 9% of their time. The simulation results show that they have over 81% idle time per year. The idle time allows the AGVs to be used not only as transport vehicles but also as energy storage in this company to reduce peak loads in production.

In future work, an economic analysis will be conducted to analyze different implementation strategies whether the AGVs' number should be reduced to save investment costs or whether the AGVs should be implemented in the simulation to reduce the peak load costs.

Table 3: AGV simulation results in one year

AGV status	AGV1	AGV2
working	8,9 %	8,5 %
idle	81,2 %	81,7 %
charging	9,9 %	9,9 %

Table 4 indicates the simulation results of the charging stations in one year. It is observed that the second charging station was nearly not used by AGVs.

Therefore, reducing the number of charging stations to one for this use case would be conceivable. However, it has to be investigated whether it is economical to have a second charging station when the AGVs discharge their batteries in peak times to support the company grid to reduce peak loads in production.

Table 4: Charging station simulation results in one year

CS status	CS 1	CS 2
working	20,5 %	0,2 %
idle	79,5 %	99,8 %

The simulation results of the AGVs on the highest (26.09.2021) and lowest energy consumption day (02.07.2021) are illustrated in Table 5. The results show that the AGVs on the day with the highest energy consumption work significantly longer than average working time and are charged more. For this day, it can be examined whether the AGVs have additional availability to minimize arising peak loads in the production. The results on the day with the least energy consumption highlight that the AGVs tend to have above-average idle time. It must be verified whether the charging times can be increased to enable the AGVs to pull more energy in off-peak times to use later.

Table 5: AGV simulation results during the highest and lowest energy consumption day

	Highest energy consumption day		Lowest energy consumption day	
	AGV1	AGV2	AGV1	AGV2
AGV Status				
working	15,9 %	17,2 %	6,7 %	5,5 %
idle	74,0 %	67,8 %	83,2 %	84,4 %
charging	10,1 %	15 %	10,1 %	10,1 %

The simulation results of Table 6 support the interpretation of Table 4. When it is cost-effective to utilize the second charging station for peak power reduction with AGV batteries, the charging station can be deployed. Otherwise, it is recommended to reduce the number of charging stations to one. However, (Weeber et al. 2020) show that machine availability is key to energy-efficient manufacturing. Because machine availability is dependent on the supply of materials by AGVs, in case of doubt, more charging stations than needed are appropriate.

Table 6: Charging station simulation results in the highest and lowest energy consumption day

CS status	highest energy consumption day		lowest energy consumption day	
	CS 1	CS 2	CS 1	CS 2
working	25,4 %	0 %	21,9 %	0 %
idle	74,6 %	100 %	78,1 %	100 %

CONCLUSION

A production simulation model has been developed with AGV operation to generate production data for the validation of the approach. It aims to apply AGV batteries as energy storage devices to mitigate peak loads in a manufacturing company. To realize this simulation model, first, the applied approach for the simulation modeling is presented. Then, the related work of different researchers is introduced, which created different AGV simulation concepts in production. After making sure that a simulation for the problem statement described has not yet been developed, a new concept of the simulation model with a case study is described. Considering the assumptions of the new simulation model concept, a simulation model is developed and presented in this paper. The simulation model results indicate that the generated data from the simulation model can be used to validate the described approach above. The results also show that the AGV's energy consumption can vary widely. In addition to the parameters of energy consumption and associated costs, the increasingly important energy flexibility issue must be considered. As a next step, the generated data will be entered into a software system to calculate whether using the AGV batteries as energy storage to minimize peak loads is cost-effective.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of the Kopernikus-Project "SynErgie" by the Federal Ministry of Education and Research of Germany (BMBF) and the project supervision by the project management organization Projektträger Jülich (PtJ).

REFERENCES

- Bauernhansl, T.; A. Schatz; and J. Jäger. 2014. "Komplexität bewirtschaften – Industrie 4.0 und die Folgen". *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb*, 109(5), 347–350. <https://doi.org/10.3139/104.111140>
- Eley, M. 2012. *Simulation in der Logistik: Einführung in Die Erstellung ereignisdiskreter Modelle unter Verwendung des Werkzeuges Plant Simulation* (1st ed.). Springer-Lehrbuch Ser. Springer Berlin / Heidelberg. <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=971208>
- Hall, A. D. and R.E. Fragen. 1956. "Definition of System". *General Systems*, 1(1), 18–28.
- Vom Brocke, J.; A. Simons; B. Niehaves; K. Riemer; and A. Clevén. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process". In *17th European Conference on Information Systems (ECIS)*. https://www.researchgate.net/publication/259440652_Reconstructing_the_Giant_On_the_Importance_of_Rigour_in_Documenting_the_Literature_Search_Process. Accessed 17.01.2022
- KUKA AG. 2016, June 9. "KUKA Mobile Plattform 1500". <https://www.kuka.com/de-de/produkte-leistungen/mobilit%C3%A4t/mobile-plattformen/kmp-1500>. Accessed 19.01.2022

- Kunst, A. 2018. *Relevanz von autonomen Transportsystemen in der Logistikbranche in Deutschland 2018* | Statista. <https://de.statista.com/prognosen/943349/expertenbefragung-zu-autonomen-transportsystemen-in-der-logistikbranche>. Accessed 17.01.2022
- Kurnik, C. W.; F. Stern; and J. Spencer. 2017. *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. <https://doi.org/10.2172/1406991>
- Ndiaye, M. A.; S. Dauzère-Pérès; C. Yugma; L. Rullière; and G. Lamiable. 2016. "Automated transportation of auxiliary resources in a semiconductor manufacturing facility". In *2016 Winter Simulation Conference (WSC)*. Arlington, Virginia
- Weeber, M.; J. Wanner; P. Schlegel; K. Birke; and A. Sauer. 2020. "Methodology for the Simulation based Energy Efficiency Assessment of Battery Cell Manufacturing Systems". <https://www.semanticscholar.org/paper/Methodology-for-the-Simulation-based-Energy-of-Cell-Weeber-Wanner/cc821d0c2a93616456f81d68a678d872c259b13a>
- Ma, N.; C. Zhou; and A. Stephen. 2021. "Simulation model and performance evaluation of battery-powered AGV systems in automated container terminals". *Simulation Modelling Practice and Theory*, 106, 102146. <https://doi.org/10.1016/j.simpat.2020.102146>
- Mousavi, M.; H.J. Yap; and S.N. Musa. 2017. "A Fuzzy Hybrid GA-PSO Algorithm for Multi-Objective AGV Scheduling in FMS". *International Journal of Simulation Modelling*, 16(1), 58–71. [https://doi.org/10.2507/IJSIMM16\(1\)5.368](https://doi.org/10.2507/IJSIMM16(1)5.368)
- Pfeilsticker, L.; E. Colangelo; and A. Sauer. 2019. "Energy Flexibility – A new Target Dimension in Manufacturing System Design and Operation". *Procedia Manufacturing*, 33, 51–58. <https://doi.org/10.1016/j.promfg.2019.04.008>
- Roesch, M.; D. Bauer; L. Haupt; R. Keller; T. Bauernhansl; G. Fridgen; G. Reinhart; and A. Sauer. 2019. "Harnessing the Full Potential of Industrial Demand-Side Flexibility: An End-to-End Approach Connecting Machines with Markets through Service-Oriented IT Platforms". *Applied Sciences*, 9(18), 3796. <https://doi.org/10.3390/app9183796>
- Li, S.; J. Yan; and L. Li. 2018. "Automated Guided Vehicle: the Direction of Intelligent Logistics". *Undefined*. <https://www.semanticscholar.org/paper/Automated-Guided-Vehicle%3A-the-Direction-of-Li-Yan/8b852eb668d7e5de0047ee6897c8176d695da4c2>
- SimPlan AG. 2021, September 2. *Simulation mit Plant Simulation*. <https://plant-simulation.de/>. Accessed 17.01.2022
- Stegmüller, D. and M. Zürn. 2014. "Wandlungsfähige Produktionssysteme für den Automobilbau der Zukunft". In T. Bauernhansl, M. ten Hompel, & B. Vogel-Heuser (Eds.), *Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration* (pp. 103–119). Springer Vieweg. https://doi.org/10.1007/978-3-658-04682-8_5
- Bangsow, S. 2021. *AGV modelling using tracks*. https://www.bangsow.eu/detail_en.php?id=851. Accessed 17.01.2022
- Ullrich, G. and T. Albrecht. 2019. *Fahrerlose Transportsysteme: Eine Fibel - mit Praxisanwendungen - zur Technik - für die Planung* (3., vollständig überarbeitete Auflage). Springer Vieweg. <https://doi.org/10.1007/978-3-658-27472-6>
- VDI-Gesellschaft Produktion und Logistik. *VDI 3633 Blatt 1:2014-12: Simulation of systems in materials handling, logistics and production - Fundamentals* (2014-12). <https://www.beuth.de/en/technical-rule/vdi-3633-blatt-1/149034959?webservice=vdin>. Accessed 14.01.2022
- VDI-Gesellschaft Produktion und Logistik. 2005-10. *VDI 2510:2005-10: Automated Guided Vehicle Systems (AGVS) (VDI 2510)*. Beuth Verlag. <https://www.beuth.de/de/technische-regel/vdi-2510/78228504>. Accessed 14.01.2022
- VDI-Gesellschaft Produktion und Logistik. 2014-12. *VDI 3633 Blatt 1:2014-12: Simulation of systems in materials handling, logistics and production - Fundamentals*. Beuth Verlag. <https://www.beuth.de/en/technical-rule/vdi-3633-blatt-1/149034959?webservice=vdin>. Accessed 14.01.2022
- Zhan, X.; L. Xu; J. Zhang; and A. Li. 2019. "Study on AGVs battery charging strategy for improving utilization". *Procedia CIRP*, 81, 558–563. <https://doi.org/10.1016/j.procir.2019.03.155>
- Zhu, Z.; Z. Gao; J. Zheng; and H. Du. 2018. "Charging Station Planning for Plug-In Electric Vehicles". *Journal of Systems Science and Systems Engineering*, 27(1), 24–45. <https://doi.org/10.1007/s11518-017-5352-6>

AUTHOR BIOGRAPHIES

OZAN YESILYURT got his bachelor's and master's degree in electrical engineering and information technology at the Technical University of Munich. Since 2016, he has been working as a research fellow in the Competence Center DigITools at Fraunhofer IPA.

MARIUS KURRLE obtained his Bachelor of Science in Technical Business Administration, focusing on logistics and production systems at the University of Stuttgart in 2020. He is currently pursuing his Master of Science in the same study program and works as a student assistant in the Competence Center DigITools at the Fraunhofer IPA.

ANDREAS SCHLERETH received his bachelor's and master's degree in industrial engineering and management from Karlsruhe Institute of Technology. Since 2018, he has been working as a research fellow in the Competence Center DigITools at the Fraunhofer IPA.

ALEXANDER SAUER is director of the Fraunhofer Institute for Manufacturing Engineering and Automation IPA and head of the Institute for Energy Efficiency in Production EEP at the University of Stuttgart. He specializes in resource-efficient production and digitization for sustainable production.

MIRIAM JÄGER is currently pursuing her Bachelor of Engineering in Industrial Engineering - Product Engineering at the University Furtwangen. During her internship at Fraunhofer IPA, she worked on literature research.

Digital Twins for Lighting Analysis: Literature Review, Challenges, and Research Opportunities

Muhammad Umair Hassan¹, Stavroula Angelaki⁴, Claudia Viviana Lopez Alfaro²,
Pierre Major³, Arne Styve¹, Saleh Abdel-Afou Alaliyat¹, Ibrahim A. Hameed¹,
Ute Besenecker⁴, Ricardo da Silva Torres^{1,5}

¹NTNU – Norwegian University of Science and Technology, Ålesund, Norway

²United Future Lab Norway, Ålesund, Norway

³AugmentCity, Ålesund, Norway

⁴KTH Royal Institute of Technology, Stockholm, Sweden

⁵Wageningen University & Research, Wageningen, The Netherlands

KEYWORDS

Digital Twin, lighting analysis, light modelling, simulation, lighting design, literature review.

ABSTRACT

Light modelling, simulation, and photometric calculations are by now common tasks in the lighting design process. These practices contribute to the definition and comparison of suitable layout arrangements and help predict the impact of lighting devices. Those tasks demand the use of tools to support the simulation of different scenarios, the analyses of their pros and cons according to different criteria (e.g., health and safety, perception, aesthetics, energy consumption, and costs), and decision-making. Digital twins have emerged as relevant technologies to simulate and visualize different “what-if” scenarios associated with physical entities and processes. In this paper, we investigate the state-of-the-art research concerning the use of digital twins for supporting lighting analysis in the urban/outdoor context. We also present and discuss challenges and research opportunities related to the design, implementation, and validation of digital twins in this domain.

INTRODUCTION

Light is vital for enabling vision and fostering well-being. In both indoor and outdoor environments, good lighting has been associated with health (physically and emotionally), productivity, and comfort [Mattsson et al., 2020], [Lowden and Kecklund, 2021]. Appropriate lighting design has also been associated with energy efficiency of the built environment [Han et al., 2019]. Another essential aspect is the role lighting plays in visual perception, including facilitating orientation and the aesthetic experience of objects and spaces [Besenecker et al., 2018]. Studies related to light pollution and its associated negative environmental impacts on wildlife have also gained a lot of attention recently [Straka et al., 2021], [Ditmer et al., 2021]. Determining the proper lighting qualities and configurations for a specific site is a

challenging task, as it often requires dealing with conflicting scenarios and perspectives: human and non-human needs, architectural considerations, energy efficiency, and intervention and maintenance costs. In addition, the needs for a lighting condition typically change over time (time of day, seasons, weather condition, user needs). In this sense, the use of suitable technologies to support modelling, designing, planning, and simulating different lighting scenarios is of paramount importance.

Digital Twins (DTs) are promising technologies for supporting the design, development, and simulation of lighting interventions, as well as of analyzing possible light effects considering different settings. According to [Stanford-Clark et al., 2019], a DT “*is a dynamic virtual representation of a physical object or system, usually across multiple stages of its lifecycle. It uses real-world data, simulation or machine learning models, combined with data analysis, to enable understanding, learning, and reasoning. Digital twins can be used to answer what-if questions and should be able to present the insights in an intuitive way.*” In fact, DT has emerged as an integrative technology for modelling and simulating data and processes, opening a plethora of opportunities for supporting the decision-making process based on the evaluation of realistic scenarios [Mylonas et al., 2021], [Pylidianis et al., 2021], [Verdouw et al., 2021], [Shahat et al., 2021], [Neethirajan and Kemp, 2021]. Successful usage of DT to support decision-making has been reported in several applications, including, for example, manufacturing [Leng et al., 2021], [Li et al., 2022], ship design and simulation [Fonseca and Gaspar, 2019], [Coraddu et al., 2019], agriculture and livestock farming [Pylidianis et al., 2021], [Verdouw et al., 2021], [Neethirajan and Kemp, 2021], and smart cities [Major et al., 2021], [Shahat et al., 2021], [Mylonas et al., 2021]. In those applications, DTs have been used for monitoring, diagnostics, and prognostics to optimize process performance and utilization. Often, sensory data are combined with historical data, human expertise, simulation, and data-driven learning to improve the outcome of prognos-

tics for addressing existing needs. The goal is to support the design, development, and implementation of effective, efficient, and sustainable operations.

How to construct suitable digital twins to support effective lighting analysis? That is the question that guides our work. In this paper, we review and summarize the recent literature aiming to characterize to what extent digital twins have been utilized to support lighting analysis. We also discuss challenges and research opportunities in the area. A special attention is given to the characterization of light modelling and lighting design in the context of urban applications.

In short, the contributions of this paper are two-fold:

1. It provides an overview of the state-of-the-art research involving digital twin light modelling for design. Those studies are characterized according to their service categories, technology readiness level (TRL), and a digital twin conceptual model; and
2. It discusses challenges and open research opportunities in this research field, especially in the context of urban applications.

BACKGROUND CONCEPTS

This section presents relevant concepts related to lighting analysis. This section also provides an overview of background concepts related to digital twins in the context of smart city applications.

Light Modelling, Lighting Design, Photometric Calculations

Lighting practice increasingly uses digital simulations in the design process using various tools and software packages. There are several different objectives to use software tools, such as, for example, aesthetic presentation/visualization, photometric light level calculation and predictions of brightness distribution, or daylight and indoor climate simulations. To illustrate the workflow, the most common options are summarised in Figure 1. The starting point is usually a computer-aided design (CAD) drawing of the environment or building that will be imported into a modelling/simulation software. The digital model is then used through the software choosing one of the following options:

1. **Visualization for Presentation:** Refinement of materials, colours, textures, and lighting to be rendered, in most cases, using a separately obtained plugin. This option results in images or animations often used to illustrate and present a design proposal.
2. **Calculation for Light Distribution and Predictions:** Performing photometric light calculations by either using the available plugin extensions for a specific software or by importing the model to a standalone lighting software. The result is not a photorealistic image but a three-dimensional representation of the space with calculated values regarding the light-dark distribution. Depending on the modelling

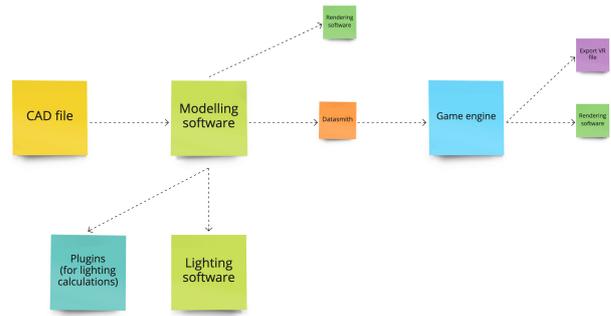


Fig. 1: Summary of options for lighting simulation.

software and the plugin used, the results are more optimized for daylight or electric light calculations. In the case of daylighting, tools have emerged to include feedback loops for algorithmic design [Xie and Sawyer, 2021].

3. **Interactive Simulation:** Importing the model into a game engine that allows the user to make changes to the lighting and, through real-time rendering, visualize the results. This path allows exploration of multiple lighting alternatives in a time-efficient way, and it allows for virtual reality (VR) immersion or photorealistic images using a rendering software.

While there is an abundance of tools offered for each of the three workflows, they typically cater to a specific part of a designer’s process, and are not always compatible. Therefore, it is difficult for a user to understand whether different, complementary software solutions can be used (conveniently) together throughout a design process [Kort et al., 2003]. The investigations and developments of efficient scenarios are necessary. At this point there appear to be two overall strands of workflow. One is related either to the production of models to generate visualizations for presentations of design options, or to the calculations of lighting-related quantities to assess the feasibility and code compliance of a lighting installation. The other, more recent development, is the use of game engines to create an immersive, real-time interactive virtual environment. There are advantages and disadvantages to using each, and in various projects, these alternatives are combined [Mackey and Roudsari, 2018]. Figure 2 illustrates some of these common tools and processes for lighting simulation.

The starting point for any process is the three-dimensional digital model. The model has to be designed with enough detail, at the scale desired, to provide conditions close to reality while being optimized for ease of use. Object materials, finishes, textures, and colours are important factors for the realistic reflections and distribution of light in a space. Most modelling software has integrated rendering engines, but the results are usually not of high aesthetic quality, and separately obtained plugins are used for

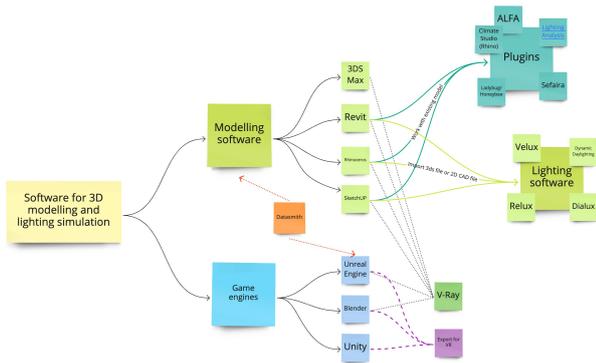


Fig. 2: Common tools and processes in lighting simulation.

the process of rendering a scene that looks photorealistic. Lighting criteria can be chosen, for example directionality/distribution, colour of light and the level of brightness, but these options are not necessarily based on real luminaires. To enable that, .ies data files [Committee, 2019] are available from light fixture manufacturers that can be imported into the software tools. These files are based on lab measurements and include data on the light output/amount and optical distribution of a specific actual luminaire. This information is the basis for any software used for photometric light calculation. It has to be noted that the .ies file data is based on white light perceived by humans; what is not available are information on spectrum and the related properties (that are apparent in an actual light installation). There have been efforts to address that shortcoming with the development of data file formats that include spectral information [Committee, 2020b], and some software packages have emerged that include aspects of spectral information. In addition to simulating man-made light sources, software packages also include daylight simulators, typically a combination of an artificial sun and sky.

To perform daylight and electric light analyses and calculations based on different lighting standards (e.g., EN standards or IES/ANSI standards) [Mandal et al., 2019], [Committee, 2020a], specific lighting engineering software packages have been developed. They are more accurate in terms of lighting calculations but do not offer the possibility of aesthetically and realistically illustrating a scene.

The use of game engines to develop VR and augmented reality (AR) simulations enable the evaluation of lighting interactively. The user can either import the model or build the scene within the game engine. Files can be created that are compatible with VR headsets to provide immersive experiences, and interactively test different settings for, e.g., brightness, distribution, or colour appearance of the lighting.

Questions about the opportunities and limitations

of simulating lighting in VR has entered research practice [Bellazzi et al., 2021]. Light interventions and proposals can often be developed quicker and cheaper in VR than in real-world mock-ups. Research projects have emerged to assess to what extent findings based on VR and AR are applicable in natural environments, comparing outcomes from both [Rockcastle et al., 2021], [Kort et al., 2003], [Chamilothori, 2019], and even to test how spectral data beyond vision could be visualized [Lalande et al., 2021]. In addition to using computer generated simulations in the lighting design process, there is emerging interest and technological capability to establish feedback connections for lighting between the virtual and the physical environment.

Digital Twins: from Manufacturing to Smart Cities

This section provides an overview of relevant research initiatives related to the design, use, and validation of digital twins in different applications.

Digital twins were originally developed to improve product life cycle in the manufacturing processes [Stark et al., 2019], [Schleich et al., 2017], [Haag and Anderl, 2018], [Leng et al., 2021], [Li et al., 2022]. One of the first definitions was coined by Michael Grieves who introduced a DT as “*virtual representation of what has been produced*” [Grieves, 2014]. In his vision, a DT would be essential to reduce costs and improve productivity, as well as support innovation and ensure the quality of products. This vision of digital twins were later redefined as replications of living as well as non-living entities to transmit data between physical and virtual objects [El Saddik, 2018]. More recent definitions, such as the one of [Stanford-Clark et al., 2019], emphasizes the dynamic aspect of DTs. According to this view, sensing technologies are used to monitor the state of physical entities, leading to regular updates of the states of their digital counterparts. [Stanford-Clark et al., 2019] also highlight the use of simulation and data-driven methods (e.g., machine learning) to support understanding, learning, and reasoning based on digital twins. In this sense, the use of visualization (e.g., of raw sensed data, and of simulation and machine learning results) and visual analytics tools is also relevant to support the assessment of different what-if scenarios.

In the context of manufacturing activities, [DeRoy et al., 2017], presented a building block of a digital twin for 3D printing machines by proposing (1) heat and material flow simulations, (2) grain structure and texture evaluation measures, and (3) residual stress and distortion calculations. The digital twins for the manufacturing industry are not there to replace real-time experiments but to reduce the number of such experiments needed to evaluate the product performance after it is built. Similarly, digital twins for lighting analysis could be explored to

assess different intervention options according to different criteria (e.g., comfort and energy use) possibly reducing costs associated with planning processes.

Future smart cities will depend on digital developing systems that can address the expanding computational demands of their population. [Shahat et al., 2021] provided an overview of the potentials of city digital twins. According to their review, the most common themes addressed in the literature include Data Management (Data processing, Interoperability, Software fusion, Open-source software); Visualization (Navigation, 3D real-time experience, Multi-spatial and temporal scales, Unified platform, Behavior modeling, Network dynamics, and Personalized information systems); Situational Awareness (Monitoring, Tracking, Localization, Face recognition, and Analysis); Planning and Prediction (Policy evaluation, Simulation, and What-if scenarios); and Integration and Collaboration (Multiple domains integration, Stakeholders’ participation, Citizens’ engagement, and Open platforms).

In this context, [Austin et al., 2020], for example, combined machine learning and semantic models to architect smart city digital twins. They designed an approach to provide city stakeholders with an enhanced level of situational awareness and decision-making support for urban infrastructure management. [Austin et al., 2020] envisaged the knowledge representation and reasoning (KRR) strategy to entail domain and meta domain data ontologies. To better facilitate the decision-making and management of cities, a knowledge discovery technique applied for the smart city digital twins was introduced by [Mohammadi and Taylor, 2020].

Digital twins for lighting analysis are expected to play an important role in smart city applications to foster sustainability (e.g., energy consumption, citizen well-being).

DIGITAL TWIN MODEL FOR LIGHTING ANALYSIS

This section describes the conceptual model adopted for describing digital twins employed for lighting analysis. Figure 3 depicts the main elements of this model. Lighting is a powerful tool for shaping the experience in built environments; it enables orientation, and can encourage behavioural patterns. Recent developments in sensing, control and solid state lighting technology provide the foundation to connect lighting analysis and control with digital twin applications. We adopt the conceptual model introduced by [Verdouw et al., 2021] for describing a digital twin application (elements on the left in the figure).

In the adopted conceptual model, the digital twins often are automatically updated (close to real time) according to sensed data. In the figure, the module Sensor collects information about the status of physical entities (e.g., lighting devices). Sensed data and

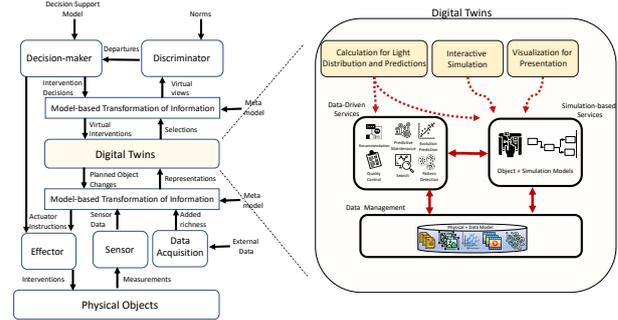


Fig. 3: Digital twin conceptual model for lighting analysis. Adapted from [Verdouw et al., 2021].

data obtained from external sources (e.g., weather condition) are transformed to suitable representations. Those representations are used to synchronize changes observed in physical objects into their digital counterparts. Data-driven and simulation results are compared to norms (e.g., lighting standards) by the discriminator module. The decision-making module utilizes this information (e.g., deviations from the norms) to “recommend” suitable intervention actions on digital and physical objects. Interventions on the physical world are performed by the Effector module.

The part highlighted on the right connects the the main areas of lighting analysis with data-driven and simulation-based services. In this model, we assume that all data needed by those services are available in data management systems implemented in the digital twin layer.

LITERATURE OVERVIEW

This section presents a literature review concerning the use of digital twins in the context of light analysis.

Methodology

The selection of literature studies followed typical methods employed in academic work: definition of terms and the search string, digital library sources, inclusion and exclusion criteria, as well as data analysis procedures.

Search string

Preliminary searches demonstrated that the term “digital twin” has not been extensively used in the context of studies in the area of light analysis. For this reason, we also considered the terms “simulation” and “visualization.” In the context of light analysis, we focused on studies dedicated to “light modelling,” “lighting design,” and “photometric analysis.” The final search string was defined as: (“digital twin” OR “simulation” OR (“visualization” OR “visualisation”)) AND (“indoor” OR “outdoor” OR “urban”) AND (“light modelling” OR “lighting design” OR “light design” OR “lighting analysis” OR “light analysis” OR “photometric analysis”).

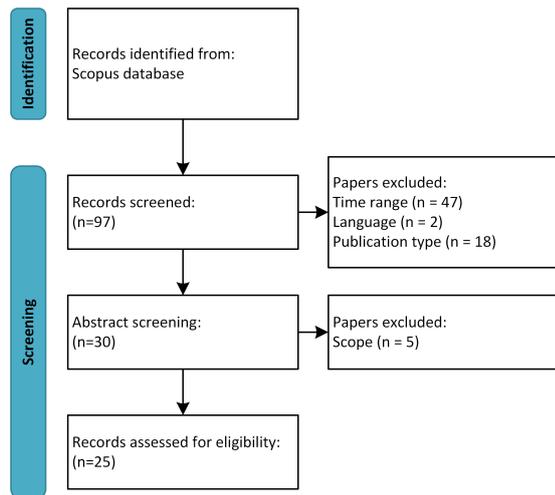


Fig. 4: Paper selection process.

Source

We searched for papers found in the Scopus database¹. Scopus is a well-known and widely-used digital library, which indexes relevant published literature in several domains. The search for relevant literature was performed on Jan. 27th, 2022.

Inclusion and exclusion criteria

As the main inclusion criterion, the paper should discuss research questions related to light analysis (light modelling, lighting design, and photometric analysis). Papers were excluded from our analysis if they did not have an abstract, were published as an abstract or extended abstract, were written in a language other than English, and were not accessible on the Web. Our analysis focused on recently published studies. Therefore, papers published before 2017 were also excluded.

Figure 4 summarizes the paper selection process. We ended up with 25 articles after applying the inclusion and exclusion criteria.

Data analysis procedures

The goal of our analysis is to assess to what extent digital twins have been utilized to support lighting analysis. We followed a similar methodology to the one employed by [Pylaniadis et al., 2021]. Services associated with the different usage of digital twins were classified according to the service categorization provided by [Tao et al., 2019], [Cimino et al., 2019]. The following categories were considered: real-time monitoring, energy consumption analysis, system failure analysis and prediction, optimization/update, behaviour analysis/user operation guide, technology integration, and virtual maintenance. The goal is to identify potential gaps concerning the use of digital twins. Next, we categorized the studied cases according to their technology readiness

¹<https://www.scopus.com> (As of Jan. 2022).

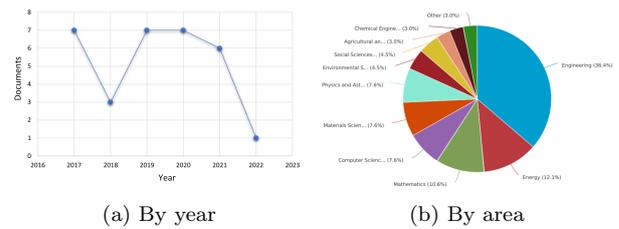


Fig. 5: Distribution of articles.

level (TRL) [Commission, 2014]. Three categories were considered: *concept* (levels 1–2), *prototype* (levels 3–6), and *deployed* (levels 7–9). The goal is to characterize the maturity of existing initiatives in the area. Finally, we assessed retrieved documents according to the digital twin conceptual model introduced in Figure 3.

Results & Discussion

The distribution of documents by year and subject area are available in Figures 5a and 5b, respectively. Except for 2018, around seven papers have been published each year, including topics related to digital twin and lighting analysis. We can observe that a high diversity in terms of subject areas, with a predominance of papers being published in Engineering (36.4%). Also, Energy (12.1%), Mathematics (10.8%), Computer Science (7.6%), Material Sciences (7.6%), and Physics (7.6%) are areas well covered in the studies.

Table I presents the papers retrieved in our search. As we can observe, most of the studies do not include explicitly terms related to “digital twin.” The diversity of areas illustrated in Figure 5b, is also evident by the wide scope of objectives and benefits handled by the different studies. Studies covered a wide variety of applications including light pollution [Tereci and Ozata, 2020], light comfort assessment [Michael et al., 2018], and energy efficiency [Selim et al., 2020], [Makaremi et al., 2019]. In another direction, some studies explored the use of immersive technologies in different lighting analysis [Natephra et al., 2017], [Krupinski, 2020]. As expected, we found studies covering both indoor and outdoor lighting analysis. Examples of typical physical twins range from office to urban environments.

Finally, regarding the TRL, most of the studies have been associated with low-level outcomes (concept and prototype levels). Those results suggest that further applied research towards the validation of the proposed methods in operational environments, including competitive commercial usage scenarios, seem to be needed.

Recall that most of the research initiatives are not associated with realistic digital twins (from the perception perspective). The literature also lacks studies focusing on supporting the creation of such digital twins (digital authoring tools).

TABLE I: Selected documents.

Title	Objective	Benefits	Physical twin	TRL
Development and Testing of a Modular Sunlight Transport System Employing Free-Form Mirrors [Whang et al., 2022]	To propose an inexpensive method of designing the modular natural-light lighting system in office areas based on the development of an auxiliary lighting system.	A low-cost and easy-to-design natural light illumination system for assisted lighting of office spaces.	Office Area	prototype
Modeling Natural Light Availability in Skyscraper Farms [Eaton et al., 2021]	To investigate the availability of natural light in skyscraper farms in dense urban formations compared with plant factories.	Skyscraper farming in dense urban environments.	Skyscraper Farm	concept
Daylighting simulation of a heritage building by comparing matrix methods and solar models [Saraiva et al., 2021]	To observe the optimal approaches to predict indoor lighting by simulating the low-cost monitoring devices.	Modelling and simulation of indoor lighting of heritage sites to support engineers and designers in their architect ideas.	Joanna Library, University of Coimbra, Portugal	prototype
Non-visual effects of office light environment: Field evaluation, model comparison, and spectral analysis [Zeng et al., 2021]	To study the non-visual effects of light in office environments.	Insights into developing simulation guidelines for office spaces.	Office Areas	prototype
The Effectiveness of Virtual Reality Simulation on the Qualitative Analysis of Lighting Design [Lee and Lee, 2021]	To analyze the effectiveness of virtual reality simulation in the landscape light design.	Use of VR-based simulation in urban lighting designs.	Urban Plaza	prototype
Optimization of luminaire layout to achieve a visually comfortable and energy efficient indoor general lighting scheme by Particle Swarm Optimization [Mandal et al., 2021]	To provide an optimized solution for lighting designs by determining the layouts of indoor areas.	The proposed method can be used in any indoor environment with any luminaire type.	Office Area	concept
Design Aids for Supplementary Lighting Design in India [Lala et al., 2020]	To design an aid tool that verifies the availability of sufficient daylight in office environments during the early stages of a building design.	Optimization of daylight design in indoor environments.	Office Area	prototype
Virtual Reality System and Scientific Visualisation for Smart Designing and Evaluating of Lighting [Krupinski, 2020]	To analyze the opportunities of using VR applications in light modelling for indoor and outdoor environments.	Use VR simulations to design and analyze lighting conditions in indoor and outdoor environments.	Indoor and Outdoor Areas	concept
Making the switch from task illumination to ambient illumination standards: Principles and practicalities, including energy implications [Cuttle, 2020]	To study the ambient illumination effect over task illumination in indoor environments.	Cost-efficient use of lighting energy in indoor areas.	Indoor Environment	concept
A New Trend for Indoor Lighting Design Based on A Hybrid Methodology [Selim et al., 2020]	To propose a method for installing energy-efficient lighting solutions in indoor environments.	High accuracy, faster, and economic model for light modelling in indoor environments.	Indoor Environment	concept
Analysis and Strategies for Zonal Lighting Design of Konya Mevlana Museum and Mevlana Culture Centre Axis [Tereci and Ozata, 2020]	To simulate the zone lighting and check its effect compared to single building lighting.	A model that prevents light pollution at night	Outdoor Environment	concept
Concerning the Concept of Light-Colour Arrangement of the Urban Environment in the Central Part of Tumen [Shechetkov et al., 2020]	To analyze the outdoor lighting by visualizing the general illumination plans of the city under design.	Visualizing the under development city designs for lighting analysis.	Outdoor Environment	concept
A Study on the Improvement of the Evaluation Scale of Discomfort Glare in Educational Facilities [Lee and Lee, 2019]	To provide an adequate natural lighting design to educational facilities that enhances students' health learning performance.	Improve the evaluation scale of discomfort glare in classrooms.	Educational Facility	prototype
Space Syntax Analysis Applied to Urban Street Lighting: Relations between Spatial Properties and Lighting Levels [Laccese et al., 2019]	To analyze the correlation between spatial properties of the urban roads and lighting levels which further helps design the lighting systems for urban areas.	To help city planners in designing light conditions on the roads.	Pontedera roads, Italy	prototype
A Novel Approach for the Definition of an Integrated Visual Quality Index for Residential Buildings [Zanon et al., 2019]	To propose an integrated index to evaluate the visual quality of a residential building.	Visual comfort of residents in a building.	Indoor Environment	prototype
Effects of surface reflectance and lighting design strategies on energy consumption and visual comfort [Makaremi et al., 2019]	To investigate the effects of different lighting design conditions in indoor environments that help in reducing the lighting energy use and improve the visual comfort level.	Indoor surface reflection helps in improving energy efficiency and visual comfort in indoor environments.	Indoor Environment	prototype
Daylighting performances and visual comfort in Le Corbusier's architecture. The daylighting analysis of seven unrealized residential buildings [Iommi, 2019]	To discover and evaluate the daylighting performance in unrealized housing projects designed by Le Corbusier.	A simulation of daylighting to assess and compare performance of different projects.	Indoor Environment	prototype
The impact of room surface reflectance on corneal illuminance and rule-of-thumb equations for circadian lighting design [Cai et al., 2018]	To assess the impact of room surface reflectance on corneal illuminance.	Proposal of rule-of-thumb equations for circadian lighting design.	Indoor Environment	concept
Environmental assessment of an integrated adaptive system for the improvement of indoor visual comfort of existing buildings [Michael et al., 2018]	To propose a prosthetic renovation model that can be integrated into the building to improve building inhabitants' visual comfort.	A new prosthetic moveable lighting design for optimizing natural light performance, reducing glare issues, and minimizing the energy consumption in residential buildings.	Indoor Environment	prototype
Integrating building information modeling and virtual reality development engines for building indoor lighting design [Natephra et al., 2017]	To propose an authoring tool by integrating the Building Information Management (BIM) tool and Unreal game engine to design and analyze lighting conditions.	Use VR-based simulations to visualize real-time lighting conditions to facilitate users to simulate various lighting scenarios.	Indoor Environment	prototype
Application of a coelostat daylighting system for energy savings and enhancement of indoor illumination: A case study under clear-sky conditions [Oh et al., 2017]	To examine the effectiveness of coelostat daylighting systems for enhancing the indoor visual environment of offices.	The proposed system may be used to analyze daylighting performance of buildings.	Indoor Environment	concept
Energy Optimized Envelope for Cold Climate Indoor Agricultural Growing Center [Hachem-Vermette and MacGregor, 2017]	To examine the lighting conditions based on the envelope model for indoor farming.	Development of low-energy indoor farming environment for the cold climate zones.	Indoor Agriculture Facility	prototype
Integrated Lighting Efficiency Analysis in Large Industrial Buildings to Enhance Indoor Environmental Quality [Katunský et al., 2017]	To investigate and analyze the efficiency and adequacy of integrated lighting in industrial buildings for enhancing indoor environmental quality.	Integrated lighting solution for the enhancement of visual comfort in industrial facilities.	Indoor Industrial Environment	prototype
Particle Swarm Optimization for Outdoor Lighting Design [Castillo-Martinez et al., 2017]	To propose a new particle swarm optimization algorithm to facilitate light designers in configuring the best optimal and energy-efficient lighting parameters.	Use of DIALux software for creating energy-efficient outdoor lighting simulations.	Outdoor Environment	prototype
The Impact of Shading Type and Azimuth Orientation on the Daylighting in a Classroom-Focusing on Effectiveness of Façade Shading, Comparing the Results of DA and UDI [Lee et al., 2017]	To evaluate different patterns and characteristics of varying façade shading types and their impact on daylight metrics.	Insights to design better façade shading for indoor environments that facilitates human comfort.	Educational Facility	concept

TABLE II: Distribution of documents according to different service categories.

Service Category	Documents
Real-time monitoring	[Whang et al., 2022], [Eaton et al., 2021], [Saraiva et al., 2021], [Zeng et al., 2021], [Tereci and Ozata, 2020], [Cai et al., 2018], [Lee and Lee, 2019], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017]
System failure analysis	[Zeng et al., 2021]
Optimization/update	[Mandal et al., 2021], [Michael et al., 2018]
Technology integration	[Zanon et al., 2019], [Lee et al., 2017]
Energy consumption analysis	[Cuttle, 2020], [Selim et al., 2020], [Makaremi et al., 2019], [Tai and Jang, 2018], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017], [Hachem-Vermette and MacGregor, 2017], [Katunský et al., 2017], [Castillo-Martinez et al., 2017]

Table II categorizes the different studies according to the different services addressed in the paper. As it can be observed, there are studies related to all service categories. It is worth mentioning the high number of studies covering real-time monitoring and energy consumption analysis.

Table III depicts the categorization of documents according to the digital twin conceptual model presented in Figure 3. We evaluated 25 documents based

on seven control functions for light modelling: (i) sensor, (ii) discriminator, (iii) decision-maker, (iv) effector, (v) calculation for light distribution and predictions, (vi) interactive simulation, and (vii) visualization for presentation. We found that most of the works have measured the actual performance of the lighting conditions, whereas about half of the selected documents have compared the actual performance of lighting conditions with the desired norms. The works by [Mandal et al., 2021], and [Natephra et al., 2017] selected the appropriate interventions for decision-making by providing the error ratios while [Mandal et al., 2021], [Selim et al., 2020], and [Natephra et al., 2017] implemented the effector function for correcting the light modelling in the environment. Most of the studies were found to be calculating the light distributions and predicting the lighting conditions for the environment. Also, interactive simulations are provided by most of the works. The visualization for presentation is also illustrated in over half of the evaluated studies.

TABLE III: Categorization of documents according to the digital twin conceptual model.

Functions	Documents
Sensor	[Eaton et al., 2021], [Saraiva et al., 2021], [Zeng et al., 2021], [Mandal et al., 2021], [Lala et al., 2020], [Krupinski, 2020], [Cuttle, 2020], [Selim et al., 2020], [Lee and Lee, 2019], [Leccese et al., 2019], [Zanon et al., 2019], [Makaremi et al., 2019], [Iommi, 2019], [Cai et al., 2018], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017], [Katunský et al., 2017], [Castillo-Martinez et al., 2017], [Lee et al., 2017]
Discriminator	[Whang et al., 2022], [Eaton et al., 2021], [Saraiva et al., 2021], [Mandal et al., 2021], [Lala et al., 2020], [Krupinski, 2020], [Cuttle, 2020], [Leccese et al., 2019], [Natephra et al., 2017], [Lee et al., 2017]
Decision Maker	[Mandal et al., 2021], [Natephra et al., 2017]
Effector	[Mandal et al., 2021], [Selim et al., 2020], [Natephra et al., 2017]
Calculations for Light Distribution and Predictions	[Whang et al., 2022], [Eaton et al., 2021], [Saraiva et al., 2021], [Mandal et al., 2021], [Cuttle, 2020], [Selim et al., 2020], [Shehepetkov et al., 2020], [Zanon et al., 2019], [Makaremi et al., 2019], [Cai et al., 2018], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017], [Hachem-Vermette and MacGregor, 2017], [Katunský et al., 2017], [Castillo-Martinez et al., 2017], [Lee et al., 2017]
Interactive Simulation	[Whang et al., 2022], [Eaton et al., 2021], [Saraiva et al., 2021], [Lee and Lee, 2021], [Mandal et al., 2021], [Lala et al., 2020], [Krupinski, 2020], [Cuttle, 2020], [Tereci and Ozata, 2020], [Lee and Lee, 2019], [Leccese et al., 2019], [Zanon et al., 2019], [Iommi, 2019], [Cai et al., 2018], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017], [Hachem-Vermette and MacGregor, 2017], [Katunský et al., 2017], [Castillo-Martinez et al., 2017], [Lee et al., 2017]
Visualization for Presentation	[Whang et al., 2022], [Saraiva et al., 2021], [Zeng et al., 2021], [Lala et al., 2020], [Cuttle, 2020], [Shehepetkov et al., 2020], [Lee and Lee, 2019], [Leccese et al., 2019], [Zanon et al., 2019], [Iommi, 2019], [Cai et al., 2018], [Michael et al., 2018], [Natephra et al., 2017], [Oh et al., 2017], [Katunský et al., 2017]

APPLICATIONS

As we could observe, most of the recent literature has not been associated with high-level TRL (e.g., “deployed”). This section presents examples of real-world applications that could benefit from the utilization of digital twins for lighting analysis.

Digital Twins for Urban Lighting Planning

Cities have to be livable and planned for circadian (day and night) and seasonal activities. To increase the use of green areas, local governments need to invest in infrastructures that improve accessibility and attractiveness. Artificial light infrastructure is one of the most common investments. When designed well, its benefits include the increase of visits (number of people and hours of use), reduction in criminality and accidents, and a positive impact in the togetherness of the communities surrounding these areas [Feng and Murray, 2018].

For holistic urban planning, i.e., cross-silo and cross-disciplinary analysis of city plans, one has to consider the cities in dimmed light environments. In this scenario, to involve and engage all the stakeholders (e.g., the decision makers, neighbors, project managers, and nature preservation activists) in the planning process, the use of physical and/or virtual participatory co-creation immersive arenas (digital twins), as the one depicted in Figure 6, would be of paramount importance. The goal would be to raise awareness, identify and anticipate conflicts at early planning stages, tentatively find pathways to reach solutions and consensus, and in the long run save time and avoid delays in the implementation of interventions.

Digital Twins for Walkability and Wildlife Impact Assessment

In the last years, digital twins have expanded their application to other areas of human activities, such



Fig. 6: Stakeholders in a co-creation arena for lighting intervention planning in Ålesund, Norway.

as urban planning and governance. Digital applications and 3D models complement the analytical tools of geographical information systems (GIS) by improving the communication of results and the interaction among stakeholders. We claim that the use of twin technologies associated with GIS can be a “game changer” in the design of artificial lighting infrastructure in green areas. Through these tools it would be possible to have an integrative approach to evaluate the impacts of different design scenarios of artificial lighting infrastructure, including the assessment of simulations that accommodate the (possibly conflicting) needs and interests of different stakeholders (human and non-human).

Artificial lighting infrastructure in parks and outdoor areas are important, especially for countries that have prolonged dark seasons (e.g., Nordic countries) and have a strong tradition of outdoor activities throughout the year. Proper lighting in parks, playgrounds, and green areas would allow people of all ages to use them and interact in the society for longer periods and with a greater sense of security. Good lighting infrastructure also plays a relevant role in the quality of the public spaces in urban areas and thus impacts the wellness of its citizens. Urban planners are today oriented to create spaces that promote the health of its residents and thus different indices such as “walkability” are used in the planning process. Walkability is an indicator that integrates different factors in order to evaluate how friendly a region is to walking activities [Beiler and Phillips, 2016]. Its relevance relies on the fact that residents of more walkable areas often have better health and thus quality of life [Pineo and Rydin, 2018]. New lighting technologies can help to improve walkability of an area by increasing availability (e.g., hours of use), improving security and enhancing the appreciation of citizens.

Unfortunately, there is also strong evidence of negative impacts in the use of artificial lighting in green areas [Falchi et al., 2011]. The most obvious is the increase in energy resources that leads to the economic cost for local governments. A consequence is also the increase in CO2 emissions [Nejat et al., 2015] that is

today one of the main concerns in sustainable policies around the world. Negative impacts of artificial lighting have also been observed on the physiology of animals, including humans [Chepesiuk, 2009]. The impact of light pollution on wildlife and biodiversity has become an important concern during the last decades [Longcore and Rich, 2004]. Several studies have shown their impact on animal behaviour, physiology, and use of space [Gaston et al., 2013]. However there is still little information on its impact and thus possible ways to its mitigation [Gaston et al., 2012].

CHALLENGES AND RESEARCH OPPORTUNITIES

This section presents relevant challenges related to the design, development, and use of digital twins for lighting analysis, considering their use in real-world applications (high-level TRLs).

Realistic Digital Twins: Typical urban planning, design and architecture projects create appealing static renderings or non-interactive 3D animation of a prospective built environment in its spacial urban context to present design options to clients and stakeholders. Advanced light modeling is used in this case typically for aesthetic, artistic purposes and might or might not be photometrically accurate. Creating dynamic interactive virtual environments for real-time experience is a manual intensive work, simulating lighting that is close to reality in such dynamic environments adds additional complexity. The purpose is to truthfully represent reality with a trade off on model size impacting the visual quality. This requires the production of 3D models that are a truthful representation of reality for each real world object and its virtual avatar or asset in the scene. The challenge is to be as realistic and complex as necessary while reducing as much as possible the number of vertices needed for representing each asset, to save memory and bandwidth, and increase processing speed during real-time visualisation. The assets (built, road, and green infrastructure) are often mapped by cadastral services or GIS authorities with multispectral spectrometry and light detection and ranging sensor (LiDAR) to identify tree canopy, vegetation type, and building category. Moreover, commercial game engine platforms, such as Unity and Unreal, offer a wide range of advanced lighting and shading features, allowing scientists, developers, and urban planners to represent the scenes for realistic zenithal and azimuthal settings linked to the latitude and longitude of the city, day of year, and hour of day.

Authoring tools: The literature lacks studies focusing on the creation of a 3D authoring tool for digital twins of urban scenes. Some studies have been conducted earlier with the focus on using authoring tools in other applications. [Hwang and Park, 2018], for example, developed a virtual reality-based authoring tool for smart factories to support smart

manufacturing processes. An educational authoring tool to create virtual labs was proposed by [Ververidis et al., 2019]. [Pan and Mitchell, 2020], in turn, proposed a collaborative mixed reality authoring tool for character animations, while [Gordillo et al., 2017] developed an authoring tool to create useful and reusable learning objects. Those studies could be explored as starting points for future research.

A hybrid approach would be to first create a library of assets based on artificial intelligence (AI); for instance, based on machine learning algorithms employed for 3D reconstruction and retrieval. Later, an AI algorithm would place the assets in a scene. Interaction controls could be provided for customizing the scene according to users' needs.

CONCLUSIONS

In this paper, we have characterized recent initiatives towards the design, implementation, and validation of digital twins in the context of lighting analysis. Among the main findings, we could observe that a wide variety of applications have been addressed, ranging from the use of presentation strategies to support the analysis of different lighting design possibilities to the development of interactive simulators for guiding the definition of the most efficient lighting configurations for indoor and outdoor environments. In fact, the predominant service categories covered by the recent literature refers to real-time monitoring and energy consumption analysis. Most of the outlined research has been focused on low TRLs as well. This paper has also presented and discussed relevant applications and research challenges related to the use of digital twins for lighting analysis. We especially advocate for the use of such technologies in the context of lighting intervention planning aiming at urban operations. This is especially of paramount importance nowadays, given the impact of lighting infrastructure on citizens and wildlife.

Future work will focus on extending our survey to compare and characterize existing tools and libraries to support lighting analysis. Finally, we have been involved with the development of digital twin authoring tools to support lighting intervention planning and analysis.²

ACKNOWLEDGEMENTS

This work has been conducted in the context of the NORDARK project, funded by NordForsk. This work was also partially funded by the NFR SMART-PLAN (310056) and Twin Fjord (320627) projects.

REFERENCES

- [Austin et al., 2020] Austin, M., Delgoshai, P., Coelho, M., and Heidarinejad, M. (2020). Architecting smart city digital twins: Combined semantic model and machine learning approach. *Journal of Management in Engineering*, 36(4):04020026.

²<https://nordark.org/> (As of Feb. 2022).

- [Beiler and Phillips, 2016] Beiler, M. R. O. and Phillips, B. (2016). Prioritizing pedestrian corridors using walkability performance metrics and decision analysis. *Journal of Urban Planning and Development*, 142(1):04015009.
- [Bellazzi et al., 2021] Bellazzi, A., Bellia, L., Chinazzo, G., Corbisiero, F., D’Agostino, P., Devitofrancesco, A., Fragliasso, F., Ghellere, M., Megale, V., and Salamone, F. (2021). Virtual reality for assessing visual quality and lighting perception: A systematic review. *Building and Environment*, page 108674.
- [Besenecker et al., 2018] Besenecker, U., Krueger, T., Pearson, Z., Bullough, J. D., and Gerlach, R. (2018). The experience of equivalent luminous colors at architectural scale. *Cultura e Scienza del Colore - Color Culture and Science*, 10:13–20.
- [Cai et al., 2018] Cai, W., Yue, J., Dai, Q., Hao, L., Lin, Y., Shi, W., Huang, Y., and Wei, M. (2018). The impact of room surface reflectance on corneal illuminance and rule-of-thumb equations for circadian lighting design. *Building and Environment*, 141:288–297.
- [Castillo-Martinez et al., 2017] Castillo-Martinez, A., Almagro, J., Gutierrez-Escobar, A., Del Corte, A., Castillo-Sequera, J., Gómez-Pulido, J., and Gutiérrez-Martínez, J.-M. (2017). Particle swarm optimization for outdoor lighting design. *Energies*, 10(1).
- [Chamilothori, 2019] Chamilothori, K. (2019). *Perceptual effects of daylight patterns in architecture*. PhD thesis, École Polytechnique Fédérale de Lausanne.
- [Chepesiuk, 2009] Chepesiuk, R. (2009). Missing the dark: health effects of light pollution. *Environmental Health Perspectives*, 117(1):A20–A27. 19165374[pmid].
- [Cimino et al., 2019] Cimino, C., Negri, E., and Fumagalli, L. (2019). Review of digital twin applications in manufacturing. *Computers in Industry*, 113:103130.
- [Commission, 2014] Commission, E. (2014). Technology readiness levels (trl); extract from part 19 - commission decision c(2014)4995. Technical report, European Commission.
- [Committee, 2019] Committee, T. I. C. (2019). Ies standard file format for the electronic transfer of photometric data and related information. Technical report, Illuminating Engineering Society.
- [Committee, 2020a] Committee, T. I. C. (2020a). Directional positioning of photometric data. Technical report, Illuminating Engineering Society.
- [Committee, 2020b] Committee, T. I. C. (2020b). Ies standard format for the electronic transfer of spectral data. Technical report, Illuminating Engineering Society.
- [Coraddu et al., 2019] Coraddu, A., Oneto, L., Baldi, F., Cipollini, F., Atlar, M., and Savio, S. (2019). Data-driven ship digital twin for estimating the speed loss caused by the marine fouling. *Ocean Engineering*, 186:106063.
- [Cuttle, 2020] Cuttle, C. (2020). Making the switch from task illumination to ambient illumination standards: Principles and practicalities, including energy implications. *Lighting Research and Technology*, 52(4):455–471.
- [DebRoy et al., 2017] DebRoy, T., Zhang, W., Turner, J., and Babu, S. (2017). Building digital twins of 3d printing machines. *Scripta Materialia*, 135:119–124.
- [Ditmer et al., 2021] Ditmer, M. A., Stoner, D. C., and Carter, N. H. (2021). Estimating the loss and fragmentation of dark environments in mammal ranges from light pollution. *Biological Conservation*, 257:109135.
- [Eaton et al., 2021] Eaton, M., Harbick, K., Shelford, T., and Mattson, N. (2021). Modeling natural light availability in skyscraper farms. *Agronomy*, 11(9).
- [El Saddik, 2018] El Saddik, A. (2018). Digital twins: The convergence of multimedia technologies. *IEEE MultiMedia*, 25(2):87–92.
- [Falchi et al., 2011] Falchi, F., Cinzano, P., Elvidge, C. D., Keith, D. M., and Haim, A. (2011). Limiting the impact of light pollution on human health, environment and stellar visibility. *Journal of Environmental Management*, 92(10):2714–2722.
- [Feng and Murray, 2018] Feng, X. and Murray, A. T. (2018). Spatial analytics for enhancing street light coverage of public spaces. *LEUKOS*, 14(1):13–23.
- [Fonseca and Gaspar, 2019] Fonseca, I. A. and Gaspar, H. M. (2019). A prime on web-based simulation. In *Proceedings of the 33rd International ECMS Conference on Modelling and Simulation, ECMS 2019 Caserta, Italy, June 11-14, 2019*, pages 23–29.
- [Gaston et al., 2013] Gaston, K. J., Bennie, J., Davies, T. W., and Hopkins, J. (2013). The ecological impacts of nighttime light pollution: a mechanistic appraisal. *Biological Reviews*, 88(4):912–927.
- [Gaston et al., 2012] Gaston, K. J., Davies, T. W., Bennie, J., and Hopkins, J. (2012). Review: Reducing the ecological consequences of night-time light pollution: options and developments. *Journal of Applied Ecology*, 49(6):1256–1266.
- [Gordillo et al., 2017] Gordillo, A., Barra, E., and Quemada, J. (2017). An easy to use open source authoring tool to create effective and reusable learning objects. *Comput. Appl. Eng. Educ.*, 25(2):188–199.
- [Grieves, 2014] Grieves, M. (2014). Digital twin: Manufacturing excellence through virtual factory replication.
- [Haag and Anderl, 2018] Haag, S. and Anderl, R. (2018). Digital twin – proof of concept. *Manufacturing Letters*, 15:64–66. Industry 4.0 and Smart Manufacturing.
- [Hachem-Vermette and MacGregor, 2017] Hachem-Vermette, C. and MacGregor, A. (2017). Energy optimized envelope for cold climate indoor agricultural growing center. *Buildings*, 7(3).
- [Han et al., 2019] Han, H. J., Mehmood, M. U., Ahmed, R., Kim, Y., Dutton, S., Lim, S. H., and Chun, W. (2019). An advanced lighting system combining solar and an artificial light source for constant illumination and energy saving in buildings. *Energy and Buildings*, 203:109404.
- [Hwang and Park, 2018] Hwang, Y. and Park, S. (2018). Development of vr based authoring tool for smart factory. In Duy, V. H., Dao, T. T., Zelinka, I., Kim, S. B., and Phuong, T. T., editors, *AETA 2017 - Recent Advances in Electrical Engineering and Related Sciences: Theory and Application*, pages 1078–1087, Cham. Springer International Publishing.
- [Iommi, 2019] Iommi, M. (2019). Daylighting performances and visual comfort in le corbusier’s architecture. the daylighting analysis of seven unrealized residential buildings. *Energy and Buildings*, 184:242–263.
- [Katunský et al., 2017] Katunský, D., Dolníková, E., and Doroudiani, S. (2017). Integrated lighting efficiency analysis in large industrial buildings to enhance indoor environmental quality. *Buildings*, 7(2).
- [Kort et al., 2003] Kort, Y. A. d., Ijsselstein, W. A., Kooijman, J., and Schuurmans, Y. (2003). Virtual laboratories: Comparability of real and virtual environments for environmental psychology. *Presence: Teleoperators & Virtual Environments*, 12(4):360–373.
- [Krupinski, 2020] Krupinski, R. (2020). Virtual reality system and scientific visualisation for smart designing and evaluating of lighting. *Energies*, 13(20).
- [Lala et al., 2020] Lala, S., Jain, K., Kumar, A., Kumar, A., Rajasekar, E., and Kulkarni, K. (2020). Design aids for supplementary lighting design in india. *Journal of The Institution of Engineers (India): Series A*, 101(4):643–656.
- [Lalande et al., 2021] Lalande, P., Demers, C. M., Lalonde, J.-F., Potvin, A., and Hébert, M. (2021). Spatial representations of melanopic light in architecture. *Architectural Science Review*, 64(6):522–533.
- [Leccese et al., 2019] Leccese, F., Lista, D., Salvadori, G., Beccali, M., and Bonomolo, M. (2019). Space syntax analysis applied to urban street lighting: Relations between spatial properties and lighting levels. *Applied Sciences (Switzerland)*, 9(16).
- [Lee and Lee, 2021] Lee, J.-H. and Lee, Y. (2021). The effectiveness of virtual reality simulation on the qualitative analysis of lighting design. *Journal of Digital Landscape Architecture*, 2021(6):195–202.
- [Lee et al., 2017] Lee, K., Han, K., and Lee, J. (2017). The impact of shading type and azimuth orientation on the daylighting in a classroom-focusing on effectiveness of façade shading, comparing the results of da and udi. *Energies*, 10(5). cited By 11.
- [Lee and Lee, 2019] Lee, S. and Lee, K. (2019). A study on the improvement of the evaluation scale of discomfort glare in educational facilities. *Energies*, 12(17).
- [Leng et al., 2021] Leng, J., Wang, D., Shen, W., Li, X., Liu, Q., and Chen, X. (2021). Digital twins-based smart manu-

- facturing system design in industry 4.0: A review. *Journal of Manufacturing Systems*, 60:119–137.
- [Li et al., 2022] Li, L., Lei, B., and Mao, C. (2022). Digital twin in smart manufacturing. *Journal of Industrial Information Integration*, 26:100289.
- [Longcore and Rich, 2004] Longcore, T. and Rich, C. (2004). Ecological light pollution. *Frontiers in Ecology and the Environment*, 2(4):191–198.
- [Lowden and Kecklund, 2021] Lowden, A. and Kecklund, G. (2021). Considerations on how to light the night-shift. *Lighting Research & Technology*, 53(5):437–452.
- [Mackey and Roudsari, 2018] Mackey, C. and Roudsari, M. S. (2018). The tool (s) versus the toolkit. In *Humanizing Digital Reality*, pages 93–101. Springer.
- [Major et al., 2021] Major, P., Li, G., Hildre, H. P., and Zhang, H. (2021). The use of a data-driven digital twin of a smart city: A case study of Ålesund, norway. *IEEE Instrumentation Measurement Magazine*, 24(7):39–49.
- [Makaremi et al., 2019] Makaremi, N., Schiavoni, S., Pisello, A., and Cotana, F. (2019). Effects of surface reflectance and lighting design strategies on energy consumption and visual comfort. *Indoor and Built Environment*, 28(4):552–563.
- [Mandal et al., 2019] Mandal, P., Dey, D., and Roy, B. (2019). Optimization of luminaire layout to achieve a visually comfortable and energy efficient indoor general lighting scheme by particle swarm optimization. *Leukos*.
- [Mandal et al., 2021] Mandal, P., Dey, D., and Roy, B. (2021). Optimization of luminaire layout to achieve a visually comfortable and energy efficient indoor general lighting scheme by particle swarm optimization. *LEUKOS - Journal of Illuminating Engineering Society of North America*, 17(1):91–106.
- [Mattsson et al., 2020] Mattsson, P., Johansson, M., Almén, M., Laike, T., Marcheschi, E., and Ståhl, A. (2020). Improved usability of pedestrian environments after dark for people with vision impairment: an intervention study. *Sustainability*, 12(3).
- [Michael et al., 2018] Michael, A., Gregoriou, S., and Kalogirou, S. (2018). Environmental assessment of an integrated adaptive system for the improvement of indoor visual comfort of existing buildings. *Renewable Energy*, 115:620–633.
- [Mohammadi and Taylor, 2020] Mohammadi, N. and Taylor, J. E. (2020). Knowledge discovery in smart city digital twins. In *53rd Hawaii International Conference on System Sciences (HICSS)*.
- [Mylonas et al., 2021] Mylonas, G., Kalogeras, A., Kalogeras, G., Anagnostopoulos, C., Alexakos, C., and Muñoz, L. (2021). Digital twins from smart manufacturing to smart cities: A survey. *IEEE Access*, 9:143222–143249.
- [Natephra et al., 2017] Natephra, W., Motamedi, A., Fukuda, T., and Yabuki, N. (2017). Integrating building information modeling and virtual reality development engines for building indoor lighting design. *Visualization in Engineering*, 5(1).
- [Neethirajan and Kemp, 2021] Neethirajan, S. and Kemp, B. (2021). Digital twins in livestock farming. *Animals*, 11(4).
- [Nejat et al., 2015] Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M., and Abd. Majid, M. Z. (2015). A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries). *Renewable and Sustainable Energy Reviews*, 43:843–862.
- [Oh et al., 2017] Oh, S., Dutton, S., Selkowitz, S., and Han, H. (2017). Application of a coelostat daylighting system for energy savings and enhancement of indoor illumination: A case study under clear-sky conditions. *Energy and Buildings*, 156:173–186.
- [Pan and Mitchell, 2020] Pan, Y. and Mitchell, K. (2020). Posemmr: A collaborative mixed reality authoring tool for character animation. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 758–759.
- [Pineo and Rydin, 2018] Pineo, H. and Rydin, Y. (2018). *Cities, health and well-being*. Royal Institution of Chartered Surveyors (RICS).
- [Pylianidis et al., 2021] Pylianidis, C., Osinga, S., and Athanasiadis, I. N. (2021). Introducing digital twins to agriculture. *Computers and Electronics in Agriculture*, 184:105942.
- [Rockcastle et al., 2021] Rockcastle, S., Danell, M., Calabrese, E., Sollom-Brotherton, G., Mahic, A., Van Den Wymelenberg, K., and Davis, R. (2021). Comparing perceptions of a dimmable led lighting system between a real space and a virtual reality display. *Lighting Research & Technology*, page 1477153521990039.
- [Saraiva et al., 2021] Saraiva, N., Rodrigues, E., Gaspar, A., and Costa, J. (2021). Daylighting simulation of a heritage building by comparing matrix methods and solar models. *Solar Energy*, 224:685–696.
- [Schleich et al., 2017] Schleich, B., Anwer, N., Mathieu, L., and Wartzack, S. (2017). Shaping the digital twin for design and production engineering. *CIRP Annals*, 66(1):141–144.
- [Selim et al., 2020] Selim, F., Elkholly, S., and Bendary, A. (2020). A new trend for indoor lighting design based on a hybrid methodology. *Journal of Daylighting*, 7(2):137–153.
- [Shahat et al., 2021] Shahat, E., Hyun, C. T., and Yeom, C. (2021). City digital twin potentials: A review and research agenda. *Sustainability*, 13(6).
- [Shchepetkov et al., 2020] Shchepetkov, N., Kapeleva, S., Bugaev, D., Matovnikov, G., and Kostareva, A. (2020). Concerning the concept of light-colour arrangement of the urban environment in the central part of tyumen. *Light and Engineering*, 28(1):34–42.
- [Stanford-Clark et al., 2019] Stanford-Clark, A., Frank-Schultz, E., and Harris, M. (2019). What are digital twins? Technical report, IBM.
- [Stark et al., 2019] Stark, R., Fresemann, C., and Lindow, K. (2019). Development and operation of digital twins for technical systems and services. *CIRP Annals*, 68(1):129–132.
- [Straka et al., 2021] Straka, T. M., von der Lippe, M., Voigt, C. C., Gandy, M., Kowarik, I., and Buchholz, S. (2021). Light pollution impairs urban nocturnal pollinators but less so in areas with high tree cover. *Science of The Total Environment*, 778:146244.
- [Tai and Jang, 2018] Tai, N.-C. and Jang, H.-W. (2018). Design and development of an unmanned aerial vehicle to capture real-world illumination for image-based lighting for dense urban environment. *Computer-Aided Design and Applications*, 15(2):157–163.
- [Tao et al., 2019] Tao, F., Zhang, H., Liu, A., and Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, 15(4):2405–2415.
- [Tereci and Ozata, 2020] Tereci, A. and Ozata, O. (2020). Analysis and strategies for zonal lighting design of konya mevlana museum and mevlana culture centre axis. *Light and Engineering*, 28(3):22–30.
- [Verdouw et al., 2021] Verdouw, C., Tekinerdogan, B., Beulens, A., and Wolfert, S. (2021). Digital twins in smart farming. *Agricultural Systems*, 189:103046.
- [Ververidis et al., 2019] Ververidis, D., Chantas, G., Migkotzidis, P., Anastasovitis, E., Papazoglou-Chalikias, A., Nikolaidis, E., Nikolopoulos, S., Kompatsiaris, I., Mavromanolakis, G., Thomsen, L. E., Liapis, A., Yanakakis, G., Müller, M., and Hadiji, F. (2019). An authoring tool for educators to make virtual labs. In Auer, M. E. and Tsiatsos, T., editors, *The Challenges of the Digital Transformation in Education*, pages 653–666, Cham. Springer International Publishing.
- [Whang et al., 2022] Whang, A.-W., Chen, Y.-Y., Leu, M.-Y., Tseng, W.-C., Lin, Y.-Z., Chang, H.-W., Tsai, C.-H., Liang, Y.-C., Zhang, X., Lin, C.-T., Huang, T.-C., Chang, C.-M., and Chen, H.-C. (2022). Development and testing of a modular sunlight transport system employing free-form mirrors. *Energies*, 15(2).
- [Xie and Sawyer, 2021] Xie, J. and Sawyer, A. O. (2021). Simulation-assisted data-driven method for glare control with automated shading systems in office buildings. *Building and Environment*, 196:107801.
- [Zanon et al., 2019] Zanon, S., Callegaro, N., and Albatici, R. (2019). A novel approach for the definition of an integrated visual quality index for residential buildings. *Applied Sciences (Switzerland)*, 9(8).
- [Zeng et al., 2021] Zeng, Y., Sun, H., Lin, B., and Zhang, Q. (2021). Non-visual effects of office light environment: Field evaluation, model comparison, and spectral analysis. *Building and Environment*, 197.

On the Use of Graphical Digital Twins for Urban Planning of Mobility Projects: a Case Study from a new District in Ålesund, Norway

Pierre Major¹, Ricardo da Silva Torres^{2,3}, Andreas Amundsen⁴,
Pernille Stadsnes¹, Egil Tennfjord Mikalsen¹
¹AugmentCity, Ålesund, Norway

²Wageningen University & Research, Wageningen, The Netherlands

³NTNU – Norwegian University of Science and Technology, Ålesund, Norway

⁴United Future Lab Norway, Ålesund, Norway

KEYWORDS

Digital Twin, Urban Planning, Planning Support System, Mobility.

ABSTRACT

Urban planning is a complex task often involving many stakeholders of varying levels of knowledge and expertise over periods stretching years. Many urban planning tools currently exist, especially for mobility planning. However, the use of such tools often relies on ad-hoc modelling of “*expensive*” domain experts, which hampers to incorporate new knowledge and insights into the planning process over time. Another issue refers to the lack of interactivity, i.e., stakeholders can not easily change configurations of simulations and visualize the impact of those changes. This paper presents and discusses the benefits of using a graphical digital twin to overcome such shortcomings. We demonstrate how a digital-twin-based approach improves the current planning practices from two perspectives. The first refers to automating the configuration and data integration in models, making the tool flexible and scalable to large-scale planning involving multiple cities on a national level and supporting automatic updates of employed models when input data is updated. The second refers to supporting interaction with the model through a user interface that allows stakeholders to perform actions, leading to insightful what-if scenarios and therefore better-informed decisions. We demonstrate the effectiveness of using graphical digital twins in a compelling real usage case study concerning urban mobility planning in Ålesund, Norway. Finally, this paper also outlines recommendations and further research opportunities in the area.

INTRODUCTION

While cities host currently 50% of the world’s population, they consume more than 80% of the world energy and contribute to more than 60% of the greenhouse gases (GHG)¹. Climate Change Mitiga-

tion (CCM) and Climate Change Adaptation (CCA) require profound societal changes, especially in terms of fostering greener mobility.

CCA and CCM demand a rapid response. However, urban mobility planning is a complex and time-consuming process, often involving the analysis of large volumes of data, processed through different tools, and involving interests and needs of several stakeholders (e.g., politicians, planners, citizens) [Fiore et al., 2019].

Urban planning often involves multi-disciplinary teams of experts with their own language and expertise [Brömmelstroet, 2010]. That situation may lead to vocabulary mismatches. Reports utilized in the decision-making processes have a high level of abstraction, being more convenient to a knowledgeable audience (e.g., urban planners, technical experts, and politicians). Domain experts tend to use jargon, metrics, and procedures, with difficult understanding for other stakeholders. Processing and analyzing such reports are also time-consuming tasks. There is also a lack of a common platform where insights from the various domain experts can be shared. Another issue refers to the fragmentation of plans, reflecting geographical and political particularities. The dynamic evolution of cities also poses additional challenges for the planning process. Often, this evolution is unevenly distributed in space and time, which may lead to tension both in planning and in realization phases.

Knowledge database, in the form of data collections, simulated analysis, and insight synthesis, comes very early in urban planning projects. Available data are often manually collected and manually fed into simulation models. Thus the model insights are static and are outdated when plans are matured, changed, or even when unforeseen events force behaviour change and durably transform mobility patterns, e.g., corona or energy crises. Furthermore, few pathways are explored during the analysis. The “business-as-usual” is compared against the evaluation of a discrete (in the mathematical sense) set of alternatives.

In short, the existing urban planning tools are not adequate for supporting effective decision-making.

¹<https://bit.ly/unhabitatemissionreport>(As of Feb.2022)

The employed procedures lack transparency, and there is a poor connection between the tools and the actual planning process [Brömmelstroet, 2010]. There is, therefore, a need for tools to “play with,” aiming to generate strategies early on in the planning process; avoid using static representations in the dynamic planning cycles [Aspen and Amundsen, 2021]; and promote the early involvement of citizens, which may prevent conflicts and delays in later implementation phases.

Digital twin technologies have emerged as a promising alternative to address such challenges. Here, a digital twin is not seen as a control tool, but as a planning support system (PSS), i.e., as a tool that supports fast and better-informed decision-making towards sustainable city planning [Shahat et al., 2021], [Mylonas et al., 2021], [Ramu et al., 2022]. Digital twins have been used for anchoring the Sustainable Development Goals (SDGs) in concrete measures at the strategic and zoning plan level by supporting the analysis of different what-if scenarios and promoting active communication with other stakeholders, including citizens.

This paper presents and discusses the benefits of using a graphical digital twin to support urban mobility planning. We demonstrate that our digital-twin-based approach may improve the current practices in two ways:

1. Automating the configuration and data integration in models, providing i) scalability to multiple cities on a national level, and ii) automatic updates of the models when input data is updated.
2. Interaction with the model through a user-interface allowing stakeholders to perform actions, such as adding new infrastructure or changing capacity of a road segment.

BACKGROUND CONCEPTS AND RELATED WORK

Urban planning and Digital Twins

[Brömmelstroet, 2010] outlined relevant challenges and bottlenecks towards the implementation of an effective planning support system. Examples include the focus given on technical aspects, which lead to a lack of transparency and loose connections with the planning process.

Digital twins have emerged as a promising technology for addressing such challenges in planning activities. A digital twin can be defined as “*a dynamic and interactive virtual representation of a physical object or system, usually across multiple stages of its lifecycle. It uses real-world data, simulation or machine learning models, combined with data analysis, to enable understanding, learning, reasoning, and communication to stakeholders. Digital twins can be used to answer what-if questions and should be able to present the insights in an intuitive way.*” [Stanford-Clark et al., 2019]. In this context, digital twins

function as planning support systems, which could be utilized for land use and transportation planning.

The use of the term digital twin in an urban planning context has been used for a wide range of applications [Ketzler et al., 2020], [Deng et al., 2021], [Shahat et al., 2021], [Mylonas et al., 2021], [Ramu et al., 2022], including, for example, disaster simulation, land use analysis, and garbage management. Next, we discuss existing initiatives related to the use of digital twins for mobility analysis.

Digital Twins for Urban Mobility Analysis

Various forms of mobility analysis are essential input to planning on a regional and municipal level. In Norway, for example, on a regional level, these analyses are typically performed by consultants or the road authorities, which maintain a regional transportation model. A regional planning process can span years, and the models are typically run in the first part of the process for a limited set of scenarios. The results are static in the sense that they must be operated by experts, and there are limited possibilities for exploring different scenarios so that stakeholders can better understand the possibilities and limitations of the models. The planning process, on the other hand, is often dynamic: new data, expanding knowledge base, and other external factors may change strategies and goals.

Recent advances in establishing large scale traffic-models for cities based on open data-sets [Sánchez-Vaquerizo, 2022] and the availability of open-source models (e.g., MATSIM²) have fostered the development of digital twins for urban mobility modelling and assessment. One example is the work of [Chao et al., 2020], who investigated the benefits of state-of-the-art traffic visualization for the purpose of autonomous driving, not used for urban planning. [Major et al., 2021] also demonstrated the usefulness of graphical digital twin to visualize insights from different urban topics (e.g., mobility and energy) in a city. Their work, however, has not considered a dynamical scenario creation.

[Aspen and Amundsen, 2021] showed the usefulness of a systems’ theory approach in instrumenting the SDGs into master plan planning strategies. In this paper, we investigate zoning plans and interventions at a lower level. [Metze, 2020] reviewed the literature over environmental visualisation. Many traits are similar to sustainable urban planning, such as cognitive tainted connotative interpretation of visual structures by stakeholders, i.e., the expert documentation is not always adapted to decision-makers and stakeholders from the civil society.

²<https://www.matsim.org/> (As of Feb. 2022).

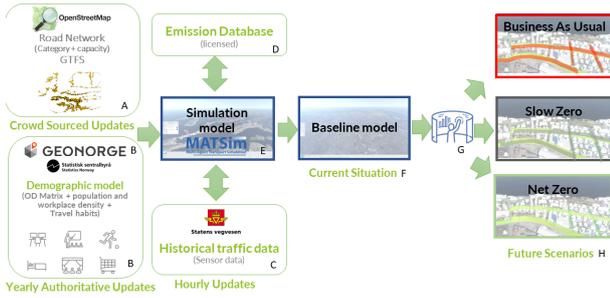


Fig. 1: Digital-twin-based mobility model assessment pipeline for interactive exploration of realistic what-if scenarios.

PROPOSED DIGITAL TWIN FOR MOBILITY ASSESSMENT

Digital-twin-based Mobility Model Assessment

Figure 1 illustrates the digital-twin-based mobility model assessment pipeline for interactive exploration of realistic what-if scenarios. The pipeline is generic, but, in this paper, its use is illustrated with data of Ålesund, Norway.

In the figure, we represent the main data sources, including the frequency of the data updates and information about which updates are automatic. The pipeline also includes a traffic simulation software that produces a *baseline model*. Finally, through a graphical interface, users interact with the model, performing analysis on different future scenarios. The whole mobility complexity is hidden from the users as all the parameters that can be generically entered for the whole country are pre-entered or automatically updated.

The architecture of the proposed digital twin is centered around the concept of an Origin-Destination (OD) matrix. An OD matrix can be seen as a list or matrix, which establishes the connection between the location where people work with the location where they live. These matrix values can be seen, therefore, as a proxy of the probability of traffic between two locations. The locations can be defined in terms of grid elements or geographical units. In addition, the travel habits, i.e., the probability of using a mode for a certain activity (detailed later in Fig. 4) is also a key concept in the mobility model implemented in the digital twin.

The key elements of the digital-twin pipeline are the following:

- The crowd-sourced open-source Open Street Map (OSM) (module represented with A in Figure 1) is imported to generate the network using road type, direction, and capacity. Crowd-sourcing implies that the network is constantly updated. Furthermore, the bus network information in the form of a General Transit Feed Specification (GTFS) time table are of-

ten openly published.

- Demographics data, such as the demographic and workforce densities and the OD Matrix are updated every year by the Norwegian Statistics Central Bureau (Figure 1 B). The travel habits surveys, which are updated and published every four years, are the default values for the simulation.

- Norwegian Roads Authority (SVV) hourly traffic measurements at key locations are dynamically updated daily, with traffic direction, and vehicle length and category (e.g., bike, car, trailer) – Figure 1 C. The data is automatically downloaded for the validation of the “current situation” scenario.

- The HBEFA database of emissions³ is used to quantify the CO₂ emissions for each scenarios based on the modal distribution and the energy mix of the country (Figure 1 D).

- The agent-based mobility engine MATSIM (Figure 1 E), an open-source library with an active contributor base, is used to simulate the scenarios, such as the “current situation” or “baseline model.” It also supports the creation of alternative pathways by changing some of its parameters.

- The “base model” is the result of the calibrated MATSIM model (Figure 1 F), validated with the traffic measurements from C. It serves as base for the model when users (e.g., urban planners) – Figure 1 G – play with different parameters.

- An infinity of what-if scenarios can be simulated. Here, three are selected (Figure 1 H): “*business-as-usual*”, which refers to population growth without changing the lifestyle and travel habits; “*Slow Zero*”, which relates to some change in habits towards a more sustainable living (e.g., living in denser areas); and “*Net-Zero*” paradigm shift, which refers to initiatives towards zero carbon emissions (e.g., closing motorways).

The proposed pipeline is flexible to support configuration automation of different mobility models. This eases the inclusion of new knowledge and plans and thus dramatically simplifies the identification of suitable mobility models considering different scenarios. Furthermore, the pipeline is generic and scalable to handle data of other regions (e.g., other Norwegian cities).

User Interface Functionality: Functional View

This section presents and discusses how the user can easily create new scenarios via the digital twin interface. Figure 2 shows a screenshot which summarizes the main available features: creating, choosing, duplicating, and deleting (if needed) scenarios. To create scenarios, one can “play” with the following parameters:

- Adding and editing network links in the networks, i.e. add new roads, bridges, and tunnels to the existing infrastructure and specifying speed limit, link

³<https://www.hbefa.net/e/index.html/> (As of Feb. 2022).

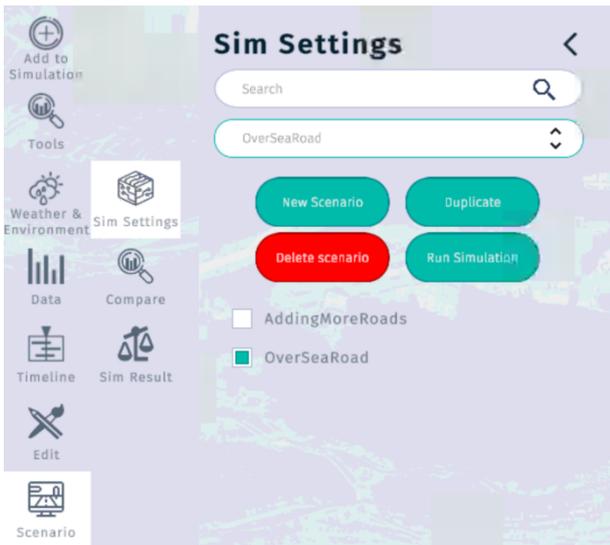


Fig. 2: Screenshot of the interface used for creating scenarios.

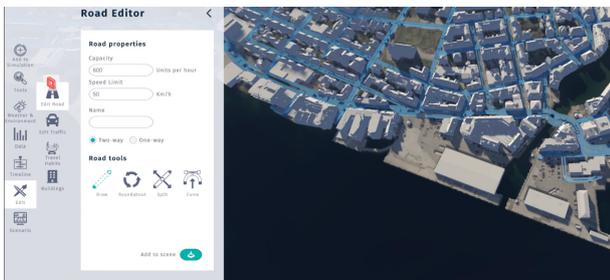


Fig. 3: Screenshot of the road editor interface.

capacity, and directionality, as illustrated in Figure 3. This can be utilized to support the creation of complex scenarios, such as closing a specific road for vehicle traffic.

- The flexibility to change the travel habits by activity type (e.g. work, school, service, shopping, care, etc.) is illustrated in Figure 4. This feature allows, for example, to simulate and visualise the effect of active mobility on congestion and on CO₂ emissions.
- Demographic and work force statistics can be altered on the 250m × 250m grid unit level. This allows modelling scenarios where new residential or multi-purpose districts are created.
- Travel habits can be modified by activity to reflect anticipated societal behaviour change, see Figure 4.
- Some modifications are only possible in the background at the configuration file level, and are not yet part of the scenario building features. Examples include the GTFS public transport time table, which must be edited offline; and Modifications on the OD matrix.

Once a scenario is created, the simulation can be run to estimate the impact on traffic and emissions. The results can then be shown in the 3D graphical digital twin.

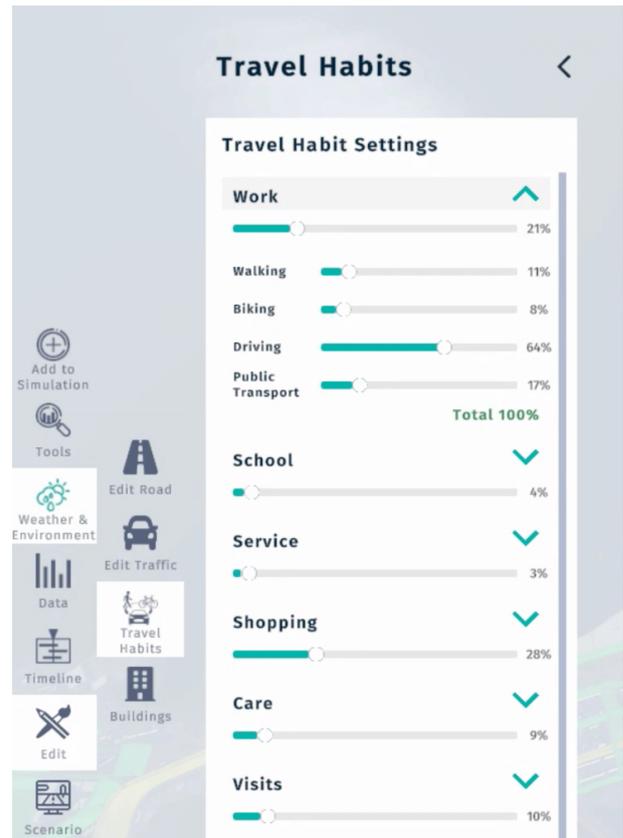


Fig. 4: Screenshot of the interface to collect inputs related to travel habits per activity. The travel habits are expanded for the “work“ activity.

Additionally, specific features are provided to support the assessment of plans. For example, the tool allows to hide/show the buildings from cadastral register and import new BIM models in IFC or FBX formats.

USAGE SCENARIOS

To demonstrate the benefits of our digital-twin-based approach, we present usage scenarios related to the creation of a zoning plan in Ålesund, Norway.

Figure 5 illustrates four currently parallel projects in the city centre of Ålesund. Albeit they are confined to a bounding box of 600 m width, each of them has a different maturity, project timeline (start, concept, phases, stop), and political and citizen involvement phase or policy. Project (A) relates to the construction of a new road and a bridge. The goal is to allow crossing the sound westward/eastward. The new district (B), still partly in the conceptual phase, is affected by the bridge (A). There are also plans to transform the current cargo port and the bus and speedboat terminal into a modern multipurpose district. Project (C), currently in the detail phase, is the bus terminal relocation project⁴). This project does not take into account the speedboat terminal

⁴<https://bit.ly/zoningplan2018> (As of Feb. 2022).



Fig. 5: Urban planning project portfolio.

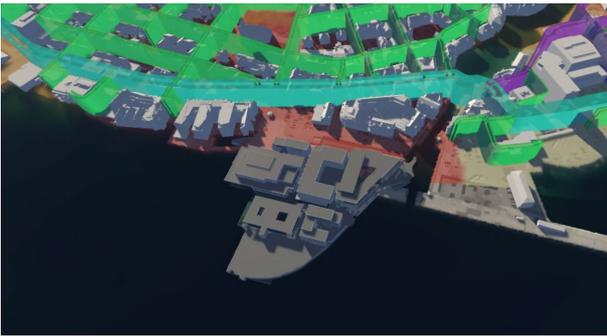


Fig. 6: Example of a case related to future districts in Ålesund, Norway. The figure illustrates removed hangars, added BIM of the new district, the road infrastructure, and information related to population density in the area.

that crosses the fjord. Finally, the new tunnel (D) under the city centre is expected to flatten the curve of peak traffic and revitalize the city centre by avoiding congestion.

Figure 6 illustrates how the visualisation helps stakeholders anchor their attention and contextualize the *future of city district* by removing BIM models (compare with Fig. 3), adding new BIM models, visualising the population density grid, and the road network.

Future built infrastructure can easily be modified to reflect alternative zoning plans⁵. Figure 7, for example, mirrors the network conditions of one such scenarios.

The mobility is only part of the big picture in urban planning and needs to be contextualized with other relevant insights. Figure 8 illustrates how the mobility model provides insights that can be superposed with other types of information, such as a heatmap representing the distance to schools dynamically calculated by open route service⁶. Regions highlighted in blue are associated with low-range values, while those in red related to higher values. The heatmap could be used, for example, to encode the distance to municipal and medical services, wildlife

⁵<https://bit.ly/brosundtunnel> (As of Feb. 2022).

⁶<https://openrouteservice.org/>, (As of Feb. 2022).



Fig. 7: Example of a case related to the inclusion of a new infrastructure. Two new links are added to the network: a bridge and a tunnel (highlighted in yellow).

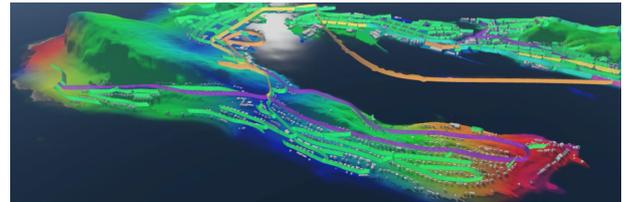


Fig. 8: Example of a case related to the assessment of the road network and distances to school.

impact assessment, and response time from emergency services. This exemplifies the kind of new knowledge that could be generated in later project phases. Using the same tool early on from the master plan to the zoning plan ensures continuous update of the insights regarding relevant variables considered in the planning process over time.

LESSONS LEARNED AND RECOMMENDATIONS

The developed digital twin has been validated in the context of the assessment of complex mobility scenarios. Obtained feedback from stakeholders has been positive towards the adoption of such technologies during the life cycle of mobility planning processes.

A key recommendation refers to the use of dynamic graphical digital twins to plan for automation of knowledge base using open source tools and open data sources. This may foster the continuous coordination of the insights and strategies between the master plan and their affiliated zoning plans throughout iteration and refinements, especially when the blueprints becomes more concrete.

CONCLUSIONS

This paper introduced ongoing research dedicated to the design, implementation, and validation of a digital-twin-based mobility assessment tool to support urban planning operations. The proposed digital twin synchronizes seamlessly the current physical world (city) and the digital twin in an urban mobility planning context, both through automation/integration from different sources and supporting for interaction from many stakeholders along the planning process.

This paper also demonstrates how the digital twin can be used to prototype mobility solutions both at the conceptual and detailed levels, allowing the creation and maturation of scenarios and plans along the urban planning project. Furthermore, the tool is scalable to other cities and regions (especially in Norway). Finally, it is connected to authoritative national databases, allowing not only the visualisation and analysis of past and present data, but also the validation of different models, considering newer calibrations and simulations of future scenarios.

Future work will address the challenges of how digital twins can be improved towards becoming more integrated into the planning process from master plan to zoning plan. Investigating how the tool can transform the planning process and how the process can be improved by the tool will be addressed in future work. Further research also concerns investigating how urban planners can use the tool as a co-creation platform across silos, including public/private organisations, municipal departments, in inter-municipal planning projects. Finally, to gain trust in the predictions, further research should focus on validation, accountability, and visualisation of parameter sensitivity.

ACKNOWLEDGEMENTS

This work has been conducted in the context of the Norwegian Research Council's SMARTPLAN (310056), Data-driven co-creation project (321102), and Twin Fjord (320627) projects.

REFERENCES

- [Aspen and Amundsen, 2021] Aspen, D. M. and Amundsen, A. (2021). Developing a participatory planning support system for sustainable regional planning — a problem structuring case study. *13*, 13.
- [Brömmelstroet, 2010] Brömmelstroet, M. T. (2010). Equip the warrior instead of manning the equipment: Land use and transport planning support in the netherlands. *Journal of Transport and Land Use*, 3(1):25–41.
- [Chao et al., 2020] Chao, Q., Bi, H., Li, W., Mao, T., Wang, Z., Lin, M. C., and Deng, Z. (2020). A survey on visual traffic simulation: Models, evaluations, and applications in autonomous driving. *Computer Graphics Forum*, 39:287–308.
- [Deng et al., 2021] Deng, T., Zhang, K., and Shen, Z.-J. M. (2021). A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *Journal of Management Science and Engineering*, 6(2):125–134.
- [Fiore et al., 2019] Fiore, S., Elia, D., Pires, C. E., Mestre, D. G., Cappiello, C., Vitali, M., Andrade, N., Braz, T., Lezzi, D., Moraes, R., Basso, T., Kozievitch, N. P., Fonseca, K. V. O., Antunes, N., Vieira, M., Palazzo, C., Blanquer, I., Meira, W., and Aloisio, G. (2019). An integrated big and fast data analytics platform for smart urban transportation management. *IEEE Access*, 7:117652–117677.
- [Ketzler et al., 2020] Ketzler, B., Naserentin, V., Latino, F., Zangelidis, C., Thuvander, L., and Logg, A. (2020). Digital twins for cities: A state of the art review. *Built Environment*, 46:547–573.
- [Major et al., 2021] Major, P., Li, G., Hildre, H. P., and Zhang, H. (2021). The use of a data-driven digital twin of a smart city: A case study of Ålesund, norway. *IEEE Instrumentation Measurement Magazine*, 24(7):39–49.
- [Metze, 2020] Metze, T. (2020). Visualization in environmental policy and planning: a systematic review and research

- agenda. *Journal of Environmental Policy and Planning*, 22:745–760.
- [Mylonas et al., 2021] Mylonas, G., Kalogeras, A., Kalogeras, G., Anagnostopoulos, C., Alexakos, C., and Muñoz, L. (2021). Digital twins from smart manufacturing to smart cities: A survey. *IEEE Access*, 9:143222–143249.
- [Ramu et al., 2022] Ramu, S. P., Boopalan, P., Pham, Q.-V., Maddikunta, P. K. R., Huynh-The, T., Alazab, M., Nguyen, T. T., and Gadekallu, T. R. (2022). Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions. *Sustainable Cities and Society*, 79:103663.
- [Shahat et al., 2021] Shahat, E., Hyun, C. T., and Yeom, C. (2021). City digital twin potentials: A review and research agenda. *Sustainability*, 13(6).
- [Stanford-Clark et al., 2019] Stanford-Clark, A., Frank-Schultz, E., and Harris, M. (2019). What are digital twins? Technical report, IBM.
- [Sánchez-Vaquerizo, 2022] Sánchez-Vaquerizo, J. A. (2022). Getting real: The challenge of building and validating a large-scale digital twin of barcelona's traffic with empirical data. *ISPRS International Journal of Geo-Information*, 11.

PIERRE MAJOR, currently as Head of Research for AugmentCity and OSC Ocean, he received his M.Sc. degree in Electrotechnique and Information Technology from the Swiss Federal Institute of Technology of Zürich (ETHZ) in 2005 and his Industrial Ph.D. on "data-driven models for fast virtual prototyping" at the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU), Ålesund Norway, in 2021. His domains of interest are virtual prototyping of demanding offshore operations and graphical digital twins of systems such as cities or ships.

RICARDO DA SILVA TORRES is Professor in Data Science and Artificial Intelligence at Wageningen University and Research. Dr. Torres holds also a position as Professor in Visual Computing at the Norwegian University of Science and Technology (NTNU) since 2019. He used to hold a position as a Professor at the University of Campinas, Brazil (2005 - 2019). Dr. Torres received a B.Sc. in Computer Engineering from the University of Campinas, Brazil, in 2000 and his Ph.D. degree in Computer Science at the same university in 2004. Dr. Torres has been developing multidisciplinary eScience research projects involving Multimedia Analysis, Multimedia Retrieval, Machine Learning, Databases, Information Visualisation, and Digital Libraries. Dr. Torres is author/co-author of more than 200 articles in refereed journals and conferences and serves as a PC member for several international and national conferences. Currently, he has been serving as Associate Editor of Pattern Recognition Letters, Pattern Recognition, and IEEE Systems. He is a member of the IEEE.

ANDREAS AMUNDSEN is the digital twin coordinator for the municipality of Ålesund. He has

previously worked as a research scientist with SINTEF, focusing on various aspects of software development within the offshore and maritime sector. He has also worked several years in the private sector as an IT consultant. He holds a M.Sc. degree from the Norwegian University of Science and Technology (NTNU).

EGIL TENNFJORD MIKALSEN Egil Tennfjord Mikalsen leads the technological vision and the engineering organisation of OSC AS as the Chief Technology Officer (CTO). Mikalsen connects technology, strategy, and business. Through his leadership and technical expertise he has developed the organisation and technology to deliver high end simulation and visualisation technologies in offshore, renewables and smart cities. Mikalsen also serves as a part of the leader group in OSC AS. Mikalsen has over 15 years experience in managing teams, developing complex systems and strategy.

PERNILLE STADSNES Pernille Stadsnes is Product Manager at AugmentCity AS, daughter company of OSC AS, leading the software and product development of digital twins for cities since 2019. Miss Stadsnes is also project manager for AugmentCity in research related projects. She received a B.Eng. In Ship Design from the Norwegian University of Science and Technology (NTNU) in 2015 and a M.Sc. In Systems Engineering and Industrial Economics from University of South Eastern Norway in 2019.

**Special Student Track
on
AI, Machine Learning,
Simulation and
Visualization**

GENOR: A Generic Platform for Indicator Assessment in City Planning

Léo Leplat¹, Ricardo da Silva Torres^{1,3}, Dina Aspen¹, Andreas Amundsen²
¹NTNU – Norwegian University of Science and Technology, Ålesund, Norway
²United Future Lab Norway, Ålesund, Norway
³Wageningen University & Research, Wageningen, The Netherlands

KEYWORDS

Information visualization, data analysis, digital twin, urban mobility, walkability, bus service availability

ABSTRACT

More and more data have been generated in city planning in the past few years. Clear visualizations of these data are helpful to support information comprehension and retention for urban practitioners and policy-makers. Much research has been carried out on modeling and integrating data, and different tools have been developed for their visualization. However, such tools are generally application dependant and cannot be easily tailored to other problems. This work introduces GENOR, a generic platform for indicator assessment in city planning. It consists of a client-server application able to store and process any indicator and provide 2D and 3D map-based views for visualization. A game engine (Unity¹) was used to create the client application, and the server consists of a REST API and a Database Management System (DBMS). Two cases studies were conducted to show the use of the platform: walkability and bus service availability assessment, both in the city of Ålesund, Norway. Obtained results demonstrate that the platform is flexible as it can be tailored to different applications seamlessly.

INTRODUCTION

Over the past few years, more and more urban data have been generated and made available. These data must be stored, processed, and visualized concisely and clearly to be available to planning practitioners and policy-makers (Doraiswamy et al. 2018; You et al. 2020). Especially in the context of urban mobility planning activities, the proper assessment of the spatial distribution of different indicators (e.g., walkability, bus service availability, and demographics) is of paramount importance.

Indicator assessment platforms need to be flexible. New data sources and indicators are often added and removed based on the type of plan being considered. For example, the relevance of indicators may vary between urban and rural applications. Also, many existing visualization methods are in 2D, even

though they may provide less understanding of specific relations among data (Doraiswamy et al. 2018). Furthermore, determining which visualization is best suited for communicating a given indicator is still an open question. It might be the case that the visualization mode should be changed from one group of stakeholders to another. In this case, switching between different visualization options might be an essential feature. For example, changing the visualization from 3D to 2D might produce a more valuable interpretation of spatial relations if the third dimension is irrelevant for a given inquiry.

Indicator visualization platforms and tools already exist (B.Longva et al. 2021; Fortini and Davis 2018; Psyllidis et al. 2015; Perhac et al. 2017), but they mostly focus on one specific problem/domain and thus can not be easily tailored to different applications. Therefore, the primary goal of our study was to design and implement a generic visualization platform (GENOR) to support indicator assessment in decision-making related to urban areas. The platform handles the definition, storage, processing, and visualization of indicators associated with a region, using 2D and 3D map-based views. The goal of this platform is to provide clear indicator visualizations that support planners and policy-makers in decision-making. The platform was designed to be seamlessly customizable to different indicator-based analyses.

Two case studies are considered to validate the proposed platform, both related to urban mobility analysis for Ålesund, Norway: walkability and bus service assessment. Walkability, which can be defined by a city's attractiveness or its opportunity for walking (Weinberger and Sweet 2012), has become crucial in urban planning. Indeed, walking habits may lead to health benefits and reduce traffic congestion, air pollution, air emissions, and the dependence on fossil fuels for transportation (Hall and Ram 2018). In this scenario, the proper assessment of walkability conditions of regions is of paramount importance in decision-making associated with urban planning processes. In a second case study, we demonstrate the use of the platform in the assessment of how the availability of bus service differs across different regions of the city. That kind of analysis might be helpful to urban mobility planners in their tasks aiming to assess the effectiveness of public transportation services.

In summary, the main contributions of this work

¹<https://unity.com/> (As of Feb. 2022).

are threefold:

- A software platform that stores and processes any multidimensional indicator.
- A generic visualization tool for indicator assessment using both 2D and 3D map-based views.
- Validation of the platform for two compelling applications related to urban mobility problems: walkability and bus service availability assessment.

RELATED WORK

Much research has been carried out on analyzing and visualizing urban data. This section provides an overview of recent data modeling and integration approaches and relevant urban visualization methods.

Data modeling and integration

One stream of literature has focused on data modeling and integration, especially working with inconsistent and heterogeneous data. For example, (Fortini and Davis 2018) presented methods that enable integration and visualization of urban data coming from multiple heterogeneous sources. Their platform is divided into integration, data storage, and service providing parts. The integration module is responsible for combining information from multiple heterogeneous data sources by performing data extraction, preprocessing, and standardization on each type of data, resulting in files with only one format suited for storage. In this work, we adopted an architecture similar to theirs.

(Psyllidis et al. 2015) presented a web-based platform that supports the analysis, integration, and visualization of large-scale and heterogeneous urban data. Datasets are usually specific to one sector or domain. Therefore, establishing correlation of information from different sectors is a difficult task. Their platform aims at solving this problem with the use of a semantic enrichment and integration component in the form of a web-based interface. It allows the user to define the relations between urban systems, data sources, and the city technology enablers. Inspired by their study, we used similar ideas to create a platform as generic as possible in this work.

(Chen et al. 2018) described an urban data visualization tool that focuses on the cross-domain correlation from multiple data sources by providing selection, filtering, and aggregating features. This is done using a visual query model for cross-domain correlation and a visual analytic framework for urban data visualization, correlation, querying, and reasoning. We adopt similar strategies for handling different types of indicators.

Initiatives on urban data visualization

Several initiatives have been proposed to support the proper visualization of urban data, including for mobility and transportation analysis. For a review in the area, the reader may refer to (Andrienko et al. 2017).

(Eberhardt and Silveira 2018) presented a list of visualization techniques applied to Open Government Data. They found that map-based presentation is the most used visualization method, and the main use case of developed visualization tools refer to studies related to transportation issues. A similar vision regarding the importance of map-based visualization strategies was reported by (Prandi et al. 2021). Their work described an infrastructure to collect data and a map-based visualization method to display and understand tourist flows. Their tool helped communities foster awareness about sustainability issues but was only designed for one purpose and area.

In another venue, (Sauda et al. 2007) provided an interactive urban visualization tool that displays multiple dimensions of urban data, letting the user better understand the urban environment. Many existing visualization methods are in 2D, even though they provide less understanding of spatial relations. (Johansson et al. 2016), in turn, provided methods that highlight social values through 3D visualization. (Perhac et al. 2017) describes another urban data visualization method that uses virtual reality (VR) with a game engine (Unity) to display data in a more immersive way.

Similar to many of the initiatives presented, we adopted a map-based visualization to support the analysis of indicators. In our solution, we also explore 3D visualizations to understand spatial relations better. Following the strategy employed by (Perhac et al. 2017), we use Unity in the implementation of the proposed platform.

Several tools have been developed to address urban mobility analysis, and recently, growing attention has been devoted to walkability assessments. In general, such tools are application-dependent and thus require changes in the source code to address other indicator-based problems. For example, (B.Longva et al. 2021) recently proposed a digital twin for walkability assessment. Different from our proposal, the proposed tool was designed to only address walkability indicator assessment using one static 2D visualization method. Also, its implementation is focused on data of a specific region.

GENOR: A GENERIC PLATAFORM FOR INDICATOR ASSESSMENT

This section introduces GENOR in terms of architectural, functional, and implementation aspects.

Architectural view

Figure 1 presents the architecture of GENOR. Its architecture is divided into two entities: a client, to handle the user interaction and visualization issues, and a web server, to handle the data storage and processing.

The client side handles all the interactions with the user. Overall, the client application allows users

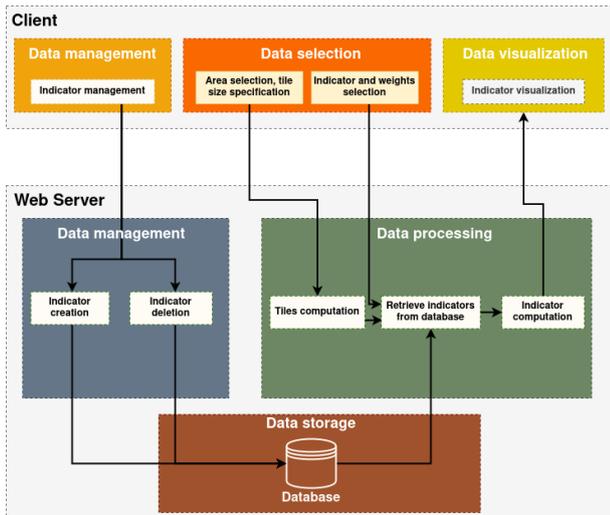


Fig. 1: Architectural view of the GENOR platform.

to define indicators and visualize computed indicator scores for a particular area.

- The data management part handles the definition of the indicators to be visualized. These indicators can also be deleted if they are no longer needed. Once the user defines or deletes indicators, the client application will send these data to the server which will process this information.
- To visualize indicators, users must first choose an area where the indicators will be displayed, and the actual indicators to visualize. If multiple indicators are selected, this part also includes features related to the assignment of weights to each indicator. The client application will send these data to the server, and the server will respond with the computed indicators.
- The data visualization process happens when the server responds to the data selection part request by sending values of indicators for specific locations. The client application will create one or more visualization methods to show these data efficiently based on selected information.

The server is responsible for storing and processing data. The user has no direct interactions with it, but the server is essential for the client application to work correctly.

- The data management part handles the definition and deletion of the indicators to be visualized. This information is then sent to the data storage module.
- The data processing part handles the computation of indicators. Based on an area and indicators selected by the user, the server communicates with the data storage module to compute the specified indicators for that particular region.
 - First, the region of interest is divided into multiple tiles whose size is given by the user. The idea is to compute one indicator value for each tile.
 - Then, the server communicates with the database to retrieve the relevant indicators of the region of



Fig. 2: Database diagram.

interest.

– Finally, based on the previous data, the server assigns a value of all indicators to each tile. For each tile, all indicators are merged into one global indicator using weights defined by the user. The server can then send a list of tiles with one value for each of them to the client application.

- The data storage module contains only a Database Management Systems (DBMS) used to store indicators. Even if it is represented within the web server, the database could be physically separated from the web server.

Usually, indicators are grouped when an analysis is made. For example, a walkability assessment combines multiple indicators, such as the number of pedestrians crossings or the average speed limit in the neighborhood. Therefore, we opted for grouping indicators by *category*. An indicator is a value assigned to an area, and a category is a group of one or more indicators.

The web server is a REpresentational State Transfer Application Programming Interface (REST API). A REST API is an API that conforms to the constraints of the REST architectural style. REST is a software architectural style commonly used to create interactive applications that use Web services. The client uses keywords to make requests: GET to retrieve resources; POST to submit new data to the server; PUT to update existing data; and DELETE to remove data.

Figure 2 presents the database diagram used in this work. A table *indicator* keeps track of the indicators. Each table row consists of a unique identifier *id* and the indicator name. A table *category*, similar to the *indicator* table, is used to keep track of the categories. To determine the indicators a category possesses, a third table *category_indicator* was created. Each row contains the identifier of a category and the identifier of an indicator.

Functional view

Figure 3 presents a functional view of the client application. Three actions can be performed from the main screen: managing indicators, computing indicators, and changing the map display.

Managing indicators

The first action the tool offers refers to the management of indicators. This means adding or deleting indicators and categories.

Adding indicators: When the user wants to add an indicator, the corresponding window opens. This

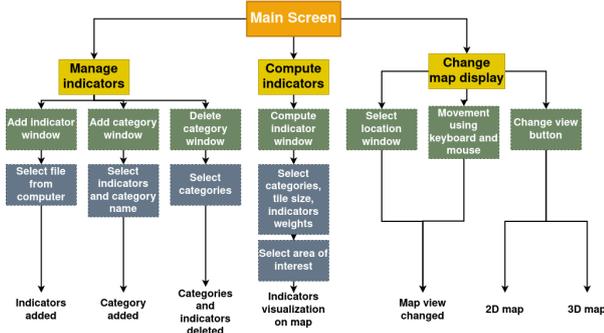


Fig. 3: Functional flow diagram of the platform.

TABLE I: Example of an indicator table.

geometry	indicator
POLYGON [(62.473940, 6.178904), (62.473940, 6.190553), (62.472451, 6.178904), (62.472451, 6.190553)]	0.5
POLYGON [(48.098141, -1.378394), (48.098141, -1.341239), (48.080142, -1.378394), (48.080142, -1.341239)]	0.7

feature allows the user to choose a local file that contains values of indicators for some regions of space. To make the platform as generic as possible, files with any vector-based spatial data format are accepted, such as Geojson,² Shapefile,³ or GeoPackage⁴ for example. The file is then sent to the server through a POST request.

The server reads the file and converts it to a generic format. Once performed, the server will request the database to create a table for each indicator present in the file and populate it with the corresponding values. Table I shows an example with two rows. The table has an indicator column that contains the indicator value, and a geometry column that contains an area corresponding to that indicator value. The coordinates follow the EPSG:4326⁵ format.

Adding categories: A category is a group of indicators. Therefore, the “add-category” window shows a list of all available indicators. To have access to this list, the client sends a GET request to the server before actually opening the window. The user can define a category name, choose one or more indicators and send this information to the server through a POST request.

With the information the client application provides, the server will then create a row with the category name to the *category* table. The category identifier, automatically created, is sent to the server. Then, for each indicator the client wants in this category, a row will be created in the *category.indicator* table with the previous category identifier and the identifier corresponding to the indicator.

Deleting categories: The delete category win-

²<https://geojson.org/> (As of Feb. 2022).

³<https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf> (As of Feb. 2022).

⁴<https://www.geopackage.org/> (As of Feb. 2022).

⁵<https://epsg.io/4326> (As of Feb. 2022).



Fig. 4: Part of page 1 of the “compute-indicator” window.

dow shows the list of available categories. This list is given by the server after a GET request by the client. Then, the user can choose one or more categories, and the selection will be sent to the server through a DELETE request.

The server will remove each category and its associated indicators from the database. This means that some rows will be removed from the *indicator*, *category*, and *category.indicator* tables, and the tables containing the indicators will be dropped.

Computing Indicators

Once some indicators and categories are defined, the user can start computing and visualizing them for the selected area. This area is divided into several tiles, where one indicator value is computed. The client application can then provide a visualization of the results.

Client side: The “compute-indicator” window (Figures 4 and 5) allows the user to choose which indicator to visualize. On the first page of the window, a list of available categories is displayed (retrieved from the server through a GET request), as well as an input to specify the size of the tiles. On the second page of the window, a list of indicators corresponding to the selected categories is displayed. The user can specify the weight associated with each of the indicators. Then, the user is asked to select an area on the map in which the indicators will be computed. Finally, all of these data (selected categories, tile size, weights of indicators, selected area) are sent to the server through a GET request.

The server computes the indicators (as described in the next part) and sends back tiles with values of indicators that can be visualized in the client application.

The current version of the platform provides two visualization methods:

- A 2D visualization, where each tile is represented by a square whose color changes depending on the indicator’s value. A high value is represented by the



Fig. 5: Part of page 2 of the “compute-indicator” window.

green color, while a low value is represented by a red color. This visualization method is commonly used by urban data visualization tools (Eberhardt and Silveira 2018).

- A 3D visualization, where some relief and 3D buildings are added to the map. Each tile is represented by a vertical bar, whose height and color depend on the indicator’s value, like the previous visualization. 3D visualizations offer more understanding of spatial relations than 2D visualizations when the user can freely move in the environment (van lammeren et al. 2010). Furthermore, 3D models are considered to be more beneficial for citizens, planners, and politicians (Ranzinger and Gleixner 1997).

Server side: When the server is asked to compute the indicators, three steps are conducted: First, the server creates the tiles by dividing the selected area based on the tile size. The centroid (or center) of the tiles is also computed. Secondly, for each indicator, the server tries to obtain the values of this indicator for that specific area from the database. Each indicator has a corresponding table inside the database, in which the values of this indicator are attached to areas. Then, a spatial join between the centroids of the tiles and the data taken from the database is made. If a centroid of a tile is located inside an area of the indicator table, then the server can assign the corresponding indicator value to this tile.

Figure 6 shows an example. An indicator table contains two rows and a table representing the tiles. Before the spatial join, the tiles have no value for the indicator. After the join, the first row gets the value 0.2 because its centroid is located inside the first polygon of the indicator table. The second row still has no value because its centroid is not located inside any polygon.

Finally, one global indicator is computed for each tile by combining all indicators, as shown in Equation 1, with ω_i the weights and v_i the indicators values.

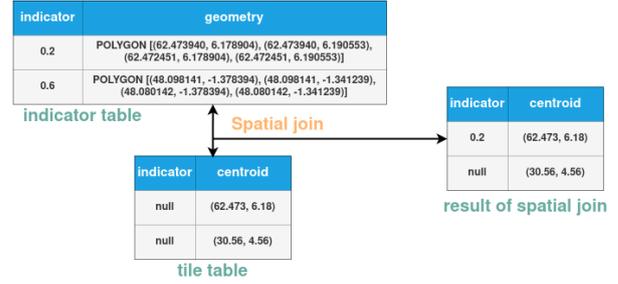


Fig. 6: Example of a spatial join.

$$globalIndicator = \sum_{i \in \{indicators\}} \omega_i \times v_i \quad (1)$$

Changing the Map Display

Two map views are available: a 2D view and a 3D view. In the 2D map it is possible to pan by using the mouse. The 3D view presents a view of projected satellite photos on terrain with elevation and 3D buildings. The mouse and the keyboard can be used to move inside the scene.

Implementation Aspects

Our implementation considered usage scenarios that rely on 2D and 3D visualization methods on top of a world map. Therefore, the client side was implemented using a Unity application. Unity is a cross-platform game engine that can be used to create 2D and 3D games and interactive simulations.

The utilization of a map provider is needed to display an interactive map. Many exist, such as ArcGIS⁶, Google Maps⁷, Geopipe⁸. We adopted Mapbox⁹ because it provides a well-documented Software Development Kit (SDK) for Unity.

The server side consists of two components: the web server and the database. The requirements for the server were to create a RESTful API that can work with Geographic Information System (GIS) data. Creating a RESTful can be done easily using different tools, such as JavaScript (Node.js) and PHP. We adopted Python-based technologies for handling GIS data, because of the availability of several GIS packages. Therefore, the web server is written in Python. To create the RESTful API, the package FastAPI¹⁰ was used, which is a modern and fast web framework for building an API using Python.

The database has a similar requirement than the web server: it had to support GIS data. We adopted Postgresql¹¹, which is a free and open-source re-

⁶<https://developers.arcgis.com/unity-sdk/> (As of Feb. 2022).

⁷https://developers.google.com/maps/documentation/gaming/overview_musk (As of Feb. 2022).

⁸<https://geopi.pe/games> (As of Feb. 2022).

⁹<https://www.mapbox.com/unity> (As of Feb. 2022).

¹⁰<https://fastapi.tiangolo.com/> (As of Feb. 2022).

¹¹<https://www.postgresql.org/> (As of Feb. 2022).

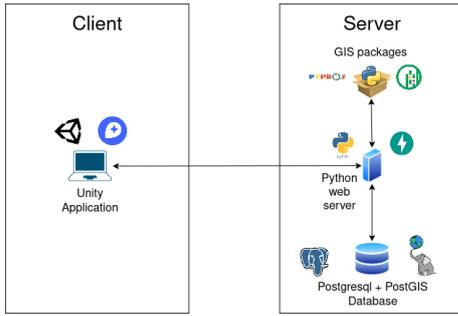


Fig. 7: Implementation diagram, illustrating the main technologies employed in the implementation of the proposed platform.

lational database management system, along with PostGIS¹², which is an open source software that supports handling geographic objects in PostgreSQL.

The implementation diagram is shown on Figure 7.

PLATFORM VALIDATION

This section presents two case studies related to the use of GENOR for the visualization of indicators related to urban mobility.

On the visualization of walkability indicators

The first case study is about the visualization of walkability (B.Longva et al. 2021) in the city of Ålesund in the context of planning processes.

Several indicators can be used to compute walkability. In this use case, we use the following indicators:

- Population density: a higher density means a more walkable area.
- Park areas: the indicator is higher when a park is nearby.
- Street connectivity: more street intersections give a higher score.
- Elevation: the highest score is at the lowest altitude.
- Speed limit: lower speed limit gives a higher score.
- Pedestrian crossings: the indicator is higher when a pedestrian crossing is nearby.

Datasets used: Global datasets have been used to compute the indicators. This means that even though this use case is limited to the city of Ålesund, it can be easily extended to any other place of the world.

The population density dataset¹³ was obtained from the Kontur¹⁴ company. Kontur is a geospatial data and real-time risk management solutions provider for humanitarian, private, and governmental organizations. The dataset is free, was released in 2020, and consists of hexagons with population counts at 400m resolution. The park areas,

¹²<https://postgis.net/> (As of Feb. 2022).

¹³<https://data.humdata.org/dataset/kontur-population-dataset> (As of Feb. 2022).

¹⁴<https://www.kontur.io/> (As of Feb. 2022).



Fig. 8: Area of interest.

street connectivity, speed limit, and pedestrian crossings datasets were obtained from OpenStreetMap¹⁵. OpenStreetMap is a free, editable map of the whole world that is being built by volunteers. Instead of downloading these raw datasets that are heavy, we used an API made available by OpenStreetMap to retrieve the data. The elevation data were obtained from Open Topo Data¹⁶. It is a free elevation API that can give access to several datasets. We chose to use the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) dataset, which is a joint effort between the Ministry of Economy, Trade, and Industry (METI) of Japan and the National Aeronautics and Space Administration (NASA) of the US. The dataset offers a 1 arc-second resolution, corresponding to a resolution of about 30m at the equator.

Indicator creation: When creating indicators, the platform expects one or more files with any vector-based spatial data format, such as Geojson, Shapefile, or GeoPackage for example. Therefore, we have to create such a file with the previous indicators.

First, we have to determine the area of interest where the indicators will be determined. Since we want to study the city of Ålesund, the area is a bounding box with the following coordinates:

- Minimum longitude: 5.938799.
- Minimum latitude: 62.436930.
- Maximum longitude: 6.420250.
- Maximum latitude: 62.536570.

This corresponds to the area in green on Figure 8.

The population dataset consists of hexagons with population counts at 400m resolution. We chose to compute one value of all indicators for each hexagon within the area of interest. The Geopandas package is used to read the population dataset by taking only the hexagons located inside the area of interest. The resulting data are then projected onto the EPSG:3857 projection system, the projected coordinate system used for rendering maps in, e.g., Google Maps and OpenStreetMap. Since we want indicators to take values between 0 and 1, the density indicator is computed as shown in Equation 2.

¹⁵<https://www.openstreetmap.org/> (As of Feb. 2022).

¹⁶<https://www.opentopodata.org/> (As of Feb. 2022).

$$densityIndicator = \frac{population}{maxPopulation} \quad (2)$$

The parks of Ålesund are given by OpenStreetMap through the Osmnx package. It is projected onto the EPSG:3857 projection system to be consistent with the hexagons. First, each hexagon's centroid (or geometric center) is computed. Then, we set the distance to park parameter for each hexagon as the distance between the centroid and the nearest park within walking distance. The walking distance is considered to be 800 meters, the farthest radial distance based on a ten minutes walk¹⁷. If a centroid has no park within walking distance, the associated parameter equals the walking distance. Finally, since the park area indicator is higher when a park is nearby, we set the park area indicator as shown in Equation 3.

$$parkAreaIndicator = 1 - \frac{distanceToPark}{walkingDistance} \quad (3)$$

As for the park areas, the number of street intersections is given by OpenStreetMap through Osmnx. For each hexagon, we count the number of intersections whose distance to the centroid is inferior to the walking distance. We set the street connectivity indicator as shown in Equation 4.

$$streetConnectIndicator = \frac{numberOfIntersect}{maxNumberOfIntersect} \quad (4)$$

The elevation data were collected from Open Topo Data. For each hexagon, the elevation of the centroid is queried. The associated indicator is written in Equation 5.

$$elevationIndicator = 1 - \frac{elevation}{maxElevation} \quad (5)$$

The speed limit data come from OpenStreetMap. For each hexagon, we make an average of the speed limit of each road whose distance to the centroid is inferior to the walking distance. The speed limit indicator is written in Equation 6.

$$speedLimitIndicator = 1 - \frac{averageSpeedLimit}{maxAverageSpeedLimit} \quad (6)$$

The pedestrian crossings indicator computation is similar to the park area indicator computation, we just check the distance to the nearest crossing instead of the nearest park.

Now that we have the values of the six indicators, we use the GeoPandas package to export them into a

¹⁷<https://www.dcla.net/blog/walkability-standards> (As of Feb. 2022).

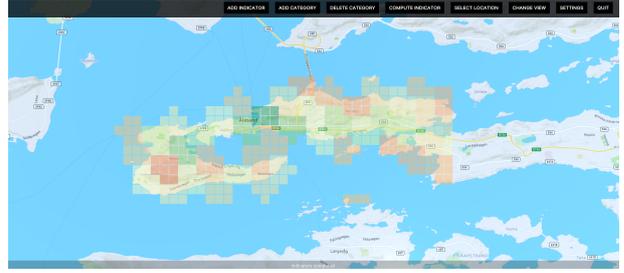


Fig. 9: 2D visualization of walkability.

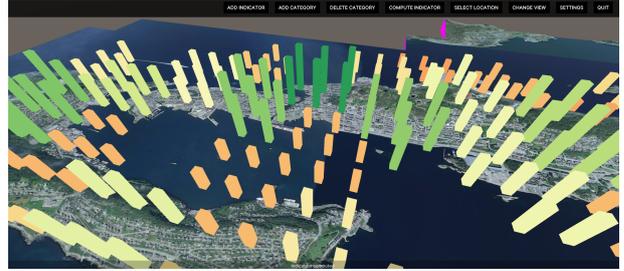


Fig. 10: 3D visualization of walkability.

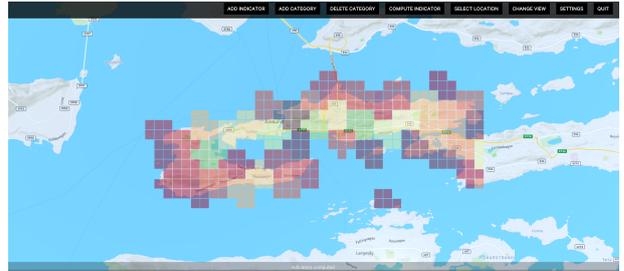


Fig. 11: 2D visualization of the population indicator.

GeoJSON file, which is an open standard format designed for representing simple geographical features. A weighted linear sum is employed to combine the different indicators.

Visualization of indicators: In the platform, the previously created GeoJSON can be uploaded via the add indicator window. Then, a *walkability* category can be defined in the add category window. Finally, the “compute-indicator” window will lead to the visualization of the indicators.

Figures 9 and 10, respectively, show the 2D and 3D results with all weights set to 1. A green color means a high value, and a red color means a low value. We can see that the walkability indicator has the highest scores in the city center and the lowest in the southwest.

Indicators can also be visualized individually. For example, Figure 11 shows only the 2D result of the population indicator. The red tiles indicate areas with virtually no houses.

On the visualization of bus service availability

The second case study is bus service availability in the city of Ålesund. The Public Transport Access

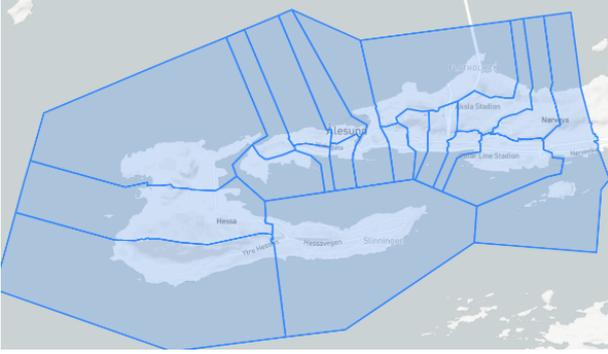


Fig. 12: Districts of Ålesund.

Level (PTAL) (Wu and Hine 2003) was computed for different districts of Ålesund and at different intervals of time.

Datasets used: Ålesund was divided into districts based on data coming from Statistics Norway¹⁸. Figure 12 shows the districts of Ålesund.

Data related to buses (bus stops, frequency) were collected from Entur,¹⁹ which operates the national registry for all public transport in Norway.

Indicator creation: To determine the PTAL for one district, we computed the PTAL for all buildings inside that district and then averaged the result. The building data were provided by Mapbox²⁰. For each location, the PTAL is computed following an algorithm described in (Wu and Hine 2003):

- Calculate the walk times from the location to the nearest service access points (bus stops in this example). Only bus stops within 640 meters are considered, because we assume people will walk up to eight minutes to a bus service²¹.
- For each bus stop, calculate the scheduled waiting time (SWT), which is half the time interval between arrivals of buses at this stop. The scheduled waiting time indicator reflects the frequency of buses arriving at a specific stop.
- For each bus stop, calculate the average waiting time (AWT), which is the SWT added to a reliability factor. The reliability factor reflects the fact that actual wait times can be longer due to buses arriving late. We set it to two minutes in this use case.
- For each bus stop, calculate the total access time (TAT), which is equal to the walk time added to the AWT.
- For each bus stop, calculate the equivalent doorstep frequency (EDF), which is a measure of what the service frequency would be like if the service was available without any walking time. It is equal to $EDF = 0.5 \times (\frac{60}{TAT})$.

¹⁸<https://kart.ssb.no> (As of Feb. 2022).

¹⁹<https://developer.entur.org/stops-and-timetable-data> (As of Feb. 2022).

²⁰<https://www.mapbox.com> (As of Feb. 2022).

²¹<https://content.tfl.gov.uk/connectivity-assessment-guide.pdf> (As of Feb. 2022).

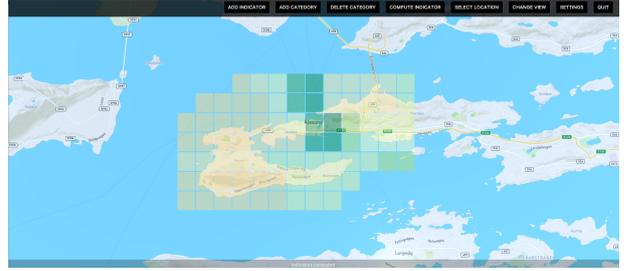


Fig. 13: 2D visualization of the bus service availability.

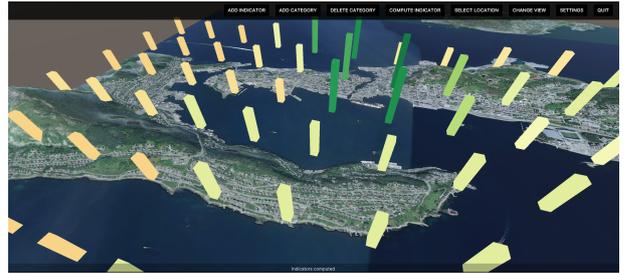


Fig. 14: 3D visualization of the bus service availability.

- Finally, calculate the Access Index (AI) by summing all the EDF. This gives a value from 0 (worst PTAL) to 40 (best PTAL).

Visualization of indicators: Figures 13 and 14 show, respectively, the 2D and 3D result with all weights set to 1. A green color means a high value and a red color means a low value, so we can see that like the walkability indicator, the bus service availability indicator is the highest in the city center, and the lowest in the south west of the city.

CONCLUSIONS

This work introduced GENOR, a generic platform for indicator assessment in city planning. It consists of a client-server architecture, where the client handles the creation, selection, and visualization of indicators with map-based views, while the server handles the management, storage, and processing of indicators.

The main asset of this platform is its capacity to be used for any problem, making it generic. Indeed, it can be used for different applications as demonstrated through two case studies: the visualization of the walkability assessment and the visualization of the bus service availability. The platform made possible the analysis of these indicators flexibly because of the possibility of defining both the area of interest and the weight applied to each indicator. These two examples were restricted to the city of Ålesund. However, the platform can work with any other city and even with areas of different scales, like countries.

Future work includes the addition of more visualization methods, such as graphs, charts, and di-

agrams, to facilitate the visualization of indicators. Comparison methods like ranking or clustering would allow multi-dimensional indicators to be easily compared. Also, methods to deal with inconsistent or missing data would make the platform more robust to real-world problems. Finally, carrying out user studies with stakeholders could be helpful to evaluate the platform in real-world usage scenarios. This could include adding additional indicators to support holistic spatial sustainability impact assessments in urban planning, exploring various multi-criteria aggregation functions to compute tile values, and techniques to elicit and determine indicator weights.

ACKNOWLEDGEMENTS

This work has been conducted in the context of the NORDARK project, funded by NordForsk, and the Smart Plan project, funded by the Research Council of Norway [grant number 310056].

REFERENCES

- Andrienko, G., Andrienko, N., Chen, W., Maciejewski, R., and Zhao, Y. (2017). Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):2232–2249.
- B.Longva, R.Torres, and D.Aspen (2021). Digital twin for walkability assessment in city planning. In *Digital Twin for Walkability Assessment in City Planning*.
- Chen, W., Huang, Z., Wu, F., Zhu, M., Guan, H., and Maciejewski, R. (2018). Vaud: A visual analysis approach for exploring spatio-temporal urban data. *IEEE Transactions on Visualization & Computer Graphics*, 24(09):2636–2648.
- Doraiswamy, H., Freire, J., Lage, M., Miranda, F., and Silva, C. (2018). Spatio-temporal urban data analysis: A visual analytics perspective. *IEEE Computer Graphics and Applications*, 38(5):26–35.
- Eberhardt, A. and Silveira, M. S. (2018). Show me the data! a systematic mapping on open government data visualization. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, dg.o '18, New York, NY, USA. Association for Computing Machinery.
- Fortini, P. M. a. and Davis, C. A. (2018). Analysis, integration and visualization of urban data from multiple heterogeneous sources. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities*, ARIC'18, page 17–26, New York, NY, USA. Association for Computing Machinery.
- Hall, C. M. and Ram, Y. (2018). Walk score® and its potential contribution to the study of active transport and walkability: A critical and systematic review. *Transportation Research Part D: Transport and Environment*, 61:310–324.
- Johansson, T., Segerstedt, E., Olofsson, T., and Jakobsson, M. (2016). Revealing social values by 3d city visualization in city transformations. *Sustainability*, 8(2).
- Perhac, J., Zeng, W., Asada, S., Arisona, S. M., Schubiger, S., Burkhard, R., and Klein, B. (2017). Urban fusion: Visualizing urban data fused with social feeds via a game engine. In *2017 21st International Conference Information Visualisation (IV)*, pages 312–317.
- Prandi, C., Nisi, V., Ribeiro, M., and Nunes, N. (2021). Sensing and making sense of tourism flows and urban data to foster sustainability awareness: a real-world experience. *Journal of Big Data*, 8.
- Psyllidis, A., Bozzon, A., Bocconi, S., and Bolivar, C. (2015). A platform for urban analytics and semantic data integration in city planning. In *A Platform for Urban Analytics and Semantic Data Integration in City Planning*.
- Ranzinger, M. and Gleixner, G. (1997). Gis datasets for 3d urban planning. *Computers, Environment and Urban Systems*, 21:159–173.
- Sauda, E., Wessel, G., Kosara, R., Chang, R., and Ribarsky, W. (2007). Legible cities: Focus-dependent multi-resolution visualization of urban relationships. *IEEE Transactions on Visualization & Computer Graphics*, 13(06):1169–1175.
- van lammeren, R., Houtkamp, J., Colijn, S., Hilferink, M., and Bouwman, A. (2010). Affective appraisal of 3d land use visualization. *Computers, Environment and Urban Systems*, 34:465–475.
- Weinberger, R. R. and Sweet, M. N. (2012). Integrating walkability into planning practice. *Transportation Research Record*, 2322:20 – 30.
- Wu, B. M. and Hine, J. P. (2003). A ptal approach to measuring changes in bus service accessibility. *Transport Policy*, 10(4):307–320. Transport and Social Exclusion.
- You, L., Zhao, F., Cheah, L., Jeong, K., Zegras, P. C., and Ben-Akiva, M. (2020). A generic future mobility sensing system for travel data collection, management, fusion, and visualization. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4149–4160.

Heuristic Techniques for Reducing Energy Consumption of Household

Sarah M. Daragmeh, Anniken Th. Karlsen and Ibrahim A. Hameed
Norwegian University of Science and Technology,
P.O. Box 1517, NO-6025, Aalesund, Norway

E-mails: smdaragm@stud.ntnu.no, anniken.t.karlsen@ntnu.no, ibib@ntnu.no

KEYWORDS

Advanced metering system, load/appliances scheduling, heuristic techniques, genetic algorithm (GA), particle swarm optimization (PSO).

ABSTRACT

Efficient energy demand management plays an essential role in smart grid, sustainable and smart cities applications and efforts to reduce CO₂ emissions. In this paper, we propose a framework for describing the household daily energy consumption and how it can be used to help residential households to perform appliance rescheduling to reduce energy consumption and hence reducing their energy bills while keeping resident's comfort. In this paper, heuristic optimization techniques such as genetic algorithm (GA) and particle swarm optimization (PSO) are used for solving the load scheduling problem. Due to its ability to deal with computational complex scenarios in less computational time using less and less computational resources, Heuristic optimization techniques are used. In the proposed model, dynamic pricing is adopted where the objective is to minimize the overall cost of electricity consumption and payments by scheduling different devices in a way that fulfil each individual's constraints and preferences. Here, MATLAB was used as the simulation platform. Simulation results showed that GA and PSO can optimize energy consumption and bills and at the same time fulfils needs and preferences of each individual customer.

INTRODUCTION

With the steady increased electricity demands in recent years, the need arises for mentoring energy usage and improving efficiency of energy use. Energy efficiency, in this regard, means using less energy to get the same job done, while cutting energy bills and reducing pollution. Smart energy metering technology is crucial for monitoring and improving energy use. Electricity smart metering become available to enormous numbers of end customers worldwide. By the end of 2020, it reached around 72% of European consumers (European Commission Joint Research Centre, 2021). In Norway, 100% of electricity consumers have received smart meters by 1 January 2019 (NVE-RME, 2022). Advanced metering system (AMS) allows the consumers to track their power usage and receive information about their electricity consumption and enables the distribution companies to move to a smart distribution system depending on current energy demands (Istad, 2019).

As power consumption is continuously increasing, the need arises for understanding consumption patterns, i.e., measurement and analysis of consumption overtime, and consumer's behavior. There are several load management strategies, which allow both utility companies and consumers

to detect and control overloads (Gaur et al. 2017). Demand-side management (DSM) in Smart Grid (SG) is a strategy that enables a more efficient and reliable grid operations. In this approach, there are two main functions: energy management and demand-side control activities for end-users.

In a residential area, every smart home is equipped with energy management controller (EMC) and smart meters to provide stable and reliable bi-directional communication between utilities and consumers. The communication between EMC and electrical appliances, sensors, local generation, and energy storage systems (ESSs) is done through home area network (HAN). After each data collection, EMC Sends it to SG domain. Figure 1 shows a simple architecture of DSM architecture.

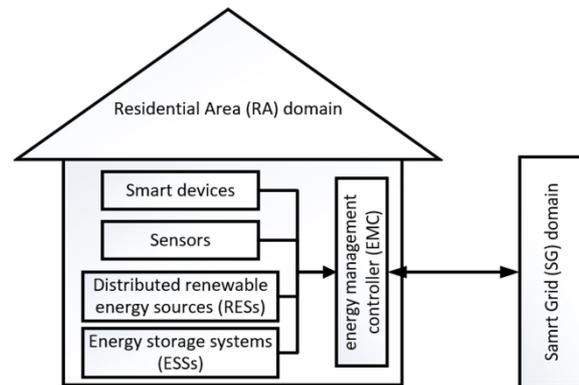


Figure 1: Simple Architecture of DSM

With an objective of contributing residents' awareness on efficient energy consumption we have investigated energy demand and how this affects electricity prices dynamically. More specifically we have looked closer at two popular heuristic optimization techniques: genetic algorithm (GA) and particle swarm intelligence (PSO) to solve the appliances' scheduling problem due to their capabilities in solving this kind of complex problems (Haupt and Haupt, 2004). With an objective of minimizing the overall cost of electricity payment by scheduling different devices according to their individual constraints, dynamic pricing was adopted. MATLAB was used as our simulation platform.

The rest of the paper is organized as follows: Related work is described in the next section. Then we explain the load scheduling model developed, followed by a discussion of simulation results. Finally, conclusions are drawn, and future work suggested.

RELATED WORKS

In recent years, smart grid played a significant role in designing sustainable systems that promotes energy efficiency and reducing CO₂ emissions. A smart grid is an electrical grid which includes a variety of operation and

energy measures including: AMSs, smart appliances, RESs, and energy efficient utilizing (Nejad et al. 2013; Saleh et al. 2015; Rahim et al. 2016). Electronic power conditioning and control of electricity production and distribution are important aspects of smart grid technology (Benysek, 2011). In the demand side, using smart home concepts using smart devices, sensors, RESs, ESSs, and EMC will lead to better demand side management, as it is shown in Figure 1. There have been many attempts to optimally schedule smart appliances in a way that enhances energy efficiency. Rahim et al. (2016), evaluated the performance of home energy management controllers designed based on a set of heuristic algorithms; GA, binary PSO, and ant colony optimization (ACO) algorithms. They solved the load scheduling problem as a non-deterministic polynomial-time NP-hard scheduling problem. Mehrshad (2013) considered the problem as a multi-objective optimization problem and provided a solution based on GA. Heuristic algorithms were widely used by many researchers to solve the appliance scheduling problem using GA (Cardenas et al. 2009; Yogyong and Audomvongseree 2011; AboGaleela et al. 2012; Chen et al. 2013; Mehrshad et al. 2013; Zhao et al. 2013; Oladeji and Olakanmi 2014; Rahim et al. 2016; Rasheed et al. 2016), PSO (Pedrasa et al. 2009; Zhou et al. 2014; Mahmood et al. 2016), and ACO (Liu et al. 2011; Hazra et al. 2012; Dethlefs et al. 2015).

In optimizing load scheduling, beside cutting energy costs for the end customers, other objectives have been considered. For instance, AboGaleela (AboGaleela et al. 2012) considered a load distribution scheme by applying one of following load control strategies: load shifting, peak clipping, valley filling, or load building over time. Minimizing the peak to average ratio (PAR) (AboGaleela et al. 2012; Zhao et al. 2013; Rahim et al. 2016; Rasheed et al. 2016), and load scheduling over multiple consumers in a defined neighborhood area (Mohsenian-Rad et al. 2010) are also considered.

In our proposed model, the objective is to minimize the energy consumption bill while keeping the resident's comfort. The proposed objective function in this model can easily be modified to accommodate other objectives and needs.

SYSTEM MODEL

Residents are an essential element of the smart energy consumption model. In this model, the aim is to increase the customers' awareness of their energy consumption by analyzing their historical energy consumption data that is recorded by AMS. Then, a load scheduling model is designed incorporating the use of heuristic optimization algorithms such as GA and PSO with the aim to enable the customers to efficiently control their energy consumption. The proposed method will be described in detail in the below sections.

Energy consumption model

Let $A = \{a_1, a_2, a_3, \dots, a_m\}$ be the set of appliances in the house, where m is the total number of appliances. Then, by dividing the day to small time slots (e.g., hours), the daily energy consumption of an appliance can be calculated using the equation:

$$E(a, t) = \{E(a, t_1) + E(a, t_2) + \dots + E(a, t_{t_{max}})\} \quad (1)$$

Where $E(a, t_1)$ is the energy consumption of the appliance a in the time slot t_1 .

The total consumption demand for the all the appliances in one day is then calculated as follows:

$$E = \sum_{t=1}^{t_{max}} \sum_{i=1}^m E(a_i, t) \quad (2)$$

Where m is the number of appliances, t is the time in hours and t_{max} is 24 hours. The energy consumption of each appliance depends on the it's characteristics and the user lifestyle. To manage the energy consumption through appliance scheduling; appliances are classified into two categories: shiftable and non-shiftable appliances (see Table 1).

Load categorization

The power consumption pattern of different types of consumers depends on the kind of appliances, which is used in the consumer's house. In general, electrical appliance can be categorized into schedulable (i.e., shiftable) and non-schedulable (i.e., non-shiftable) appliances. Non-shiftable appliances are used in a specific period with non-changed power level. These appliances include essential equipment such as: lights, cooker, kettle, ventilation, etc. In contrast, shiftable appliances such as washing machine, electrical vehicle and clothes dryer can be moved to another time to use. For example, we can charge the electrical vehicle during the night to avoid the peak hours to reduce the energy costs. Table 1 summarizes most of the used appliances in a typical Norwegian house/apartment categorized into shiftable (S) and non-shiftable (NS) appliances.

Table 1: Household Electrical Load

	Power (KW)	Quantity	Load type
Television	0.1	1	NS
PC	0.1	2	NS
Phone	0.05	2	NS
Bulbs (inside and outside)	0.025	10	NS
Iron	1.5	1	NS
Ventilation	0.5	1	NS
Refrigerator	0.160	1	S
Water heater	3	1	S
Space heater	2	1	S
Washing machine	1.5	1	S
Dish washer	3	1	S
Clothes dryer	4	1	S
Electrical car	4	1	S
Coffee machine	1.5	1	NS
Oven	3	1	NS
Freezer box	0.175	1	S
Microwave	0.8	1	NS
Cook top	3	1	NS
Hoover	0.7	1	NS
Hair dryer	0.75	1	NS

In our proposed model, we selected only four appliances in the scheduling optimizing problem, as it is shown in Table 2. We selected these four appliances for simplicity and as a proof-of-concept, but other shiftable appliances can be easily added to the model.

Table 2: Parameters of shiftable appliances

Appliance	Start time (h)	End time (h)	Power (KW)	Operational time (h)
Washing machine	7am	7pm	1.5	2
Clothes dryer	9am	9pm	4	2
Dish washer	6am	10pm	3	2
Electrical car	16pm	6am	4	4

Energy price model

Based on the daily energy demand, the time periods in the day can be classified as peak or non-peak hours (Rahim et al. 2016). During peak hours, the cost of the energy consumption is the highest. There are several tariff models that can be used to define electrical energy prices for a full day or for shorter periods during the day. Real-time electricity prices (RTEP) can change hourly reflecting the utility cost of supplying energy to consumers at that specific time. In Norway, this is called “spotpris” or spot price that follows the prices in the Nordpool which change hourly (Nordpoolgroup.com). In contrast to RTEP, ToU tariff model is defined for electricity prices depending on the time of a day and it is pre-defined in advance. Critical peak pricing (CPP) is a variant of ToU, and the price is considerably raising in the high demand (e.g., peak hours) (Oladeji and Olakanmi, 2014). In our model, we used ToU by considering the energy demand side and historical spot prices reference to Aalesund¹ region.

The total energy cost for each time slot t is the summation of the energy consumed by the ON appliances at this time slot, multiplied by the price at this time slot.

Problem statement

To reduce the energy consumption cost, the user can schedule the shiftable appliances to perform their jobs on non-peak hours. The non-shiftable devices must operate at any time depending on their characteristics or the user’s needs and preferences. Then the reduction of electricity bill is not possible with non-shiftable appliances. But the electricity usage cost can still be reduced by scheduling the shiftable appliances. In this model, heuristic algorithms such as GA and PSO, will be used to solve the scheduling problem in a way that minimizes the defined cost function.

Objective/cost function

The overall objective/cost function is to minimize the electricity bill by scheduling the shiftable devices to perform their jobs at optimal time where energy cost is minimum. The multi-objective cost function has two parts: minimizing the electricity bill and minimizing the waiting time to keep the user’s comfort. Each shiftable appliance has start time (st), end time (et), and operation time (ot) as it is shown in Figure 2.

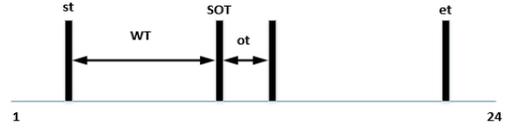


Figure 2: Parameters of appliance

The total electrical energy consumption cost at each hour is the total energy consumption at this hour multiplied by the electricity price at this hour as follows:

$$Cost_i = p_i \sum_{k=1}^m E(a_k, t_i) \quad (3)$$

where $Cost_i$ is the total electricity consumption cost of an hour i , p_i is the electricity price of an hour i , $E(a_k, t_i)$ is the energy consumption by an appliance k at an hour i . m is the total number of appliances. Then the daily cost is the summation of the cost of each hour as follows:

$$Cost = \sum_{i=1}^{24} Cost_i \quad (4)$$

To keep residents’ comfort, we consider the user’s wish to switch on the appliance at the given start time (st), as it is shown in Figure 2. Then, we schedule the shiftable appliances in a way that minimizes the waiting time, as it is shown in Eq. (5):

$$WT = \sum_{k=1}^m (SOT_k - st_k) \quad (5)$$

where WT is the total waiting time for all the shiftable appliances, SOT_k is the start operation time for device k , and st_k is the given possible start time by the user for the appliance k .

Then the objective function to select better or optimized solution can be modeled as follows:

$$\min \left(w_1 \left(\sum_{i=1}^{24} Cost_i \right) + w_2 \left(\sum_{k=1}^m (SOT_k - st_k) \right) \right) \quad (6)$$

where w_1 and w_2 are weights of two parts of objective function and their values are between 0 and 1, and $w_1 + w_2 = 1$.

Genetic Algorithm

Genetic algorithm (GA) is the most popular heuristic technique. GA is an optimization and search technique based on the principles of genetic and natural selection (Haupt and Haupt, 2004). A GA allows a population composed of many individuals to evolve under specified selection rules to a state that maximize the fitness (i.e., minimizes the cost function). Genetic algorithms (GAs) were invented by John Holland in 1960s and were developed by him and his students in 1960s and 1970s (Holland, 1975). GA belongs to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution such as selection (reproduction), crossover (recombination) and mutation (altering). The evolution process starts from a population of individuals generated

¹ Aalesund, is a municipality in Møre og Romsdal County, western cost of Norway.

randomly within the search space and continues for generations. In each generation, fitness of every individual is evaluated, and multiple individuals are randomly selected from the current population based on their fitness and modified by recombination and mutation operation to form a new population. Then this new population will be used for the next generation of the evolution. In general, the search process ends when either a maximum number of generations have been produced or a fitness level has been reached for the population. The flowchart of GA is shown in figure 3.

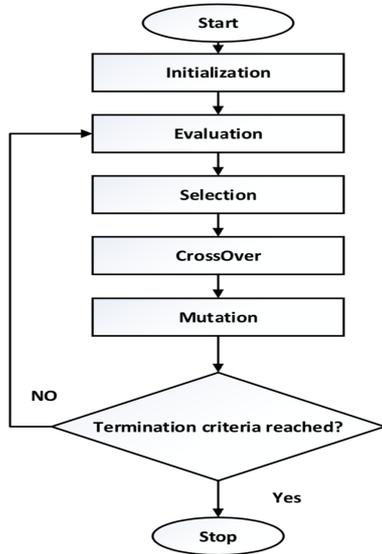


Figure 3: Flowchart of GA

GA for appliances' scheduling problem

In this scheduling problem, the objective is to find the optimal start operation time for the shiftable appliances. Then, the chromosome length is the number of shiftable appliances, and the variables are the start operation times (*SOPs*) for the appliances as follows:

$$\text{chromosome} = [SOT_1, SOT_2, \dots, SOT_m] \quad (7)$$

Where m is the number of shiftable appliances. and the *SOTs* take only integer values.

Particle Swarm Optimization

Particle swarm Optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution regarding a given measure of quality. Kennedy and Eberhart introduced PSO in 1995 (Kennedy and Eberhart, 1995). PSO was originally used to solve non-linear continuous optimization problems, but more recently it has been used in many practical, real-life application problems. For example, PSO has been successfully applied to track dynamic systems (Eberhart and Shi, 2001) and evolve weights and structure of neural networks (Zhang et al. 2000). PSO draws inspiration from the sociological behavior associated with bird flocking. It is a natural observation that birds can fly in large groups with no collision for extended long distances, making use of their effort to maintain an optimum distance between themselves and their neighbors.

The PSO methodology operates by placing a group of individual particles into a continues search space, wherein

each particle is initialized with a random position and a random initial velocity in the search space. The position and velocity are updated synchronously in each iteration of the algorithm. Each particle adjusts its velocity according to its own flight experience and the other's experience in the swarm in such a way that it accelerates towards positions that have high fitness values in previous iterations. In other words, each particle keeps track of its coordinates in the solution space that are associated with the best solution that has achieved so far by itself. This value is called personal best (*pbest*), Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called (*gbest*). So, the basic concept of PSO lies in accelerating each particle toward its *pbest* and the *gbest* locations, with a random weighted acceleration at each time step. Figure 4 shows the flow chart of a standard PSO algorithm.

The modification of the particle's position can be mathematically modeled according to following equations:

$$\vec{v}(k+1) = w\vec{v}(k) + c_1\vec{R}_1(\vec{pbest} - \vec{s}_i(k)) + c_2\vec{R}_2(\vec{gbest} - \vec{s}_i(k)) \quad (8)$$

Where,

$\vec{v}(k)$ is the velocity of a particle at iteration k .

\vec{R}_1 and \vec{R}_2 are random numbers in the range of $[0,1]$ with the same size of the swarm population.

c_1 and c_2 are learning factors which will be fixed through whole the process.

w is the inertia weight, and it is calculated as:

$$w = w_{start} - \frac{w_{start} - w_{end}}{K} k \quad (9)$$

Then the new position for the particles is the addition of the position at k iteration and the distance that the particles will fly with the new velocity $\vec{v}(k+1)$. The position is updated by:

$$\vec{s}_i(k+1) = \vec{s}_i(k) + \vec{v}(k+1) \quad (10)$$

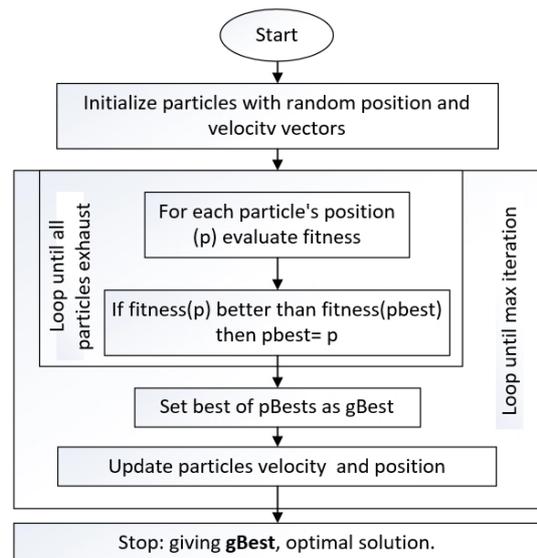


Figure 4: Flowchart of general PSO algorithm

PSO for appliances' scheduling problem

As it is described in GA, PSO can also be applied to find the optimum start operation time SOP for each shiftable appliance. The particle position vector includes the start operation times as the following:

$$\text{particle position} = [SOT_1, SOT_2, \dots, SOT_m] \quad (11)$$

SIMULATIONS AND RESULTS

In order to study the power consumption patterns; we analyzed the data we got from Mørenett². We got data for 1112 meters in Sunnmøre region, Norway. Table 3 summarizes the data; the consumptions are in hourly rates from 18 November 2018 to 25 November 2019. Figure 5 shows the total consumption for all the 1112 meters/consumers.

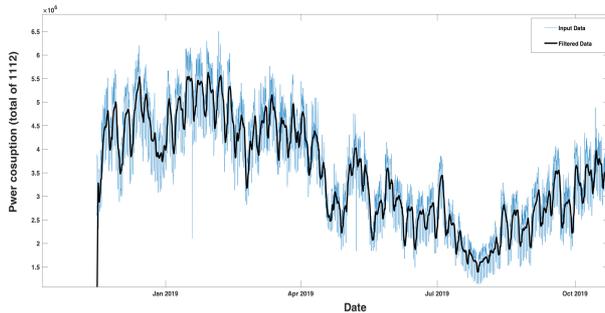


Figure 5: Energy consumption of 1112 meters

Table 3: Data from Mørenett

Apartments and houses	960
Industry	114
Cabin	38
Total	1112

Scheduling scenario

For scheduling, we have selected the week 47 (18 – 24 November 2019) to test and validate the proposed scheduling model. We have selected one of the customer's meters, then we added 4 appliances (washing machine, clothes dryer, dish washer and vehicle charger) randomly within time limits (Table 2). Figure 6 shows the original power consumption during week 47 and the consumption after adding the shiftable appliances.

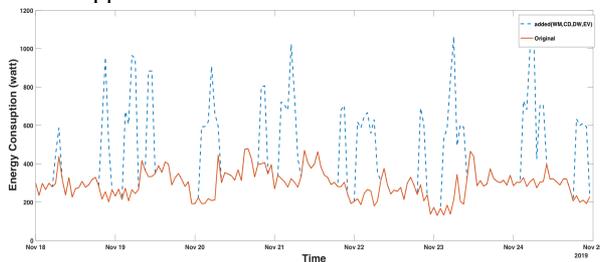


Figure 6: Week 47 (sample), added shiftable appliances

Scheduling by GA

For scheduling, we got the “spotpris” or spot prices for this region in week 47, as it is shown in Figure 7. Then, we

counted the bill for this end-user for week 47 which was 300.67 NOK. After that, we optimize it by finding an optimized/better scheduling.

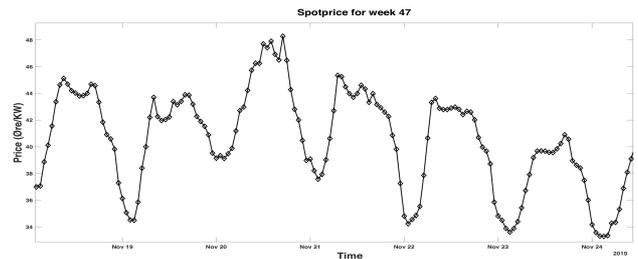


Figure 7: Week 47 spotprice from Nordpool AS (Nordpoolgroup.com)

We used GA with defined parameters in Table 4. The objective function defined in Eq. (6) is used with $w_1 = 0.7$ and $w_2 = 0.3$. The choice of weight values reflects the importance of energy cost compared to end-users' comfort.

Table 4: GA Parameters

Number of optimisation variables	4
Upper limit on optimisation variables	[19,21,22,28]
Lower limit on optimisation variables	[7,9,6,16]
Maximum iteration	100
Population size	100
Selection rate	0.8

Figure 8 shows our results. Upper part shows the original energy consumption and the optimized one. The middle figure shows the prices on hourly-based. The bottom figure shows the consumption cost for both original consumption and optimized consumption by applying scheduling.

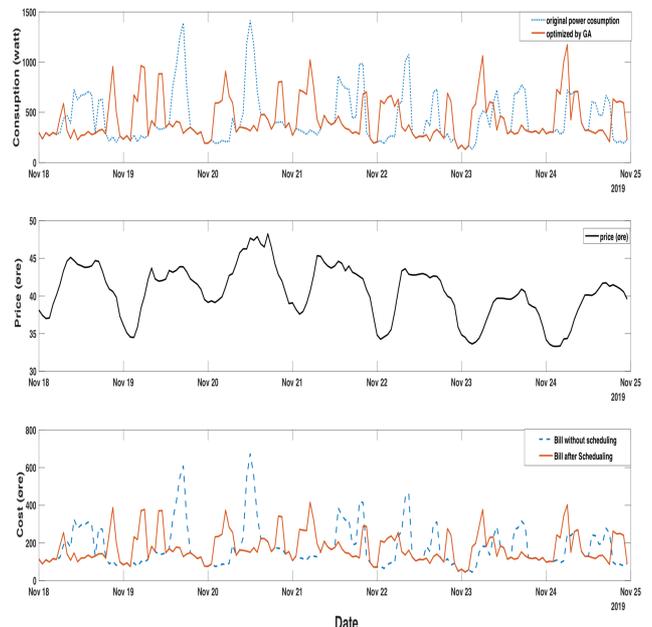


Figure 8: GA week-47 2019

² Mørenett is a network company in Sunnmøre and parts of Nordfjord

Table 5: Result for week 47-2019

Weekly bill without scheduling (NOK)	Weekly bill with scheduling (NOK)
300.67	291.17

Daily price model

To test our scheduling optimization model, we have designed a daily price model shown in Figure 9. We have considered the historical prices and the daily usage patterns.

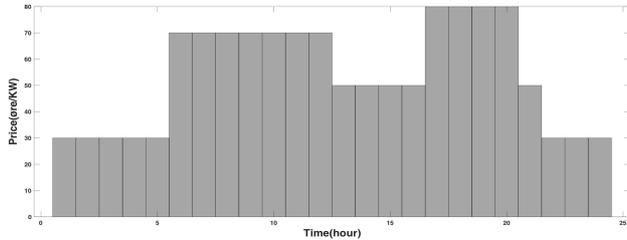


Figure 9: Daily prices

We used the shiftable appliances with defined parameters in Table 2. The schedule for each appliance is optimized by GA to minimize the objective function defined in Eq. (6). Table 6 summarizes the obtained results that shows that 9.5NOK could be saved daily, then 285NOK monthly and 3467.5NOK annually.

Table 6: Optimization Results

	Washing machine	Clothes dryer	Dish washer	Electrical vehicle	Daily cost of the shiftable appliances
Start time without GA	17	19	18	16	21.6 (NOK)
Start time with GA	7	13	13	22	12.1 (NOK)
Monthly saved = 285 NOK					
Annually saved = 3467.5 NOK					

Scheduling by PSO

In this section, we applied PSO algorithm in the same manner as it is in the previous section. A customized PSO toolbox has been developed from scratch in MATLAB environment since existing PSO in the optimization toolbox can't be modified to incorporate integer PSO. The PSO parameters are summarized in Table 7.

Table 7: PSO Parameters

Swarm size	200
Dimension of the problem	4
Maximum iteration	100
c1 (cognitive parameter)	1.5
c2 (social parameter)	1.5
C (constriction factor)	1
Inertia start	0.9
Inertia end	0.4
Upper limit on optimisation variables	[19,21,22,28]
Lower limit on optimisation variables	[7,9,6,16]
Maximum velocity	3

PSO provided very similar results to that obtained by GA. Figure 10 shows the convergence of PSO algorithm which shows that it can converge faster than GA.

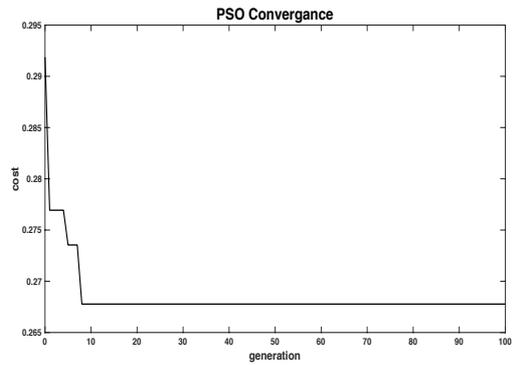


Figure 10: PSO convergence

CONCLUSIONS

The objective of this article is to increase the user awareness of energy efficiency and how optimization algorithms can be very beneficial in appliance scheduling in a way that minimizes consumption and at the same time keep customer's satisfaction. Demand side management is an essential part in design smart grids and sustainable energy systems. Residents can reduce the energy consumption considerably by using smart appliances that can be operated optimally. Also, scheduling the shiftable appliances by taking into consideration the varying electricity prices and the resident's comfort; would you reduces the electricity bill and consequently, reduce the electricity prices leading to smart grids and sustainability.

We have successfully used GA and PSO for scheduling four shiftable appliances. For GA, we used the MATLAB optimization toolbox, while we have developed a PSO-based optimization toolbox in MATALB that can handle integer decision variables. The simulation results for the defined scenario showed a cut in electricity bill up to 285NOK monthly on average.

Future work

This work can be extended in many different ways. For instance, developing a visualization tool that can be used by residents to increase their awareness of how to manage their energy consumption. The objective function can be extended to include different aspects of the problem such as including load control strategies (e.g., load shifting), and minimizing the peak to average ratio. Additionally, solving the problem as a multi-objective optimization problem should be investigated.

ACKNOWLEDGMENT

We would like to thank Mørenett for their help by providing the energy consumption data.

Also, we would like to express a special thanks to Professor Ricardo Torres for providing guidance and feedback throughout this project work.

REFERENCES

AboGaleela, M.; El-Sobki, M. and El-Marsafawy, M. 2012. "A two-level optimal DSM load shifting formulation using genetics algorithm case study: residential loads,". In Proc. of IEEE PES

- Power Africa 2012 Conference and Exposition, Johannesburg, South Africa.
- Benysek, G.; Kazmierkowski, M.; Popczyk, J. and Strzelecki, R. 2011. "Power electronic systems as a crucial part of Smart Grid infrastructure - a survey". *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 59(4), pp.455-473.
- Cardenas, J. J.; Garcia, A.; Romeral, J. L. and Andrade, F. 2009. "A genetic algorithm approach to optimization of power peaks in an automated warehouse," in *Industrial Electronics, 2009. IECON '09. 35th Annual Conference of IEEE*, Porto, Portugal.
- Chen, C.; Lan, M.; Huang, C.; Hong, Y. and Low, S. H. 2013. "Demand response optimization for smart home scheduling using genetic algorithm". In *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, Manchester, UK.
- Dethlefs, T.; Preisler, T. and Renz, W. 2015. "Ant-Colony based Self-Optimization for Demand-Side-Management", in: *Conference in SmartER Europe (E-World Energy and Water)*, Essen, pp. 1–8.
- Eberhart, R. and Shi, Y. 2001, "Tracking and optimizing dynamic systems with particle swarms", *Proc. Congress on Evolutionary Computation 2001*, Seoul, Korea
- Energyusecalculator.com. 2022. "Energy Use Calculator – Calculate electricity usage and energy cost of any device". [online] Available at: <http://energyusecalculator.com/index.htm> [Accessed 8 February 2022].
- Enerwe.no. 2017. "Så mye svinger strømprisen i løpet av døgnet". [online] Available at: <https://enerwe.no/sa-mye-svinger-strømprisen-i-lopet-av-dognet/144438> [Accessed 8 February 2022].
- European Commission Joint Research Centre, 2021. "Smart Metering deployment in the European Union. [Online]. Available at: <https://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters>. [Accessed 08 February 2022].
- Gaur, G.; Mehta, N.; Khanna, R. and Kaur, S. 2017. "Demand side management in a smart grid environment," *IEEE International Conference on Smart Grid and Smart Cities (ICSGSC)*, 2017, pp. 227-231, doi: 10.1109/ICSGSC.2017.8038581.
- Hazra, J.; Das, K. and Seetharam, D. P. 2012. "Smart Grid Congestion Management through Demand Response", in: *IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, Tainan, pp. 109–114.
- Haupt, R. and Haupt, S., 2004. "Practical genetic algorithms", 2nd Ed, J. Wiley, 2004.
- Holland, J. 1975, "Adaptation in Natural and Artificial Systems", Ann Arbor: University of Michigan Press.
- Istad, m., 2019. "Data from HAN ports fitted to smart meters (AMS) can provide you with valuable information" - #SINTEFblog. [online] #SINTEFblog. Available at: <https://blog.sintef.com/sintefenergy/han-port-smart-meters-ams/> [Accessed 10 February 2022].
- Kennedy, J. and Eberhart, R. 1995, "Particle Swarm Optimisation", *Proceedings of IEEE International Conference on Neural Networks IV*, pp. 1942–1948.
- Liu, B.; Kang, J.; Jiang, N. and Jing, Y. 2011. "Cost control of the transmission congestion management in electricity systems based on ant colony algorithm", *Energy Power Eng.* 3 pp17–23.
- Mahmood, D.; Javaid, N.; Alrajeh, N.; Khan, Z.A.; Qasim, U. and Imran Ahmed, M. I. 2016. "Realistic scheduling mechanism for smart homes", *Energies* 9 202.
- Mehrshad, M.; Tafti, A. D. and Effatnejad, R. 2013. "Demand-side management in the smart grid based on energy consumption scheduling by NSGA-II". *International Journal of Engineering Practical Research*, vol. 2, no. 4.
- Mohsenian-Rad, A.; Wong, W. S. and Jatskevich, J. 2010. "Autonomous demand side management based on game-theoretic energy consumption scheduling for the future smart grid". *IEEE Trans. on Smart Grid*, vol. 1, no. 3, pp. 320-331.
- Molla, T. 2020. "Smart Home Energy Management System". In B. Khan, H. Alhelou, & G. Hayek (Eds.), *Handbook of Research on New Solutions and Technologies in Electrical Distribution Networks* (pp. 191-206). IGI Global. <https://doi.org/10.4018/978-1-7998-1230-2.ch011>
- Nejad, M. F.; Saberian, A. M. and Hizam, H., et al. 2013. "Application of smart power grid in developing countries". *IEEE 7th International Power Engineering and Optimization Conference (PEOCO)* (PDF). IEEE. pp. 427–431. doi:10.1109/PEOCO.2013.6564586. ISBN 978-1-4673-5074-7.
- Nordpoolgroup.com. 2022. "Market data | Nord Pool". [online] Available at: <https://www.nordpoolgroup.com/Market-data1/Dayahead/Area-Prices/NO/Hourly/?view=chart> [Accessed 8 February 2022].
- NVE-RME, 2022. "Smart metering (AMS)". [Online]. Available at: <https://2021.nve.no/norwegian-energy-regulatory-authority/retail-market/smart-metering-ams/>. [Accessed 08 February 2022].
- Oladeji, O. and Olakanmi, O. O. 2014. "A genetic algorithm approach to energy consumption scheduling under demand response," *2014 IEEE 6th International Conference on Adaptive Science & Technology (ICAST)*, Ota, 2014, pp. 1-6. doi: 10.1109/ICASTECH.2014.7068096.
- Pedrasa, M. A. A.; Spooner, T. D. and MacGill I.F. 2009. "Scheduling of Demand Side Resources Using Binary Particle Swarm Optimization", *IEEE Trans. Power Syst.* 24 1173–1181.
- Rahim, S.; Javaid, N.; Ahmad, A.; Khan, S.; Khan, Z.; Alrajeh, N. and Qasim, U. 2016. "Exploiting heuristic algorithms to efficiently utilize energy management controllers with renewable energy sources". *Energy and Buildings*, 129, pp.452-470.
- Rasheed, M. B.; Javaid, N.; Awais, M.; Khan, Z. A.; Qasim, U.; Alrajeh, N.; Iqbal, Z. and Javaid, Q. 2016. "Real time information based energy management using customer preferences and dynamic pricing in smart homes", *Energies* 9 542.
- Saleh, M. S.; Althabani, A.; Esa, Y.; Mhandi, Y. and Mohamed, A. A. 2015. "Impact of clustering microgrids on their stability and resilience during blackouts". *International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*. pp. 195–200.
- Strøm fra Glitre Energi. 2019. "Timesprising - Strøm fra Glitre Energi". [online] Available at: https://www.glitreenergi.no/strom/timesprising/?fbclid=IwAR2khAh4rbKgf-iZmbc35iy9nb0te7jEz6p_a6SVBwhgnWqXyyxpK_Ovqlo [Accessed 8 February 2022].
- Yogyong, W.; and Audomvongseeree, K. 2011. "Optimal fuel allocation for generation system using a genetic algorithm" in *Proc. of the 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Khon Kaen, Thailand.
- Zhang, C.; Shao, H. and Li, Y. 2000, "Particle Swarm Optimisation for Evolving Artificial Neural Network", In the 2000 IEEE International Conference on Systems, Man, and Cybernetics, vol.4, pp.2487-2490.
- Zhao, Z.; Lee, W. C.; Shin, Y. and Song, K. 2013. "An optimal power scheduling method for demand response in home energy management system", *IEEE Trans. Smart Grid* 4. 1390–1400.
- Zhou, Y.; Chen, Y.; Xu, G. and Zhang, Q. 2014, "Home Energy Management with PSO in Smart Grid" in: *Industrial Electronics (ISIE)*, IEEE 23rd International Symposium, Istanbul, pp. 1666–1670.

AUTHOR BIOGRAPHIES

SARAH M. DARAGMEH is a project manager at Caverion Norge AS, and she is a master student in the Electrical Power Engineering program at the Norwegian University of Science and Technology (NTNU), Norway. She has also a MSc degree in Simulation and Visualization from NTNU, Norway. She obtained her bachelor's degree in electrical engineering from An Najah University, Palestine. Her research interest includes Optimization, Machin Learning and Smart Grid.

ANNIKEN Th. KARLSEN is an Assoc. Professor at the Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering at the Norwegian University of Science and Technology (NTNU). Karlsen teaches and researches within Technology Management and Digital Transformation and is Head of the Sustainable Digital Transformation research group. She has, among others, a PhD degree in information science from the University of Bergen and a MSc degree in Information Technology from the University of Aalborg, Denmark. Karlsen has done empirical research within several sectors, including maritime, marine, offshore, food, consultant, health and banking.

IBRAHIM A. HAMEED is a Professor at the Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway. Hameed is Deputy Head of research and innovation within the same department. Hameed is an IEEE senior member and elected chair of the IEEE Computational Intelligence Society (CIS) Norway section. Hameed has a Ph.D. degree in Industrial Systems and Information Engineering from Korea University, Seoul, South Korea and a PhD degree in Mechanical Engineering from Aarhus University, Aarhus, Denmark. His current research interest includes Artificial Intelligence, Machine Learning, Optimization, and Robotics.

LEARNED PARAMETERIZED CONVOLUTIONAL APPROXIMATION OF IMAGE FILTERS

Olga Chaganova
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia;
Moscow Institute of Physics and Technology
(National Research University)
Institutskiy per. 9, Dolgoprudny, 141701, Russia
E-mail: o.chaganova@visillect.com

Anton Grigoryev
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: me@ansgri.com

KEYWORDS

Convolutional neural networks, grayscale morphology, edge detection, computational efficiency, image processing

ABSTRACT

Multilayer neural networks are considered universal approximators applicable to a wide range of problems. There are quite detailed theoretical and applied studies for fully connected networks, while for convolutional networks the results are more scarce. In this paper, we tested the approximating capability of deep neural networks with typical architectures like ConvNet, ResNet, and UNet as applied to classical image processing algorithms. Canny edge detector and grayscale morphological dilation with the disk structuring element were selected as target algorithms. As we have seen, even relatively lightweight neural models are able to approximate a filter with fixed parameters. Since image processing algorithms are parameterized, we considered different approaches to the parameterization of the neural networks and discovered that even the simplest one, which is an adding parameters in the input image channels, works well for filter with low parameter count. Also, we measured an inference time of a neural network approximation and a classical implementation of the grayscale dilation with the disk structuring element. Starting from a certain radius, a neural network works faster than the algorithm even on one core of the CPU without fine-tuning the architecture for performance, thus confirming the viability of ConvNets as a differentiable approximation technique for optimization of classical-based methods.

INTRODUCTION

Nowadays, artificial neural networks (ANNs) are considered universal approximators for analytic functions of any complexity. Theoretical studies have been conducted for feedforward neural networks; in particular, Hornik et al. (1989) proved that standard perceptron is capable of approximating any Borel measurable func-

tion from one finite-dimensional space to another to any desired degree of accuracy. Similar studies for convolutional networks showed that a deep convolutional neural network can approximate any continuous function to an arbitrary accuracy when its depth is large enough (Zhou 2019). But there is still a lack of studies about the neural networks capability of approximating algorithms, particularly image processing algorithms. Many convolutional network architecture families, such as ConvNet, ResNet, UNet, etc., are successfully used in various image processing tasks (Li et al. 2021; Andreeva et al. 2019), including ones with existing algorithmic solutions (Karnaushkin and Sklyarenko 2022; Panfilova and Kunina 2020). However, it is not obvious that especially for tasks where specialized algorithmic solutions are known, a general-purpose neural network could solve the same task with comparable computational and representational efficiency. To research this aspect, we decided to test the ability of typical NN architectures to approximate individual image processing operations, since difficulties with this simplified problem would indicate deeper issues for use of neural networks as a general-purpose tool for image processing. We chose Canny edge detector and morphological dilation as examples of typical image processing operations since the former is comparatively well-researched from the neural approximation point of view (Fernández et al. 2011), and for the latter there is no computationally efficient algorithm anyway.

The rest of the paper is structured as follows: the next section provides a brief theoretical overview of existing work on approximating classical algorithms with neural networks and approaches to parameterization of the neural networks. The following section includes a description of the experiments and a discussion of the results. Finally, the paper is concluded with a summary and pointers to the possible future work.

RELATED WORK

In this section, we will describe the main existing results on the approximation of classical filters, which can

be considered as baselines for further research. Since classical algorithms have the important property of natural parametrization, which means they can be customized for specific requirements, we will also consider approaches to the parametrization of neural networks.

Neural network approximation of classical algorithms

The first works on the research topic appeared even in the 20th century. For example, De Ridder et al. (1999) investigated the application of neural networks to nonlinear image processing using Kuwahara filter for image smoothing as a target. The authors experimented with several architectures of a network, including single- and double-layer perceptrons with different sizes of the hidden layers and a specially designed modular neural network. Although the authors managed to train the model to smooth the images, the result was still too far from true Kuwahara filter smoothing. The reason for this could be due to unsuitable model architecture and loss function.

Fernández et al. (2011) attempted to approximate the Canny and Sobel edge detectors with single-layer perceptron. To take into account the spatial dependence in the images, a pixel and its eight adjacent pixels were fed to the input of the model, which almost corresponds to a convolution with a kernel size of 3x3. The authors did not manage to achieve a good quality of the approximation, which can probably be explained by the inability of the single-layer model to learn complex dependencies in the data. However, these studies can be taken as a baseline for subsequent experiments.

Fernández et al. (2011); De Ridder et al. (1999)) used neural networks with typical architecture: just a multilayer perceptron. A totally different approach was proposed by Zhukovskiy et al. (2018). The authors developed a complete reproduction of the Canny, Niblack, and Harris filters using only neural network operations: convolution, pulling, concatenation, etc., so the algorithm's parameters do not need to be selected manually because their optimal values will be determined during training. But this approach has its disadvantage since it requires creating a new architecture for each filter, which is very time-consuming and requires a deep understanding of the algorithm's inner workings. It also does not correlate with the aim of our study.

Parameterization approaches

In the previous works, the ability of neural networks to approximate a particular filter with fixed parameters was investigated, and those researches are of more theoretical interest. Even though in real-world applications this approach is acceptable, because a filter with specified parameters is often used, it is necessary to be able to customize the filter parameters when creating and optimizing prototypes of the final system. For example, Karnaushkin and Sklyarenko (2022) used Canny edge detector to develop a computer vision-based method of pre-alignment of a channel optical waveguide and a lensed fiber, and the values of the algorithm parameters

were being tuned during the development stage. Sometimes it is necessary to give a user the ability to specify the system parameters, as in the case of an ANN-based image pre-compensation system, which must adjust to the parameters of a particular person's eye (Yu et al. 2021) and thus cannot be retrained for each possible set of parameters.

Model parameterization can be achieved in several ways:

(1) by adding parameters values as metadata to the input image channels. This approach was used by Tziolas et al. (2020). In order to enrich the data and improve the quality of the model predicting the clay content of the soil, not only remotely sensed values, which constitute the spectral input channel, but also geographic coordinates, which constitute the auxiliary input channels, are fed as an input. This is the simplest way to provide dependence of the model on auxiliary parameters, which does not entail a significant increase in the size of the model, but in some cases, it might be not enough to learn more complex dependencies between parameters and expected output.

(2) by parameterizing the convolution kernel. Traditional convolutional kernels are shared for all examples in a dataset, which can significantly decrease the quality of the model in case the data are characterized by a large variability. Conditional convolutions (CondConv), proposed by Yang et al. (2020), and Dynamic convolutions (DynamicConv), proposed by Chen et al. (2020)), share a common idea: instead of using a single convolution kernel per layer, which is the same for any input, these types of convolutions aggregates multiple convolution kernels, which are input dependent. DynamicConv aggregates several kernels based upon their attentions while CondConv parameterizes the convolutional kernels as a linear combination with example-dependent scalar weights computed using a routing function with learned parameters. Another example of this approach is the Adaptive convolutional neural network (ACNN), presented in Kang et al. (2017). In ACNN the filter weights are generated with a learned sub-network, which is a simple multilayer perceptron and input of which is the side information or metadata. Replacing traditional convolution layers by convolutions with adaptive weights can disentangle the variations related to the side information and extract discriminative features related to the current context, but it can significantly increase the size of the model and slow down its inference time.

(3) by adding extra input layers for metadata to the neural network. Wei (2020) and Gessert et al. (2020) used a multiple-input neural network, which consists of two branches: the first one is a single- or multilayer perceptron that process a vector of metadata and the second one is a convolutional neural network that process an input image. The outputs of these sub-networks are then concatenated and processed with the common part of the network. Ellen et al. (2019) developed this idea and proposed several schemes for metadata incorporation. While in the previous works metadata was added

to a convolutional neural network, it can be added to a graph neural network too (Mudiyansele et al. 2021).

For our experiments, we have chosen two classical algorithms: Canny edge detector and grayscale morphology, namely, dilation with the disk structuring element. Both of them are often used in industrial recognition systems (e.g. Panfilova and Kunina (2020), Panfilova et al. (2021), Erlygin and Teplyakov (2021)). There is a baseline on an approximation of Canny operator (Fernández et al. 2011), but there was a feedforward neural network used, while research for convolutional neural networks was not conducted. Also, we did not manage to find any published studies for approximation of morphological dilatation. Moreover, there is no computationally efficient implementation of dilation with the disk structuring element compared, at least, to the rectangular structuring element, for which fast window size-independent algorithm (vHGW) is available (Limonova et al. 2020).

METHODS AND RESULTS

In this section, we will describe the results of experiments on approximation Canny edge detector and grayscale dilation with the disk structuring element. Using them as target filters allowed us to investigate two different approaches to approximating filters: regression and classification. Since Canny detects edges on the image, it might be considered as a classification of each pixel whether it is a part of an edge or not. On the other hand, morphological filters do not „classify“ each pixel; instead, they transform an input image, and this transformation might be considered as the image-to-image regression.

Approximation with fixed parameters

At first, we wanted to test the ability of the neural networks to approximate a particular filter with specified parameters. There are three key parameters of Canny edge detector (Canny 1986): standard deviation for Gaussian kernel σ , lower bound (low threshold) and upper bound (high threshold) for hysteresis thresholding. We specified these parameters as $\sigma = 1$, low threshold = 0.1, high threshold = 0.2. For morphological dilation, we approximated the filter with two sizes of the radius: $R_1 = 5$ as a baseline and $R_2 = 20$ to be sure that the receptive field of the network is large enough to approximate dilation with a big radius.

We tried three architectures: ConvNet, which is a stack of convolutional layers with ReLU as activation function, ResNet, and UNet. These architectures might be considered as the gold standard of convolutional networks because they are well-known and often used in many cases of image processing and computer vision (Li et al. 2021). We considered only the most typical neural network architectures without optimizing them for a particular filter. The reason is that our study aims to investigate the universality of standard neural network models for solving classical image processing problems for which there already exists a well-working analytical algorithm.

As we mentioned earlier, the problem of Canny edge detector approximation can be formulated in terms of a pixel-by-pixel classification, thus we used the binary cross-entropy loss function. The metrics are Matthew’s correlation coefficient, or MCC, (Cheng et al. 2021), and Intersection over Union, or IoU, (Zheng et al. 2019). For morphological dilation approximation, we used the mean squared error loss function and mean absolute error as a metric. Also, we compared different models using the loss function value on the test dataset.

For each filter, all three networks were trained with the same hyperparameters. Particularly, we used Adam optimizer with learning rate 0.001 and parameters $\beta_1 = 0.9, \beta_2 = 0.999$; cosine annealing scheduler with warm restarts with a period of 8 epochs; 50 epochs for training. As the dataset, we used Linnaeus5 with images of size 128x128 pixels split into training, validation, and test sets with 6000, 1000, and 1000 images, respectively.

All three models showed comparable results for morphology approximation, but for Canny detector approximation, the best quality was achieved with ResNet (table 1), that is why in the following experiments we used only this architecture. Our implementation of ResNet reproduces the classic implementation and differs only in the number of layers and the number of filters in convolutional layers. It consists of 7 residual blocks, each of them containing 2 convolutional layers with 12 filters. A visualization of the ResNet work is shown in the fig. 1 and fig. 2 for Canny edge detector and grayscale morphology with the disk structuring element, respectively.

Model	Num. of parameters	Metrics		
		MCC	IoU	Test BCE
ConvNet	25.8k	0.9185	0.9230	0.0462
ResNet	29.1k	0.9286	0.9322	0.0425
UNet	28.7k	0.8291	0.8481	0.1020

TABLE 1: Metrics for Canny edge detector approximation

Moreover, we compared inference times of the ResNet (fig. 3) with the skimage.morphology (v.0.17.2) implementation and noticed that starting from a certain radius, neural network approximation works faster than the classical algorithm even in single-threaded inference mode. It should be noted that the inference time of the neural network is constant and depends only on technical conditions of implementation, since the input image size is fixed and independent of the approximated filter parameters.

In fully parallel GPU inference, the ResNet is faster than the CPU implementation in all cases, and while such performance comparison is unfair, it illustrates one of the benefits of NN approximations, which is the availability of highly parallelized implementations. It is worth mentioning that we did not apply any optimization techniques to speed up inference of the ResNet, although there is a more computational efficient implementation of it (e.g. Lobanov and Sholomov (2019), which provides a threefold increase in the inference performance). The parameters of the testing system are:

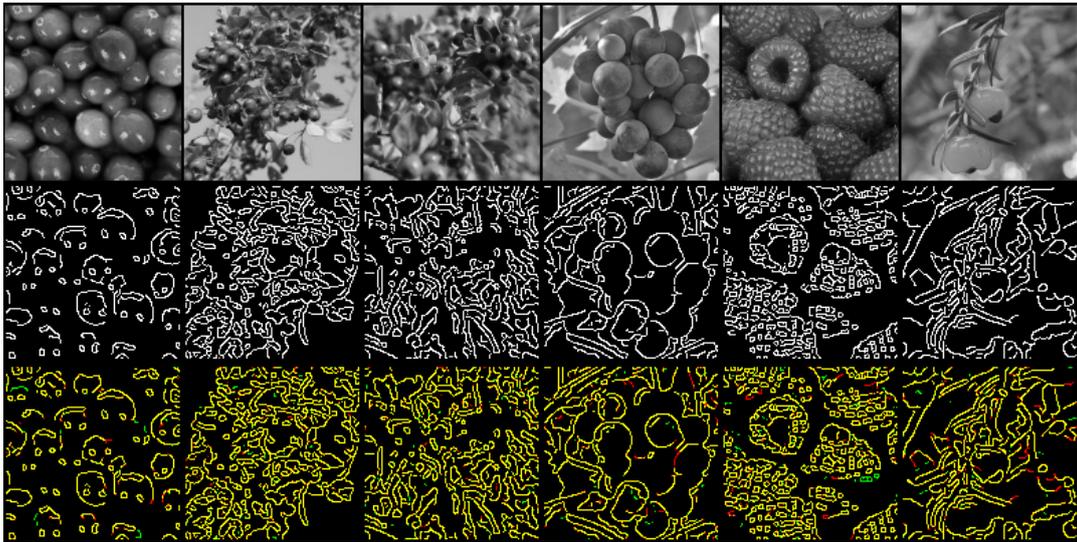


Fig. 1: Approximation of Canny edge detector with fixed parameters (1-st row: original images, 2-st row: Canny edge detector, 3-st row: pixel difference with the neural network output, where black, yellow, green, and red colours stand for true negative, true positive, false positive, and false negative pixels, respectively).

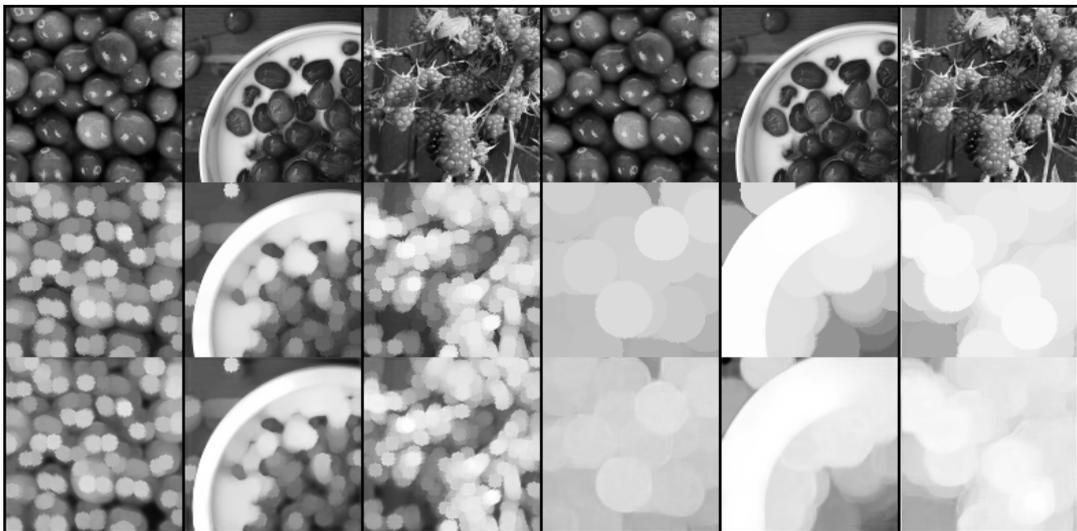


Fig. 2: Approximation of grayscale morphology with the disk structuring element and fixed parameters (1-st row: original images, 2-st row: morphological dilation, 3-st row: neural network). The first three images correspond to dilation with a radius of 5, while the second three correspond to dilation with a radius of 20. Two networks were trained to approximate the morphological dilation with different radius independently.

Intel(R) Xeon(R) W-2133 CPU @ 3.60GHz, 6 cores, 12 threads; GeForce RTX 2080 Ti 11 Gb GPU.

Thus, we found that typical neural network architectures are able to approximate at least some of the typical image processing algorithms, while having few enough weights to be computationally competitive with purpose-built algorithms.

Parameterized approximation

An important property of classical image processing algorithms is parameterization, therefore they can be customized for a particular task while neural networks usually implement a fixed operation. In real-world applications, it is often necessary for a developer or a user to be able to tune the parameters. That is why another aim of our research is to consider the approaches to the parameterization of neural networks. We started with

the simplest approach, which is adding filter parameters as auxiliary channels of an input image.

Canny detector has three key parameters: sigma, low threshold, and high threshold. The high threshold was parameterized via the low threshold multiplied by a coefficient k . We consider these parameters as independent uniformly distributed in the interval $[0; 3]$ for sigma, $[0.1; 0.2]$ for low threshold, and $[1.05; 2]$ for k , random variables. Since the distribution is multidimensional and its components are independent, it is crucial to sample from it with as maximum coverage of the probability space as possible. Latin hypercube sampling (LHS), proposed by McKay et al. (1979), ensures that the set of samples is a very good representative of the real variability. That is why we used it for sampling the set of parameters during training phase.

The training was organised as follows: at each epoch,

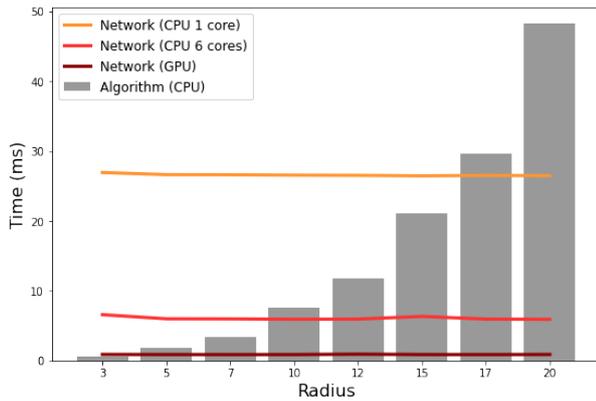


Fig. 3: Inference time of classical algorithm (disk-structured morphological dilation) and its neural approximation with (unparameterized) ResNet. The CPU algorithm implementation is single-threaded.

we sampled 1000 sets of parameters with LHS. For each image from the training set, we randomly chose a set of parameters and added its values to the channels of the image. Since there are 6000 images in the training set, each set of parameters was applied to roughly six images. In the case of continuous random variables, the number of possible sets of parameters is infinite, that is why it makes sense to increase the number of epochs so that the network can „see „as many sets of parameters as possible and learn how to approximate them. We trained our model for 200 epochs, which means there were 200000 sets of parameters. Also, we had to increase the number of convolutional kernels in the hidden layers from 12 to 16 compared to the unparameterized version. A visualization of the ResNet work is shown in the fig. 4. A dependence on the parameters is pronounced, but the quality of the approximation is slightly reduced compared to the approximation with fixed parameters. The Matthew’s correlation coefficient is 0.8917 and IoU is 0.9021.

Disk-structured morphological dilation has one parameter, which is the radius of the disk. Since the radius must be a positive integer, we consider it a discrete random variable, uniformly distributed in the interval [1; 20]. The model was trained in the same way as the Canny approximation network, but we did not use LHS because it is not necessary in case of one-dimensional discrete probability space. A visualization of the ResNet work is shown in the fig. 5. Received results showed that the simplest parameterization approach via channels of input image works well enough for some applications.

CONCLUSIONS

In this paper, we tested the hypothesis that neural networks can be effectively used to approximate and replace classical image processing operations. We chose two typical image filters, Canny edge detector and grayscale morphological dilation with the disk structuring element. This choice of algorithms allowed us to compare two different approaches to the approximation: classification and regression. As we have seen,

neural networks are equally good at both tasks, all that is required is to choose an appropriate loss function.

We used typical convolutional networks architectures like ConvNet, ResNet, and UNet with the standard loss functions (binary cross-entropy and mean squared error), not trying to optimize the architecture for a particular filter. As we have seen, even a neural network model with a relatively small number of parameters is able to approximate a filter with fixed parameters. The only requirement is the large enough number of hidden layers, which determine the size of the model’s receptive field. These results might be considered as an argument in favour of the tested hypothesis, obtained experimentally.

Since in real-world applications adjustability of filters might be crucial, we also tested the ability of a single neural network model to learn to approximate chosen filters with different parameter values. We have considered various approaches to parameterization of the neural networks, but even the simplest one, which is adding parameters in the input image channels, works well enough. There are still some artifacts, so for the better quality of the approximation, it is worthwhile to apply more complex approaches previously discussed.

Approximation of classical image processing algorithms with neural networks is not only interesting from a theoretical point of view but also leads to practical benefits:

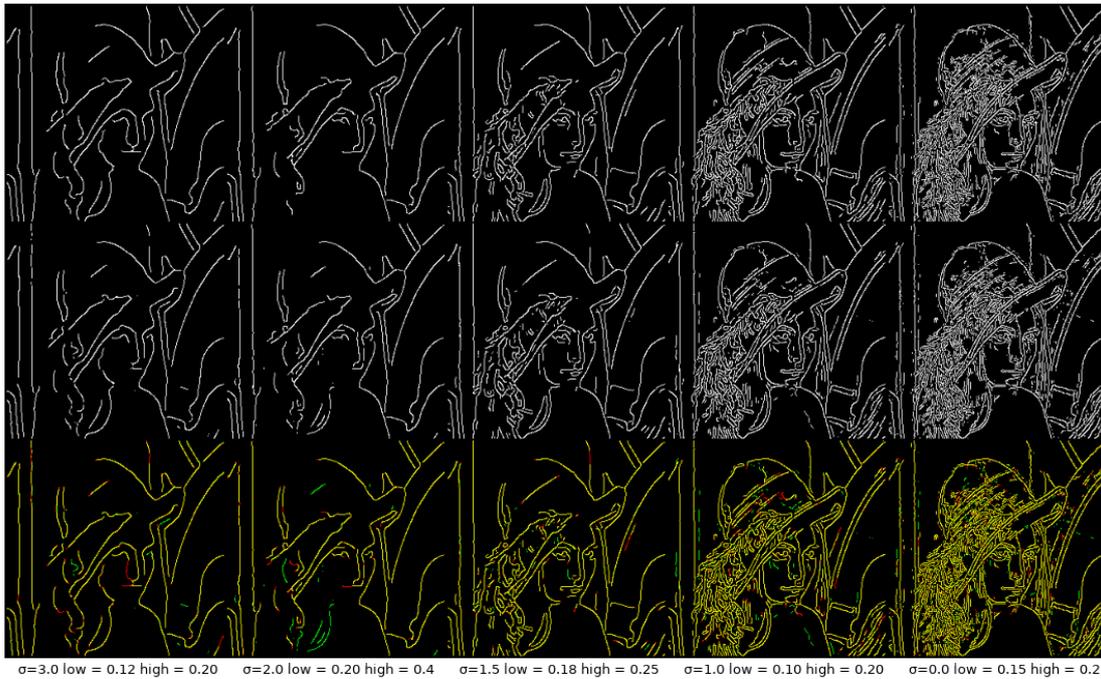
(1) a neural network is trainable, and thus can potentially achieve a better solution than a classical filter, which is deterministic and parameters of which must be chosen manually. For example, Jampani et al. (2016) proposed learnable bilateral filters, the performance of which was improved compared to a fixed parametric form;

(2) starting from the specified size of the radius, a neural network model (with ResNet-like architecture, in our case) works faster than the classical implementation of the grayscale morphology with the disk structuring element (fig. 3) even on one core of the CPU without applying any optimization techniques;

(3) despite the fact that inference speed of a classical algorithm might be lower compared to its neural network approximation in a single-core mode, neural networks can be easier parallelized both on multi-cores CPUs and GPUs, development of which, in addition, is constantly evolving in speed and power consumption (Di Febbo et al. 2018);

(4) a neural network approximation can replace the classical filter within a more complex system based on ANNs as well, allowing for end-to-end system training (Yi et al. 2016).

Thus, in this paper we have found arguments in favour of common hypothesis that neural networks are universal approximators experimentally. The direction of further research is the investigation of other approaches to parameterization of neural networks, which can help to achieve the better quality of approximation without significant increasing of the model’s parameters count.



$\sigma=3.0$ low = 0.12 high = 0.20 $\sigma=2.0$ low = 0.20 high = 0.4 $\sigma=1.5$ low = 0.18 high = 0.25 $\sigma=1.0$ low = 0.10 high = 0.20 $\sigma=0.0$ low = 0.15 high = 0.2

Fig. 4: Parameterized approximation of Canny edge detector (1-st row: Canny edge detector, 2-st row: neural network, 3-st row: pixel difference, where black, yellow, green, and red colours stand for true negative, true positive, false positive, and false negative pixels, respectively).

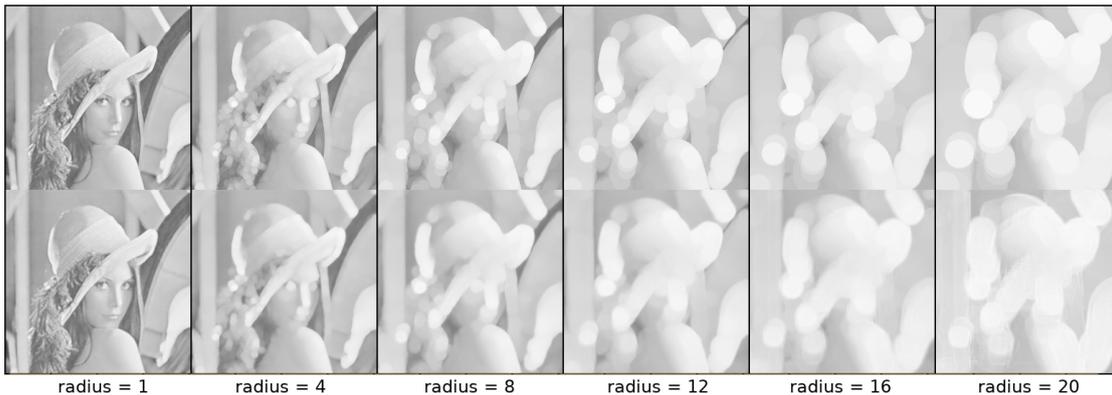


Fig. 5: Parameterized approximation of grayscale morphology with the disk structuring element (1-st row: morphological dilation, 2-st row: neural network).

ACKNOWLEDGMENT

This work is partially financially supported by Russian Foundation for Basic Research (project 18-29-26037).

REFERENCES

- Andreeva, Elena; Vladimir Arlazarov; Aleksandr Gayer; Evgeniy Dorokhov; Aleksandr Sheshkus; and Oleg Slavin. 2019. "Document recognition method based on convolutional neural network invariant to 180 degree rotation angle." *ITiVS*, (4):87–93. DOI: 10.14357/20718632190408.
- Canny, John. 1986. "A computational approach to edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698. doi: 10.1109/TPAMI.1986.4767851.
- Chen, Yinpeng; Xiyang Dai; Mengchen Liu; Dongdong Chen; Lu Yuan; and Zicheng Liu. 2020. "Dynamic convolution: Attention over convolution kernels." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Cheng, Xiaoqi; Junhua Sun; and Fuqiang Zhou. 2021. "A fully convolutional network-based tube contour detection method using multi-exposure images." *Sensors*, 21(12). ISSN 1424-8220. doi:10.3390/s21124095.
- De Ridder, Dick; Robert Duin; Piet. W. Verbeek; and Lucas J. Van Vliet. 1999. "The applicability of neural networks to non-linear image processing." *Pattern Anal. Appl.*, 2(2):111–128. ISSN 1433-7541. doi: 10.1007/s100440050022.
- Di Febbo, Paolo; Carlo Dal Mutto; Kinh Tieu; and Stefano Mattoccia. 2018. "Kcnn: Extremely-efficient hardware keypoint detection with a compact convolutional neural network." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 795–7958. doi:10.1109/CVPRW.2018.00111.
- Ellen, Jeffrey S.; Casey A. Graff; and Mark D. Ohman. 2019. "Improving plankton image classification using context metadata." *Limnology and Oceanography*:

Methods, 17(8):439–461. doi:10.1002/lom3.10324.

Erlygin, Leonid and Lev Teplyakov. 2021. “Improvement of a line segment detector based on a neural network by adding engineering features.” *Sensory systems*, 35(1):50–54. DOI: 10.31857/S0235009221010042.

Fernández, Andrea; Carlos Delgado-Mata; and Ramiro Velázquez. 2011. “Training a single-layer perceptron for an approximate edge detection on a digital image.” *Proceedings - 2011 Conference on Technologies and Applications of Artificial Intelligence, TAAI 2011*. doi:10.1109/TAAI.2011.40.

Gessert, Nils; Maximilian Nielsen; Mohsin Shaikh; René Werner; and Alexander Schlaefer. 2020. “Skin lesion classification using ensembles of multi-resolution efficientnets with meta data.” *MethodsX*, 7:100864. ISSN 2215-0161. doi:https://doi.org/10.1016/j.mex.2020.100864.

Hornik, Kurt; Maxwell Stinchcombe; and Halbert White. 1989. “Multilayer feedforward networks are universal approximators.” *Neural Networks*, 2(5):359–366. ISSN 0893-6080. doi:10.1016/0893-6080(89)90020-8.

Jampani, Varun; Martin Kiefel; and Peter V. Gehler. 2016. “Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4452–4461.

Kang, Di; Debarun Dhar; and Antoni Chan. 2017. “Incorporating side information by adaptive convolution.” In I. Guyon; U. V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30 (Curran Associates, Inc., 2017).

Karnaushkin, Pavel and Maksim Sklyarenko. 2022. “Computer vision-based method of pre-alignment of a channel optical waveguide and a lensed fiber.” *Computer Optics*, 46(1):71–82. Doi:10.18287/2412-6179-CO-919.

Li, Zewen; Fan Liu; Wenjie Yang; Shouheng Peng; and Jun Zhou. 2021. “A survey of convolutional neural networks: Analysis, applications, and prospects.” *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. doi:10.1109/TNNLS.2021.3084827.

Limonova, Elena; Arseny Terekhin; Dmitry Nikolaev; and Vladimir Arlazarov. 2020. “Fast implementation of morphological filtering using ARM NEON extension.” *arXiv preprint arXiv:2002.09474*.

Lobanov, Mikhail and Dmitry Sholomov. 2019. “On the acceleration of the convolutional neural network architecture based on reset in the task of road scene objects recognition.” *ITiVS*, 69(3):57–65. DOI: 10.14357/20718632190305.

Mckay, M.; Richard Beckman; and William Conover. 1979. “A comparison of three methods for selecting vales of input variables in the analysis of output from a computer code.” *Technometrics*, 21:239–245. doi:10.1080/00401706.1979.10489755.

Mudiyanselage, Thosini Bamunu; Nipuna Senanayake; Chunyan Ji; Yi Pan; and Yanqing Zhang. 2021. “Covid-19 detection from chest x-ray and patient metadata using graph convolutional neural networks.”

Panfilova, Ekaterina and Irina Kunina. 2020. “Using window hough transform for detecting elongated boundaries in an image.” *Sensory systems*, 34(4):340–353. DOI: 10.31857/S0235009220030075.

Panfilova, Ekaterina; Oleg Shipitko; and Irina Kunina. 2021. “Fast hough transform-based road markings detection for autonomous vehicle.” In *ICMV 2020*, volume 11605 (Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, Washington 98227-0010 USA, 2021). DOI: 10.1117/12.2587615.

Tziolas, Nikolaos; Nikolaos Tsakiridis; Eyal Ben-Dor; John Theocharis; and George Zalidis. 2020. “Employing a

multi-input deep convolutional neural network to derive soil clay content from a synergy of multi-temporal optical and radar imagery data.” *Remote Sensing*, 12(9). ISSN 2072-4292. doi:10.3390/rs12091389.

Wei, Chih-Chiang. 2020. “Comparison of river basin water level forecasting methods: Sequential neural networks and multiple-input functional neural networks.” *Remote Sensing*, 12(24). ISSN 2072-4292. doi:10.3390/rs12244172.

Yang, Brandon; Gabriel Bender; Quoc V. Le; and Jiquan Ngiam. 2020. “Condconv: Conditionally parameterized convolutions for efficient inference.”

Yi, Kwang Moo; Eduard Trulls; Vincent Lepetit; and Pascal Fua. 2016. “Lift: Learned invariant feature transform.” In Bastian Leibe; Jiri Matas; Nicu Sebe; and Max Welling, editors, *Computer Vision – ECCV 2016*, 467–483 (Springer International Publishing, Cham, 2016). ISBN 978-3-319-46466-4.

Yu, Xunbo; Hanyu Li; Xinzhu Sang; Xiwen Su; Xin Gao; Boyang Liu; Duo Chen; Yuedi Wang; and Binbin Yan. 2021. “Aberration correction based on a pre-correction convolutional neural network for light-field displays.” *Opt. Express*, 29(7):11009–11020. doi:10.1364/OE.419570.

Zheng, Zhen; Bingting Zha; Youshi Xuchen; Hailu Yuan; Yanliang Gao; and He Zhang. 2019. “Adaptive edge detection algorithm based on grey entropy theory and textural features.” *IEEE Access*, PP:1–1. doi:10.1109/ACCESS.2019.2927655.

Zhou, Ding-Xuan. 2019. “Universality of deep convolutional neural networks.” *Applied and Computational Harmonic Analysis*, 48. doi:10.1016/j.acha.2019.06.004.

Zhukovskiy, Aleksandr; Elena Limonova; and Dmitry Nikolaev. 2018. “Exact implementation of common image processing algorithms using fully convolutional networks.” *Trudy ISA RAN*, 68(Special issue S1):108–116. DOI: 10.14357/20790279180512.

AUTHOR BIOGRAPHIES



o.chaganova@visillect.com.

OLGA CHAGANOVA was born in Orenburg, Russia. Currently she is a M.Sc. student in Applied Mathematics and Physics at Moscow Institute of Physics and Technology. Since 2021, she works as a junior research fellow at the Vision Systems Lab of the Institute for Information Transmission Problems. Her research interests include deep learning and computer vision. Her email address is



me@ansgri.com.

ANTON GRIGORYEV was born in Petropavlovsk-Kamchatskiy, Russia. Having graduated from Moscow Institute of Physics and Technology, he has been developing industrial computer vision systems with the Vision Systems Lab at the Institute for Information Transmission Problems since 2010. His research interests

Modeling and Simulation for Performance Evaluation of Computer-based Systems

CAUSAL ANALYSIS GRAPH MODELING FOR STRATEGIC DECISIONS

Alexander H. Levis
George Mason University
4400 University Dr.
Fairfax, VA 20105 USA
alevis@gmu.edu

Amy Sliva
Kings College
133 N. River St
Wilkes-Barre, PA. 18711 USA
amysliva@kings.edu

KEYWORDS:

Influence Nets, Bayesian Nets, Gambella, South China Sea

ABSTRACT

The use of causal analysis graphs for developing and evaluating strategies in complex problems is illustrated through two case studies: agricultural production in Gambella, Ethiopia and the crisis in the South China Sea. A Timed Influence net tool called Pythia is used to analyze and evaluate possible courses of action for each case.

INTRODUCTION

Causal modeling, planning, and forecasting activities support analysts trying to understand and answer questions relevant to national and global security. Reasoning over complex causal graphs and quantitative models could be dramatically improved by automated systems that help analysts configure scenarios of interest and focus on the most uncertain parts of the model. Ideally, Natural Language Processing can be used over a variety of data and narrative sources to construct the basic causal analysis graph. Once the graph has been defined, parameters that designate the influence of a cause on an effect are specified, usually by Subject Matter Experts, and time delays added to reflect the sojourn or processing time at a node and the propagation delay between nodes. The resulting graph, called a Timed Influence Net, can then be used to (a) assess the effect of a selected set of actions, or Course of Action (COA), on the outcomes, (b) to develop optimal COAs, and (c) analyze the sensitivity of the results to the individual actions and to the influence parameters. A tool called Pythia has been used to model and analyze a wide diversity of strategic problems. The workings of the tool are described in the next section. In the subsequent two sections two very different examples are presented: reducing famine in Ethiopia and avoiding conflict in the South China Sea.

PYTHIA: A TIMED INFLUENCE NET TOOL

Pythia provides an environment to build graph-based probabilistic cause-and-effect models and to perform several analyses on them. It was developed by the System Architectures Laboratory at George Mason University to aid decision making and

Course of Action development and evaluation in complex situations. (Haider and Levis, 2008; Levis, 2014, Wagenhals and Levis, 2007) The process embodied in *Pythia* consists of four steps. The first step is the determination of the desired effects: the effects that are of interest whether they are desirable outcomes to be achieved or undesirable outcomes to be avoided. To determine how these effects can be accomplished and what could inhibit their accomplishment, an influence net model is constructed in which complex probabilistic influences between causes and effects and between effects and actions are indicated. The process for constructing the Influence Net starts with the effects on the right and works backwards toward the left. The process continues until the nodes that would influence the outcomes (or effects) are events that are controllable or scenario dependent. These large actions can then be decomposed further to the left until they become specific tasks or, in the terminology of Influence Nets, actionable events. If time is introduced, it is possible to indicate the time phasing of the actions and observe the probability of achieving the desired effects change over time. The various influences (links) in the Influence Net have processing delays associated with them; an event can take place at time t but its influence may not be felt until $t + \delta t$. Also, the actionable events (the root nodes of the Influence Net) may not all take place at the same time, but at different times. *Pythia* provides for entering delays in the influence links and delays in the actionable events. This creates a Timed Influence Net that produces, when executed, not just the final probabilities but probability profiles over time. This enables the creation and evaluation of Courses of Action in which the various actions can be distributed on the timeline so that the best probability profiles can be achieved. The influence net model is then used to carry out sensitivity analyses to determine which actionable events, alone and in combination, appear to produce the desired effects.

Graphical Interface Features

The main graphical interface features are: Actionable events (root nodes) drawn as rectangles (Fig.1, top left) while non-actionable events are drawn as rounded rectangles (top right) thus making them visually distinct. Different link styles are used in *Pythia* to distinguish between positive (blue with

pointed arrow head; bottom left) and negative (red with round arrow head; bottom right) influences.

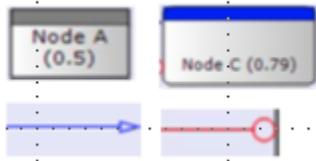


Fig. 1: Pythia symbols.

The underlying analytical framework for Influence Nets is Bayesian Nets. Consequently, Conditional Probability Tables (CPT) need to be constructed from the influence values. The Causal Strength (CAST) logic algorithm is used. (Haider, S., & Levis, A. H. 2008). The set of values for the influences is shown in Fig. 2.

If the premise (Parent) is TRUE this will influence the cosequence (Child)	If the premise (Parent) is FALSE this will influence the cosequence (Child)
99 (Significantly More Likely)	99 (Significantly More Likely)
90 (More Likely)	90 (More Likely)
66 (Moderately More Likely)	66 (Moderately More Likely)
33 (Slightly More Likely)	33 (Slightly More Likely)
00 (No Impact)	00 (No Impact)
-33 (Slightly Less Likely)	-33 (Slightly Less Likely)
-66 (Moderately Less Likely)	-66 (Moderately Less Likely)
-90 (Less Likely)	-90 (Less Likely)
-99 (Significantly Less Likely)	-99 (Significantly Less Likely)

Figure 2: Link Properties Window.

A color-coding scheme for the nodes is used that assists a user in estimating the likelihood of occurrence of a particular event in an Influence Net. The coloring scheme for the nodes ranges from Darker Blue ($p > 0.88$) or Significantly More Likely to Darker Red. ($p < 0.11$) or Significantly Less Likely.

Once a Timed Influence Net is completely specified by a user, *Pythia* computes the marginal probabilities of all the events. When a COA is specified, *Pythia* generates probability profiles of selected events. A profile shows the likelihood of occurrences of events over a period of time. The period is determined from the COA and temporal information available in the form of link and node delays.

GAMBELLA, ETHIOPIA

To demonstrate the technical approach, a use case is presented based on the food security and migration

situation in the Gambella region of Ethiopia (Fig. 3). This use case has many features that illustrate causal analysis graph modeling, including temporal dynamics with seasonal variations based on the agricultural cycle, and several possible decision points for interventions. The model is based on data, (Dallal, et al., 2021) but it is a simplified representation of the actual conditions in Gambella. It focuses on the temporal effects of weather/rainfall and the impact of various interventions, including:

- (1) Increase in food imports (current purchases)
- (2) Long-term infrastructure investments (roads, irrigation, storage, processing facilities, etc.)
- (3) Investment in cropland development
- (4) Investment in better seeds (current purchases)
- (5) Investment in chemical fertilizers (current purchases)

The following source nodes represent the scenario: (1) Weather; (2) Increases in Agricultural Labor; (3) Displaced Persons in camps; (4) Direct Food Aid

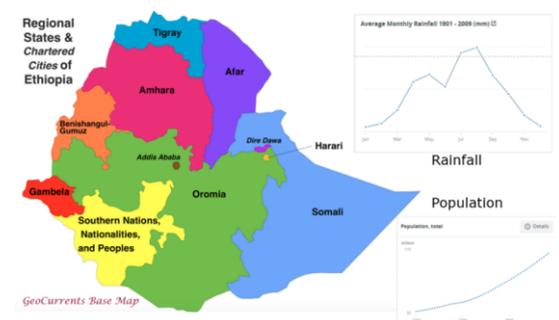


Fig. 3: On Ethiopia and Gambella. Rapid population growth: 51M (1993) to 108M (2018)

The model captures external events in a timeline (monthly) modeled on historical and scientific observations. Exogenous decisions are included in the timeline based on reasonable expectations, given a humanitarian goal. A textbook Solow-Swan model was used to construct parts of the scenario.

The model is shown in Fig. 4, which contains the model parameters (i.e., the influence parameters) described in the CAST format indicating the conditional influence of the parent nodes on their children. The crop season scenario over 12 months is shown in Fig. 5. External events and decisions (e.g., policy interventions) are applied to the model as observations (evidence), and are represented as tables indicating the state of each event in each month. The model calculates posterior marginals for crop yield increase, crop production increase, adequate urban food, and adequate rural food using a Bayesian Net algorithm.

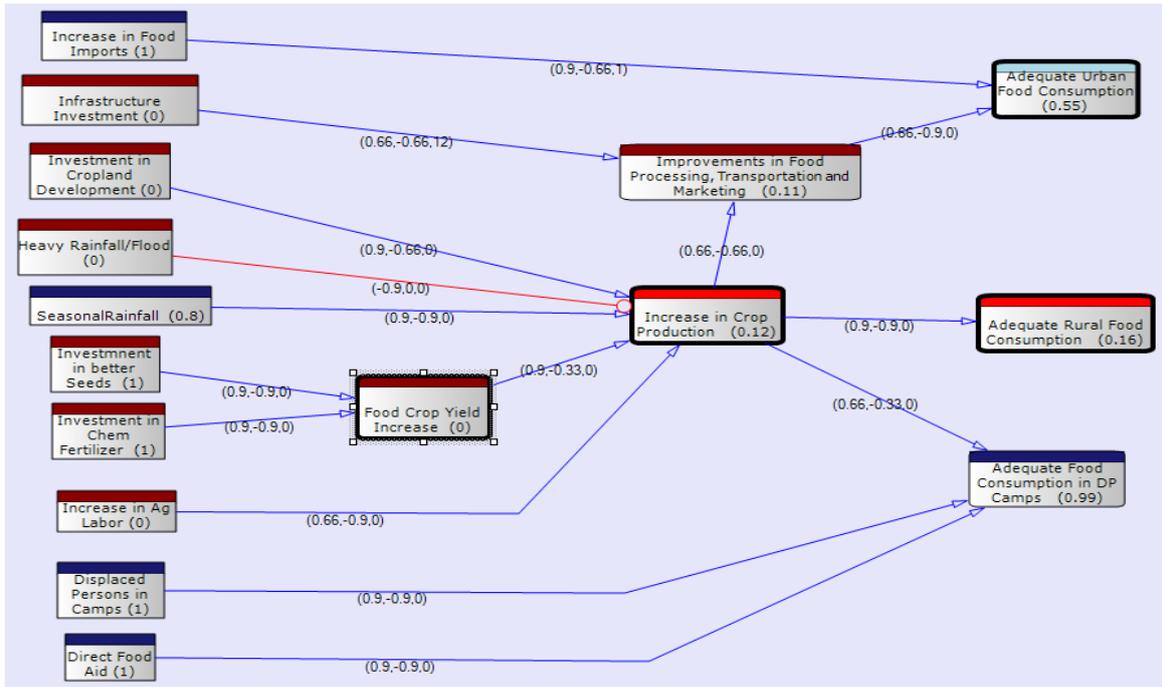


Fig. 4: Causal Analysis Graph (Influence net) for Gambella, Ethiopia test case

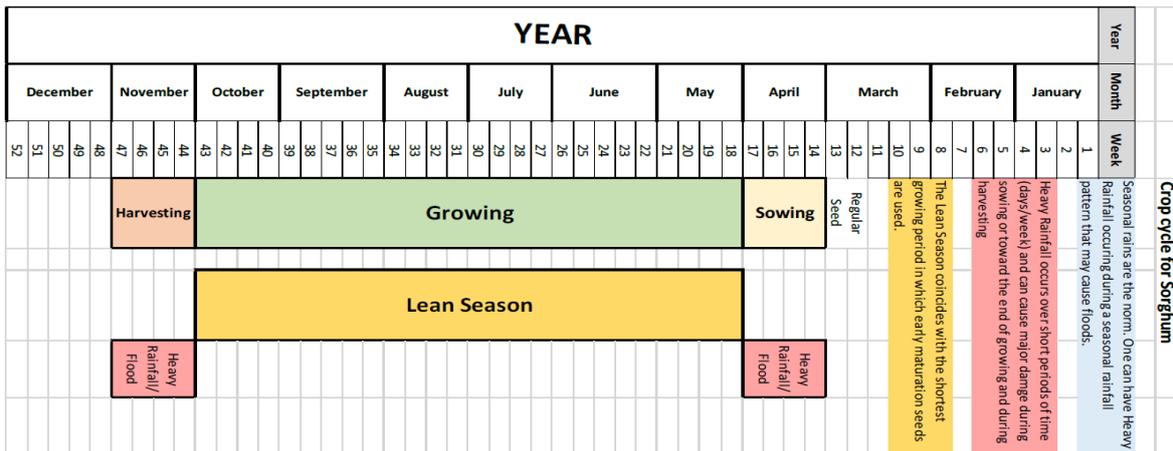


Fig. 5: Crop season scenario for Gambella test case

In this scenario, there is heavy rainfall during both the harvesting and sowing period. Using Pythia analytic algorithms, the objective is to predict the distributions over model variables over time and assess the stability of possible interventions under different conditions. Figure 6 shows the trajectories of certain model variables of interest over time. For example, in month 4, an increase in the probability that food yield will rise is observed (this is during the growing season), whereas this value drops in month 11 when flooding impacts potential crop yields.

To illustrate the behavior of the demonstration model, the scenario shown in Fig. 7 was developed. Here, the policy being explored is the impact of increasing food imports and investment in seeds, while the scenario includes heavier than average rainfall.

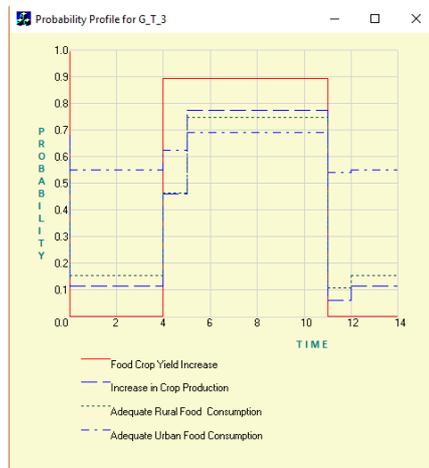


Fig. 6: Probability profiles of the bold nodes in the Gambella model. The probabilities shown on the model graph are the final probabilities

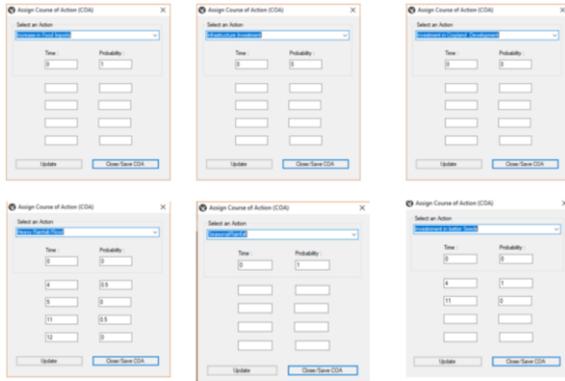


Fig. 7: Evaluation Scenario

Table 1 contains the results from running the Gambella model on the above scenario over a 12-month prediction period. What is shown is not only the seasonal variation, but the way in which the model is responding to the different weather conditions at different times of the year (i.e., heavy rainfall only matters if it disrupts the sowing or harvest).

Table 1: Runtime results for Gambella demonstration model.

Month	p(Crop Yield Increase)	p(Crop Production Increase)	p(Adequate Urban Food)	p(Adequate Rural Food)
0	0.01	0.12	0.55	0.16
1	0.01	0.12	0.55	0.16
2	0.01	0.12	0.55	0.16
3	0.01	0.12	0.55	0.16
4	0.9	0.4	0.61	0.41
5	0.9	0.78	0.69	0.75
6	0.9	0.78	0.69	0.75
7	0.9	0.78	0.69	0.75
8	0.9	0.78	0.69	0.75
9	0.9	0.78	0.69	0.75
10	0.9	0.78	0.69	0.75
11	0.01	0.12	0.55	0.16
12	0.01	0.12	0.55	0.16

Multiple scenarios were executed (a) covering a 12-month period and (b) a ten-year period. The ten-year period was used to analyze the impact of long-term investments in land reclamation, population changes and policy changes such as the one in which refugees were allowed to leave the camps and be employed. Key drivers in the Gambella case (and in Ethiopia in general) are the population increase, urbanization, the need to put more cropland into agricultural production, improvement in the productivity of the agricultural sector, and investment in food processing, transportation and markets to serve the urban population.

SOUTH CHINA SEA: THE SECOND THOMAS SHOAL CASE 2014

Background: The Second Thomas Shoal (2TS) is located south-east of Mischief Reef in the north-eastern part of the Spratly Islands. There are no settlements north or east of it. It is a tear-drop shaped atoll, 11 nautical miles (20 km; 13 mi) long North-South and fringed with coral reefs. The coral rim surrounds a lagoon which has depths of up to 27 meters (89 ft) and is accessible to small boats from the East.

A CNN report of April 22, 2020 stated that the Philippines has filed a diplomatic protest over China's creation of two new districts of Sansha City, the

southernmost city of Hainan province, which cover features in the disputed South China Sea, including the Philippine-claimed Spratly Islands, Scarborough Shoal and Fiery Cross Reef.

Because over the years there have been many diverse crises in the Spratly Islands area, a different approach was taken in developing the Timed Influence Net (TIN) model. A generic model was created first that contained all identified actions and reactions of the three principal actors. These were determined from reviewing published reports as well as newspaper articles from the Philippines and the US reporting on the various crises. These included documents that were available from the Center for the Study of Terrorism and Responses to Terrorism at the University of Maryland (START) (Wilkenfeld and Ellis, 2021) and from the DoD Strategic Multi-layer Assessments (SMA) Program. This generic Influence Net was very complex and was not operational. It was used as the basis from which specialized models were extracted to analyze specific Use Cases. The model is static (no time delays) and the strengths of the relationships are not inserted.

In this Influence Net model three Actors are considered: The Philippines (PH), the People's Republic of China (PRC), and the United States (US). Each actor has a set of available actions (See Table 2). What is interesting and challenging in is that many of these actions can be either initiating actions or responses to another actor's actions. The sequencing can be modeled using appropriate delays. For example, the Philippines may initiate the transport of building materials and the PRC then blocks access to the shoal by deploying Navy or Coast Guard assets. Conversely, the PRC may block access to the shoal by sea (through Navy or Coast Guard assets) thus forcing the Philippine Navy to resupply the Sierra Madre through air drops. This complexity is reflected in the many relationships that are represented by links in the model. However, only some of these relationships are active for any particular scenario.

Table 2: Possible Actions by the three Actors.

 Philippines (PH)	PH resupplies the Sierra Madre by Air drops
	PH resupplies the Sierra Madre by sea
	PH reinforces Sierra Madre military presence
	PH repairs the Sierra Madre
	PH brings in replacement ship
	PH brings in building materials
	PH sends message to PRC
	PH appeals to ASEAN PH requests US support
 China (PRC)	PRC Navy blocks access to Second Thomas Shoal (2TS)
	PRC Coast Guard blocks access to Second Thomas Shoal (2TS)
	PRC Coast Guard escort fisherman to 2TS area
	PRC endangers air drops to the Sierra Madre
 USA	US aids PH in Air drops
	US supports PH in ASEAN
	US conducts FON cruises in 2TS area

Technical Approach: Two objectives (outcomes) are considered for the two illustrative Use Cases that were analyzed: (a) The Philippines maintains military presence in the Second Thomas Shoal (2TS). (b) The PRC continues to take provocative actions.

The general technical approach consists of three steps. In this paper, only the first two steps are described. **Step 1:** Use Case development. Since there are many possible scenarios, a set of distinct Use Cases were developed. Each Use Case was expressed as a subset of the general model. This is accomplished primarily by setting the influences on the non-active links to 0 and by adding or subtracting some specialized nodes. **Step 2:** Static analysis. Consider the Use Cases developed in Step 1 and populate the model with the appropriate influence values on the active links for each scenario. Run the Static propagation algorithm to determine the marginal probabilities of the two final effects or outcomes. Conduct sensitivity analysis with respect to the initiating actions and with respect to influences. Run the SAF optimization algorithm (Haider and Levis, 2008)) to find the optimal Courses of Action. Document the results. **Step 3:** Dynamic analysis. Introduce time delays in the nodes to indicate when they become active to reflect each one of the two Use Cases. Add delays on the links to reflect times it takes for assets to execute their actions. Consider different time dependent courses of action and execute them to obtain the probability profiles of nodes of interest (i.e., $p(t)$ vs. t). Investigate the effect of time delays in the courses of action on the probability profiles. Document the results.

Use Case 1 Narrative: Philippines resupplies Sierra Madre by sea. The Philippine Navy is attempting to resupply its base at the 2TS (the Sierra Madre) by sea. A nearby unit of the PRC Navy moves to intercept and block the resupply ship. PH appeals to ASEAN and requests US support. The US supports the PH complaint at ASEAN and directs a unit of the US Navy to conduct a Freedom of Navigation (FON) exercise near 2TS where the PRC Navy units are moving. The question that is posed is twofold: What is the probability as a result of this set of events that PH will maintain its presence in 2TS and the PRC continues to take provocative actions?

Goal in Context: Explore the effect of PRC reaction to PH action

Scope: US actions in this situation

Pre-Condition: PRC Navy is monitoring the situation at 2TS

Success End Condition: PH maintains presence in 2TS and PRC decreases provocative actions

Minimal Guarantees: Direct confrontation between US and PRC Navies is avoided

Primary Actor: The Philippines (PH)

Trigger Event: The Sierra Madre needs supplies

Main Success Scenario

Step	Entity	Action Description
1	PH	Initiates Resupply of 2TS by sea
2	PRC	Navy vessels move to block access to 2TS
3	PH	Appeals to ASEAN
4	PH	Requests US support
5	US	Supports PH in ASEAN
6	US	Conducts FON in 2TS
7	PH	Provides supplies to 2TS

Scenario Variations

Step	Variable	Possible Variations
3	ASEAN	PH does not appeal to ASEAN
4	US Support	PH does not request US help
6	FON mission	US decides not to conduct FON mission

The model for the main success scenario of Use Case 1 is shown in Fig.8.

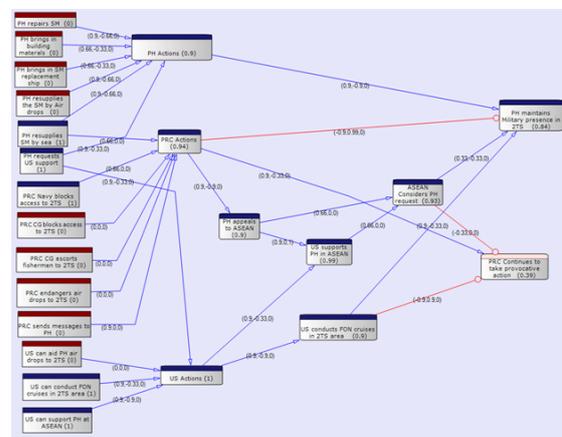


Fig. 8: Influence Net of Use Case 1 for the main success scenario. Results from Static analysis.

If the actions depicted as square nodes with dark blue stripes are taken, then the probability that the Philippines will maintain military presence in the Second Thomas Shoal (2TS) is 84% and the probability that the PRC will continue to take provocative actions drops to 39%.

Consider now the first variation: PH does not appeal to ASEAN and, consequently, the US does not support PH at ASEAN. In that case, the probability that the Philippines will maintain military presence in the Second Thomas Shoal (2TS) remains essentially the same (83%) but the probability that the PRC will continue to take provocative actions increases to 53%. In the second variation, the PH does not require support by the US and the US does not take any action. The only action taken by the Philippines is to appeal to ASEAN where the US supports the appeal. This changes drastically the results. The probability that the Philippines will maintain military presence in the Second Thomas Shoal drops to 16% but probability that the PRC will continue to take provocative actions increases to 89%.

To confirm the interpretation of these results one can conduct a sensitivity analysis of the outcomes with respect to the initiating actions. The results for the third variation are shown in Table 3. It is clear that none of the actions by PH and US with ASEAN have much impact on PRC's objectives. Conversely, sensitivity analysis of the PH objective to maintain military presence at 2TS is very much dependent on requesting US help. See Table 4.

Table 3: Sensitivity analysis of “PRC continues to take provocative actions” to active inputs.

Actions Name	LowerProbability	UpperProbability	Difference
PH resupplies the SM by Air drops	0.889	0.889	0
PH resupplies SM by sea	0.814	0.889	0.075
PH repairs SM	0.889	0.889	0
PH brings in SM replacement ship	0.889	0.889	0
PH brings in building materials	0.889	0.889	0
PH requests US support	0.89	0.889	-0.001
PRC Navy blocks access to 2TS	0.814	0.889	0.075
PRC sends messages to PH	0.889	0.924	0.035
PRC CG blocks access to 2TS	0.889	0.889	0
PRC CG escorts fishermen to 2TS	0.889	0.889	0
PRC endangers air drops to 2TS	0.889	0.889	0
US can aid PH air drops to 2TS	0.889	0.889	0
US can conduct FON cruises in 2TS area	0.889	0.889	0
US can support PH at ASEAN	0.89	0.889	-0.001

Table 4: Sensitivity analysis of “PH maintains military presence on 2TS” to active inputs.

Actions Name	LowerProbability	UpperProbability	Difference
PH resupplies the SM by Air drops	0.602	0.657	0.056
PH resupplies SM by sea	0.216	0.602	0.386
PH repairs SM	0.602	0.657	0.056
PH brings in SM replacement ship	0.602	0.646	0.044
PH brings in building materials	0.602	0.646	0.044
PH requests US support	0.165	0.602	0.437
PRC Navy blocks access to 2TS	0.643	0.602	-0.042
PRC sends messages to PH	0.602	0.582	-0.019
PRC CG blocks access to 2TS	0.602	0.602	0
PRC CG escorts fishermen to 2TS	0.602	0.602	0
PRC endangers air drops to 2TS	0.602	0.602	0
US can aid PH air drops to 2TS	0.602	0.602	0
US can conduct FON cruises in 2TS area	0.602	0.602	0
US can support PH at ASEAN	0.599	0.602	0.003

Dynamic analysis: Introduction of sequencing of actions and delays (of the order of days) produced insignificant variations in the probability profiles of the objective nodes over time.

Use Case 2 Narrative: The PRC takes provocative actions. The PRC takes the initiative by sending fishermen to fish in 2TS waters. The fishing boats are escorted by the PRC Coast Guard. The Philippines react by sending a formal message to the PRC and making an appeal to ASEAN. To avoid direct confrontation with the PRC Coast Guard, the Philippines attempts to resupply the Sierra Madre at 2TS by air drops. The PRC however endangers the air drops by having helicopters from the Coast Guard interfere with the flights of the PH helicopters. PH asks for help from the US and the US supports the ASEAN appeal, conducts a FON mission and overflies the air drop zone to deter the PRC from interfering with the resupply. All parties try to avoid a direct military confrontation. Finally, the PRC fishermen depart.

Characteristic Information

Goal In Context: Explore the effect of PH reaction to PRC action

Scope: US actions in this situation
 Pre-Condition: PRC Coast Guard is near 2TS
 Success End Condition: PH maintains presence in 2TS and PRC decreases provocative actions
 Minimal Guarantees: Direct confrontation between US and PRC Navies is avoided
 Primary Actor: PRC
 Trigger Event: PRC fishermen in 2TS waters

Main Success Scenario

Step	Entity	Action Description
1	PRC	Coast Guard cutters escort PRC fishermen
2	PH	Sends formal protest to PRC
3	PH	Resupplies the 2ts by Air Drops
4	PRC	Endangers air drops to 2TS
5	PH	Requests US support
6	PH	Appeals to ASEAN
7	US	Supports PH in ASEAN
8	US	Aids PH air drops to 2TS
9	US	Conducts FON missions in 2TS area
10	PRC	Fishermen complete mission and leave

Scenario Variations

Step	Variable	Possible Variations
3	PH	PH does not attempt to resupply 2TS through air drops
5	PH	PH does not request US support but US still supports PH in the ASEAN appeal but does not conduct a FON mission and does support the air drops.

The model for the main success scenario of Use Case 2 is shown in Fig. 9.

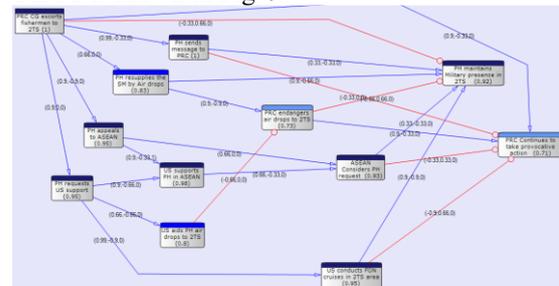


Fig. 9: Influence Net of Use Case 2 for the main success scenario. Results from Static analysis.

The influence net shows that the Philippines will send a message to PRC with probability 1, but the probability of initiating resupply by air drops is only 83%. The US supports the appeal to ASEAN with probability 98%, but the probability of aiding the air drops by sending aircraft in the area is only 80%, indicating that there is concern of an accident precipitating a crisis. The final outcome is that, as a result of all these actions, the Philippines will maintain its military presence in 2TS (at 92%) and the PRC will continue provocative actions (at 71%) since the fishermen completed their actions without major incident.

An interesting result is obtained from sensitivity analysis. There is a single initiating action: PRC CG escorts fishermen to 2TS. If this action is taken, the PRC is emboldened to continue provocative actions (probability 71%). But if this action is not taken, then the probability of continuing provocative actions drops to 42%. (Table 5)

Table 5: Sensitivity analysis of “PRC continues to take provocative actions” to “PRC CG escorting fishermen to 2TS”.

PRC Continues to take provocative action			
Actions Name	Lower Probability	Upper Probability	Difference
PRC CG escorts fishermen to 2TS	0.423	0.71	0.286

In the first variation the Philippines does not attempt to resupply 2TS by air drops but takes all the other actions and so does the US. The results are somewhat surprising. The probability that PH will maintain military presence in 2TS is 72% but the probability of the PRC continuing to take provocative action drops to 39%. One possible interpretation is that the fact that PH does not resupply 2TS may mean that the military presence appears to be well established and not threatened by the presence of the PRC Coast Guard and the fishermen. On the other hand, the immediate US Navy Freedom of Navigation mission has a strong deterrence effect.

The second variation is a kind of worst-case scenario. The Philippines only send a message to the PRC and appeal to ASEAN. The US supports the appeal but takes no other action since PH has not requested help. The result is clear. The PRC is emboldened by the fact that neither PH nor the US take any serious action (No FON mission, no resupply) and the probability of continuing to carry out provocative actions to make PH abandon the 2TS goes to 95%. The probability that PH will maintain its military presence drops to 7%.

Both models (each a subset of a more general causal model) shows that if PH and the US do not take strong highly visible actions such as resupplying the military presence in the 2TS and the US Navy conducting FON missions in the area, the ability of PH to maintain a military presence in 2TS becomes very problematic. An emboldened PRC can increase the pressure if it perceives that there is no serious reaction.

CONCLUSION

Causal Analysis Graphs and their implementation as a Timed Influence Net provide a useful and rapid approach to examining complex strategic situations and determining the consequences of alternative courses of action.

ACKNOWLEDGEMENT

Part of this work was performed under DARPA contract number W911NF-18-C-0015 to Charles River Analytics, Inc. and part under the Minerva Research Initiative, Award No. N00014-18-1-2369.

REFERENCES

- Dalal, M., Kane, S., Blumstein, D., Pfeffer, A., Tittle, J., Levis, A. H., Ihler, A., and Nodianos, N. 2021. Sensitivity Analysis of Uncertainty in Causal Environments”, Report # R1708764, Charles River Analytics, Inc., Cambridge, MA, USA.
- Haider, S., and A. H. Levis. 2008. "Modeling time-varying uncertain situations using dynamic influence nets." *International Journal of Approximate Reasoning* 49.2: 488- 502.
- Levis, A. H. (Aug 2014). “Pythia 1.8 User Manual, v. 1.03”, System Architectures Laboratory, George Mason University, Fairfax, VA.
- Wagenhals, L. W. and Levis, A. H. 2007. “Course of Action Analysis in a Cultural Landscape Using Influence Nets.” *Proceedings of the IEEE Symposium On Computational Intelligence for Security and Defense Applications*, Honolulu, HI.
- Wilkenfeld, J. and Ellis, D. 2021. “Escalation Management in the Gray Zone”, Final Report, START/ICONS, Univ. of Maryland, College Park, MD. USA



Dr. Alexander H. Levis is University Professor Emeritus of Electrical and Computer Engineering at George Mason University, Fairfax, VA, USA. He was educated at Ripon College where he received the AB degree (1963) in Mathematics and Physics and then at MIT where he received the BS (1963), MS (1965), ME (1967), and Sc.D. (1968) degrees with control systems as his area of specialization. For the last fifteen years, his areas of research have been multi-formalism modeling to address strategic issues and resilient architecture design. Dr. Levis is a Life Fellow of IEEE, a Fellow of AAAS, and INCOSE, and an Associate Fellow of AIAA. He has over 300 publications documenting his research.



Dr. Amy Sliva is an Assistant Professor of Computer Science at King’s College, Wilkes-Barre, PA, USA. Dr. Sliva received a BS from Georgetown University (2005) and an MS (2007) and PhD (2011) from the University of Maryland in Computer Science, and an MPP (2010) from University of Maryland in International Security. Dr. Sliva was previously a Senior Scientist at Charles River Analytics and an Assistant Professor of Computer Science and Political Science at Northeastern University. She has over 15 years of experience developing large-scale data analytics and models of human behavior for decision making.

PREDICTING PERFORMANCE OF HETEROGENEOUS AI SYSTEMS WITH DISCRETE-EVENT SIMULATIONS

Vyacheslav Zhdanovskiy
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia;
Moscow Institute of Physics and Technology
(National Research University)
Institutskiy per. 9, Dolgoprudny, 141701, Russia
E-mail: zhdanovskiy.vd@phystech.edu

Lev Teplyakov
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: teplyakov@visillect.com

Anton Grigoryev
Institute for Information Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: me@ansgri.com

KEYWORDS

Heterogeneous computing; Discrete-event simulations; Artificial intelligence; Video analytics; Software architecture

ABSTRACT

In recent years, artificial intelligence (AI) technologies have found industrial applications in various fields. AI systems typically possess complex software and heterogeneous CPU/GPU hardware architecture, making it difficult to answer basic questions considering performance evaluation and software optimization. Where is the bottleneck impeding the system? How does the performance scale with the workload? How the speed-up of a specific module would contribute to the whole system? Finding the answers to these questions through experiments on the real system could require a lot of computational, human, financial, and time resources. A solution to cut these costs is to use a fast and accurate simulation model preparatory to implementing anything in the real system. In this paper, we propose a discrete-event simulation model of a high-load heterogeneous AI system in the context of video analytics. Using the proposed model, we estimate: 1) the performance scalability with the increasing number of cameras; 2) the performance impact of integrating a new module; 3) the performance gain from optimizing a single module. We show that the performance estimation accuracy of the proposed model is higher than 90%. We also demonstrate, that the considered system possesses a counter-intuitive relationship between workload and performance, which nevertheless is correctly inferred by the proposed simulation model.

INTRODUCTION

In recent years, there has been a significant growth of interest in machine learning and artificial intelli-

gence (AI) technologies. Analysts expect the AI market to grow to more than USD 660 billion by 2028 (GrandViewResearch.com 2021). Nowadays, AI technologies are used in various fields, such as urban services (Wang and Sng 2015), retail (Weber and Schütte 2019), medicine (Chen et al. 2021), etc.

One of the key technologies that allowed for the progress is deep learning. In particular, deep neural networks made it possible to achieve almost human-like object recognition quality in problems like image classification (Rawat and Wang 2017), object detection (Arnold et al. 2019; Liu et al. 2020) and image segmentation (Guo et al. 2018). In order to achieve this accuracy and still provide reasonable recognition speed, deep neural networks have to exploit special hardware providing fast matrix multiplication, most commonly — GPUs. This poses a problem for software developers, because they have to design the architecture of their AI-based applications with heterogeneous CPU/GPU computations in mind.

In detail, software developers have to utilize the parallelism provided by modern multi-core CPUs along with the GPU acceleration. Meanwhile the GPU is used for the neural network inference, the CPU is used to prepare the neural networks's input and postprocess its output, run various computer vision algorithms like tracking, localization, keypoint detection. The CPU also handles all the additional modules delivering AI results to the end user — API calls, business logic, database management, etc. In order to deliver on all these tasks on advanced heterogeneous hardware with the given time, memory and energy constraints, software developers have to design rather complex architectures (Sutter et al. 2005).

The complex software architecture, in turn, complicates performance evaluation and software optimization of such systems (Voss et al. 2019c). In particular,

it is hard to locate the bottleneck module and to predict, how its optimization will affect the performance of the entire system. It may lead to lots of human and financial resources as well as time resources being spent on optimization without any significant increase in performance of the entire system. For a new module to be designed and implemented, it might be hard to map the system’s constraints to the constraints of a specific module. It may turn out - after the resources are spent on implementing a new module and its integration into the system! - that the module results in unaffordable slowdown of the system.

An elegant way to avoid the aforementioned problems is to design a simulation model of the system and infer the feasibility of optimizations on the model prior to implementing them in the code. Discrete-event simulations of software have been used before in other areas, for example, planning, evaluation and optimization of Hadoop clusters (Bian et al. 2014; Wang et al. 2014; Liu et al. 2016; Chen et al. 2016). However, to the best of our knowledge, no one yet has tried to apply this approach to applications with heterogeneous CPU/GPU computing, such as AI systems. Moreover, such a simulation model can be used for other tasks like capacity planning. This can be achieved by evaluating the model on a collected set of suitable hardware configurations (Korobov et al. 2020).

In terms of performance modelling, different AI applications have their own unique specialties. The key difference is which hardware components to consider. For example, video analytics (Wang and Sng 2015) and self-driving cars (Badue et al. 2021) are mostly CPU/GPU intensive, meanwhile AI on pure-cloud solutions also rely heavily on the network performance, which requires to consider the I/O subsystem in the model.

In this work, we consider a video analytics system as a typical example of AI application utilizing CPU/GPU computing. The possibility of applying our research to other AI applications is discussed in Section III. While the GPU is used to infer the neural network responsible for complex object detection tasks such as human detection, the CPU is used for preprocessing the video frames, postprocessing the neural network output and running classic computer vision algorithms such as ORB keypoint detector and descriptor (Rublee et al. 2011). The considered system has the following functionality:

- processing video feed from multiple video cameras;
- person detection using YOLOv4 (Bochkovskiy et al. 2020);
- person 3D localization;
- single- and multi-camera person tracking.

We design a discrete-event simulation model which is low-cost both in terms of its development and simulation speed and can easily be adopted by the software developers. We use it to estimate:

1. the performance scalability with the increasing number of video cameras;
2. the performance gain from optimizing a single sys-

tem module;

3. the performance impact caused by integrating of a new module in the system.

The rest of the paper is structured as follows: Section I describes the software architecture of the video analytics system, Section II provides a description of the proposed simulation model, Section III contains numerical results. Section IV concludes the paper.

SYSTEM DESCRIPTION

Flow graph paradigm

A common design pattern to efficiently implement parallelism in heterogeneous and parallel systems is to use the flow graph paradigm (Grigoryev et al. 2015; Badue et al. 2021; Huang et al. 2021). With it, the algorithm is represented as a data flow graph (Voss et al. 2019b), see Fig. 1. Each graph node (vertex) receives a

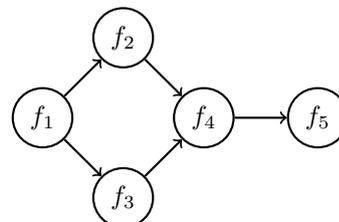


Fig. 1: Example of a flow graph. Nodes f_2 and f_3 can process messages broadcasted from f_1 in parallel. f_4 can process messages from f_2 and f_3 independently or wait until messages from both nodes are present — it depends on the desired behaviour.

message from its predecessors, processes it and broadcasts the output to the successors. Input and output nodes are the exception:

- input nodes either produce an input message by itself or receive it somewhere from outside the graph;
- output nodes do not broadcast the result, but instead store it or deliver it outside the graph.

The considered video analytics system has multiple places where parallelism can be utilized. In particular:

- frames from different video streams can be processed in parallel;
- a single frame can be processed in parallel by multiple data-independent detectors, for example, by the neural network and the keypoint detector.

Flow graphs allow to efficiently utilize these types of parallelism by executing different nodes of the graph in parallel at the same time.

There are multiple frameworks implementing this paradigm. Notable examples include oneTBB¹ and Taskflow² (Huang et al. 2021). In this work, we use oneTBB.

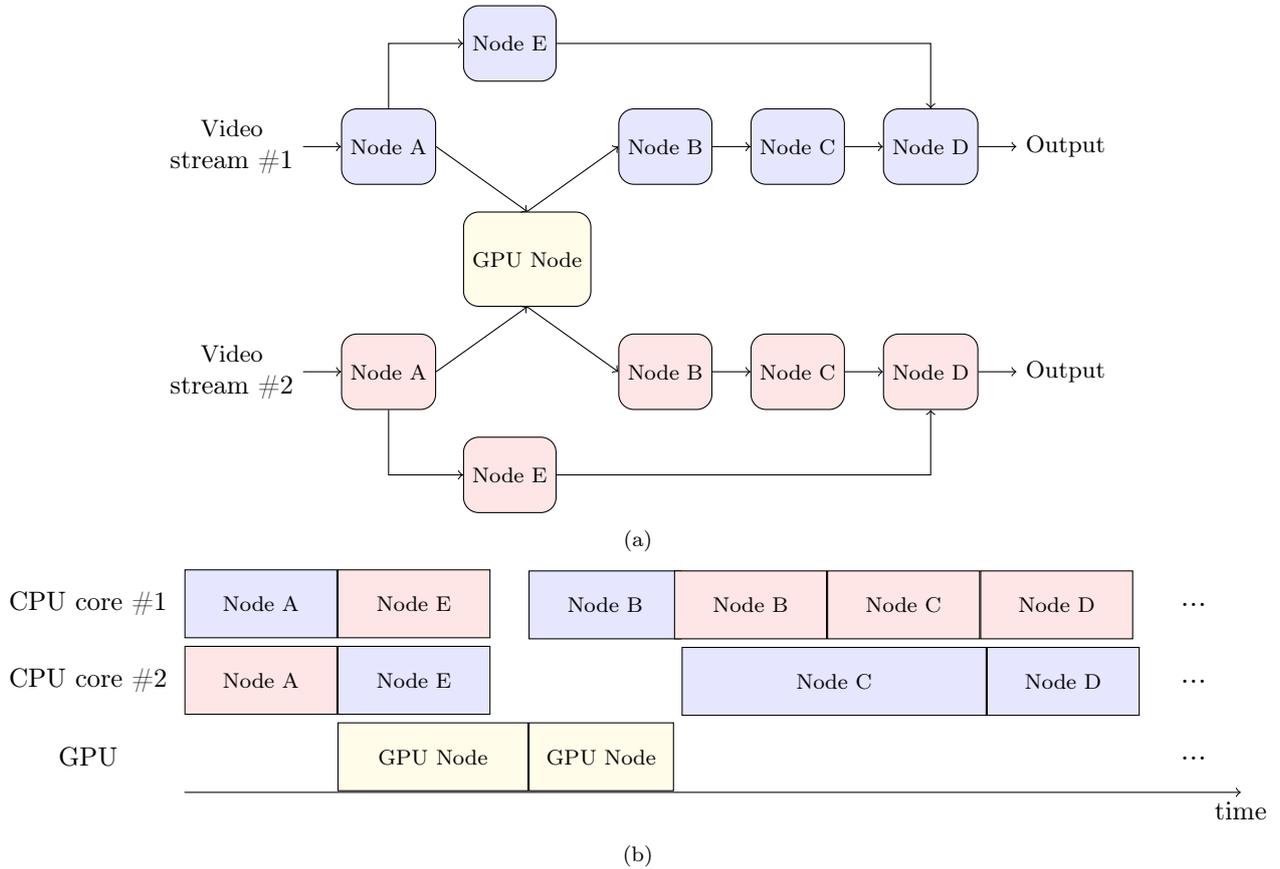


Fig. 2: Software architecture of the studied video analytics system represented as a flow graph (a). Sample timing diagram of the system execution (b). Note that **Node E** is processed in parallel with the **GPU Node**, meanwhile **Node B** has to wait until **GPU Node** finishes processing its task. **Node E** is the CPU-intensive module we consider in Section III. It is turned off for all experiments except “Integration of a new module in the system”.

Software architecture of the video analytics system

In this section, we provide a brief description of the software architecture used in the considered video analytics system.

Overall, the whole system is represented by a flow graph, see Fig. 2a. Each video stream from a single camera is represented by a separate graph component. We shall notice that we deliberately show a less detailed version of the real graph for the sake of simplification. All graph components are executed in a single thread pool.

Meanwhile many flow graph nodes can be executed in parallel, some of them can not because their functions are not thread safe. In particular, this is true with the neural network nodes. Because neural networks are usually implemented in a way (NVIDIA 2022) that each neural network instance can be executed on GPU by only one context at a particular moment, multiple video streams have to put their tasks for the neural network inference in a shared queue and then wait for their completion, see Fig. 2b. We implement this function-

ality with oneTBB’s *async nodes* (Voss et al. 2019a).

SIMULATION MODEL

In this section, we describe the details of the proposed simulation model. Overall, the system flow graph has a near-direct representation in the model. We implemented the model in Python using the SimPy³ library.

Algorithm 1 Flow graph node simulation for a basic node

```

Q – input queue (from predecessors)
S – output queue (to successors)
C – pool of free CPU cores
P – distribution of the node’s running time
while not all frames are processed do
    m ← Q.pop() (blocks if Q is empty)
    c ← C.pop() (blocks if C is empty)
    t ∼ P
    wait(t)
    S.push(m)
    C.push(c)
end while

```

¹Formerly TBB; more details at: <https://www.intel.com/content/www/us/en/developer/tools/oneapi/onetbb.html>

²More details at: <https://taskflow.github.io/>

³More details at: <https://simpy.readthedocs.io/en/latest/index.html>

Algorithm 2 Flow graph node simulation for a GPU async node

```
Q — input queue (from predecessors)
S — output queue (to successors)
C — pool of free CPU cores
G — GPU
P — distribution of the node’s running time
while not all frames are processed do
  m ← Q.pop() (blocks if Q is empty)
  c ← C.pop() (blocks if C is empty)
  G.lock() (blocks if G is busy)
  C.push(c)
  t ∼ P
  wait(t)
  G.release()
  c ← C.pop() (blocks if C is empty)
  S.push(m)
  C.push(c)
end while
```

The source code of the implementation is available at GitHub⁴. It contains only ≈ 400 LoC of Python (including the profiling trace parsers), meanwhile the real system contains ≈ 15000 LoC of C++ (not including the dependencies).

We consider a CPU with N logical cores in the model. Each flow graph node waits until the input message is present, see Algorithm 1. Then it waits until there is a free CPU core, then the node occupies the CPU core for the execution time. The execution time is sampled from independent empirical distributions measured by profiling the real system. Input messages that have not yet been processed are stored in a queue. Effectively, this represents the work of oneTBB’s thread pool.

The GPU neural network node is modelled in a slightly different way, see Algorithm 2. Like a basic flow graph node, it waits for an input message and a CPU core. Then it waits until the GPU is free, then locks it for the execution time, while yielding the CPU core for some another flow graph node. When the GPU is done with processing the message, the node waits again for a free CPU core in order to broadcast its output to the successors. Effectively, this represents the mechanism shown in Fig. 2b.

The flow graph nodes’ execution times are sampled from empirical distributions measured by profiling the real system. Fig. 3 contains an example of such distribution for the neural network inferencer. The oneTBB flow graph nodes are profiled using Intel Flow Graph Tracer and Flow Graph Analyzer (Voss et al. 2019c). The non-oneTBB activities like the video decoding and the neural network inference are sampled by our in-house tracing library.

We deliberately use flow graphs in the model instead of some well-known formalisms like Petri Nets (Peterson 1977) and Queuing Networks (Chandy et al. 1975). It makes designing the simulation model easier because

⁴More details at: <https://github.com/iitpvisionlab/heterogeneous-ai-system-simulator>

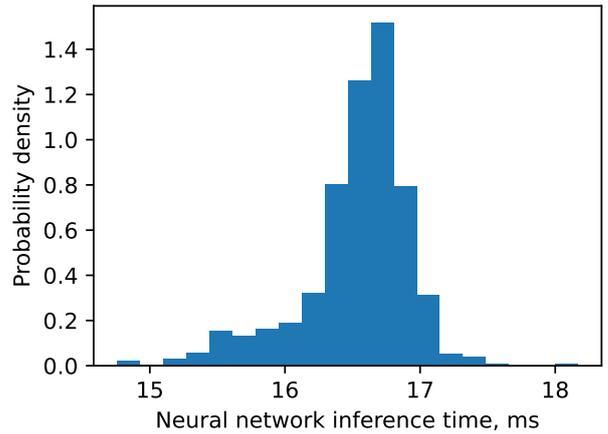


Fig. 3: Empirical distribution of neural network inference time.

we can explicitly use the same graph structure and synchronization primitives used as the real system. This approach also has potential for the model to be automatically generated by parsing the profiling traces. Moreover, our approach is easier to be adopted by the software developers who are unlikely to be experts in modelling and simulation.

We shall notice that we do not consider the overhead caused by communications between the nodes, as the execution time of each node significantly exceeds the average communication time in our case. However, our simulation model can easily be upgraded by introducing additional timings between consecutive graph nodes. This will allow to model other AI applications like AI on pure-cloud solutions, where there is significant communication overhead cause by the network. The communication overhead can also be important in self-driving cars, where the TCP/IP stack is often used for communication even within a single machine, e.g. the ROS framework(Quigley et al. 2009).

EVALUATION

In this section we evaluate the proposed simulation model. In order to do it, we compared the performance metrics predicted by the simulator with those measured on the real video analytics system. In our case, the performance metrics is the average frames per second (FPS) per each video stream.

The hardware specifications are listed in Table 1.

TABLE 1: Testbed hardware specifications

Parameter	Value
CPU	Intel Core i7-7800X, 3.5 GHz, 6 Cores, 12 Threads
GPU	NVIDIA GeForce GTX 1080 Ti, 11 GB VRAM
RAM	64 GB

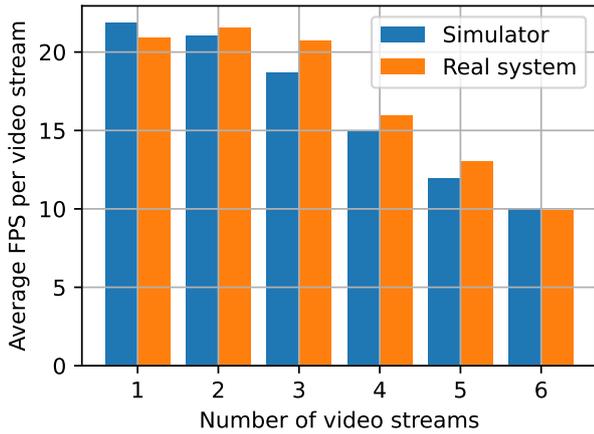


Fig. 4: FPS per video stream depending on the number of video streams.

Estimating the model accuracy and the system scalability

In this experiment, we varied the number of video streams in order to estimate the accuracy of the simulation model. The results presented in Fig. 4 demonstrate, that the performance prediction error rate of the simulation model is less than 10%, which is accurate enough to rely on the model’s prediction in planning the scalability of the system. It is also lower than the error rate threshold of 20% used in a related work (Bian et al. 2014).

On small number of video streams (up to three) the system experiences almost no performance drop due to efficient parallelism. However, when the number of video streams increases, the performance drops significantly, because the neural network becomes the bottleneck that hinders the parallelism: threads stand idle waiting for a task to execute. When the number of video streams exceeds the number of CPU logical cores, the performance drops even more because there is not enough free threads to execute appearing tasks.

Integration of a new module in the system

To study an impact of a new module on the system, we conducted the following experiment. We added a CPU-intensive module (Node E in Fig 2a) in the system and measured the overall slowdown on both the real model and the simulator. The results, see Fig 5, show that the module produces approximately 2.5X slowdown on 1 video stream, meanwhile producing no noticeable slowdown on 12 video streams. The result is counter-intuitive: the increase in workload makes the overhead of the module unnoticeable instead of it growing linearly with the workload. This occurs because the neural network, while being comparably fast on one stream, becomes the bottleneck on high number of video streams.

This experiments demonstrates, that for systems with high degree of parallelism the impact of adding a new module or changing the existing one on system’s performance could be nontrivial.

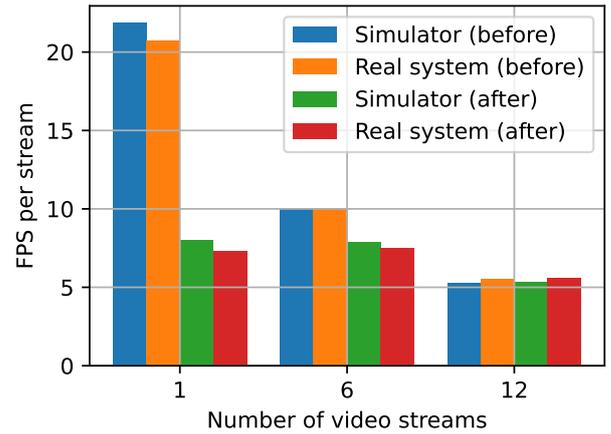


Fig. 5: Performance impact caused by integrating a new CPU-intensive module

Estimating the feasibility of software optimizations

In this section we study the feasibility of optimizations in the considered system with the help of the proposed simulation model by the example of the following optimization.

The quality of service (QoS) of the considered video analytics system highly depends on tracking algorithms. Tracking accuracy, in turn, depends on many parameters of the algorithm. In order to improve QoS, it is needed to find the optimal values of the parameters.

The simplest way to find the optimal values is the grid search. However, for the high-dimensional search space of parameters it is computationally unfeasible. More sophisticated optimization methods like Bayesian optimization allow to mitigate but not fully alleviate this problem, still requiring many runs of the system.

Fortunately, all the required runs use the same input data with different values of parameters. Therefore, in theory, it is possible to save a lot of time by caching neural network predictions and reusing them on each run instead of inferring the real network.

However, the cache may not provide the desired performance gain, because some other module can become the bottleneck. Therefore, prior to implementing and testing the optimization (which takes about one week for a software engineer), we study the feasibility of the optimization on the simulation model (which takes a couple of hours).

First, we used the developed simulation model to estimate the performance gain of implementing the neural network cache without considering the overhead of the cache itself. Effectively, we set the execution time of the GPU node to zero. The speed-up appeared to be 13.8x (Table 2, “ideal” cache).

Then we implemented a LevelDB-based⁵ cache module detached from the system, benchmarked its performance and used the data in simulation model to ac-

⁵More details at: <https://github.com/google/leveldb>

TABLE 2: Overall system speedup on 6 video streams from implementing the cache

Experiment	Overall system speedup
Real system	11.3x
Simulator, “ideal” cache	13.8x
Simulator, “real” cache	12.0x

count for the overhead. The speed-up appeared to be 12.0x (Table 2, “real” cache).

Finally, we integrated the developed cache module into the system. The real achieved speed-up was equal to 11.3x (Table 2, real system), close to the predicted value of 12.0x.

The experiment demonstrates, that with a negligible overhead in time spent on simulating each step of the implementation, it allows to correctly estimate the limit of optimization’s impact. Moreover, it could save a lot of time, if it had emerged that a specific optimization step is not worth implementing.

CONCLUSIONS

In this work, we proposed a discrete-simulation model to predict the performance of a heterogeneous CPU/GPU video analytics system. The proposed model can easily be adopted by the software developers who are not experts in simulation and modelling. We showed that the accuracy of performance estimation using the proposed system is higher than 90% in each experiment.

We used the simulation model to predict performance scale with workload; to estimate the impact of a new module on the whole system, which demonstrated counter-intuitive results yet correctly predicted by the simulator; to predict the feasibility of an optimization.

We believe such a simulation model should become a workplace tool for software designers and it could save lots of resources by easily inferring the feasibility of optimizations and modifications prior to doing some costly work on a real system.

Possible future research includes taking into consideration the hardware configuration to predict, for example, an optimal hardware price — quality of service trade-off of a system.

REFERENCES

Arnold, Eduardo; Omar Y Al-Jarrah; Mehrdad Dianati; Saber Fallah; David Oxtoby; and Alex Mouzakitis. 2019. “A survey on 3D object detection methods for autonomous driving applications.” *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795.

Badue, Claudine; R nik Guidolini; Raphael Vivacqua Carneiro; Pedro Azevedo; Vinicius B Cardoso; Avelino Forechi; Luan Jesus; Rodrigo Berriel; Thiago M Paixao; Filipe Mutz; et al. 2021. “Self-driving cars: A survey.” *Expert Systems with Applications*, 165:113816.

Bian, Zhaojuan; Kebin Wang; Zhihong Wang; Gene Munce; Illia Cremer; Wei Zhou; Qian Chen; and Gen Xu. 2014. “Simulating Big Data clusters for system planning, evaluation, and optimization.” In *2014 43rd International Conference on Parallel Processing*, 391–400.

Bochkovskiy, Alexey; Chien-Yao Wang; and Hong-Yuan Mark Liao. 2020. “YOLOv4: Optimal speed and accuracy of object detection.” *arXiv preprint arXiv:2004.10934*.

Chandy, K. Mani; Ulrich Herzog; and Lin Woo. 1975. “Parametric analysis of queuing networks.” *IBM Journal of Research and Development*, 19(1):36–42.

Chen, Jianguo; Kenli Li; Zhaolei Zhang; Keqin Li; and Philip S Yu. 2021. “A survey on applications of artificial intelligence in fighting against COVID-19.” *ACM Computing Surveys (CSUR)*, 54(8):1–32.

Chen, Qian; Kebin Wang; Zhaojuan Bian; Illia Cremer; Gen Xu; and Yejun Guo. 2016. “Cluster performance simulation for Spark deployment planning, evaluation and optimization.” In *International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, 34–51.

GrandViewResearch.com. 2021. “Artificial intelligence market size, share trends analysis report by solution, by technology (deep learning, machine learning, natural language processing, machine vision), by end use, by region, and segment forecasts, 2021 - 2028.” <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market>. Accessed: 18.01.2022.

Grigoryev, Anton; Timur Khanipov; Ivan Koptelov; Dmitry Bocharov; Vassily Postnikov; and Dmitry Nikolaev. 2015. “Building a robust vehicle detection and classification module.” In *ICMV 2015*.

Guo, Yanming; Yu Liu; Theodoros Georgiou; and Michael S Lew. 2018. “A review of semantic segmentation using deep neural networks.” *International journal of multimedia information retrieval*, 7(2):87–93.

Huang, Tsung-Wei; Dian-Lun Lin; Chun-Xun Lin; and Yibo Lin. 2021. “Taskflow: A lightweight parallel and heterogeneous task graph computing system.” *IEEE Transactions on Parallel and Distributed Systems*, 33(6):1303–1320.

Korobov, Nikita; Oleg Shipitko; Ivan Konovalenko; Anton Grigoryev; and Marina Chukalina. 2020. “SWaP-C based comparison of onboard computers for unmanned vehicles.” In *ER(ZR)-2019*, volume 154, 573–583 (Springer, Singapore, 2020).

Liu, Jun; Bianny Bian; and Samantika Subramaniam Sury. 2016. “Planning your SQL-on-Hadoop deployment using a low-cost simulation-based approach.” In *2016 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 182–189.

Liu, Li; Wanli Ouyang; Xiaogang Wang; Paul Fieguth; Jie Chen; Xinwang Liu; and Matti Pietik inen. 2020. “Deep learning for generic object detection: A survey.” *International journal of computer vision*, 128(2):261–318.

NVIDIA. 2022. “NVIDIA TensorRT Documentation.” <https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html#threading>. Accessed: 28.01.2022.

Peterson, James L. 1977. “Petri nets.” *ACM Computing Surveys (CSUR)*, 9(3):223–252.

Quigley, Morgan; Ken Conley; Brian Gerkey; Josh Faust; Tully Foote; Jeremy Leibs; Rob Wheeler; Andrew Y Ng; et al. 2009. “Ros: an open-source robot operating system.” In *ICRA workshop on open source software*, volume 3, page 5.

Rawat, Waseem and Zenghui Wang. 2017. “Deep convolutional neural networks for image classification: A comprehensive review.” *Neural computation*, 29(9):2352–2449.

Rublee, Ethan; Vincent Rabaud; Kurt Konolige; and Gary Bradski. 2011. “ORB: An efficient alternative to SIFT

- or SURF.” In *2011 International conference on computer vision*, 2564–2571.
- Sutter, Herb et al. 2005. “The free lunch is over: A fundamental turn toward concurrency in software.” *Dr. Dobbs’s journal*, 30(3):202–210.
- Voss, Michael; Rafael Asenjo; and James Reinders, *Beef Up Flow Graphs with Async Nodes*, 513–534 (Apress, Berkeley, CA, 2019a). ISBN 978-1-4842-4398-5. doi:10.1007/978-1-4842-4398-5_18.
- Voss, Michael; Rafael Asenjo; and James Reinders, *Flow Graphs*, 79–107 (Apress, Berkeley, CA, 2019b). ISBN 978-1-4842-4398-5. doi:10.1007/978-1-4842-4398-5_3.
- Voss, Michael; Rafael Asenjo; and James Reinders, *Flow Graphs: Beyond the Basics*, 451–511 (Apress, Berkeley, CA, 2019c). ISBN 978-1-4842-4398-5. doi:10.1007/978-1-4842-4398-5_17.
- Wang, Kebin; Zhaojuan Bian; Qian Chen; Ren Wang; and Gen Xu. 2014. “Simulating Hive cluster for deployment planning, evaluation and optimization.” In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, 475–482.
- Wang, Li and Dennis Sng. 2015. “Deep learning algorithms with applications to video analytics for a smart city: A survey.” *arXiv preprint arXiv:1512.03131*.
- Weber, Felix Dominik and Reinhard Schütte. 2019. “State-of-the-art and adoption of artificial intelligence in retailing.” *Digital Policy, Regulation and Governance*.

AUTHOR BIOGRAPHIES

VYACHESLAV ZHDANOVSKIY



was born in Verkhnyaya Pyshma, Russia. He obtained his B.Sc. in Applied Physics and Mathematics in 2020 from Moscow Institute of Physics and Technology (MIPT). Currently he is a M.Sc. student in Computer Science and Engineering at MIPT. Since 2020, he works at the Vision Systems Lab at the Institute for Information Transmission Problems. His research interests include heterogeneous and parallel computing, computer vision and distributed systems. His e-mail address is zhdanovskiy.vd@phystech.edu.

LEV TEPLYAKOV



was born in Arkhangelsk, Russia. He obtained his B.Sc. and M.Sc. in Applied Physics and Mathematics from Moscow Institute of Physics and Technology (MIPT) in 2017 and 2019 correspondingly. Since 2016, he has been developing industrial computer vision systems with the Vision Systems Lab at the Institute for Information Transmission Problems. His research interests include heterogeneous and parallel computing, object detection and tracking. His e-mail address is teplyakov@visillect.com.

ANTON GRIGORYEV



was born in Petropavlovsk-Kamchatskiy, Russia. Having graduated from Moscow Institute of Physics and Technology, he has been developing industrial computer vision systems with the Vision Systems Lab at the Institute for Information Transmission Problems since 2010. His research interests are image processing and enhancement methods, autonomous robotics and software architecture. His e-mail address is me@ansgri.com.

Epistemic Games with Conditional Believes for Modelling Security Threats Defence in Cloud Computing Systems

Lukasz Gaża
Cracow University of Technology
31-155 Cracow, Poland

Agnieszka Jakóbiak
Cracow University of Technology
31-155 Cracow, Poland

KEYWORDS

Epistemic Games, Cloud, Security.

ABSTRACT

We presented Epistemic Games with Conditional Believes model for automating security decisions in Cloud Computing systems. The model assumes attack-defence scenarios. The game stages model Cloud provider and Cloud attacker rivalry to maximise their payoffs. The paper presents the methodology for including the believes about opponent's rationality. The presented model allows considering the attack on Cloud system in a realistic way. The proposed solution has been tested by the experimental analysis on Cloud Sim simulator. Presented model enables finding strategies for the Cloud provider to protect assets from cyber-security attacks.

I. INTRODUCTION

The aim of the presented study is to examine the possibility of building the automatic decision system based on Epistemic Games for security decision making in Cloud Computing systems. This topic is important due to the their very rapid development and the fact that the complexity of such systems forces their users to automatise a lot of the decision making process. Game theory supports modelling strategic reasoning, considering rational game players who are benefiting from maximising their profits with interaction with other players. Automatising of security decisions in Cloud systems is the crucial process in securing such complex systems. The presented research is a continuation of our previous development of game theory based modelling of the competition between Cloud providers and Cloud attackers, [21],[15],[14]. The current research is a try to reformulate the assumptions about the Cloud attackers. It incorporates the believes about the utility functions for the Cloud attackers, instead. Epistemic games include into the decision-making process the decision-maker's beliefs about the state of the environment and opponents rationality. It models the situation when each game player is maximising the subjective expected utility assuming not the opponents utility functions, but the defenders belief that such functions will be used.

Our main contributions are:

- to adapt the epistemic approach for modelling the uncertainty of the attacker utility function;
- to introduce the model of belief as the the behaviour scheme of the computer system attacker;
- to propose the payoff functions for both Cloud defender and Cloud attacker.

The paper is organised as follows. Section (II) is describing the game modelling in the context of security decisions. Section (III) is presenting our model incorporating both payoff functions. In section IV we presented the results of the simulation based on Cloud Sim testing environment and the Python coded model. The paper ends with Section (V), which contains conclusions based on the conducted experiments and obtained results. Ideas for future work and potential improvements are also discussed there.

II. RELATED WORK

Game theoretic models were successfully used for modelling security related decisions. In [16], authors used Nash game for securing wireless network system. A zero-sum multi-stage two-player competitive game was used in [11] for securing a network of computers. In [23] Bayesian game was introduced and in [12], authors used stochastic game models. Multiple adversaries were considered in [6] and The bi-level game-theoretic model was used. In [9] authors used Bayesian attacker detection games with incomplete information for modelling the interaction between nodes in wireless networks with channel uncertainty and the concept of Nash equilibrium. Stackelberg games were applied for choosing the security levels of virtual machines [20]. A Cloud system defence was modelled in [15],[14] but we found the game model assumptions to be too strong to model the realistic attacks. For the best of our knowledge epistemic approach was not used so far for modelling the uncertainty of the attacker utility function. A novelty of the proposed approach is to use the concept of belief as the behaviour scheme of the computer system attacker. An additional novelty is the formulation of the payoff functions considering the probabilities of successful attacks in case when considered asset is protected or not. The main differences between the cited solutions and the proposed paper are the usage of Epistemic Game Theory instead of traditional game-based approaches. It enables to model the Cloud Computing related security decisions considering the fact that some information is

uncertain and therefore may be provided as the belief.

III. THEORETICAL MODEL

Let us denote by $i = 1, 2, \dots, N$ the set of players. For each player let C_i be a set of possible choices. For a particular player number i a choice combinations for his opponents is a list $(c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_N)$ where $c_1 \in C_1, \dots, c_{i-1} \in C_{i-1}, c_{i+1} \in C_{i+1}, \dots, c_N \in C_N$. A belief for a player i about his opponent's choices is a probability distribution b_i over the set $C_1 \otimes C_{i-1} \otimes C_{i+1} \otimes C_N$ that defines for every opponents choice $(c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_N)$ some probability $b_i(c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_N) \geq 0$ such that

$$\sum_{i=1}^N b_i(c_1, \dots, c_{i-1}, c_{i+1}, c_N) = 1 \quad (1)$$

A utility function for a player i assigns a number $u_i(c_1, \dots, c_N)$ to every combination of (c_1, \dots, c_N) and represents the outcome that the player i derives.

A dynamic game is defined by [18]:

- non-terminal history x that consists a set of choices that have been made by players in the past and resulted in x ;
- the beginning of the game is a non-terminal history denoted by \emptyset ;
- terminal history represents the situation when the game ends. Every terminal history z consists a set of choices that leads to z ;
- the set of non-terminal histories is denoted by X and set of terminal histories is denoted by Z ;
- for every non-terminal history $x \in X$ let the $I(x)$ denotes the set of players who must choose at x and call it the set of active layers at x . For a player i the set X_i denotes the set of non-terminal histories where player i is active;
- for every player i a collection of sets of information that i has about the opponent's past choices is denoted by H_i . Every information set $h \in H_i$ consist of a set of non-terminal histories x_1, x_2, \dots, x_k that have been realised without the knowledge which one. If $h = x$ then player is sure that history x has been realised;
- a set of available choices for player i and information set h is denoted by $C_i(h)$ meaning that the player i is able to make any choice from $C_i(h)$ when the game reaches the information set h .

When game reaches the terminal state $z \in Z$ every single player is rewarded by utility $u_i(z)$. Here we are considering only a perfect recall game when every player remembers his choices and his opponent's previous choices.

Complete choice plan is called a strategy. A strategy for a player i is a function s_i that assigns to his information set $h \in H_i$ available choice $s_i(h) \in C_i(h)$ unless h can not be reached due to some choice $s_i(h')$ at earlier information set $h' \in h_i$. In this case no choice needs to be done at h . We denote by S_i a set of all strategies for player i .

The chosen strategy combination $(s_1, \dots, s_{i-1}, s_{i+1}, s_N)$ leads to h if there is s_i strategy for player i that together with this strategy would lead to h . Strategy s_i

leads to information set h if there is some strategy for the opponents that together with this strategy would lead to h . Those assumptions result in the definition of the conditional belief for a player i at h about the opponent's strategies that is a probability distribution $b_i(h)$ over the set of the opponent's strategy combinations assigning a positive probability only to strategy combinations that leads to h .

Among all strategies, we may consider dominant strategies:

- a strictly dominant strategy is that strategy that always provides greater utility to a the player, taking into account all the other player's strategies;
- a weakly dominant strategy is strategy that results at least the same utility for all the other player's strategies, [19].

Lets consider the information set h for player i . In such a case if player holds a conditional belief $b_i(h)$ and given the strategy s_i that leads to h this strategy is optimal if:

$$u_i(s_i, b_i(h)) \geq u_i(s'_i, b_i(h)) \quad (2)$$

for every other strategy s'_i that leads to h . Similarly, a conditional belief vector $b_i = [(b_i(h))]_{h \in H_i}$ for player i about his opponent's strategies concatenates at every information set $h \in H_i$ conditional beliefs $b_i(h)$ about the opponent's strategies.

A belief hierarchy in dynamic games defines for every player belief about the opponent's choices and the opponent's belief hierarchies. The hierarchy is formulated as follows:

- First order belief: the belief that player has about the opponent's strategies; Let us denote the belief hierarchy by $t_i^{s_i}$ (type) indication belief of the player i starting at his choice s_i .
- Second order belief: the belief that player has about the belief that the opponents have about their opponent's strategies;
- Third order belief: the belief that player has about the belief that the opponents has about the belief of this player opponent's strategies;
- and so on.

For every player i we denote by T_i the set of types that are considered for this player. Then, an epistemic model specifies for every player i a set T_i of possible types. Additionally, every type t_i for player i specifies for every information set $h \in H_i$ a probability distribution $b_i(t_i, h)$ over the set:

$$(S_1 \otimes T_1) \otimes \dots \otimes (S_{i-1} \otimes T_{i-1}) \otimes (S_{i+1} \otimes T_{i+1}) \otimes \dots \otimes (S_n \otimes T_n) \quad (3)$$

of his opponent's strategy - type combination. This probability distribution assigns only positive probability to the opponent's strategy combinations that leads to h . b_i represents the conditional belief that type t_i has at h about the opponent's strategies and types.

To define the beliefs about the future strategies, we introduce:

- Two information sets h and h' are simultaneous if there is a history that is present in both h and h' ;

- Information set h' follows information set h if there is a history x in h and a history x' in h' such that history x' follows history x ;
- Information set h' weakly follows information set h either h' follows h or is simultaneous with h ;

If we consider a type t_i for player i , and information set h for player i and an information set h' for player j , then we say that type t_i believes at h that the opponent j will chose rationally at h' if his conditional belief $b_i(t_i, h)$ at h assigns only positive probability to strategy-type pairs (s_j, t_j) for player j where strategy s_j is optimal for type t_j at information set h' . Analogically, we say that type t_i believes at h in opponent's future rationality if t_i believes at h that j will choose rationally at every information set h' for player j that weakly follows h . Type t_i expresses the common belief in future rationality if t_i express k -fold belief in future rationality for every order belief k .

Now, we can define choosing the rational strategy under belief in future rationality, as follows:

- Player i can rationally choose some strategy s_i under common belief in future rationality if there is some epistemic model and some type t_i for player i in this model such that t_i expresses common belief in future rationality and strategy s_i is optimal for type t_i at every information set $h \in H_i$ that s_i leads to.

Consider an information set $h \in H_i$ for player i . Let $S_i(h)$ be a set of strategies of player i that leads to h and $S_{-i}(h)$ be a set of his opponent's strategy combinations that leads to h . The pair

$$\Gamma^0(h) = (S_i(h), S_{-i}(h)) \quad (4)$$

is called the full decision problem for player i at h . Similarly, a reduced decision problem for player i at h is a pair

$$\Gamma^1(h) = (D_i(h), D_{-i}(h)) \quad (5)$$

where $D_i(h) \subset S_i(h)$ and $D_{-i}(h) \subset S_{-i}(h)$.

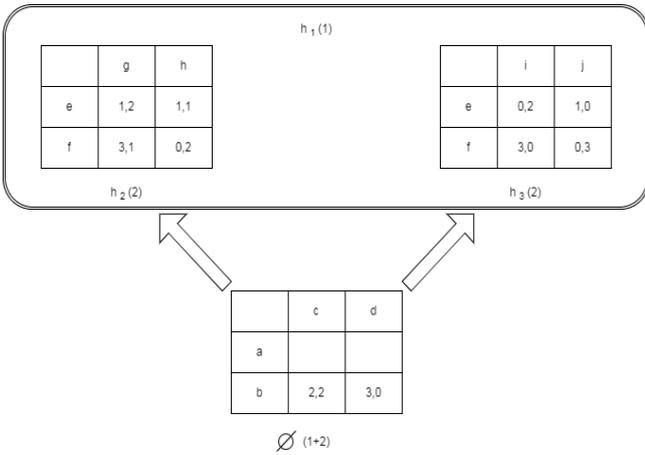


Fig. 1: The graph of game for two players

Algorithm for finding the strategies that can be rationally chosen under belief of future rationality is the backward dominance algorithm, [18]:

1. STEP 1: For every full decision problem $\Gamma^0(h)$ eliminate for every player i strategies that are strictly dominated at some decision problem $\Gamma^0(h')$ that weakly dominates $\Gamma^0(h)$ and at which player i is active. Denote resulted reduced decision problem by $\Gamma^1(h)$ for every information set h . If player i is active at h and the full decision problem is given by $(S_i(h), S_{-i}(h))$ when we are removing strategy from $(S_i(h))$. If player j is active at h but not player i and the full problem is a pair $(S_j(h), S_{-j}(h))$ in such case we are removing from $S_{-j}(h)$ every strategy combination that contains strategy s_i for player i .
2. STEP 2: For every reduced decision problem $\Gamma^1(h)$ eliminate for every player i those strategies that are strictly dominated at some reduced decision problem $\Gamma^1(h')$ that weakly follows $\Gamma^1(h)$ and at which player i is active. Denote resulted reduced decision problem by $\Gamma^2(h)$ for every information set h .
3. STEP 3: continue until no more strategies may be eliminated.

Theorem: For every $k \geq 1$ the strategies that can rationally be chosen by a type that expresses up to k -fold belief in future rationality are strategies in Γ^{k+1} that survived the first $k + 1$ steps of the backward dominance algorithm presented above. Additionally, the strategies that can rationally be chosen by a type that expresses common belief in future rationality are strategies that survived full backward dominance algorithm. Those strategies are in Γ^k for every k , [18].

The example of the game is presented in the fig 1. In the first stage player 1 may choose a or b, player 2 c or d. In the first case the game has another stage. In second case the game ends with a payoff for player 1 equal to 2 if player 2 chosen c, a payoff for player 1 equal 3 if player 2 chosen d. The relevant payoffs for player two are 2 and 0. If the game is continued the game moves to information set $h_1(1)$ for player 1 and information sets $h_2(2)$ and $h_3(2)$ for player 2. In the second stage player 1 may choose e or f, player 2 g or h if combination (a,c) was chosen in first stage. In the second stage player 1 may choose e or f, player 2 g or h but the player 2 may chose g or h if combination (a,c) was chosen in first stage and he may choose i or j if combination (a,d) was chosen in first stage. The payoffs are given by two tables in side rounded box. $X = \emptyset, (a, c), (a, d)$, $Z = \{(b, c), ((a, c), (f, h)), \dots, (b, d)\}$, not-terminal histories $\emptyset, (a, c), (a, d)$, $I(\emptyset) = \{player1, player2\}$, $I((a, c)) = \{player1, player2\}$, $I((a, d)) = \{player1, player2\}$, $H_1 = \{\emptyset, h_1\}$, $H_2 = \{\emptyset, h_2, h_3\}$, $C_1(\emptyset) = \{a, b\}$, $C_1(h_1) = \{e, f\}$, $C_2(\emptyset) = \{c, d\}$, $C_2(h_2) = \{g, h\}$, $C_2(h_3) = \{i, j\}$, $u_1(b, c) = 2$, $u_2(b, d) = 0$, $u_1((a, d), (f, i)) = 3$, $u_2((a, d), (f, i)) = 0$, [18].

IV. NUMERICAL SIMULATION

In the proposed model, the attacker is a player 2 and the Cloud defender is player 1, (see Table 1).

The numerical experiment follows our research presented in [14]. *Assets* is the set of Cloud system components to be protected:

Player 1	Player 2
Cloud provider Defender	Malicious individual, hacker Attacker

TABLE 1: Players roles inside the cloud.

$$a \in Assets \quad (6)$$

Single attack is targeted into a specific asset. All considered attacks against the asset a are gathered in the form of the set:

$$Attacks^a = \{attack_1^a, \dots, attack_m^a\} \quad (7)$$

where m is the number of considered attacks. A countermeasure is an action taken to protect the asset. A set of considered countermeasures against the asset a is denoted by:

$$controls^a = \{c_1^a, \dots, c_n^a\}. \quad (8)$$

If we denote by $P^a(attack_i^a, c_j^a)$ be the probability of a successful attack on asset a by using threat number $i \in \{1, 2, \dots, m\}$ that are protected by the countermeasure $j \in \{1, 2, \dots, n\}$.

Therefore the j -th pure strategy, [19] s_j^1 for the Defender is applying the countermeasure c_j^a . Then,

$$s_j^1 = 1, s_{-j}^1 = 0 \quad (9)$$

if c_j was chosen by player 1. Additionally, the i -th pure strategy s_i^2 strategy for the Attacker is choosing the threat number i , that is

$$s_i^2 = 1, s_{-i}^2 = 0 \quad (10)$$

if a_i^a was chosen by player 2.

Val^a is income that player 1 gains from a protected asset a when it is working. $CostDef_{c_j^a}$ is cost of applying for asset a control number $j \in \{1, 2, \dots, n\}$. By $Gain^a$ let us denote reward for the player 2 for the successful attack into asset a , and by $CostAttack_{attack_i^a}^a$ let us denote the cost of such attack.

The payoff player 1 was modelled as:

$$u_1(s_1^a, b_1(h))^a = \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} s_1^j \beta_1^i [P^a(attack_i^a, c_j^a) * (-Val^a - CostDef_{c_j^a}^a) + (1 - P^a(attack_i^a, c_j^a)) * (Val^a - CostDef_{c_j^a}^a)] + \dots \quad (11)$$

$$+ \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} (1 - s_1^j) (\beta_1^i [\bar{P}^a(attack_i^a, c_j^a) * (-Val^a) + (1 - \bar{P}^a(attack_i^a, c_j^a)) * (Val^a)]) \quad (12)$$

where $\bar{P}^a(attack_i^a, c_j^a)$ is the probability of the successful attack number i on asset a considering the fact that the countermeasure c_j was chosen for this asset. Val^a indicates the value obtained by the Cloud provider from the asset working properly. This form of the payoff function assumes the worst case scenario. If the attack was successful, the task must be performed

ones again therefore the provider lost the computational cost and paid for the protection resulting in $-Val^a - CostDef_{c_j^a}^a$, see eq.(11). If the Attack was unsuccessful, he invested in protection but his asset was working producing the income: $Val^a - CostDef_{c_j^a}^a$. If asset was not protected the provider may expect $-Val^a$ in case of successful attack, and Val^a income in case of not successful attack. Those costs are lower, but the probabilities of being attacked when unprotected are higher.

Analogously, the payoff for player 2 was as:

$$u_2(s_2, b_2(h))^a = \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} \beta_2^j s_2^i [P^a(attack_i^a, c_j^a) * (Gain^a - CostAttack_{attack_i^a}^a) + (1 - P^a(attack_i^a, c_j^a)) * (-CostAttack_{attack_i^a}^a)] + \dots \quad (13)$$

$$+ \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} \beta_2^j (1 - s_2^i) [\bar{P}^a(attack_i^a, c_j^a) * (Gain^a - CostAttack_{attack_i^a}^a) + (1 - \bar{P}^a(attack_i^a, c_j^a)) * (-CostAttack_{attack_i^a}^a)] \quad (14)$$

$$b_1(h) = (\beta_1^1, \beta_1^2, \dots, \beta_1^m) \quad (15)$$

$$\beta_1^1 + \beta_1^2 + \dots + \beta_1^m = 1 \quad (16)$$

$$b_2(h) = (\beta_2^1, \beta_2^2, \dots, \beta_2^n) \quad (17)$$

$$\beta_2^1 + \beta_2^2 + \dots + \beta_2^n = 1 \quad (18)$$

$$s_1^a = (c_1^a, c_2^a, \dots, c_n^a) \quad (19)$$

and

$$s_2^a = (attack_1^a, attack_2^a, \dots, attack_m^a) = (a_1^a, a_2^a, \dots, a_m^a) \quad (20)$$

For simulating the attacks to the Cloud infrastructure, the numerical test was performed on a CloudSim environment [1]. The cloud infrastructure model is presented in Table 2 discussed in [14].

Asset number a	VM type	Speed GFLOPS	Energetic profile $min^a : max^a$ in Watts
-1	1	0.02	90:105
0	1	0.02	90:105
1-20	1	0.02	90:105
21-40	2	0.05	93:110
41-60	3	0.1	100:120
61-80	4	0.2	150:170
81-100	5	0.3	200:230

TABLE 2: SimGrid VMs used for simulation, $Val^a = (max^a - min^a)/2$

The $Attacks^a$ were chosen according to the Cloud Security Alliance list of 7 most dangerous threats for cloud systems, see [7], see Table 3. :

1. $a_1^a, attack_1^a$: Task injection
2. $a_2^a, attack_2^a$: Denial-of-service attack (DoS attack)
3. $a_3^a, attack_3^a$: Task modification
4. $a_4^a, attack_4^a$: Distributed DoS attack (DDoS)
5. $a_5^a, attack_5^a$: Tasks loss
6. $a_6^a, attack_6^a$: Energy denial-of-service attack (eDOS) see [?].
7. $a_7^a, attack_7^a$: Unknown attack: asset not working.

The tested $controls^a$ were selected from the cloud controls matrix [2][5]:

1. c_1^a : RSA digital signature with 1024 bit key for each task batch
2. c_2^a : "anti-virus" job to check input connections into the asset

3. c_3^a : "firewall" job to monitor tasks
4. c_4^a : escaping- closing infected VM, opening the new one
5. c_5^a : task integrity monitoring by SHA-2 hashing for each task batch
6. c_6^a : energy cupping - scaling up the VM, see [20].

During tests on SimGrid environment we simulated the execution of tasks and measured the energy consumed by simulated VMs. All Virtual Machines were monitored during task execution, idle time and scaling and escaping (cloning).

Asset	Input protection	Inner protection	Output protection
-1-100	c_1^a, c_2^a, c_3^a	c_4^a, c_5^a, c_6^a	c_1^a, c_2^a, c_3^a
-1	RSA verific. of user	escaping VM	RSA sign.
0	RSA verific. of task collector	escaping VM	RSA sign.
1-100	RSA verific. of task scheduler	escaping VM	RSA sign.

TABLE 3: Asset protection scheme

The energy expenditure is presented in Table 2 and see Table 4.

type	P_1^i	P_b^i	P_o^i	P_c^i	$CostDef_c^a$					
					c_1^a	c_2^a	c_3^a	c_4^a	c_5^a	c_6^a
1	90	106.8	63	27	20	18.4	0.53	1.06	54	90
2	93	114.6	65	28	18	14	0.57	1.14	56	93
3	100	130	70	30	12	12	0.65	1.3	60	100
4	150	200	105	45	10	3.4	1.0	2.0	90	150
5	200	290	140	60	6.5	2	1.45	2.9	120	200

TABLE 4: Measured power for a 10 GFLOPs workload [Watts] for all considered VM types, P_1^i power consumed in idle state, P_b^i power consumed in busy mode, P_o^i power consumed opening new VM, P_c^i power consumed for closing VM, the last six columns defines the values of $CostDef_c^a$

The payoff function for the player 2 was simulated in points, assuming bigger utility in case of successful attack on more powerful assets (see table 5).

Asset nr	Gain	att_1^a cost	att_2^a cost	att_3^a cost	att_4^a cost	att_5^a cost	att_6^a cost	att_7^a cost
-1	10	1	2	2	1	2	5	9
0	60	2	4	2	1	2	5	50
1-20	10	3	6	2	1	2	10	9
21-40	20	10	10	12	1	12	15	18
41- 60	30	10	10	12	2	12	20	25
61-80	40	10	10	12	3	12	25	35
81-100	50	10	10	12	4	12	30	45

TABLE 5: Gain and Cost of Attacks $attack_1^a$ - $attack_7^a$ on simulated Cloud in points

The probability of successfully attack was modelled based on [3] and are presented in Table 6 and Table 7.

Counterme.	at_1^a	at_2^a	at_3^a	at_4^a	at_5^a	at_6^a	at_7^a
c_1^a RSA	0	0.95	0.9	0.8	0	0.8	0.5
c_2^a Anti virus	0.95	0.9	0.6	0	0.8	0.8	0.5
c_3^a Firewall	0.95	0	0.6	0.8	0.8	0.8	0.5
c_4^a Escape	0.95	0	0.6	0	0	0	0
c_5^a SHA-2	0.95	0.9	0.6	0.8	0.8	0.8	0.5
c_6^a Cupping	0.95	0.9	0.6	0.8	0.8	0	0.5

TABLE 6: $P^a(attack_i^a, c_j^a)$, of successful attacks $attack_1^a - attack_7^a$ on assets protected by countermeasures $c_1^a - c_6^a$, for the sake of presentation simplicity assumed constant

The below Python code was used to simulate attack on different assets using the available attack methods and available control methods. The defender's gain is presented in Table 8 where the cost of the defence was

Counterme.	at_1^a	at_2^a	at_3^a	at_4^a	at_5^a	at_6^a	at_7^a
c_1^a RSA	0.5	1	0.95	0.8	0.2	1	0.6
c_2^a Anti virus	0.95	0.9	0.6	0.2	1	1	0.6
c_3^a Firewall	0.95	0	0.6	0.9	1	1	0.6
c_4^a Escape	0.95	0	0.6	0.5	0.2	0	0.2
c_5^a SHA-2	1	0.95	0.8	0.9	1	1	0.6
c_6^a Cupping	0.95	0.9	0.6	0.9	1	0	0.6

TABLE 7: Probabilities of successful attacks on assets that not protected, $\bar{P}^a(attack_i^a, c_j^a)$, for the sake of presentation simplicity assumed constant

defined in Table 4. The attacker's gain calculated using the below code is presented in Table 9 where the cost of the attack was defined in Table 5. The relation between attacks and defences was modelled by probabilities of successful attacks on protected assets defined in Table 6 and probabilities of successful attacks on not protected assets defined in Table 7.

```

def u_1(a, s1, beta_1):
    res = 0
    for i in range(num_attacks):
        for j in range(num_controls):
            res += s1[j] * beta_1[i] * (P_a[j][i] *
                (-Val[a] - CostDef[a][j])) + (1 -
                P_a[j][i]) * (Val[a] - CostDef[a][j]))
    for i in range(num_attacks):
        for j in range(num_controls):
            res += (1 - s1[j]) * beta_1[i] *
                (P_a_bar[j][i] * (-Val[a]) + (1 -
                P_a_bar[j][i]) * Val[a])
    return res

def u_2(a, s2, beta_2):
    res = 0
    for i in range(num_attacks):
        for j in range(num_controls):
            res += beta_2[j] * s2[i] * (P_a[j][i] *
                (Gain[a] - CostAttack[a][i])) + (1 -
                P_a[j][i]) * (-CostAttack[a][i]))
    for i in range(num_attacks):
        for j in range(num_controls):
            res += beta_2[j] * (1 - s2[i]) *
                (P_a_bar[j][i] * (Gain[a] -
                CostAttack[a][i]) + (1 - P_a_bar[j][i])
                * (-CostAttack[a][i]))
    return res

```

Asset	c_1^a	c_2^a	c_3^a	c_4^a	c_5^a	c_6^a
-1	-34.51	-33.61	-15.8	-16.06	-69.47	-106.08
0	-34.51	-33.61	-15.8	-16.06	-69.47	-106.08
1-20	-34.51	-33.61	-15.8	-16.06	-69.47	-106.08
21-40	-34.44	-31.24	-17.88	-18.14	-73.53	-111.23
41- 60	-31.35	-32.28	-21.01	-21.3	-80.63	-121.44
61-80	-29.35	-23.68	-21.36	-22.0	-110.63	-171.44
81-100	-35.52	-32.42	-32.0	-32.9	-150.94	-232.16

TABLE 8: Defender's gain for controls c_1^a - c_6^a

Asset	att_1^a	att_2^a	att_3^a	att_4^a	att_5^a	att_6^a	att_7^a
-1	25.46	26.30	25.77	24.6	24.0	24.94	24.74
0	218.75	223.79	220.62	213.6	210.01	215.66	214.46
1-20	15.46	16.30	15.77	14.6	14.0	14.94	14.74
21-40	16.92	18.60	17.54	15.2	14.0	15.89	15.49
41- 60	31.38	33.89	32.31	28.8	27.0	29.83	29.23
61-80	82.83	86.19	84.08	79.4	77.0	80.77	79.97
81-100	114.29	118.49	115.85	110.0	107.0	111.71	110.72

TABLE 9: Attacker's gain for attacks $attack_1^a$ - $attack_7^a$

The results indicate that:

- the losses for Cloud provider due to attack differs according to the considered asset type,

- if the stronger computing units (assets 61-100) are attacked the cloud defender losses are more severe than in case of less power full ones (assets -1-60),
- in the future the Cloud defender should change the controls number 5 and 6 into stronger ones for all the assets in his system, see tab 8.

The simulations also show that, assuming the Cloud attacker rationality:

- the attacker will concentrate his efforts on asset number 1, that is the scheduling unit,
- the most beneficial attack type will be attack number 2. into scheduling unit.

Considering the above results the Cloud defender should invest in protecting the unit that is scheduling the tasks in his system.

V. CONCLUSIONS

In this paper we presented Epistemic Game theory based model with Conditional Believes to build automating system for security decision making process in Clouds. The model considers the separate payoff functions for modelling both attack and defence scenarios. It relates to the different objectives of Cloud defender and Cloud attacker. The model uses the belief concept to represent the rationality of the decision making process. The behaviour of the Cloud defender and Cloud attacker is calculated by using numerical optimisation for the mathematical model presented in eq. (1)-(20). The main result of modelling process is finding the best strategies for the Cloud provider to protect from cyber-security attacks. In the future, we would like to incorporate more advanced game models, that allow mixing Stackelberg Games with Epistemic Game theory with Conditional Believes.

REFERENCES

- [1] Cloudsim, <http://www.cloudbus.org/cloudsim/>.
- [2] Security Guidance for Critical Areas of Focus in Cloud Computing V2.1. Technical report, 2009.
- [3] OWASP, Top. 10 2010. The Ten Most Critical Web Application Security Risks. Technical report, 2010.
- [4] NIST Special Publication 800-144: Guidelines on Security and Privacy in Public Cloud Computing. Technical report, 2011.
- [5] CSA Controls Matrix v.3. Technical report, 2013.
- [6] A. H. Anwar, G. Atia, and M. Guirguis. Game theoretic defense approach to wireless networks against stealthy decoy attacks. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 816–821, Sept 2016.
- [7] J. Archer, A. Boehme, D. Cullinane, P. Kurtz, N. Puhmann, and J. Reavis. Top Threats to Cloud Computing V1.0. Technical report, 2010.
- [8] N. Basilio, A. Lanzi, and M. Monga. A security game model for remote software protection. In *2016 11th International Conference on Availability, Reliability and Security (ARES)*, pages 437–443, Aug 2016.
- [9] O. Dianat and M. Orgun. Modelling bayesian attacker detection game in wireless networks with epistemic logic. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaborate-Com)*, pages 210–215, 2012.
- [10] S. Dlugosz. *Multi-layer Perceptron Networks for Ordinal Data Analysis*. Logos Verlag, 2008.
- [11] E. Eisenstadt and A. Moshaiov. Novel solution approach for multi-objective attack-defense cyber games with unknown utilities of the opponent. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1(1):16–26, Feb 2017.
- [12] M. P. Fanti, M. Nolich, S. Simié, and W. Ukovich. Modeling cyber attacks by stochastic games and timed petri nets. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002960–002965, Oct 2016.
- [13] M. T. Hagan, H. B. Demuth, and M. Beale. *Pareto-Nash-Stackelberg Game and Control Theory*. Springer International Publishing, UK, 2018.
- [14] A. Jakóbi, F. Palmieri, and J. Kołodziej. Stackelberg games for modeling defense scenarios against cloud security threats. *Journal of Network and Computer Applications*, 110:99 – 107, 2018.
- [15] A. Jakóbi. Stackelberg game modeling of cloud security defending strategy in the case of information leaks and corruption. *Simulation Modelling Practice and Theory*, 103:102071, 2020.
- [16] Y. Li, D. E. Quevedo, S. Dey, and L. Shi. A game-theoretic approach to fake-acknowledgment attack on cyber-physical systems. *IEEE Transactions on Signal and Information Processing over Networks*, 3(1):1–11, March 2017.
- [17] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.
- [18] A. Perea. Epistemic game theory: Reasoning and choice. *Epistemic Game Theory: Reasoning and Choice*, pages 1–561, 01 2012.
- [19] S. Tadelis. *Game Theory: An Introduction*. Princeton University Press, 2013.
- [20] A. Wilczyński and A. Jakóbi. Using Polymatrix Extensive Stackelberg Games in Security-Aware Resource Allocation and Task Scheduling in Computational Clouds. *Journal of Telecommunications and Information Technology*, 1, 2017.
- [21] A. Wilczyński, A. Jakóbi, and J. Kołodziej. Stackelberg security games: Models, applications and computational aspects. 3:70–79, 2016.
- [22] X. Yang, X. He, J. Lin, W. Yu, and Q. Yang. A game-theoretic model on coalitional attacks in smart grid. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 435–442, Aug 2016.
- [23] J. Zhu, B. Zhao, and Z. Zhu. Leveraging game theory to achieve efficient attack-aware service provisioning in eons. *Journal of Lightwave Technology*, 35(10):1785–1796, May 2017.

AUTHOR BIOGRAPHIES

ŁUKASZ GAŻA He received his M.Sc. in the field of Applied Physics with Computer Modelling at the Tadeusz Kosciuszko Cracow University of Technology. Since 2019 he is a Research and Teaching Assistant at the Tadeusz Kosciuszko Cracow University of Technology. His e-mail address is lukasz.gaza@pk.edu.pl.

AGNIESZKA JAKÓBIK She received her M.Sc. in the field of Stochastic Processes at the Jagiellonian University, Poland and a PhD degree in Artificial Neural Networks at the Tadeusz Kosciuszko Cracow University of Technology, Poland. Since 2009 she is an Assistant Professor at the Tadeusz Kosciuszko Cracow University of Technology, email: ajakobik@pk.edu.pl.

SIMULATING THE PROGRAMMABLE NETWORKS FOR HLA COMPATIBLE HIGH-PERFORMANCE SIMULATORS

Kayhan M. İmre
Department of Computer Engineering,
Hacettepe University
Ankara, TURKEY
E-mail: ki@hacettepe.edu.tr

KEYWORDS

Parallel discrete event simulation, network simulation, high performance programmable networks.

ABSTRACT

This paper explores a parallel discrete event simulator that simulates a programmable fat tree network. The programmable networks can be programmed to perform application specific tasks. The task explored in our research is a time management functionality offloaded to the network switches. Specifically, the network switches used for constructing the fat tree run Greatest Available Logical Time (GALT) computation. In this paper, this switch-based GALT computation is compared against two node-based GALT computations using the simulator developed.

INTRODUCTION

IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) IEEE Std 1516™-2010 defines a standard for distributed simulation (IEEE Standard 1516.1-2010). HLA has well defined managements functions for implementing distributed simulations or integrating individual simulations as a single distributed simulation. HLA standard does not exclude high performance parallel simulations that require low latency interactions and high-level parallelism. The implementation of Run Time Infrastructure (RTI) of HLA is the key element for achieving high performance execution of an HLA compliant parallel simulation. One of the bottlenecks for achieving high performance RTI execution is advancing the simulation time of the joined federates. The federates are the parallel processes of an HLA compliant parallel simulation. The time regulating federates send time advance requests (TAR) to RTI, and in return, RTI responds such a request with time advance grant (TAG) callback. Each time regulating federate provides a time advancement plus a lookahead value. Behind the scenes, the RTIs join in a distributed computation to find the minimum of all $time + lookahead$ values. This minimum value is named as Greatest Available Logical Time (GALT) that is no federate gets time advance grant beyond this value. As a consequence, the distributed GALT calculation becomes a global synchronization mechanism that harnesses the progress of federates.

Improving GALT calculation time will increase the parallelism by decreasing the time spent for synchronization. This paper presents a simulation application developed as a part of a research that explores programmable networks for improving the performance of parallel applications. The simulation developed is used for evaluating network offloaded GALT computation and some other functions of RTI. Programmable switches are becoming available for performing application specific functions (Dang et al. 2015; Jin et al. 2017; Kaur et al. 2021). In the rest of the paper, the implementation of the simulator is presented. In the evaluation section, three different GALT algorithms are compared against each other using the simulator implementation.

SIMULATING PROGRAMMABLE NETWORK

The fat tree topology is one of the preferred topologies in high performance computing domain. This topology can be built from high speed standard network switches (Al-Fares et al. 2008) but constructing the fat tree topology from programable switches can improve the network beyond its connectivity rich characteristics. Some critical distributed computations with lightweight processing but requiring coordination of many participants are good candidates for implementing on the distributed network switches. The Greatest Available Logical Time (GALT) calculation of High Level Architecture (HLA) is one of those candidates. The performance of GALT calculation has a direct effect on the overall performance of a simulation application. The GALT calculation can be considered as a synchronization point that the joining federates of an HLA compliant simulation wait for the time advancement requests to be granted, and then, the federates can continue their local computations. The simulator presented in this paper is intended for investigating the behavior of the GALT calculation when it is implemented on a programable network. Along with switch-based GALT calculations, two other node-based GALT calculations were also implemented using the simulator to provide an equal ground for assessing performance characteristics.

The first implementation choice was to make whether the simulator be sequential or parallel, and the latter was chosen to make implementation simple and scalable. The logical processes of the parallel implementation are implemented as processes of MPI-based (Message Passing Interface) parallel application. One-to-one mapping of logical processes to MPI processes provides

a good encapsulation. Additionally, one-to-one mapping of logical processes is achieved by mapping a logical process to a node to be simulated. A node can be either a programmable network switch or a computer. In Figure 1, a fat tree constructed from 4-port switches is depicted, and the numbers are MPI process identifications (i.e. ranks). There are total of 36 logical processes implement Chandy/Misra/Bryant null message algorithm to simulate the high-level behavior of the overall system (Chandy and Misra 1979; Bryant 1977; Fujimoto 2000). There are four logical processes (LPs) for simulating the core switches, eight LPs for the aggregate switches, eight LPs for the edge switches, and finally, there are sixteen computers represented by their LPs.

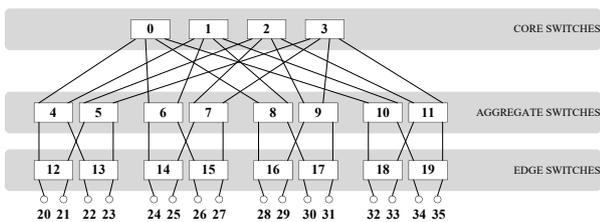


Figure 1: Fat tree for 16 compute nodes and MPI ranks.

In Figure 2, a larger fat tree constructed from 6-port switches is depicted. Each logical process can decide its role in the simulation by controlling MPI process rank, and then, each chooses its role in the simulation. There are four different roles and four corresponding LP types; Core switch, aggregate switch, edge switch and compute node (computer). Each LP can also locate its place in the topology using its own rank, and identify the MPI ranks that the LP exchanges messages (simulation events) with.

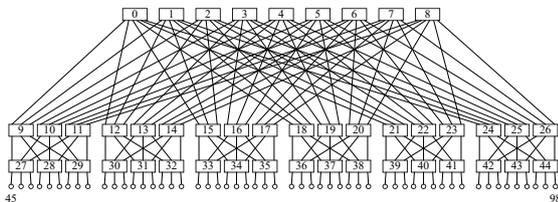


Figure 2: Fat tree for 54 compute nodes and MPI ranks.

After identifying the other LPs to communicate with, each LP assigns a pair of event FIFOs to each interaction with another LP, which also corresponds a communication link in the simulated fat tree. In Figure 3, up and downlink FIFOs assigned to communication links are shown for aggregate and edge switch LPs. In Figure 4, FIFOs are shown for core switch LPs have downlinks only. Finally, the LPs corresponding to compute nodes have only single links connected to edge switches (Figure 5). The events exchanged between LPs are used for simulating the communication packets transmitted over the communication links. The event data structure contains a payload field to transmit the real contents of the message packets simulated. This enables our simulator to run the real algorithms with real message contents while simulating the network behavior of the fat

tree. The FIFOs inside the dash-lined boxes belong to LP's logic that simulates communication link queues.

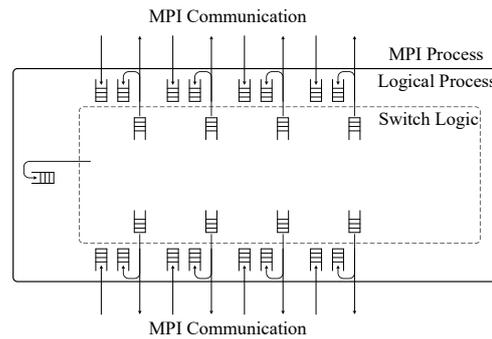


Figure 3: Logical Process FIFOs for aggregate and edge switches.

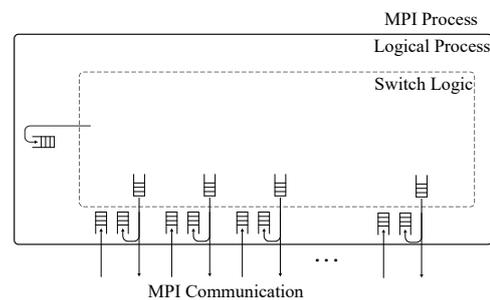


Figure 4: Logical Process FIFOs for core switches.

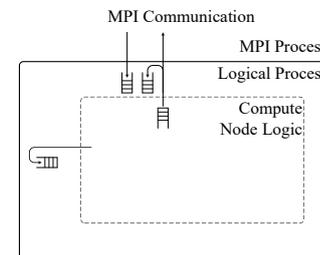


Figure 5: Logical Process FIFOs for compute nodes.

Finally, in each figure, the single FIFO standing on the left-hand side not attached to any link is used for storing events related to local processing timing. For example, in the compute node, the real application with real data can be executed but execution timing is simulated by the events that mark start and end time of a real computation. Such timings are not easy to measure using system clock but the instrumented code can predict execution timings and feed the FIFOs with proper events.

In Figure 6, the “main” function of the MPI process is given as an excerpt to show general structure of the simulator. After constructing the fat tree topology logically, each MPI process assigns up/down link FIFO pairs to communication links. Initially, each MPI process sends a null message to each link using “MPI_Isend” function which is an asynchronous message send function. In the main simulation loop, synchronous message receive function is used for waiting a message from a specific process (implementing an LP). After

making sure that there is no empty FIFO related to communication links, the algorithm finds the event with the minimum time, and consumes the events with the same time stamps. Since this is a typical implementation of Chandy/Misra/Bryant null message algorithm, there is no need to explain how the rest handles the null messages.

To simulate the programmability of communication switches, event triggered codes are placed in the “Consume_Event” function, and those codes implement application specific algorithms such as GALT calculation. The GALT calculation has several steps executed on the switches in a distributed manner. GALT calculation is triggered from a compute node by sending a time advance request (TAR) event to the edge switch LP. Upon receiving such event from a downlink, the edge switch LP checks the values received from downlinks previously, and if a new minimum value is calculated a new event is send to every uplink to inform the connected aggregate switch LPs. The aggregate switch LPs perform similar operation to inform the core switch LPs. When a core switch LP is reached, all the values from downlinks are checked. If a new minimum value is reached, this value is marked as the GALT value. In the next several steps, the GALT value is broadcasted downwards using downlinks until the event carrying the GALT value reaches a compute node. The compute node gets this event, creates a new event called “time advance grant” (TAG), and inserts the new event into the local application FIFO to finalize the GALT calculation.

```

void main(int argc, char* argv[])
{
    // Definitions and Initializations . . .
    // The construction of the Fat tree . . .
    for (int i = 0; i < ulcount; i++) // Initiate uplink communications and
        Null_Event(Time, i, UPLINK); // send NULL Events with current time
    for (int i = 0; i < dlcount; i++) // Initiate downlink communication and
        Null_Event(Time, i, DOWNLINK); // send NULL Events with current time
    while (continue_simulation) {
        for (int i = 0; i < ulcount; i++) // Wait for queues to become non-empty
            if (ulist[i].head == NULL) { // Uplink
                MPI_Recv(&urecv_buff[i][0], RECV_BUFF_SIZE, MPI_CHAR,
                    ul[i], ucount_recv[i], MPI_COMM_WORLD, &status[i]);
                // Add event to FIFO . . .
            }
        for (int i = 0; i < dlcount; i++) // Wait for queues to become non-empty
            if (dlist[i].head == NULL) { // Downlink
                MPI_Recv(&drecv_buff[i][0], RECV_BUFF_SIZE, MPI_CHAR,
                    dl[i], dcount_recv[i], MPI_COMM_WORLD, &status[i]);
                // Add event to FIFO . . .
            }
        // Find Min . . .
        Time = min;
        for (int i = 0; i < ulcount; i++) // Consume events with current time
            Consume_Event(Time, i, UPLINK); // Uplink FIFO
        for (int i = 0; i < dlcount; i++) // Consume events with current time
            Consume_Event(Time, i, DOWNLINK); // Downlink FIFO
        Consume_Event(Time, 0, APPLICATION); // Application FIFO
        for (int i = 0; i < ulcount; i++) // Send events with time + 1
            Null_Event(Time + 1, i, UPLINK); // Send if Uplink is not busy
        for (int i = 0; i < dlcount; i++)
            Null_Event(Time + 1, i, DOWNLINK); // Send if Downlink is not busy
    }
    MPI_Finalize();
}

```

Figure 6: The “main” function of the MPI process implements LPs.

EVALUATION

In this section, the performance evaluations of three different GALT calculation approaches are presented. Firstly, three different GALT calculations are

implemented using the simulator presented in this paper. The first one as explained in the paper is switch based GALT calculation. The second GALT calculation approach is a conventional one that uses compute nodes to reach the result. This approach benefits from fat tree topology characteristics by calculating local GALT values in the subtrees, recursively, and then, achieving the final GALT value in a compute node. The compute node holding the GALT value broadcasts this value to every other compute node using broadcast capability provided by the network. The third GALT calculation approach is broadcast-based, each compute node broadcasts its TAR request to everyone. Then, each compute node individually calculates the GALT value from the messages received.

The first evaluation of three approaches are carried out using a quiet network to measure pure performances of each GALT calculation approach. In the second part of the evaluation, the federates running on compute nodes are logically allocated in a two-dimensional space, and using data distribution management of HLA, each federate receives events from eight neighbors. While the federates are communicating using this publish-subscribe mechanism, the GALT calculation is triggered to measure its performance under different network traffic loads. The evaluation of three different GALT calculation approaches under various network loads are presented in the second part.

In Figure 7, the results of three different GALT calculation in a quiet network for three different network sizes are presented. For switch size 4, there are 16 compute nodes, and fat tree-based approach (FTAR) is the worst, broadcast-based approach (BTAR) is better, and programmable switch-based approach (TAR) is the best. For switch size 6, there are 54 compute nodes, the broadcast-based approach becomes the worst. For switch size 8, there are 128 compute nodes, the broadcast-based approach gets much worse, while TAR and FTAR keep their performances steady for all network sizes.

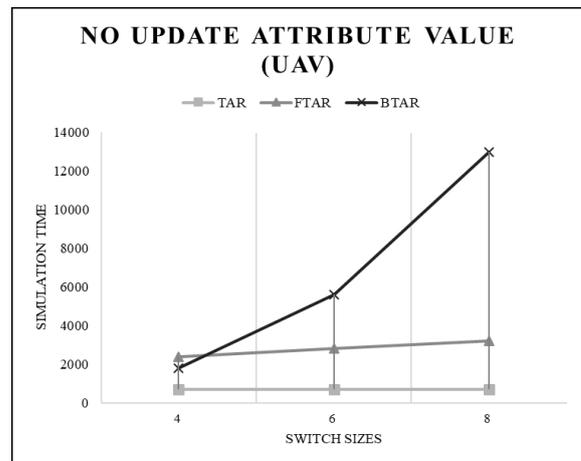


Figure 7: Three GALT algorithms for three different network sizes.

In the second part of the evaluation, additional network traffic is generated to see how it affects GALT calculations. In Figure 8, each compute node (i.e. federate running on it) sends an event to each neighbor causing each compute node receiving eight events from eight neighbors. The performances of switch-based (TAR) and fat tree-based (FTAR) calculations scale well while broadcast-based (BTAR) calculation gets much worse. In Figure 9, each compute node (i.e. federate running on it) sends five events to each neighbor causing each compute node receiving forty events from eight neighbors. The performances of TAR and FTAR are nearly the same.

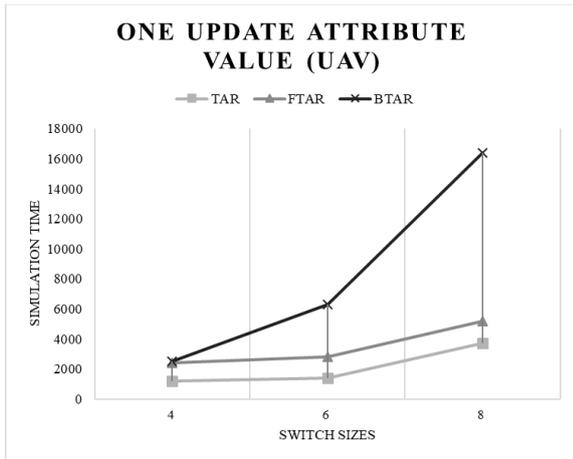


Figure 8: Three GALT algorithms for three different network sizes and light network traffic.

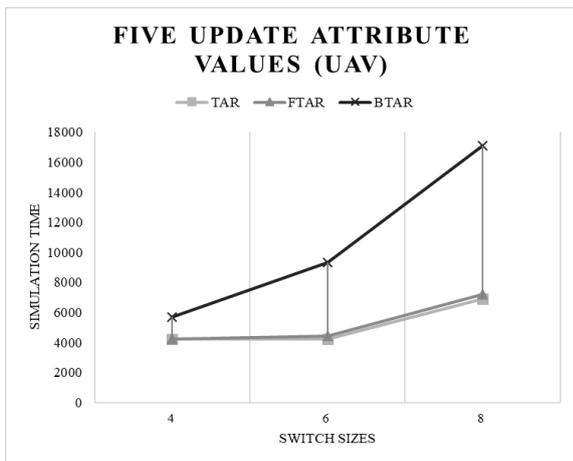


Figure 9: Three GALT algorithms for three different network sizes and heavy network traffic.

From Figure 10 to Figure 12, the performances of TAR and FTAR approaches are evaluated. Similar to previous measurements, different communication loads are tested for switch size 8 with 128 compute nodes only. In three figures, performance results for three different software related overhead values are presented. The software related overhead includes message preparation cost and handling costs of network software layers. Using some experimental values ($\alpha=100$, $\alpha=400$ and $\alpha=800$), two GALT calculation approaches (TAR and FTAR) are

compared. Since switch-based approach avoids such software related overheads, it performs much better when such overheads are high. As a general trend, when the network traffic gets heavier, the performance difference between switch-based (TAR) and fat tree-based (FTAR) calculations gets narrower.

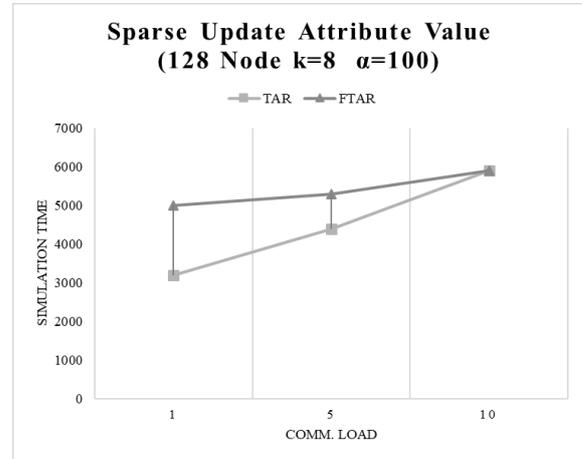


Figure 10: TAR and FTAR approaches with low software overhead.

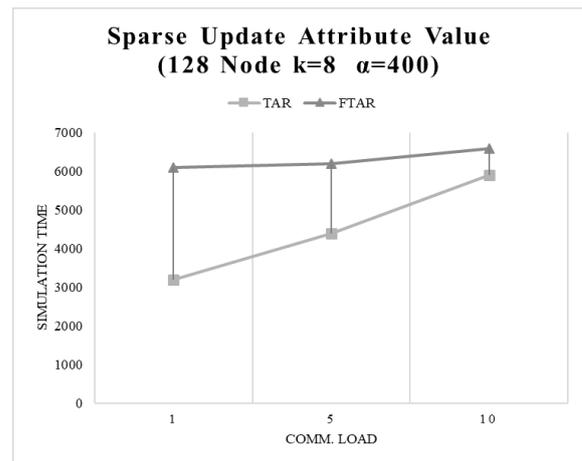


Figure 11: TAR and FTAR approaches with moderate software overhead.

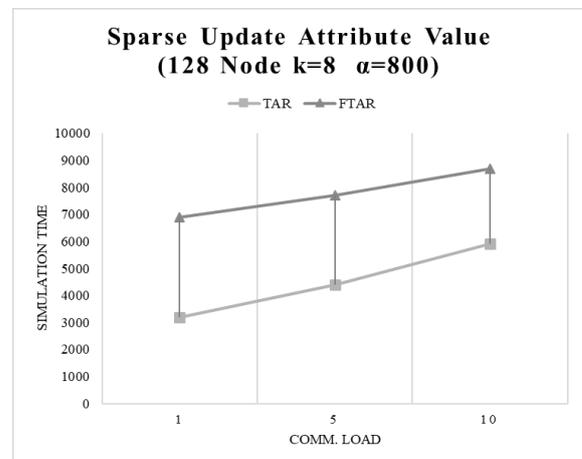


Figure 12: TAR and FTAR approaches with high software overhead.

CONCLUSION

The simulator presented in this paper simulates a simulation infrastructure that offloads some of its functionality to a programmable network. The programmable fat tree network performs GALT calculation on the programmable switches. The preliminary performance results of three different GALT calculation approaches are projected using the simulator presented in this paper. These preliminary experiments show that offloading time advancement calculation to the network switches will help to increase the performance of HLA compliant parallel and distributed simulations. The simulator is developed using MPI parallel programming library to accommodate real workloads running as a part of compute node LPs. The simulation events not only contain simulation related data but also real application data encapsulated in the payload field. This system can easily scale up using parallel computers both to shorten the execution time and to overcome memory limitations.

REFERENCES

- Al-Fares M.; Loukissas A.; and Vahdat A. 2008. "A scalable, commodity data center network architecture". SIGCOMM Comput. Commun. Rev. 38, 4 (October 2008), 63–74.
- Bryant, R.E. 1977. "Simulation of Packet Communications Architecture Computer Systems". MIT-LCS-TR-188, Massachusetts Institute of Technology.
- Chandy K. and Misra J. 1979. "Distributed Simulation: A Case Study in Design and Verification of Distributed Programs". IEEE Transactions on Software Engineering. Vol. 5. No. 5. 440–452.
- Dang H.T.; Sciascia D.; Canini M.; Pedone F.; and Soulé R. 2015. "NetPaxos: consensus at network speed". In Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research (SOSR '15). Association for Computing Machinery, New York, NY, USA, Article 5, 1–7.
- Fujimoto R. M; 2000. "Parallel and Distributed Simulation Systems", Wiley Interscience.
- IEEE Standard 1516.1-2010. "Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Federate Interface Specification."
- Jin X.; Li X.; Zhang H.; Soulé R.; Lee J.; Foster N.; Kim C.; and Stoica I. 2017. "NetCache: Balancing Key-Value Stores with Fast In-Network Caching". In Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17). Association for Computing Machinery, New York, NY, USA, 121–136.
- Kaur S.; Kumar K.; and Aggarwal N. 2021. "A review on P4-Programmable data planes: Architecture, research efforts, and future directions", Computer Communications, Volume 170, 109-129.

AUTHOR BIOGRAPHY



KAYHAN M. İMRE received his B.Sc., M.Sc., degrees in Computer Engineering from Hacettepe University, Ankara, Turkey and his Ph.D. degree in Computer Science from University of Edinburgh, Scotland, in 1985, 1987 and 1993 respectively. He is an Associate Professor at the Computer Engineering

Department, Hacettepe University, Ankara. His research interests are in parallel processing, parallel and distributed simulation and real-time systems.

Agent and Evolutionary-based Modelling and Simulation of a Simplified Living System

Adrian Sośnicki
Velis Real Estate Tech
Cracow, Poland

Daniel Grzonka
Department
of Computer Science
Cracow University
of Technology
Cracow, Poland

Łukasz Gaża
Department
of Computer Science
Cracow University
of Technology
Cracow, Poland

KEYWORDS

Living Systems, Artificial Life, Agent Systems, Simulation, Modelling, Genetic Algorithms, Behavioural Systems.

ABSTRACT

Modelling artificial life has been an issue explored for several decades. However, science continues to surprise us with novel approaches to this problem. The aim of this paper is to innovative model a simplified living environment based on the agent paradigm and genetic algorithms. This paper also proposes a novel way of defining agent systems and artificial life embedded in a genetic approach. In the modelled and implemented environment there is one species of fauna and a simple species of flora that serves as food for the fauna. The fauna is implemented using agents inscribed in a genetic representation. The experimental part of the work includes calls to the simulator and the study of the dependencies resulting from the simulation mechanisms.

I. INTRODUCTION

Many technologies take their inspiration from the natural environment. This is certainly the right direction — nature has had billions of years to develop organisms adapted to the prevailing and often changing conditions. The method identified with adaptation is natural selection, which has helped form the genetic code of living beings for millennia. It now serves the broader scientific community. Virtual living beings, however, do not seek food or shelter, but function maxima or the shortest paths in a graph (i.e. the equivalents of optimal solutions).

The process of natural selection is somewhat like a random process, but the possible solution converges much more quickly towards the sub-optimal one, and adapts when the environment changes. A solution always exists — often very surprisingly — even for a complex and constantly changing environment.

This claim can be put to the test by constructing a simulator in which virtual entities try to survive. If the simulated biosphere is equipped with the right tools, natural mechanisms will allow it to find a way to develop and persist. Exploring such a virtual environ-

ment is an engaging activity whose solutions are often surprising. It shows how a genetic algorithm — and therefore low-level natural intelligence — creates high-level intelligence. And all this without prior knowledge of your own environment.

The rest of the paper is structured as follows. Section II discusses the related work. We briefly define the general agent paradigm in Section III. Section IV introduces genetic representation of simulated living system. Next section (V) presents formal model of our novel agent-based ecosystem. Evaluation of the proposed living system is presented in Section VI. Finally, Section VII concludes conducted experiments.

II. RELATED WORK

The topic of artificial life modelling is very widely covered in the literature. There are many publications devoted to the issues of agent modelling, as well as the construction of living systems environments. However, there is no universal standard, defining a general approach to solving the problem. A few selected works addressing these topics are presented below.

A fundamental example of an artificial life simulation is Conway's Game of Life, which was invented in 1970 by British mathematician John Conway and popularised by Martin Gardner [8]. It is one of the first and most famous examples of a cellular automaton. The game is played on an infinite plane divided into square cells. Each cell can be in one of two states: it can be either alive or dead. A dead cell with exactly three living neighbours becomes alive in the following unit of time. A living cell with 2 or 3 living neighbours remains alive; it dies with a different number of neighbours.

M. Ling in [14] describes the new Python library that allows for modelling of artificial life simulation and Digital Organism Simulation Environment (DOSE). The presented approach is based on GA and biological hierarchy. It starts from genetic sequence that create the whole population. Genetic code is based on the structure of 3-nucleotide codons in naturally occurring DNA, and is built from 3-character instruction set that accepts no operand. In addition, the context of a 3D world consisting of ecological cells is introduced to simulate a physical ecosystem.

Another interesting solution is MUTANT: a multi-

agent toolkit for artificial life simulation by S. Calderoni and P. Marcenac [2]. It is generic platform that allows scientists in various fields of research to easily build simulation environment. The platform includes a model of self-adaptive agent with genetic evolving capabilities as learning mechanisms. It implements tools for agents design behaviour's programming environment's description and observation of running simulations. MUTANT is being developed in Java with the aim of being directly usable throughout the Internet.

J. Csonto et al. in [5] develop artificial life simulation based on multi-agent simulation system that could at least partially substitute the real experiments with real algae cells. Proposed simulator is based on real biological parameters of alga *Chlorella kessleri*, and use partial implementation of other mathematical models of algae population growth whereby it is possible to simulating the process of absorbing heavy metals from contaminated water. Model implementation is done in Swarm—multi agent object based simulation system and it's libraries.

There are also a number of review articles that attempt to provide a history and organise the nomenclature associated with Artificial-Life Ecosystems [7], characterise use cases for the idea of ALife (Artificial Life) [12], deals with the intersection of Artificial Intelligence (AI) and virtual worlds, focusing on AI agents and exploring the potential implications toward the human-level AI [17], representing a bottom-up approach to modelling complex life systems by using agents [15], or discuss major challenges to building live simulations covering various aspects [21].

III. THE AGENT PARADIGM

The idea of an agent paradigm (agent-oriented programming) dates back to 1990, with Yoav Shoham's publication of [20] dedicated to this issue. On the other hand, as early as in the 1970s, the roots of the software agent can be found in Carl Hewitt's actor model [1], [10]. Currently, the most popular definition of an agent is given by M. Wooldridge [24]:

An agent is a computer system that is embedded in a certain environment, and that is capable of autonomous actions in order to achieve ordered, specified goals.

Central to this definition is the feature of agent autonomy. It allows the agent to maintain full control over its internal state and the actions it takes [13]. Another feature is the perception mechanism, which allows the agent to observe the environment and the other agents in it. Based on the state of the environment and the state of other agents, it can make decisions [9].

An important element of an agent system is the environment. K. Cetnarowicz in [3] gives two types of components present in the environment — these are: (i) agents — a component of the environment, perceivable by the other actors in the environment; (ii) resources — components that do not have the ability to take initiatives, but can change according to their own established algorithm.

A formal notation of the abstract architecture of in-

telligent agents was proposed by M. Wooldridge [22], [24]. It defines an agent, an environment and functions representing the agent's various activities in the environment. According to the author, an intelligent agent is one that is capable of flexibly autonomous actions taken to achieve the goals [26]. By „flexible actions” he means three elements: (i) proactiveness – the ability to take initiative to achieve mandated goals; (ii) reactivity – the ability to perceive the environment and to react in good time to changes in it; (iii) social ability – the ability to interact with other agents (and even humans) to achieve mandated objectives.

As described earlier, agents are part of the environment in which they are situated. They can interact with it by performing certain actions. The environment reacts to these actions by changing its state. In [24] the following concepts and definitions concerning the environment and the agents that operate within it are presented.

Assume a finite set:

$$E = \{e, e', \dots\} \quad (1)$$

represents the set of states of the environment. A finite set of actions:

$$AC = \{\alpha, \alpha', \dots\} \quad (2)$$

represents the ability of agents to operate in the environment.

At each moment in time, the environment is in one of the states e , starting from an initial state e_0 . Based on this state, the agent performs the appropriate action. As a result, the environment changes its current state to one of many theoretically possible states. This entire process can be referred to as the agent's movement (or history) and can be written as a sequence:

$$r : e_0 \xrightarrow{\alpha_0} e_1 \xrightarrow{\alpha_1} e_2 \xrightarrow{\alpha_2} e_3 \xrightarrow{\alpha_3} \dots \xrightarrow{\alpha_{u-1}} e_u. \quad (3)$$

The above discussion assumes that the environment is deterministic — so its state depends only on the history of actions taken by agents and changes to its own states. The environment may be non-deterministic, which means that the outcome of executing actions in certain states may be uncertain.

Formally, the environment can be defined by a triple:

$$Env = \langle E, e_0, \tau \rangle, \quad (4)$$

where:

E — a finite set of environmental states,

e_0 — initial state of the environment,

τ — state transformation function.

If an agent does not refer to its experience when choosing an action, it is a purely reactive agent. It makes decisions based only on the current state of the environment. Such an agent can be defined by a function:

$$Ag : E \rightarrow AC. \quad (5)$$

In the context of this paper, agents should be considered in the context of reactive agents.

IV. GENETIC REPRESENTATION

A. Flora

A.1 Environmental map

Flora creates a map of the environment — a rectangular matrix that is also a map of the vegetation. It is navigated by agents. In the context of the taxonomy presented in this paper, the environment is accessible, deterministic, episodic, static and discrete.

A.2 Plant height

Each field on the map has a numerical value, ranging from zero to a specified maximum. A field with a value of zero is treated as barren land where nothing grows. A field with a value above zero is treated as having plants, which are food for the animals. In addition, if the plant is sufficiently mature, it will, in each cycle of the simulation, spread to one of up to 8 neighbouring fields with no plant.

The flora grows slowly at first. Its growth rate gradually increases and reaches a maximum when the plant is halfway to its maximum size. Then the rate of development drops until it reaches zero when the plant is fully grown. This is described by the function:

$$\Delta R = \frac{4\Delta R_{max}R(R_{max} - R)}{R_{max}^2}, \quad (6)$$

where:

ΔR — plant growth;

ΔR_{max} — maximum growth per cycle;

R_{max} — maximum plant size;

R — plant height.

The actual growth of the plant is in each cycle randomly between 50% and 100% of the value resulting from the function.

B. Fauna

The most important simulated living being is a representative of the fauna modelled as an agent based on a genetic algorithm, hereafter referred to in the paper as agent, individual or animal. It moves through the environment, feeds on the vegetation growing on it and reproduces. The simulated animals are hermaphrodites, so they need any partner to reproduce. This occurs instantly and results in one offspring.

B.1 Energy

The most important parameter of each agent is its energy — a characteristic based on the energy profile presented in [3]. This is a numerical value that regulates the animal's behaviour. If it is low, the agent will seek food which, when consumed, will be converted into energy. When it is high enough, the individual will be ready to reproduce. If it falls to zero the individual dies.

B.2 Chromosome

Each animal has a chromosome consisting of 40 bits. It defines six traits of an agent. The value of each trait is the decimal number decoded from the corresponding chromosome fragment plus one. Traits have a value of 1—16 or 1—256, depending on whether the gene is four-bit or eight-bit.

Eight-bit genes:

- Maximum Energy — The maximum level of energy an agent can have. The energy of an individual in the initial population or at birth is a fraction of the maximum energy;
- Life expectancy — The number of simulation cycles an animal will live before it dies of old age;
- Willingness to reproduce — An individual will only seek a reproductive partner if it achieves at least this much energy. This value is independent of the actual cost of reproduction;
- Wanted robustness of plant — An individual will only seek a reproductive partner if it achieves at least this much energy.

Four-bit genes:

- Visual range — Determines how many fields the animal can see around it, including diagonal fields. The agent can see the exact amount of food and features of other individuals within this range;
- Jaw size - Determines the maximum size of one portion of food.

In addition, each animal has a trait called fitness, with a value between 0 and 1. This value determines the closeness of the animal's genes to perfection. It is not a function of the evaluation of the individual, but only an intuitive measure of the quality of useful genes. It is calculated from the formula:

$$f = \frac{\sum_{i=1}^n \frac{G_i}{G_{imax}}}{n}, \quad (7)$$

where:

f — value of adaptation (fitness);

G_i — the value of i-this gene belonging to the adaptation genes;

G_{imax} — the maximum value of i-th gene belonging to the adaptation genes;

n — the number of genes belonging to the adaptation genes.

Selected genes make up adaptation. Individuals with a higher fitness value have an advantage over those with a lower one.

Fitting genes

Maximum energy – the higher it is, the more energy can be stored by an individual. It is also more likely that this value will be higher than the cost of reproduction and the desire to reproduce.

Life expectancy – long-lived animals exist longer in the environment, giving them more chances to seek food and reproduce.

Visual range – long range vision allows animals to move deliberately towards suitable food or partners.

Jaw size – this parameter determines the amount of food consumed during one cycle. A larger jaw size minimises the loss of valuable energy (feeding is an activity that costs energy) and saves time that the individual can spend searching for food or a partner.

Other genes

Willingness to reproduce – this gene defines a threshold value of energy for an individual to decide to reproduce. A low value will therefore encourage the individual to reproduce at a low energy level. This will accelerate population growth and gene transfer. But too low will expose the individual to life-threatening low energy levels shortly after reproduction. A high value will allow individuals to store more energy, giving them a safe energy reserve just after reproduction.

Wanted robustness of plant – if an animal seeks out a plant of low maturity, it will widen the range of food available to it, giving it an advantage in the population. This naturally increases the animal’s energy, allowing for rapid reproduction and therefore a rapid increase in population size. On the other hand, a high value will allow vegetation to grow faster on the map and increase the chance of spreading in barren fields.

B.3 Algorithm of agent behaviour

Individuals act towards a simple, predetermined algorithm. An agent’s behaviour is governed by its current energy value, its desire to reproduce, and the food it seeks. Depending on these values, the individual performs a different action.

Moving – the animals move around the map in eight directions. The movement takes each of them a whole cycle and they then cover a distance of exactly one field. Moving costs energy.

An agent moves in a particular direction if there is an object it is aiming at in its line of sight. This object is food, when the individual wants to eat, or a potential partner, when the animal wants to reproduce. The agent always moves towards the nearest object that meets its requirements. If there are more such objects, it will move towards a randomly selected one. If there is no object in sight, the agent will move randomly.

Reproduction – when an animal wishes to produce offspring, i.e., enters a state of reproduction, it checks the adaptation value of other individuals in sight. It sees only those agents that are also in the state of reproduction and whose adaptation value is sufficiently high. It must be at least as high as the agent’s own adaptation minus some tolerance. In this way individuals will not interbreed with others who have received an unfavourable mutation.

The animal will move towards the nearest suitable partner to be in the same field as it. In case there are many partners available in the field, one will be chosen randomly. In the final step, the animal must be accepted — it must have a high enough adaptation value from the partner’s point of view. Otherwise, the suitor will be ignored and will have to wait out the simulation cycle without doing anything. If the courtship is successful, the parents pay the cost of reproduction in

energy. They also both go into an idle state to prevent them from reproducing again in the same simulation cycle. Exactly one offspring is born, in the same field as the parents. It is fully independent from the moment of birth.

Feeding – if the individual does not have enough energy to reproduce, it will search for food. It heads towards a field where there is sufficiently tall vegetation. The agent pays the cost of feeding itself. An agent during one cycle can eat as much as the size of its jaw. An agent eats less food if its energy requirement is lower, or if the plant is too low. The value of the map field — and thus the plant growing on it — will be reduced by the value eaten by the animal. If it is reduced to zero, the field becomes barren.

V. AGENT MODEL

As stated by D. Grzonka in [9], the constituent elements of agent systems can be: types of agents, locations of agents, strategies implementing coherent goals, executable actions, states of the environment and operators allowing to perceive and interact with the environment. Based on the definition proposed by the author, a definition of a multi-agent system inscribed in the problem under consideration can be presented:

$$MAS = \{AG, ID, TP, LOC, K, ES, ACT, ST, GL, GEN\}, \quad (8)$$

where:

AG — a set of agents belonging to a multi-agent system;

ID — a set of unique identifiers for agents;

TP — a collection of all agent types;

LOC — a set of locations where agents may be present;

K — a set of all possible states of knowledge of the agent;

ES — a set of all possible states of the environment;

ACT — a set of actions that can be performed by agents within the environment;

ST — a set of strategies implemented by agents;

GL — a set of agent objectives;

GEN — a set of all possible combinations of features recorded in chromosome form.

An agent (*ag*) is defined as follows:

$$AG \ni ag = \{id, tp, st, k, l, gn, gl, en, ae, \gamma, \beta, \delta, \alpha, \epsilon\}, \quad (9)$$

where:

id \in *ID* — a unique system-wide identifier for the agent;

tp \in *TP* — agent type;

st \in *ST* — agent’s strategy;

k \in *K* — current (temporary) knowledge of the agent;

l \in *LOC* — current location of the agent (position on the map);

gn \in *GEN* — a set of agent characteristics recorded in chromosome form;

gl \in *GL* — agent’s current objective;

en \in \mathbb{Z} — agent’s current vital energy;

ae \in \mathbb{N} — current age of the agent;

γ — observation (perception) function to monitor the state of the environment;
 β — strategy selection function;
 δ — a decision-making function which, on the basis of the strategy pursued, selects actions;
 α — an action function that, based on a selected action, executes it on the environment changing its state;
 ϵ — agent adaptation function.

A detailed definition of the model components can be found in [9].

The paper proposes one type of agents: $TP = \{Agfauna\}$, and 2 strategies. Strategies define how the fauna-agent functions in the environment and how the goals are achieved. Strategies can be written using a set:

$$ST = \{st_1, st_2\}, \quad (10)$$

where:

st_1 — prospecting for food;
 st_2 — seeking a reproductive partner.

Strategies are selected depending on the agent’s objectives:

$$GL = \{gl_1, gl_2\}, \quad (11)$$

where:

gl_1 — increase energy levels;
 gl_2 — reproduction.

The most essential element of agents, which defines their causal capabilities, are actions. In the model under consideration this will be a set consisting of 8 actions:

$$ACT = \{choose_{fauna}, choose_{flora}, move_{towards}, move_{random}, eat, reproduce, idle, die\}, \quad (12)$$

where:

$choose_{fauna}$ — selecting a location within the agent’s line of sight where there is a suitable reproducing partner;
 $choose_{flora}$ — selecting a location within the agent’s perception range where the flora corresponding to the agent is present;
 $move_{towards}$ — movement of an agent towards a location selected by one of the actions $choose$;
 $move_{random}$ — movement of the agent in a random direction in case the action $choose$ does not find a suitable location;
 eat — eating flora to recover energy;
 $reproduce$ — creation of a descendant;
 $idle$ — idle action;
 die — killing the agent.

Within a single simulation cycle, within a selected strategy, an agent may perform one or more of the actions listed above.

VI. EXPERIMENTS

This experiment focuses on the development of the agents’ features during the course of the simulation.

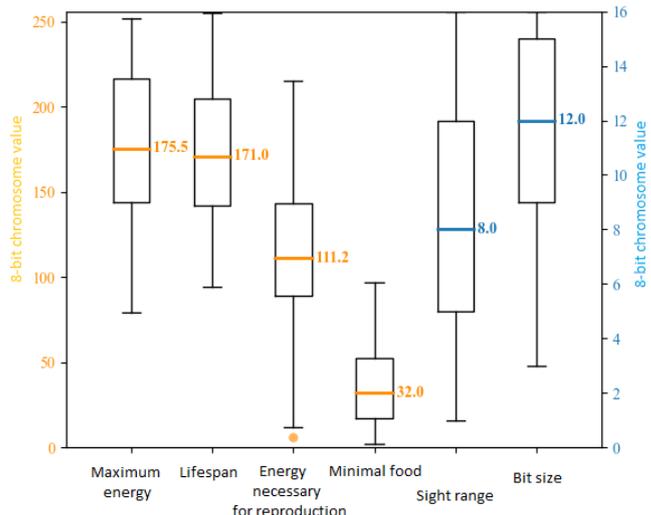


Fig. 1: Comparison of gene values at local population maxima of multiple simulation calls

Parameter	Value
Map dimensions	10x10
Number of agents in initial population	30
Maximum number of interactions	10000
Reproduction cost	50
Cost of inactivity	1
Cost of movement	3
Cost of feeding	1
Adaptation tolerance rate of partner	0.15
Agent’s initial energy at birth	30% of max
Number of chromosome crossover points	1
Mutation chance	0,1%
Maximum flora height	100
Maximum flora growth per cycle	5
Height of florets allowed to reproduce	70% of max
Size of new rhizome after flora reproduction	1
Initial map filling with flora	50%

TABLE 1: Simulation parameters

Every fixed number of cycles, numerical values of faunal traits are recorded, from which box plots will be created. All simulations were repeated 100 times. At the end of each simulation the number of cycles it took is read out. After each simulation has been called enough times, a box plot is produced indicating how long the population persists on a given set of parameters.

For each simulation called in the comparison study, the median value of each gene from the entire simulation run is recorded. After multiple simulations, a box plot of typical gene values for a given parameter set is produced.

Figure 1 presents the distribution of typical gene values over multiple simulations on the example parameter set shown below (see: tab. 1).

At the initial stage, the animals have a wide variety of gene values, as can be inferred from the zero-cycle

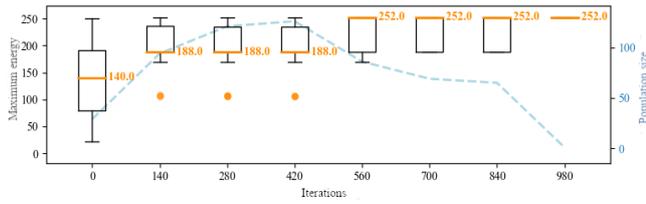


Fig. 2: Graph of faunal maximum energy gene values

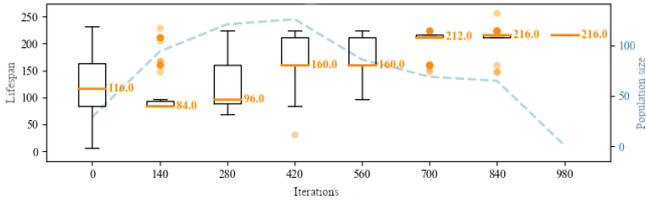


Fig. 3: Graph of fauna life expectancy gene values

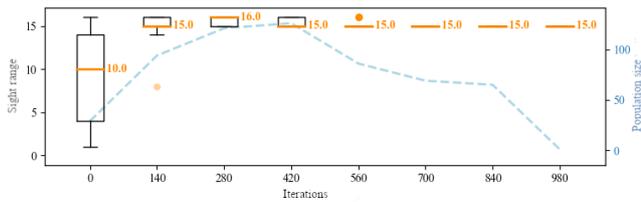


Fig. 4: Graph of fauna sight gene values

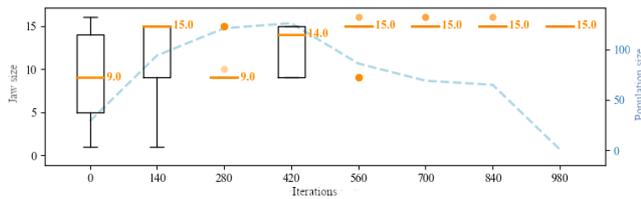


Fig. 5: Graph of fauna jaw size gene values

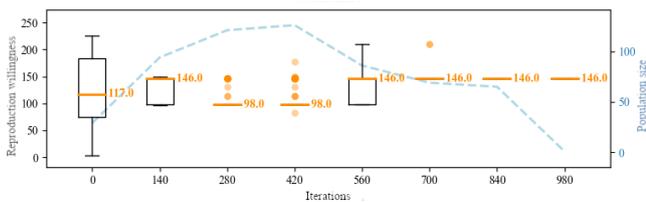


Fig. 6: Graph of faunal reproduction willingness gene values

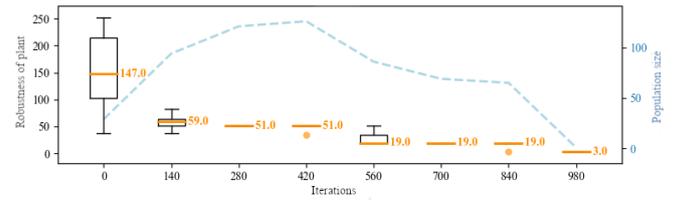


Fig. 7: Graph of the gene value for the fauna's wanted robustness of plant

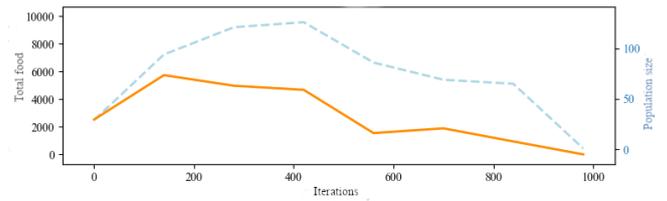


Fig. 8: Graph of total food intake

graphs in Figures 2 – 7. The ranges of the graphs tend to shorten as the faunal population increases. Animal fitness, which can be observed in Figure 9, increases with successive cycles. Agents become more and more perfect, fitness gene values increase until the population dies out.

Due to the large spread of traits, the initial population consists of individuals able to thrive in the environment and those whose gene set will prevent them from surviving. The latter group dies out in the first few dozen cycles of the simulation. The former feeds and begins to reproduce. After some time, the population begins to consist only of individuals that are able to live and reproduce. There is a significant increase in the population. Soon, however, there begins to be a shortage of food. As can be seen in Figure 8, population size is strongly correlated with available food.

Genes that are scored as significant for fitness tend to rapidly optimise upwards. An exception is the eye range gene, whose graph is shown in Figure 4. For the 10 by 10 field map, a visual range of more than five (which allows an individual to observe the entire map from the centre) was not significant for survival.

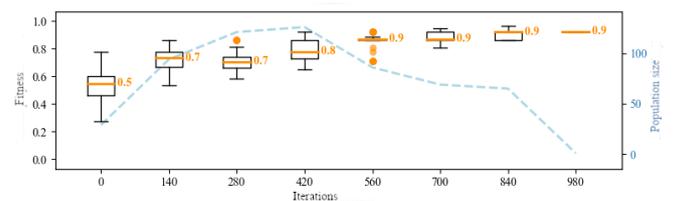


Fig. 9: Graph of faunal fitness values

VII. SUMMARY AND FUTURE WORK

In this research, a simulator of a simple natural environment inscribed in the agent paradigm has been implemented. The operation of the simulated environment was described and a formal multi-agent model of the problem was proposed. A number of experiments have been conducted, which have allowed several hypotheses to be made. The experiments made it possible to find the reason for the extinction of the simulated fauna. It was, in a way, a side effect of animal adaptation. Individuals eating lower and lower plants and enlarging their jaw gained an advantage in the environment. This combination ultimately depleted the environment of resources every time.

The simulations carried out have shown that excessive model parameterisation can adversely affect the results obtained. The developed simulator does not exhaust the whole issue of modelling natural environments. As part of future in-depth research, it would be worth extending the model to include other classical factors presented in literature and widely consolidated, such as the sex of individuals, the existence of predators, a more complex plant model (e.g. edible fruit), extended needs of individuals (e.g. sleep) or diverse terrain topography with different movement costs. It would also be useful to describe the simulator itself in more detail in future work.

REFERENCES

- [1] R. Atkinson and C. Hewitt. Synchronization in actor systems. In *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, POPL '77, page 267–280, New York, NY, USA, 1977. Association for Computing Machinery.
- [2] S. Calderoni and P. Marcenac. Mutant: a multiagent toolkit for artificial life simulation. In *Proceedings. Technology of Object-Oriented Languages. TOOLS 26 (Cat. No.98EX176)*, pages 218–229, 1998.
- [3] K. A. Cetnarowicz. *Problemy projektowania i realizacji systemów wieloagentowych*. AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, 1999.
- [4] K. A. Cetnarowicz. *Paradygmat agentowy w Informatyce: Koncepcje, podstawy i zastosowania*. Akademicka Oficyna Wydawnicza EXIT, 2012.
- [5] J. Csonto, J. Kadukova, and M. Polak. Artificial life simulation of living alga cells and its sorption mechanisms. *J. Med. Syst.*, 25(3):221–231, jun 2001.
- [6] Y. Demazeau and J.-P. Müller. Decentralized artificial intelligence. In Y. Demazeau and J.-P. Müller, editors, *Decentralized A.I. : Proc. of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Cambridge, England*, pages 3–13. North-Holland, 1990.
- [7] A. Dorin, K. B. Korb, and V. Grimm. Artificial-life ecosystems - what are they and what could they become? In *ALIFE*, 2008.
- [8] M. Gardner. Mathematical games. *Scientific American*, 223(4):120–123, 1970.
- [9] D. Grzonka. Inteligentne systemy monitoringu procesów harmonogramowania w rozproszonych środowiskach dużej skali w ujęciu wieloagentowym. *Praca doktorska*, 2018.
- [10] C. Hewitt, P. Bishop, and R. Steiger. A universal modular actor formalism for artificial intelligence. IJCAI3. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, pages 235–245, 1973.
- [11] N. R. Jennings, K. Sycara, and M. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, Jan. 1998.
- [12] K.-J. Kim and S.-B. Cho. A comprehensive overview of the applications of artificial life. *Artificial Life*, 12(1):153–182, 2006.

- [13] M. Kisiel-Dorohinicki. *Agentowe architektury populacyjnych systemów inteligencji obliczeniowej*, volume 269 of *Rozprawy. Monografie*. AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, 2013.
- [14] M. H. T. Ling. An artificial life simulation library based on genetic algorithm, 3-character genetic code and biological hierarchy. 2012.
- [15] C. M. Macal. *Agent-Based Modeling and Artificial Life*, pages 1–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
- [16] Z. Michalewicz. *Algorytmy genetyczne + struktury danych = programy ewolucyjne*. Wydawnictwa Naukowo Techniczne, 1999.
- [17] V. M. Petrović. Artificial intelligence and virtual worlds – toward human-level ai agents. *IEEE Access*, 6:39976–39988, 2018.
- [18] J. Rocha, I. Boavida-Portugal, and E. Gomes. Introductory chapter: Multi-agent systems. In *Multi-Agent Systems*. IntechOpen, 2017.
- [19] S. J. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1994.
- [20] Y. Shoham. Agent oriented programming. technical report stan-cs-90-1335. *Computer Science Department, Stanford University*, 1990.
- [21] S. Swarup and H. S. Mortveit. Live simulations. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, page 1721–1725, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems.
- [22] G. Weiss, editor. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. The MIT Press, 1999.
- [23] G. Weiss. *Multiagent Systems*. The MIT Press, 2013.
- [24] M. Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.
- [25] M. Wooldridge and P. E. Dunne. *The Computational Complexity of Agent Verification*, pages 115–127. Springer Berlin Heidelberg, 2002.
- [26] M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

AUTHOR BIOGRAPHIES

ADRIAN SOŚNICKI received his B.Eng. and M.Sc. in the field of Computer Science at Cracow University of Technology, Poland, in 2016 and 2020, respectively. Currently, he is Senior IT Business Analyst / Quality Assurance in Velis Real Estate Tech, Cracow. His main areas of interest include software engineering (programming, databases), project management and business analytics. If you want to contact him, please send an e-mail to adrian.sosnicki.skrzynka@gmail.com



DANIEL GRZONKA received his B.Eng. and M.Sc. degrees from Cracow University of Technology, Poland, in 2012 and 2013, respectively. In 2019 he received his Ph.D. degree from the Polish Academy of Sciences (in cooperation with Jagiellonian University). All degrees (with distinctions) are in Computer Science. Currently, he is an Assistant Professor at the Cracow University of Technology and Vice-Dean for Education at the Faculty of Computer Science and Telecommunications. He is



also the laureate of the prestigious START 2019 competition of the Foundation for Polish Science for the most outstanding young scientists. The main topics of his research are monitoring systems, grid and cloud computing, multi-agent systems, task scheduling problems, data mining and high-performance computing. For more information, please visit: www.grzonka.eu

ŁUKASZ GAŻA received his M.Sc. in the field of Applied Physics with Computer Modelling at the Tadeusz Kosciuszko Cracow University of Technology. Since 2019 he is a Research and Teaching Assistant at the Tadeusz Kosciuszko Cracow University of Technology. His e-mail address is lukasz.gaza@pk.edu.pl



TIME SERIES CLUSTERING WITH DIFFERENT DISTANCE MEASURES TO TELL WEB BOTS AND HUMANS APART

Grażyna Suchacka
Institute of Informatics
University of Opole
ul. Oleska 48
45-052 Opole, Poland
E-mail: gsuchacka@uni.opole.pl

KEYWORDS

Internet robot, Web bot, Web bot detection, Web session, Time series, Unsupervised classification, Clustering, Distance measure, Similarity measure

ABSTRACT

The paper deals with the problem of differentiating Web sessions of bots and human users by observing some characteristics of their traffic at the Web server input. We propose an approach to cluster bots' and humans' sessions represented as time series. First, sessions are expressed as sequences of HTTP requests coming to the server at specific timestamps; then, they are pre-processed to form time series of limited length. Time series are clustered and the clustering performance is evaluated in terms of the ability to partition bots and humans into separate clusters. The proposed approach is applied to real server log data and validated with the use of different time series distance measures and clustering algorithms. Results show that the choice of a distance measure and a clustering method significantly affects clustering efficiency. The best results for the considered scenario were achieved for distance measures based on nonparametric spectral estimators and the Euclidean distance with a complexity correction factor.

INTRODUCTION

The share of automatically generated traffic in total Web traffic has been constantly growing for many years (Bad Bot Report 2021). To cope with the presence of artificial Web agents (robots, bots) and with possible negative consequences of their activities, many studies on bot detection have been conducted. In particular, approaches based on machine learning (ML) methods have proven to be effective in distinguishing between bots and humans (i.e., human-operated Web browsers).

The majority of approaches to detect bots on Web servers have involved LM methods in the offline scenario, i.e. for completed user visits (*Web sessions*). Primarily, supervised learning has been used, e.g., decision trees, support vector machines, neural networks, ensemble methods (Iliou et al. 2019; Lagopoulos and Tsoumakas 2020; Lysenko et al. 2020; Rahman and Tomar 2021; Rovetta et al. 2017; Ustebay et al. 2019). Unsupervised learning has also been investigated (Alam et al. 2014; Suchacka and Iwański 2020; Rovetta et al. 2020; Zabihi et al. 2014). Although offline methods allows one to learn

about features of bot traffic and lay the groundwork for novel online methods, this kind of approach is unable to recognize active bots. Here the online bot detection comes into play, by observing incoming requests within active sessions and trying to infer a user type as early as possible (Doran and Gokhale 2016; Suchacka et al. 2021). Regarding time series analysis, some previous works applied supervised classification techniques for game bots (Bernardi et al. 2017), network bots (Bonneton et al. 2015), and Web bots (Chen and Feng 2013).

Research questions under consideration in this paper are stated as follows. Is it possible to separate bots from humans on a Web server by analyzing only an initial part of a stream of Web client's requests for some (possibly short) period of session duration? What similarity measures are most adequate for time series clustering in the considered scenario?

To deal with the aforementioned issues, we represent each Web session on a server as a time series whose consecutive elements correspond to numbers of requests received from a given client in subsequent one-second intervals. We consider only beginnings of the series within a certain period of session duration and apply time series clustering to distinguish between two classes of clients: 1 (bots) and 0 (humans). Two well-known clustering techniques are implemented: hierarchical clustering and partition-based clustering, each with 23 different similarity measures. This approach is applied to real data obtained from Web server access logs.

The remainder of the paper is organized as follows. The next section presents preliminaries on time series clustering and similarity measures. Then the research methodology is presented, followed by a discussion of experimental results. The last section concludes the paper and outlines prospective directions of future work.

PRELIMINARIES

Time Series Clustering

Let us consider a dataset of n time series data $D = \{F_1, F_2, \dots, F_n\}$, where F_l is a sequence of values measured in time, $l = 1, 2, \dots, n$. *Time series clustering* is the process of unsupervised partitioning of D into a set of clusters $C = \{C_1, C_2, \dots, C_k\}$ in such a way that similar time series are grouped together based on a certain similarity measure, $D = \bigcup_{i=1}^k C_i$ and the clusters are disjoint: $C_i \cap C_j = \emptyset$ for $i \neq j$, $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k$ (Aghabozorgi et al. 2015).

We consider univariate time series, which are sequences of real numbers collected regularly in time, where each number represents a value – the number of requests received from a given Web client in one-second interval. Furthermore, all pre-processed time series subject to clustering have exactly the same length.

Time Series Clustering Methods

Methods for clustering time series may be broadly classified into six groups: hierarchical, partitioning, model-based, grid-based, density-based, and multi-step clustering (Aghabozorgi et al. 2015; Kotsakos et al. 2018). We apply two algorithms for the most popular clustering types: hierarchical and partitioning clustering.

Time Series Distance Measures

Time series clustering is not easy due to such common characteristics of time series as the presence of noise, outliers, and shifts in data. A key issue is finding similar time series is applying the appropriate way of calculating the similarity of data sequences, i.e., an adoption of an adequate similarity measure (distance measure). The concept of similarity in the context of time series is complex due to the dynamic character and high dimensionality of time dependent data.

In practice, calculation of the distance between time series is often approximated with the use of various methods. There is a wide choice of similarity measures for time series clustering (Aghabozorgi et al. 2015; Montero and Vilar 2015). The most common are distance measures of three types.

- *Model-free measures.* These basic measures are suitable especially for time series with the equal length. They often operate by comparing original time series or sequences of serial features extracted from them, like correlations, autocorrelations, spectral features, or wavelet coefficients.
- *Model-based measures.* The idea of these methods consists in fitting an underlying model to each time series and calculating the similarity between the fitted models. The most commonly used models are ARIMA and ARMA.
- *Complexity-based measures.* Here the concept is to compare levels of complexity of time series by measuring the level of shared information by the compared series. The mutual information is approximated, usually using the notion of algorithmic entropy or Kolmogorov complexity.

We consider 23 different distance measures from all the three groups.

RESEARCH METHODOLOGY

This section discusses the proposed methodology to Web session time series clustering which involves the following steps:

- reading and pre-processing data from Web server logs to reconstruct user sessions and determine session features;
- assigning ground truth labels to sessions;

- transforming sessions to time series of a specified length;
- performing time series clustering with the use of various similarity measures;
- validating and comparing clustering results.

Building Time Series from Server Log Data

Extracting Sessions and Session Features

Raw data used in the study are entries recorded in a standard Web server access log of an e-commerce Web server. Each text line in a log file corresponds to a single HTTP request and contains information on the Web client's IP address, user-agent string, identifier of the requested server resource, timestamp, and other.

After pre-processing of request data Web sessions may be reconstructed. A *Web session* is defined as a sequence of HTTP requests (with more than one request) received from a Web client during a single visit to a website hosted by the server. A Web client is identified with an IP address and a user-agent string. An additional assumption is a minimum time gap between any consecutive sessions of a given client, equal to 30 minutes.

From individual fields of requests making up a session some session features may be determined, like the session duration (time interval in seconds), the session length (the total number of requests in session), the mean inter-request time, etc. Session features are useful for building a session representation for data mining approaches, as well as to identify some part of Web clients as robots based on heuristic rules.

Assigning Ground Truth Labels

The next step is session labeling with ground-truth labels. Clustering of observations from a given dataset is a task of unsupervised learning so the information about class labels is not used while generating clusters. However, this information is useful for other purposes, e.g., to perform the exploratory data analysis, to create an appropriate composition of an ultimate dataset used in experiments, or to evaluate the clustering results with the use of external indexes. Thus, the process of labeling Web sessions as bots (class *1*) or humans (class *0*) should be carried out with the greatest care and with the application of as many criteria as possible.

Our labeling procedure was described in (Suchacka and Motyka 2018) in detail. It is based on two databases of IP addresses and user-agents known to correspond to different kinds of Web browsers and bots: *Udger* database (Udger 2021) and *User agents* database (User-agents 2014). A big advantage of using *Udger* is the fact that most sessions labeled in this way have not only the class assigned (*0* or *1*) but also the Web client category, name, and version. Sessions performed by clients identified by *Udger* as bots without a client category or name were assigned to category “*Uncategorized Udger bot*”.

Some of sessions that could not be labeled with *Udger* data were flagged as bots by using additional criteria: by performing a syntactic analysis of user-agents for bot-

related keywords (category “*Unknown bot – keyword*”), observing a request for *robots.txt* file in session, or applying heuristic rules for session features. These rules included: zero image-to-page ratio, all requests with empty referrers, all responses erroneous (with 4xx HTTP codes), and all requests of type HEAD. These sessions were assigned to category “*Unknown bot – heuristics*”. Session reconstruction and labeling was accomplished using our log analyzer implemented in C#.

Representing Sessions as Trimmed Time Series

Based on request timestamps sessions were transformed into time series, each element of which denotes how many requests arrived in a given second of the session duration. The next step was to decide what sequence length should be chosen for the time series analysis.

Web sessions naturally differ in terms of duration. Some bots, for instance, tend to request only one or two requests within an extended period of time whereas other may perform extremely long-lasting sessions. On the other hand, human users may access only one page of the website if they are not interested in the contents or may spend up to several dozen minutes browsing the online store offer and selecting items they are willing to buy.

Having in mind the necessity of recognizing bot sessions as soon as possible, we aimed at considering relatively short session duration. Moreover, the shorter the sequence is, the more sessions from the original dataset are left in the ultimate dataset. Finally, the time series length equal to 100 seconds was chosen. All the time series corresponding to sessions lasting at least 100 seconds were trimmed to 100 elements whereas all shorter time series were excluded from the analysis. We excluded shorter sequences in this study because our goal was to apply many time series similarity measures, some of which are limited to time series with equal length.

A program to analyze and transform the session dataset into the time series dataset was implemented in R.

Distance Measures Used

Time series clustering was implemented in R with the use of package *TSclust* (Montero and Vilar 2020). We applied 23 similarity measures available in the package, which include a wide range of approaches to calculate the distance between time series. The measures and their shortcut names used hereafter are as follows.

1. *Model-free distance measures*

1.1. Simple measures:

- EUCL – Euclidean distance,
- FRECHET – Fréchet distance,
- DTW – Dynamic Time Warping distance,
- CORT – distance based on the first order temporal correlation coefficient.

1.2. Measures based on correlations and autocorrelations:

- COR1 – correlation-based distance,
- COR2 – correlation-based distance with the parameter allowing regulation of the fast distance decreasing,
- ACF – autocorrelation-based distance,

- PACF – distance using the partial autocorrelation function.

1.3. Measures based on periodograms:

- PER – Euclidean distance between the periodogram ordinates,
- PER_NP – Euclidean distance between the normalized periodogram ordinates,
- PER_LNP – Euclidean distance based on the logarithm of the normalized periodogram,
- PER_INT – distance based on the integrated periodograms (cumulative versions of the periodograms).

1.4. Measures based on nonparametric spectral estimators:

- SPEC_LLRLS – distance with the spectra replaced by the exponential transformation of local linear smoothers of the log-periodograms, obtained via least squares,
- SPEC_LLRLK – distance with the spectra estimated by the exponential transformation of local linear smoothers of the log-periodograms, obtained by using the maximum local likelihood criterion,
- SPEC_GLK – distance using the local maximum log-likelihood estimator computed by local linear fitting,
- SPEC_ISD – distance evaluating the integrated squared differences between nonparametric estimators of the log-spectra using local linear smoothers of the log-periodograms, obtained by using the maximum local likelihood criterion.

2. *Model-based distance measures*

- AR_PIC – Piccolo distance – Euclidean distance between autoregressive approximations of ARIMA structures,
- AR_MAH – Maharaj distance, based on autoregressive approximations of ARMA structures,
- AR_LPC_CEPS – cepstral-based distance using linear predictive coding (LPC) cepstrum for ARIMA time series.

3. *Complexity-based distance measures*

- CID – Euclidean distance with a complexity correction factor,
- PDC – distance based on divergence between permutation distributions of order patterns in m-embedding of the original series,
- CDM – compression-based distance measure,
- NCD – a simplified version of the compression-based distance measure.

Clustering Algorithms Used

Time series representing Web sessions were clustered with the use of two well-known algorithms.

The first algorithm is agglomerative hierarchical clustering with complete linkage cluster selection. The algorithm generates a hierarchy of clusters starting from each time series being a separate cluster. Then, it gradually merges the most similar pairs of clusters until

the desired number of clusters is obtained. A complete linkage cluster selection means that to decide which two clusters are closest to each other in each step, the distance between any two clusters is defined as the longest distance among all their member time series.

The second algorithm is Partitioning Around Medoids (PAM), also known as k-Medoids. The algorithm partitions all the time series into k groups so that each group contains at least one series. Each cluster has a medoid prototype, which is the most centrally located object in the cluster (i.e., the time series whose average distance to all other series in the cluster is minimum).

The hierarchical clustering was conducted with the use of *hclust* function available in package “stats” and the partitioning clustering was done with *pam* function from package “cluster”. In both cases the target number of clusters was two.

Clustering and Performance Evaluation

Bootstrap Sampling

In reality, Web sessions observed in a given time window are not balanced in terms of the client classes and categories which may negatively affect the performance of machine learning algorithms. Furthermore, it is necessary to perform clustering in reasonable time whereas some algorithms for computing time series similarity matrix have high computational complexity. To prevent a possible bias of clustering results and provide reasonable computation time, the ultimate dataset of time series used in experiments was created via the bootstrap sampling. This under-sampling method consists in drawing sample observations repeatedly with replacement from the source dataset to be used in a single experiment run. The ultimate dataset was created by drawing proportionate numbers of time series of different client categories (see subsection “Bootstrap Datasets”). The time series clustering with the use of 23 distance measures and two clustering algorithms was repeated ten times for different bootstrap subsets. The final results are averaged performance scores of all experiment runs.

Clustering Performance Measure

To evaluate the clustering quality we applied an external index $Sim(C, C')$ (Gavrilov et al. 2000; Liao 2005; Montero and Vilar 2020). Given the ground-truth cluster partition $C = \{C_1, \dots, C_k\}$ and the experimental partition $C' = \{C'_1, \dots, C'_k\}$, the similarity index expresses the agreement between these two solutions:

$$Sim(C, C') = \frac{1}{k} \sum_{i=1}^k \max_{1 \leq j \leq k} Sim(C_i, C'_j),$$

where

$$Sim(C_i, C'_j) = \frac{2|C_i \cap C'_j|}{|C_i| + |C'_j|},$$

where $|\cdot|$ denotes the cardinality of the elements in the set. The higher score is achieved, the better clustering results are. The score of 1 means that the two clusterings are the same and the value 0 means they are completely dissimilar.

RESULTS AND DISCUSSION

Data Description

Data used in the experiment were obtained from 20-hour log file for an e-commerce website, recorded in November 2019. As a result of data pre-processing, session reconstruction, and excluding single requests and admin sessions, there were 2,218 sessions in total, including 1,371 humans and 847 (38.19%) bots. The session length varied from 2 to 5,270 requests. The longest session lasted for 84,849 seconds, i.e., almost 24 hours (it was clearly a bot session).

Fig. 1 and Fig. 2 plot time series of humans and bots, respectively, taking into consideration the first five minutes of session. Traffic patterns of both groups clearly differ. Since humans navigate through the website via Web browsers, their traffic patterns reflect the way of downloading server resources by browsers: one can observe clear peaks after requesting subsequent pages with the corresponding embedded objects, separated by the users’ think time, required for browsing and analyzing the page contents (Fig. 1). On the other hand, artificial agents parse the website according to the implemented algorithms – their requests are less frequent and more regular in nature (Fig. 2).

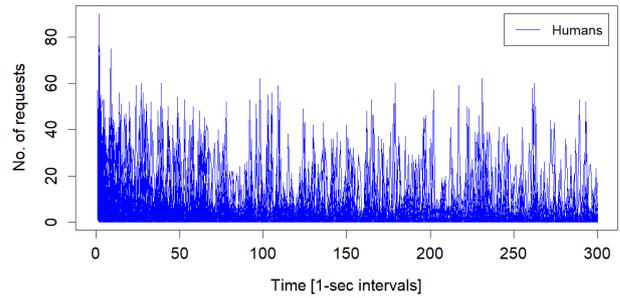


Figure 1: Visualization of Human Users’ Traffic During the First 300 Seconds of Sessions’ Duration

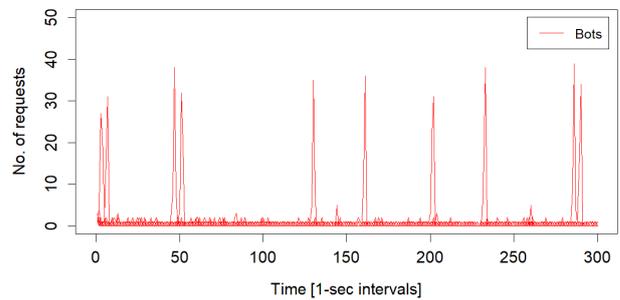


Figure 2: Visualization of Bots’ Traffic During the First 300 Seconds of Sessions’ Duration

Bootstrap Datasets

After transforming sessions to 100-second time series (and eliminating shorter sessions), the ultimate dataset to be analyzed contained 980 time series: 543 class 0 series

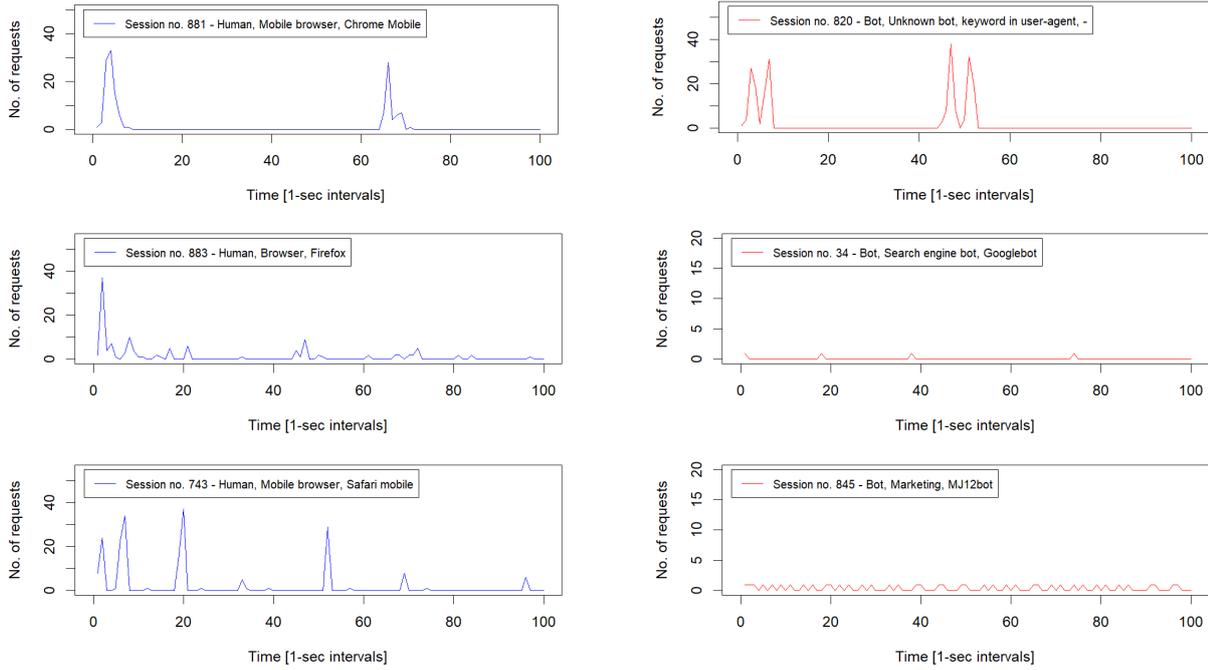


Figure 3: Visualization of Example Time Series of Class 0 (left) and of Class 1 (right)

(human sessions) and 437 class 1 series (bot sessions). Table 1 presents how many time series were from individual classes and client categories. Numbers of sessions completed via mobile and desktop browsers were roughly equal (276 and 267 time series, respectively). Regarding robot sessions, the most numerous were unknown bots identified based on heuristic rules (165 series), marketing bots (133 series), and search engine bots (120 series).

Table 1: Distribution of 100-second Time Series According to Classes and Categories

Client class	Client category	No. of sessions
0	Browser	267
0	Mobile browser	276
1	Marketing	133
1	Search engine bot	120
1	Uncategorized Udger bot	18
1	Unknown bot – heuristics	165
1	Unknown bot – keyword	1

Fig. 3 visualizes several samples of time series corresponding to sessions from both classes. It can be seen in Fig. 3-left that sessions generated by human-operated Web browsers reveal similar patterns regardless of a browser name: spikes in the number of requests correspond to successive user clicks (page views involve requests for page description files and embedded objects, like images, pdf documents, etc.). Plots in Fig. 3-right show that the Web traffic generated by bots may be more differentiated depending on a specific crawling software. Sessions no. 34 and 845 show a very different traffic characteristics than the human sessions (no. 881, 883, 743) – these are benign bots which reveal their identities

(Googlebot, MJ12bot) in user-agent fields. In contrast, session no. 820 is a bot emulating human behavior.

Each bootstrap dataset had the following composition regarding time series from different client categories:

- Browser: 20,
- Mobile browser: 20,
- Marketing: 10,
- Search engine bot: 10,
- Uncategorized Udger bot: 10,
- Unknown bot – heuristics: 10,
- Unknown bot – keyword: 1.

Clustering Results

Fig. 4 shows clustering performance scores obtained for various distance measures and clustering algorithms as a degree of agreement between the ground truth-based partition of time series and the experimental solutions. One can observe that the clustering quality highly depends both on a distance measure and a clustering method applied. In general, much better results were achieved with the partition-based method than the hierarchical one – results were higher for nearly all similarity measures applied.

Regarding the similarity measures, the lowest scores were achieved for the periodogram-based measures (PER, PER_NP, PER_LNP, PER_INT) and for the model-based measures (AR_MAH, AR_LPC_CEPS, AR_PIC). Scores of the simple measures (EUCL, FRÉCHET, DTW, CORT), as well as of the measures based on correlations and autocorrelations (COR1, COR2, ACF, PACF) are pretty similar (relatively low for the hierarchical clustering algorithm and moderate for PAM).

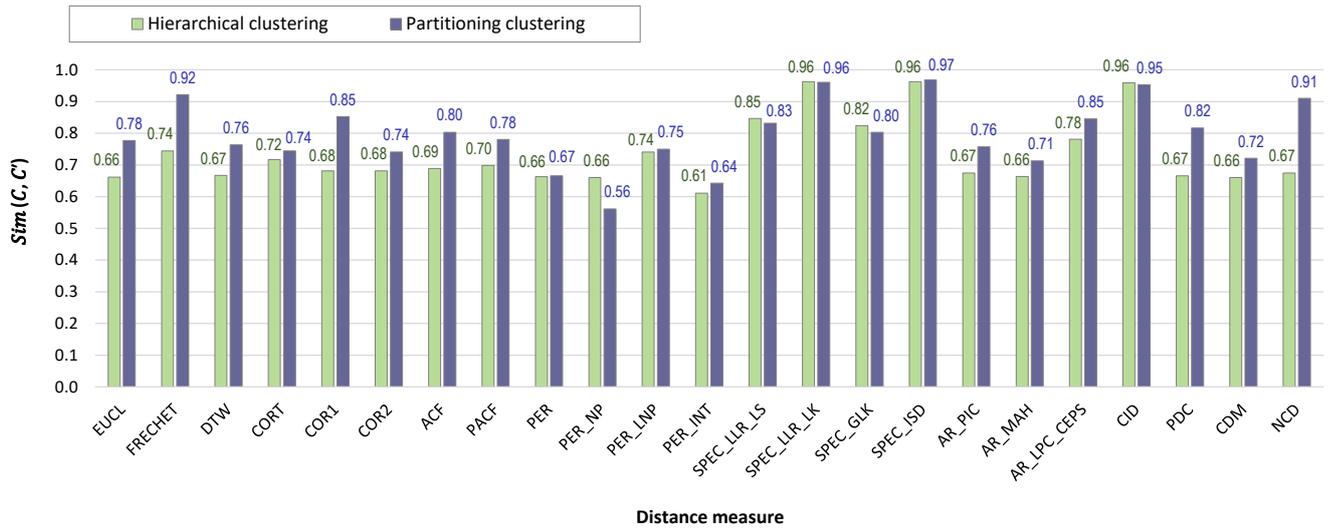


Figure 4: Clustering Performance Depending on a Distance Measure and a Clustering Method

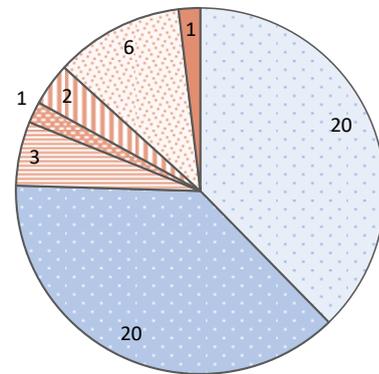
Among these solutions only the result achieved for PAM with the Fréchet distance (FRECHET) stands out positively – it should be emphasized, however, that calculating the time series similarity matrix with the Fréchet method takes the longest time, which may disqualify this method for use in real-time applications, such as the online bot detection. The efficiency of the complexity-based distance measures is difficult to generalize – in this type of approaches CID measure clearly stands out positively for both clustering algorithms.

A very effective approach to assessing the time series similarity in the scenario under consideration turned out to be the use of nonparametric spectral estimators (SPEC_LLRS_LS, SPEC_LLRS_LK, SPEC_GLK, SPEC_ISD measures). These methods, along with the Euclidean distance with a complexity correction factor (CID), achieved the highest average efficiency as regards the *Sim* index.

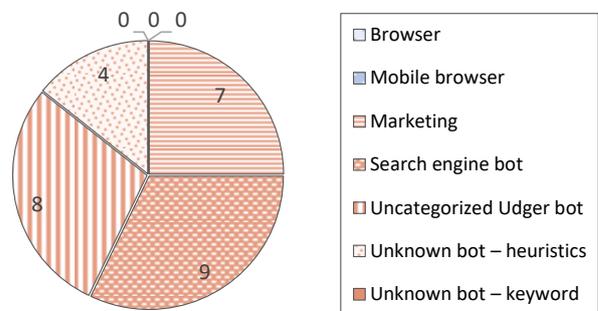
To sum up, the best clustering results of all the distance measures in the scenario under consideration were clearly achieved by three measures: SPEC_LLRS_LK and SPEC_ISD, based on nonparametric spectral estimators, and CID, the Euclidean distance with a complexity correction factor. These measures not only provided the highest efficiency of clustering time series representing bot and human sessions (performance scores of at least 0.95) but they were also insensitive to a clustering method (hierarchical or partitioning).

Let us investigate results for the best case in more detail, i.e., for SPEC_ISD distance measure and the partitioning clustering. Fig. 5 visualizes the composition of generated clusters in terms of the number of sessions from different client categories (and thereby, from different classes; class 0 is marked in shades of red, class 1 – in shades of blue). Cluster 1 is much larger – it contains 53 time series, 40 of which are of class 0 (20 Browsers and 20 Mobile browsers) and only 13 series of class 1 (from all five bot categories). Cluster 2 contains only 28 time series but all of them belong to class 1.

One can see that in this case of time series clustering it was possible to separate bots from humans up to a certain degree but not completely. This confirms that robots' online behaviors are highly differentiated, as opposed to navigational patterns demonstrated by human users, which were all gathered in one cluster. This observation clearly demonstrates that a larger number of clusters should be considered.



(a) Cluster 1 (Humans Mostly)



(b) Cluster 2 (Bots Only)

Figure 5: Composition of Clusters for the Best Case (SPEC_ISD Distance Measure, Partitioning Clustering)

CONCLUSIONS

The study discussed in this paper showed that the proposed way of clustering Web sessions of bots and humans, represented as time series of limited length, may give very good results provided that an appropriate distance measure is used. In order to draw a sound conclusion regarding the best measure, however, a larger set of experiments should be carried out because it is not sure if the same measure will give accurate answers in different scenarios.

Our prospective works include investigating time series clustering for Web session scenario taking into consideration other session features, like the number of page or image requests, the amount of data transferred to the client, as well as various lengths of time series being clustered. Other possible research direction is to increase the number of generated clusters with respect to Web client categories (i.e., clustering of multi-label time series). Furthermore, it would be undoubtedly worth performing experiments for a bigger dataset of Web sessions to embrace less common bots, like Web scrapers, attacking bots, fake crawlers, etc.

We are also planning to develop an effective online bot detection method based on time series clustering. The experimental results discussed in this paper show a big potential of the proposed approach to develop a method for identifying Web bots on the fly. By observing the initial progress of active Web sessions and comparing their request arrival patterns with prototypes of previously generated clusters, an early decision on classifying the client as bot or human might be determined.

REFERENCES

- Aghabozorgi, S.; A.S. Shirkhorshidi; and T.Y. Wah. 2015. "Time-series clustering – a decade review." *Information Systems* 53, 16-38.
- Alam, S.; G. Dobbie; Y.S. Koh; and P. Riddle. 2014. "Web bots detection using Particle Swarm Optimization based clustering". In *Proc. IEEE CEC'14*, 2955-2962.
- "Bad Bot Report 2021: The Pandemic of the Internet". 2021. Technical Report, Imperva Incapsula, <https://www.imperva.com/resources/resource-library/reports/bad-bot-report/>.
- Bernardi, M.L.; M. Cimitile; F. Martinelli; and F. Mercaldo. 2017. "A time series classification approach to game bot detection". In *Proc. of WIMS'17*, Article 6.
- Bonneton, A.; D. Migault; S. Senecal; and N. Kheir. 2015. "DGA bot detection with time series decision trees". In *Proc. of BADGERS'15*, 42-53.
- Chen, Z. and W. Feng. 2013. "Detecting impolite crawler by using time series analysis". In *Proc. of ICTAI'13*, 123-126.
- Doran, D. and S.S. Gokhale. 2016. "An integrated method for real time and offline web robot detection," *Expert Syst.* 33 (6) 592-606.
- Gavrilov, M; D. Anguelov; P. Indyk; and R. Motwani. 2000. "Mining the stock market: which measure is best? (extended abstract)". In *Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 487-496.
- Iliou C.; T. Kostoulas; T. Tsirikria; V. Katos; S. Vrochidis; and Y. Kompatsiaris. 2019. "Towards a framework for detecting advanced Web bots". In *Proc. of ARES'19*, Article no. 18.
- Kotsakos, D.; G. Trajcevski; D. Gunopulos; and C.C. Aggarwal. 2018. "Time-series data clustering". In *Data Clustering*. Chapman and Hall/CRC, pp. 357-380.
- Lagopoulos, A. and G. Tsoumakas. 2020. "Content-aware Web robot detection." *Applied Intelligence* 50(11), 4017-4028.
- Liao, T.W. 2005. "Clustering of time series data – a survey." *Pattern Recognition* 38(11), 1857-1874.
- Lysenko, S.; K. Bobrovnikova; P.T. Popov; V. Kharchenko; and D. Medzaty. 2020. "Spyware detection technique based on reinforcement learning". In *Proc. of CEUR Workshop*, vol. 2623, 307-316.
- Montero, P. and J.A. Vilar. 2015. "TSclust: An R package for time series clustering." *J. Stat. Softw.* 62(1), 1-43.
- Montero, P. and J.A. Vilar. 2020. "Package TSclust". <https://cran.r-project.org/web/packages/TSclust/TSclust.pdf>.
- Rahman, R.U. and D.S. Tomar. 2021. "Threats of price scraping on e-commerce websites: attack model and its detection using neural network." *Journal of Computer Virology and Hacking Techniques* 17(1), 75-89.
- Rovetta, S.; A. Cabri; F. Masulli; and G. Suchacka. 2017. "Bot or not? A case study on bot recognition from Web session logs". In *Quantifying and Processing Biomedical and Behavioral Signals*, SIST 103. Springer, 197-206.
- Rovetta, S.; G. Suchacka; and F. Masulli. 2020. "Bot recognition in a Web store: An approach based on unsupervised learning." *J. Netw. Comput. Appl.* 157, 102577.
- Suchacka, G.; A. Cabri; S. Rovetta; and F. Masulli. 2021. "Efficient on-the-fly Web bot detection." *Know.-Based Syst.* 223, 107074.
- Suchacka, G. and J. Iwański. 2020. "Identifying legitimate Web users and bots with different traffic profiles – an Information Bottleneck approach." *Know.-Based Syst.* 197, 105875.
- Suchacka, G. and I. Motyka. 2018. "Efficiency analysis of resource request patterns in classification of Web robots and humans". In *Proc. of ECMS'18*, 475-481.
- Udger. 2021. <https://udger.com> (access: July 12, 2021).
- User-agents. 2014. <http://www.user-agents.org> (access: September 4, 2017).
- Ustebay, S.; Z. Turgut; and M.A. Aydin. 2019. "Cyber attack detection by using neural network approaches: shallow neural network, deep neural network and autoencoder". In *Proc. of CN'19*, 144-155.
- Zabihi, M.; M.V. Jahan; and J. Hamidzadeh. 2014. "A density based clustering approach to distinguish between Web robot and human requests to a Web server." *ISC Int. J. Inf. Secur.* 6 (1), 77-89.

AUTHOR BIOGRAPHY

GRAŻYNA SUCHACKA received the M.Sc. degrees in Computer Science and in Management, as well as the Ph.D. degree in Computer Science with distinction from Wrocław University of Science and Technology, Poland. Now she is an Assistant Professor in the Institute of Informatics at the University of Opole, Poland. Her research interests include data analysis and modeling, data mining, machine learning, and Quality of Web Service with special regard to bot detection and electronic commerce support. Her e-mail address is: gsuchacka@uni.opole.pl.

Formal Verification of Neural Networks: a Case Study about Adaptive Cruise Control

Stefano Demarchi, Dario Guidotti, Andrea Pitto, Armando Tacchella

KEYWORDS

Safety Evaluation, Dependable Systems, Neural Networks, Formal Verification.

ABSTRACT

Formal verification of neural networks is a promising technique to improve their dependability for safety critical applications. Autonomous driving is one such application where the controllers supervising different functions in a car should undergo a rigorous certification process. In this paper we present an example about learning and verification of an adaptive cruise control function on an autonomous car. We detail the learning process as well as the attempts to verify various safety properties using the tool NEVER2, a new framework that integrates learning and verification in a single easy-to-use package intended for practitioners rather than experts in formal methods and/or machine learning.

INTRODUCTION

Context and Motivation. Verification of neural networks (NNs) is currently a trending topic involving different areas of AI, including machine learning, constraint programming, heuristic search and automated reasoning. A relative recent survey [HKR⁺20] cites more than 200 papers, most of which have been published in the last few years, and more contributions are appearing with impressive progression — see, e.g., [DCJ⁺19], [KHI⁺19], [WWR⁺18], [NKR⁺18], [LM17], [WPW⁺18]. The reason of this growing interest is that, while the application of NNs in various domains [LBH15] have made them one of the most popular machine-learned models to date, concerns about their vulnerability to adversarial perturbations [SZS⁺14], [GSS15] have been accompanying them since their initial adoption, to the point of restraining their application in safety- and security-related contexts. Automated formal verification — see, e.g., [LNPT18b] for a survey — offers an effective answer to the problem of establishing correctness of a NN and opens the path to their adoption in applications where they are currently not popular. One such application is autonomous driving where different functions can indeed be learned from data. Examples include advanced functions such as automatic steering and automatic speeding/braking, and more mundane ones such as traction control, launch control and anti-lock system for brakes. While it is possible to learn these functions with NNs, it is not clear whether

Stefano Demarchi, Dario Guidotti, Andrea Pitto and Armando Tacchella are with Università degli Studi di Genova, DIBRIS (Department of Informatics, Bioengineering, Robotics and Systems Engineering), Viale Causa 13, 16145 Genova. E-mail: stefano.demarchi@edu.unige.it, dario.guidotti@edu.unige.it, s3942710@studenti.unige.it (Andrea Pitto), armando.tacchella@unige.it. All authors contributed equally to the paper. The corresponding author is Armando Tacchella.

Communications of the ECMS, Volume 36, Issue 1,
Proceedings, ©ECMS Ibrahim A. Hameed, Agus Hasan,
Saleh Abdel-Afou Alaliyat (Editors) 2022
ISBN: 978-3-937436-77-7/978-3-937436-76-0(CD) ISSN 2522-2414

the rigorous certification procedures prescribed for car controllers can be passed by the learned controllers.

Objective. Our main research question is:

“Is it possible to leverage automated verification of neural networks in safety critical applications to improve confidence in the correctness of learned controllers?”

Clearly, automated verification together with standard testing techniques can provide reasonable confidence levels in a network only if the whole process that leads to a learned network is already bullet-proofed. In the following, we assume that a rigorous safety-by-design approach insures that scenario design, simulations, data-acquisitions and learning have been carried out so as to minimize errors in each phase and also in their integration. We do not expect to place blind trust in verification alone, but we expect NEVER2 and similar tools to be an essential ingredient of any safety conscious development process that involves learned controllers.

Contribution. This article is about learning and verification of a NN that replicates the function of an adaptive cruise control (ACC) similar to those used in real autonomous cars. The goal of the ACC is to maintain the vehicle at a speed set by the user and possibly adapt the speed considering other vehicles proceeding in front of it. We use our tool NEVER2 to learn and verify the controller, all in a single package. Our goal is to show how NEVER2 enables learning and verification for field practitioners who do not need to be experts in machine learning and/or verification. Our results show that NEVER2 is able to learn reasonable implementations of the ACC function (given the available data) and prove some interesting design requirements using abstraction techniques.

BACKGROUND

A. Basic Notation and Definitions

We denote n -dimensional *vectors* of real numbers $x \in \mathbb{R}^n$ — also *points* or *samples* — with lowercase letters like x, y, z . We write $x = (x_1, x_2, \dots, x_n)$ to denote a vector with its *components* along the n coordinates. We denote $x \cdot y$ the *scalar product* of two vectors $x, y \in \mathbb{R}^n$ defined as $x \cdot y = \sum_{i=1}^n x_i y_i$. The *norm* $\|x\|$ of a vector is defined as $\|x\| = \sqrt{x \cdot x}$. We denote sets of vectors $X \subseteq \mathbb{R}^n$ with uppercase letters like X, Y, Z . A set of vectors X is *bounded* if there exists $r \in \mathbb{R}, r > 0$ such that $\forall x, y \in X$ we have $d(x, y) < r$ where d is the *Euclidean norm* $d(x, y) = \|x - y\|$. A set X is *open* if for every point $x \in X$ there exists a positive real number ϵ_x such that a point $y \in \mathbb{R}^n$ belongs to X as long as $d(x, y) < \epsilon_x$. The complement of an open set is a *closed* set — intuitively, one that includes its boundary, whereas open sets do not; closed and bounded sets are *compact*. A set X is *convex* if for any two points $x, y \in X$ we have that also $z \in X \forall z = (1 - \lambda)x + \lambda y$ with $\lambda \in [0, 1]$, i.e., all the points falling on the line passing through x and y are also

in X . Notice that the intersection of any family, either finite or infinite, of convex sets is convex, whereas the union, in general, is not. Given any non-empty set X , the smallest convex set $\mathcal{C}(X)$ containing X is the *convex hull of X* and it is defined as the intersection of all convex sets containing X . A *hyperplane* $H \subseteq \mathbb{R}^n$ can be defined as the set of points

$$H = \{x \in \mathbb{R}^n \mid a_1x_1 + a_2x_2 + \dots + a_nx_n = b\}$$

where $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ and at least one component of a is non-zero. Let $f(x) = a_1x_1 + a_2x_2 + \dots + a_nx_n - b$ be the affine form defining H . The *closed half-spaces associated with H* are defined as

$$\begin{aligned} H_+(f) &= \{x \in X \mid f(x) \geq 0\} \\ H_-(f) &= \{x \in X \mid f(x) \leq 0\} \end{aligned}$$

Notice that both $H_+(f)$ and $H_-(f)$ are convex. A *polyhedron* in $P \subseteq \mathbb{R}^n$ is a set of points defined as $P = \bigcap_{i=1}^p C_i$ where $p \in \mathbb{N}$ is a finite number of closed half-spaces C_i . A bounded polyhedron is a *polytope*: from the definition, it follows that polytopes are convex and compact in \mathbb{R}^n .

B. Neural Networks

Given a finite number p of functions $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}^{n_1}, \dots, f_p : \mathbb{R}^{n_{p-1}} \rightarrow \mathbb{R}^m$ — also called *layers* — we define a *feed forward neural network* as a function $\nu : \mathbb{R}^n \rightarrow \mathbb{R}^m$ obtained through the compositions of the layers, i.e., $\nu(x) = f_p(f_{p-1}(\dots f_1(x) \dots))$. The layer f_1 is called *input layer*, the layer f_p is called *output layer*, and the remaining layers are called *hidden*. For $x \in \mathbb{R}^n$, we consider only two types of layers:

- $f(x) = Ax + b$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ is an *affine layer* implementing the linear mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$;
- $f(x) = (\sigma_1(x_1), \dots, \sigma_n(x_n))$ is a *functional layer* $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ consisting of n *activation functions* — also called *neurons*; usually $\sigma_i = \sigma$ for all $i \in [1, n]$, i.e., the function σ is applied componentwise to the vector x .

We consider the *ReLU* activation function defined as $\sigma(r) = \max(0, r)$, which finds widespread adoption. For a neural network $\nu : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the task of *classification* is about assigning to every input vector $x \in \mathbb{R}^n$ one out of m labels: an input x is assigned to a class k when $\nu(x)_k > \nu(x)_j$ for all $j \in [1, m]$ and $j \neq k$; the task of *regression* is about approximating a functional mapping from \mathbb{R}^n to \mathbb{R}^m .

C. Verification task

Given a neural network $\nu : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we wish to verify algorithmically that it complies to stated *post-conditions* on the output as long as it satisfies *pre-conditions* on the input. Without loss of generality¹, we assume that the input domain of ν is a bounded set $I \subset \mathbb{R}^n$. Therefore, the corresponding output domain is also a bounded set $O \subset \mathbb{R}^m$ because (i) affine transformations of bounded sets are still bounded sets, (ii) ReLU is a piecewise affine transformation of its input, (iii) the output of logistic functions is always

¹Input domains must be bounded to enable implementation of neural networks on digital hardware; therefore, also data from physical processes, which are potentially unbounded, are normalized within small ranges in practical applications.

bounded in the set $[0, 1]$, and the composition of bounded functions is still bounded. We require that the logic formulas defining pre- and post-conditions are interpretable as finite unions of bounded sets in the input and output domains. Formally, given p bounded sets X_1, \dots, X_p in I such that $\Pi = \bigcup_{i=1}^p X_i$ and s bounded sets Y_1, \dots, Y_s in O such that $\Sigma = \bigcup_{i=1}^s Y_i$, we wish to prove that

$$\forall x \in \Pi. \nu(x) \in \Sigma. \quad (1)$$

While this query cannot express some problems regarding neural networks, e.g., invertibility or equivalence [LNPT18b], it captures the general problem of testing robustness against *adversarial perturbations* [GSS15]. For example, given a network $\nu : I \rightarrow O$ with $I \subset \mathbb{R}^n$ and $O \subset \mathbb{R}^m$ performing a classification task, we have that separate regions of the input are assigned to one out of m labels by ν . Let us assume that region $X_j \in I$ is classified in the j -th class by ν . We define an *adversarial region* as a set \hat{X}_j such that for all $\hat{x} \in \hat{X}_j$ there exists at least one $x \in X_j$ such that $d(x, \hat{x}) \leq \delta$ for some positive constant δ . The network ν is *robust* with respect to $\hat{X}_j \subseteq I$ if, for all $\hat{x} \in \hat{X}_j$, it is still the case that $\nu(x)_j > \nu(x)_i$ for all $i \in [1, m]$ with $i \neq j$. This can be stated in the notation of condition (1) by letting $\Pi = \{\hat{X}_j\}$ and $\Sigma = \{Y_j\}$ with $Y_j = \{y \in O \mid y_j \geq y_i + \epsilon, \forall i \in [1, m] \wedge i \neq j, \epsilon > 0\}$. Analogously, in a regression task we may ask that points that are sufficiently close to any input vector in a set $X \subseteq I$ are also sufficiently close to the corresponding output vectors. To do this, given the positive constants δ and ϵ , we let $\hat{X} = \{\hat{x} \in I \mid \exists x.(x \in X \wedge d(\hat{x}, x) \leq \delta)\}$ and $\hat{Y} = \{\hat{y} \in O \mid \exists x.(x \in \hat{X} \wedge d(\hat{y}, \nu(x)) \leq \epsilon)\}$ to obtain $\Pi = \{\hat{X}\}$ and $\Sigma = \{\hat{Y}\}$. Notice that, given our definition, we consider adversarial regions and output images that may not be convex.

D. Case Study

Technically, an adaptive cruise control (ACC) is an autonomous driving function of level one², which controls the acceleration of the *ego car* — the car whereon the ACC is installed — along the longitudinal axis. An ACC has two competing objectives: keeping the ego car at the speed set by the user (*speed following mode*) and keeping a safe distance from the *exo car* in front (*car following mode*). The ACC that we consider has one output, i.e., the acceleration a suggested to the ego car in $m \cdot s^{-2}$, and 6 inputs:

- $v_p[m \cdot s^{-1}]$: the speed of the ego car.
- $S_r[m \cdot s^{-1}]$: the speed of the exo car relative to the ego car; when there is no exo car, this input has the value 0.
- $D[m]$: the actual distance between the ego car and the exo car; when there is no exo car or when the exo car is farther than $150m$ this input has the default value of $150m$.
- $TH[s]$: Minimum headway time; this is the minimum time gap between the exo car and the ego car: $TH \cdot v_p$ corresponds to D_s , i.e., the *minimum safety distance*.
- $D_0[m]$: A safety margin to be added to the minimum safety distance D_s .

²“Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles”, SAE Standards, J3016_202104.

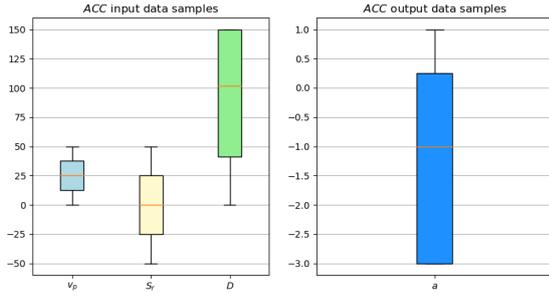


Fig. 1: Box plot for a million samples of the Adaptive Cruise Control data set ($TH = 1.5; D_0 = 5$)

In production vehicles the ACC function is implemented using classical control laws. We view the production function — called ACC_o in the following — as a black-box whose behavior should be learned by a neural network.

LEARNING

Given the goal of learning ACC_o using a NN, we should generate several instances of input-output data using, e.g., a car simulator. Since a simulator was unavailable to us at the time of this writing, we generated the dataset to learn various NNs by drawing samples from uniform distributions over the input values of ACC_o , considering the following lower and upper bounds for v_p , v_r and D :

$$0 \leq v_p \leq 50 \quad -50 \leq v_r \leq 50 \quad 0 \leq D \leq 150 \quad (2)$$

The values of TH and D_0 are kept fixed, and we obtain the corresponding output a by feeding ACC_o with the generated inputs. We generate 16 different data sets, each composed by a million samples, that feature 16 different combinations of TH and D_0 , where $TH \in \{1, 1.5, 2, 2.5\}$, while $D_0 = \{2.5, 5, 7.5, 10\}$. Figure 1 shows the distributions of input and output samples using box plots in the case $TH = 1.5$ and $D_0 = 5$.

We tested three NN architectures comprised of affine and ReLU layers: we refer to them as $Net0$, $Net1$ and $Net2$ in the following. These NNs feature increasing complexity both in terms of the number of layers and in the amount of neurons per layer. An example is shown in Figure 2 for $Net0$ on the canvas of NEVER2. The networks considered differs from each other only for the details of the hidden layers, which are the following:

- $Net0$: two affine layer of 20 and 10 neurons respectively, each followed by a ReLU layer;
- $Net1$: two affine layer of 50 and 40 neurons respectively, each followed by a ReLU layer;
- $Net2$: four affine layers of 20, 20, 20 and 10 neurons respectively, each followed by a ReLU layer;

The input of the network is in all the cases a three dimensional vector. All the networks present an output layers consisting of a linear layer of dimension 1 (without a following ReLU layer).

To learn the NNs we split the data sets in two parts, one for training and one for testing, with the ratio of 4:1. Our training phase lasts 100 epochs for each of the 16 data

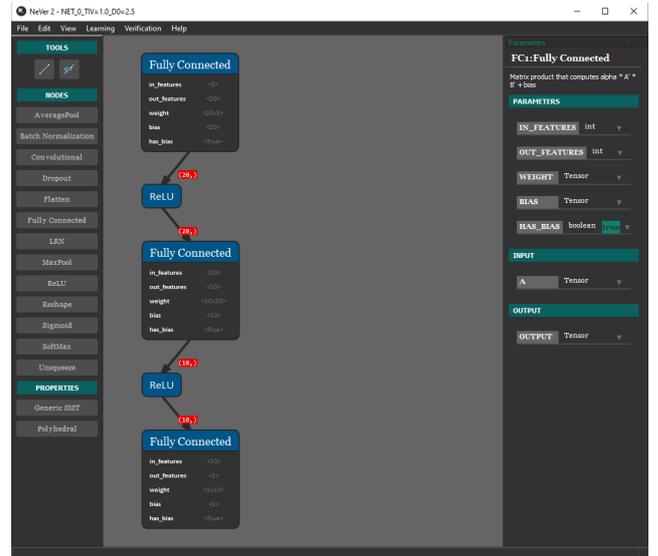


Fig. 2: NEVER2 representation of $Net0$ architecture.

sets. We consider the *Adam* optimizer [KB14] and the *ReduceLROnPlateau* scheduler. For both our loss function and our precision metric we leveraged the *Mean Squared Error (MSE) loss*. We set batch sizes to 32 for training, validation, and test sets. In our setup, we dedicated 30% of the training set to the validation process. Concerning the optional parameters, we also set the learning rate to 0.01, the weight decay to 0.0001 and the training scheduler patience to 3, i.e., the number of consecutive epochs without loss decrease that triggers training procedure abortion. All the training is performed inside NEVER2 which, in turn, is based on the PYTORCH library.³ For this reason, all the remaining parameters required by the learning algorithms are set to their default PYTORCH values.

VERIFICATION

To enable algorithmic verification of neural networks in NEVER2, we consider the abstract domain $\langle \mathbb{R}^n \rangle \subset 2^{\mathbb{R}^n}$ of polytopes defined in \mathbb{R}^n to abstract (families of) bounded sets into (families of) polytopes. We provide corresponding abstractions for affine and functional layers to perform abstract computations and obtain consistent overapproximation of concrete networks.

Definition 1: (Abstraction) Given a bounded set $X \subset \mathbb{R}^n$, an abstraction is defined as a function $\alpha : 2^{\mathbb{R}^n} \rightarrow \langle \mathbb{R}^n \rangle$ that maps X to a polytope P such that $\mathcal{C}(X) \subseteq P$.

Intuitively, the function α maps a bounded set X to a corresponding polytope in the abstract space such that the polytope always contains the convex hull of X . Depending on X , the enclosing polytope may not be unique. However, given the convex hull of any bounded set, it is always possible to find an enclosing polytope. As shown in [Zhe19], one could always start with an axis-aligned regular n simplex consisting of $n + 1$ facets — e.g., the triangle in \mathbb{R}^2 and the tetrahedron in \mathbb{R}^3 — and then refine the abstraction as needed by adding facets, i.e., adding half-spaces to make the abstraction more precise.

³<https://pytorch.org>

Definition 2: (Concretization) Given a polytope $P \in \langle \mathbb{R}^n \rangle$ a concretization is a function $\gamma : \langle \mathbb{R}^n \rangle \rightarrow 2^{\mathbb{R}^n}$ that maps P to the set of points contained in it, i.e., $\gamma(P) = \{x \in \mathbb{R}^n \mid x \in P\}$.

Intuitively, the function γ simply maps a polytope P to the corresponding (convex and compact) set in \mathbb{R}^n comprising all the points contained in the polytope. As opposed to abstraction, the result of concretization is uniquely determined. We extend abstraction and concretization to finite families of sets and polytopes, respectively, as follows. Given a family of p bounded sets $\Pi = \{X_1, \dots, X_p\}$, the abstraction of Π is a set of polytopes $\Sigma = \{P_1, \dots, P_s\}$ such that $\alpha(X_i) \subseteq \bigcup_{i=1}^s P_i$ for all $i \in [1, p]$; when no ambiguity arises, we abuse notation and write $\alpha(\Pi)$ to denote the abstraction corresponding to the family Π . Given a family of s polytopes $\Sigma = \{P_1, \dots, P_s\}$, the concretization of Σ is the union of the concretizations of its elements, i.e., $\bigcup_{i=1}^s \gamma(P_i)$; also in this case, we abuse notation and write $\gamma(\Sigma)$ to denote the concretization of a family of polytopes Σ .

Given our choice of abstract domain and a concrete network $\nu : I \rightarrow O$ with $I \subset \mathbb{R}^n$ and $O \subset \mathbb{R}^m$, we need to show how to obtain an *abstract neural network* $\tilde{\nu} : \langle I \rangle \rightarrow \langle O \rangle$ that provides a sound overapproximation of ν . To frame this concept, we introduce the notion of consistent abstraction.

Definition 3: (Consistent abstraction) Given a mapping $\nu : \mathbb{R}^n \rightarrow \mathbb{R}^m$, a mapping $\tilde{\nu} : \langle \mathbb{R}^n \rangle \rightarrow \langle \mathbb{R}^m \rangle$, abstraction function $\alpha : 2^{\mathbb{R}^n} \rightarrow \langle \mathbb{R}^n \rangle$ and concretization function $\gamma : \langle \mathbb{R}^m \rangle \rightarrow 2^{\mathbb{R}^m}$, the mapping $\tilde{\nu}$ is a consistent abstraction of ν over a set of inputs $X \subseteq I$ exactly when

$$\{\nu(x) \mid x \in X\} \subseteq \gamma(\tilde{\nu}(\alpha(X))) \quad (3)$$

The notion of consistent abstraction can be readily extended to families of sets as follows. The mapping $\tilde{\nu}$ is a consistent abstraction of ν over a family of sets of inputs $X_1 \dots X_p$ exactly when

$$\{\nu(x) \mid x \in \bigcup_{i=1}^p X_i\} \subseteq \gamma(\tilde{\nu}(\alpha(X_1, \dots, X_p))) \quad (4)$$

where we abuse notation and denote with $\tilde{\nu}(\cdot)$ the family $\{\tilde{\nu}(P_1), \dots, \tilde{\nu}(P_s)\}$ with $\{P_1, \dots, P_s\} = \alpha(X_1, \dots, X_p)$

To represent polytopes and define the computations performed by abstract layers in NEVER2, we resort to a specific subclass of *generalized star sets*, introduced in [BD17] and defined as follows — the notation is adapted from [TLM⁺19].

Definition 4: (Generalized star set) Given a *basis matrix* $V \in \mathbb{R}^{n \times m}$ obtained arranging a set of m *basis vectors* $\{v_1, \dots, v_m\}$ in columns, a point $c \in \mathbb{R}^n$ called *center* and a *predicate* $R : \mathbb{R}^m \rightarrow \{\top, \perp\}$, a generalized star set is a tuple $\Theta = (c, V, R)$. The set of points represented by the generalized star set is given by

$$\llbracket \Theta \rrbracket \equiv \{z \in \mathbb{R}^n \mid z = Vx + c \text{ such that } R(x_1, \dots, x_m) = \top\} \quad (5)$$

In the following we denote $\llbracket \Theta \rrbracket$ also as Θ . Depending on the choice of R , generalized star sets can represent different kinds of sets, but in NEVER2 we consider only those such that $R(x) := Cx \leq d$, where $C \in \mathbb{R}^{p \times m}$ and $d \in \mathbb{R}^p$ for $p \geq 1$, i.e., R is a conjunction of p linear constraints

as in [TLM⁺19]; we further require that the set $Y = \{y \in \mathbb{R}^m \mid Cy \leq d\}$ is bounded.

Given a generalized star set $\Theta = (c, V, R)$ such that $R(x) := Cx \leq d$ with $C \in \mathbb{R}^{p \times m}$ and $d \in \mathbb{R}^p$, if the set $Y = \{y \in \mathbb{R}^m \mid Cy \leq d\}$ is bounded, then the set of points represented by Θ is a polytope in \mathbb{R}^n , i.e., $\Theta \in \langle \mathbb{R}^n \rangle$. In the following, we refer to generalized star sets obeying our restrictions simply as *stars*.

The simplest abstract layer to obtain is the one abstracting affine transformations. As we have already mentioned, affine transformations of polytopes are still polytopes, so we just need to define how to apply an affine transformation to a star — the definition is adapted from [TLM⁺19].

Definition 5: (Abstract affine mapping) Given a star set $\Theta = (c, V, R)$ and an affine mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $f = Ax + b$, the abstract affine mapping $\tilde{f} : \langle \mathbb{R}^n \rangle \rightarrow \langle \mathbb{R}^m \rangle$ of f is defined as $\tilde{f}(\Theta) = (\hat{c}, \hat{V}, R)$ where

$$\hat{c} = Ac + b \quad \hat{V} = AV$$

Intuitively, the center and the basis vectors of the input star Θ are affected by the transformation of f , while the predicates remain the same. Given an affine mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the corresponding abstract mapping $\tilde{f} : \langle \mathbb{R}^n \rangle \rightarrow \langle \mathbb{R}^m \rangle$ provides a consistent abstraction over any bounded set $X \subset \mathbb{R}^n$, i.e., $\{f(x) \mid x \in X\} \subseteq \gamma(\tilde{f}(\alpha(X)))$ for all $X \subset \mathbb{R}^n$. We observe that the set $\alpha(X)$ is any polytope P such that $P \supseteq \mathcal{C}(X)$ — equality holds only when X is already a polytope, and thus $X \equiv \mathcal{C}(X) \equiv P$. Let $\Theta_P = (c_P, V_P, R_P)$ be the star corresponding to P defined as

$$c_P = 0^n \quad V_P = I^n \quad R_P = C_P x + d_P \leq 0$$

where 0^n is the n -dimensional zero vector, and I^n is the $n \times n$ identity matrix — the columns of I^n correspond to the standard orthonormal basis e_1, \dots, e_n of \mathbb{R}^n , i.e., $\|e_i\| = 1$ and $e_i \cdot e_j = 0$ for all $i \neq j$ with $i, j \in [1, n]$; the matrix $C_P \in \mathbb{R}^{q \times n}$ and the vector $d_P \in \mathbb{R}^q$ collect the parameters defining q half-spaces whose intersection corresponds to P . Given our choice of c and V , it is thus obvious that $\Theta_P \equiv P$. Recall that $f = Ax + b$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$; from definition (5) we have that $\tilde{f}(\Theta_P) = \hat{\Theta}_P$ with $\hat{\Theta}_P = (\hat{c}_P, \hat{V}_P, R_P)$ and

$$\hat{c}_P = A0^n + b = b \quad \hat{V}_P = AI^n = A$$

The concretization of $\hat{\Theta}_P$ is just the set of points contained in Θ_P defined as

$$\gamma(\hat{\Theta}_P) = \{z \in \mathbb{R}^m \mid z = Ax + b \text{ such that } C_P x \leq d_P\} \quad (6)$$

Now it remains to show that $\{f(x) \mid x \in X\} \subseteq \gamma(\hat{\Theta}_P)$. This follows from the fact that, for a generic $y \in \{f(x) \mid x \in X\}$ there must exist $x \in X$ such that $y = Ax + b$; since x satisfies $C_P x \leq d_P$ by construction of P , it is also the case that $y \in \gamma(\hat{\Theta}_P)$ by definition (6).

Algorithm 1 defines the abstract mapping of a functional layer with n ReLU activation functions in NEVER2. The function COMPUTE_LAYER takes as input an indexed list of N stars $\Theta_1, \dots, \Theta_N$ and an indexed list of n positive integers called *refinement levels*. For each neuron, the refinement level tunes the grain of the abstraction: level 0 corresponds to the coarsest abstraction that we consider — the

Algorithm 1 Abstraction of the ReLU activation function.

```

1: function COMPUTE_LAYER(input =  $[\Theta_1, \dots, \Theta_N]$ , refine =
    $[r_1, \dots, r_n]$ )
2:   output = []
3:   for  $i = 1 : N$  do
4:     stars =  $[\Theta_i]$ 
5:     for  $j = 1 : n$  do stars = COMPUTE_RELU(stars,  $j$ , refine[j],  $n$ )
6:     end for
7:     APPEND(output, stars)
8:   end for
9:   return output
10: end function

11: function COMPUTE_RELU(input =  $[\Gamma_1, \dots, \Gamma_M]$ ,  $j$ , level,  $n$ )
12:   output = []
13:   for  $k = 1 : M$  do
14:      $(lb_j, ub_j) = \text{GET\_BOUNDS}(\text{input}[k], j)$ 
15:      $M = [e_1 \dots e_{j-1} \ 0 \ e_{j+1} \dots e_n]$ 
16:     if  $lb_j \geq 0$  then  $S = \text{input}[k]$ 
17:     else if  $ub_j \leq 0$  then  $S = M * \text{input}[k]$ 
18:     else
19:       if level > 0 then
20:          $\Theta_{low} = \text{input}[k] \wedge z[j] < 0$ ;  $\Theta_{upp} = \text{input}[k] \wedge z[j] \geq 0$ 
21:          $S = [M * \Theta_{low}, \Theta_{upp}]$ 
22:       else
23:          $(c, V, Cx \leq d) = \text{input}[j]$ 
24:          $C_1 = [0 \ 0 \dots \ -1] \in \mathbb{R}^{1 \times m+1}$ ,  $d_1 = 0$ 
25:          $C_2 = [V[j, :] \ -1] \in \mathbb{R}^{1 \times m+1}$ ,  $d_2 = -c_k[j]$ 
26:          $C_3 = [\frac{-ub_j}{ub_j - lb_j} \cdot V[j, :] \ -1] \in \mathbb{R}^{1 \times m+1}$ ,  $d_3 =$ 
            $\frac{ub_j}{ub_j - lb_j} (c[j] - lb_j)$ 
27:          $C_0 = [C \ 0^{m \times 1}]$ ,  $d_0 = d$ 
28:          $\hat{C} = [C_0; C_1; C_2; C_3]$ ,  $\hat{d} = [d_0; d_1; d_2; d_3]$ 
29:          $\hat{V} = MV$ ,  $\hat{V} = [\hat{V} \ e_j]$ 
30:          $S = (Mc, \hat{V}, \hat{C}\hat{x} \leq \hat{d})$ 
31:       end if
32:     end if
33:     APPEND(output, S)
34:   end for
35:   return output
36: end function

```

greater the level, the finer the abstraction grain. In the case of ReLUs, all non-zero levels map to the same (precise) refinement, i.e., a piecewise affine mapping. Notice that, since each neuron features its own refinement level, algorithm 1 controls abstraction down to the single neuron, enabling the computation of levels with mixed degrees of abstraction. The output of function COMPUTE_LAYER is still an indexed list of stars, that can be obtained by independently processing the stars in the input list. For this reason, the **for** loop starting at line 3 can be parallelized to speed up actual implementations. Given a single input star $\Theta_i \in \langle \mathbb{R}^n \rangle$, each of the n dimensions is processed in turn by the **for** loop starting at line 5 and involving the function COMPUTE_RELU. Notice that the stars obtained processing the j -th dimension are feeded again to COMPUTE_RELU in order to process the $j + 1$ -th dimension. For each star given as input, the function COMPUTE_RELU first computes the lower and upper bounds of the star along the j -th dimension by solving a linear-programming problem — function GET_BOUNDS at line 11. Independently from the abstraction level, if $lb_j \geq 0$ then the ReLU acts as an identity function (line 13), whereas if $lb_j \geq 0$ then the j -th dimension is zeroed (line 14). The operator “star” takes a matrix M , a star $\Gamma = (c, V, R)$ and returns the star (Mc, MV, R) .

The linear-programming problem we need to solve in the

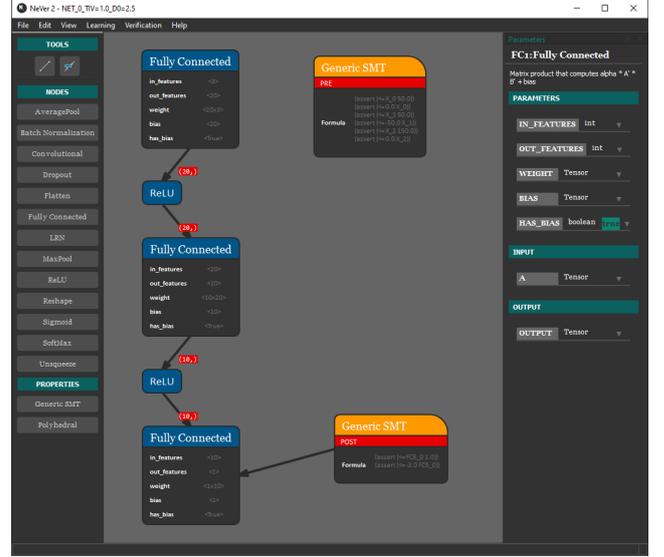


Fig. 3: Representation of the *OutBounds* property in NEVER2. Detached property blocks are treated as input pre-conditions, while property blocks linked to the NN are the output post-conditions.

GET_BOUNDS solver can be formalized as follows:

$$\begin{aligned}
 (\min/\max) z_j &= \mathbf{V}[j, :] \mathbf{x} + c[j] \\
 \text{with } \mathbf{C}\mathbf{x} &\leq \mathbf{d}
 \end{aligned}$$

The problem must be solved as minimization and maximization to provide the lower bound and the upper bound respectively. It should be noted that the complexity of the problem increases with the number of variables of the predicate of the star of interest. As consequence the computational complexity of GET_BOUNDS increases as over-approximation increases, whereas for concrete stars the complexity remains the same.

Given a ReLU mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, the corresponding abstract mapping $\tilde{f} : \langle \mathbb{R}^n \rangle \rightarrow \langle \mathbb{R}^n \rangle$ defined in algorithm 1 provides a consistent abstraction over any bounded set $X \subset \mathbb{R}^n$, i.e., $\{f(x) \mid x \in X\} \subseteq \tilde{f}(\alpha(X))$ for all $X \subset \mathbb{R}^n$.

EXPERIMENTS

We consider three properties for the ACC case study, and we verify them in NEVER2 with different NNs. The first property that we define, called *OutBounds* in the following, simply checks that the output acceleration does not exceed the bounds of the ACC_o function. Stated formally, this amounts to have NEVER2 check that, given the preconditions

$$\begin{aligned}
 0 &\leq v_p \leq 50 \\
 -50 &\leq v_r \leq 50 \\
 0 &\leq D \leq 150
 \end{aligned} \tag{7}$$

the output a satisfies the postcondition

$$-3 \leq a \leq 1. \tag{8}$$

Figure 3 shows NEVER2 canvas with the additional property specification.

The second property we consider is called *Near0*, and it is aimed at making sure that the ACC system does not output

TABLE I: NEVER2 results for the ACC data set with $TH = 1$ and $D_0 = 5$, with $\epsilon = 0$ (left) and $\epsilon = 20$ (right). CPU time is in seconds rounded to the third decimal place. The best setting for each network and property is highlighted in boldface.

$TH = 1, 5 - D_0 = 5 - \epsilon = 0$				
Network	Property	Setting	Result	CPU Time
Net0	OutBounds	over-approx	True	5.139
		mixed	True	5.055
		mixed2	True	5.112
		complete	True	6.273
	Near0	over-approx	False	5.666
		mixed	False	5.251
		mixed2	False	5.203
		complete	False	6.319
	Far0	over-approx	False	5.078
		mixed	False	4.986
		mixed2	False	5.139
		complete	False	5.186
Net1	OutBounds	over-approx	True	5.931
		mixed	True	6.662
		mixed2	True	7.309
		complete	True	51.683
	Near0	over-approx	False	5.906
		mixed	False	6.676
		mixed2	False	8.071
		complete	False	50.469
	Far0	over-approx	False	5.709
		mixed	False	5.888
		mixed2	False	6.301
		complete	False	13.041
Net2	OutBounds	over-approx	True	9.525
		mixed	True	10.482
		mixed2	True	12.525
		complete	True	26.958
	Near0	over-approx	False	9.515
		mixed	False	10.292
		mixed2	False	13.636
		complete	False	24.496
	Far0	over-approx	False	9.753
		mixed	False	9.944
		mixed2	False	12.148
		complete	False	13.27

$TH = 1.5 - D_0 = 5 - \epsilon = 20$				
Network	Property	Setting	Result	CPU Time
Net0	OutBounds	over-approx	True	5.037
		mixed	True	5.063
		mixed2	True	4.996
		complete	True	6.203
	Near0	over-approx	False	5.034
		mixed	False	5.101
		mixed2	False	4.965
		complete	False	5.345
	Far0	over-approx	False	5.008
		mixed	True	5.016
		mixed2	True	5.068
		complete	True	5.62
Net1	OutBounds	over-approx	True	5.948
		mixed	True	6.934
		mixed2	True	7.232
		complete	True	52.318
	Near0	over-approx	False	5.436
		mixed	False	5.797
		mixed2	False	5.955
		complete	False	7.667
	Far0	over-approx	False	5.344
		mixed	False	5.776
		mixed2	False	6.226
		complete	False	8.212
Net2	OutBounds	over-approx	True	9.532
		mixed	True	10.149
		mixed2	True	12.065
		complete	True	26.794
	Near0	over-approx	False	9.379
		mixed	False	9.872
		mixed2	False	11.653
		complete	True	10.696
	Far0	over-approx	False	9.453x
		mixed	False	9.848
		mixed2	False	11.558
		complete	False	10.854

positive accelerations when the vehicle ahead is too close. We frame this concept via the precondition

$$\begin{aligned}
 0 &\leq v_p \leq 50 \\
 -50 &\leq v_r \leq 50 \\
 0 &\leq D \leq 150 \\
 TH \cdot v_r + D_0 &\geq D + \epsilon
 \end{aligned} \tag{9}$$

where $\epsilon \in \mathbb{R}^+$ is a positive tolerance value in the last inequality. Notice that the input bounds are the same as *Outbound*. The last inequality stems from the fact that $TH \cdot v_r$ is the safety distance required to stop the ego car in time if the exo car brakes, and D_0 is a buffer value which, like TH , is constant for each data set. The corresponding output postcondition for *Near0* is

$$-3 \leq a \leq 0. \tag{10}$$

Intuitively, we do not want the network to output positive accelerations in this case.

Finally, the last property we consider is *Far0*, which is symmetrical with respect to *Near0*. The precondition is

$$\begin{aligned}
 0 &\leq v_p \leq 50 \\
 -50 &\leq v_r \leq 50 \\
 0 &\leq D \leq 150 \\
 TH \cdot v_r + D_0 &\leq D - \epsilon
 \end{aligned} \tag{11}$$

where $\epsilon \in \mathbb{R}^+$ is still a tolerance value and the input bounds coincide with *OutBounds* and *Near0* properties. In this case, we want to verify that when the ego car is too far from the exo car (or there is no vehicle ahead at all), the NN does not suggest negative accelerations. The output postcondition is

$$0 \leq a \leq 1. \tag{12}$$

In addition to the properties themselves, we also define different configuration for NEVER2 verification algorithms. In particular we consider 3 settings: *over-approximated*, *mixed*, and *complete*. The over-approximated setting corresponds to running algorithm 1 with *level* greater than zero whereas the complete setting amounts to choose *level* = 0. In the former case the output image of the NN computed by NEVER2 given the input preconditions is an overapproximation of the concrete one. In this case, checking whether the output image satisfies the postcondition gives us a sufficient condition only, i.e., if the inequality holds the NN is safe with respect to that property. On the other hand, if the inequality is not verified, the NN may still be safe and the check may have failed because of the loss of precision in the abstraction process. In the complete setting, on the other hand, NEVER2 computes the actual output image of the network: if the inequality in the postcondition does not hold, we are sure that the NN is not safe. However, the complete setting in algorithm 1 potentially causes the exponential blow-up in the number of stars generated, and thus the computation might simply not be feasible. The mixed setting

strikes a trade-off between complete and over-approximated setting: using an heuristic detailed in [GPT21], NEVER2 tries to concretize the least number of stars that enable proving the property without blowing the computation time. In our experiments, we consider two different sub-settings for mixed, called *mixed* and *mixed2* which differ in the number of neurons to refine, either 1 or 2, respectively.

We run our tests on a workstation featuring two Intel Xeon Gold 6234 CPU, three NVIDIA Quadro RTX 6000/8000 GPUs (with CUDA enabled), and 125.6 GiB of RAM running Ubuntu 20.04.03 LTS. For the sake of brevity, we are only going to report here a fraction of the experiments we ran in Table I for the data set with $TH = 1.5$ and $D_0 = 5$, considering $\epsilon = 0$ and $\epsilon = 20$. The results we show here are consistent with the results obtained on other data sets that we do not report. In particular, looking at Table I we can observe that:

- All the properties can be checked on all the networks in reasonable time by NEVER2: less than one minute of CPU time is required independently from the network architecture and the specific setting considered.
- The complete setting is the most expensive in computational terms; given the considerations above this should come at no surprise, but in one case, namely *Net2* on property *Near0*, the complete setting is able to prevail over the others, i.e., it certifies that the property is true; indeed mixed and over-approximated settings (shortened as over-approx in Table I) take less time, but state that the property is false because they do not manage to reach enough precision to state the correct result.
- The over-approximated setting is often faster than the other ones: 6 out of 9 cases for $\epsilon = 0$ and 7 out of 9 cases for $\epsilon = 20$; however it must be noted that its results are definite only when the property is true: 3 out of 9 cases for both values of ϵ and always for the (simplest) property *OutBounds*.
- The mixed setting is at times faster than the over-approximated one, but only in one case, namely property *Far0* on *Net0* it is able to provide a definite answer while outperforming both the complete and over-approximated settings.

Overall we can conclude that while further research is needed to improve on the capability of NEVER2 to provide definite answers with faster techniques involving over-approximation, still the tool is able to check a number of interesting properties in networks involving hundreds of neurons in a relatively small amount of CPU time. We view this as a positive result and an enabler for preliminary testing of NEVER2 at industrial settings featuring networks of comparable size to our ACC case study.

CONCLUSIONS

In this article we developed a concrete example of how our system NEVER2 can learn and verify NNs. The application we consider is, to the best of our knowledge, one of the examples of formal verification for NNs which are closest to industrial application. We have shown convincing experimental evidence that it is possible to learn an adaptive cruise control function and verify some interesting properties, all in a single package that supports the process through an easy to use graphical user interface. In future works, we

intend to deepen our research and find more applications that require NNs to be learned and verified, possibly with more complex architectures to stress NEVER2 capabilities.

REFERENCES

- [BD17] Stanley Bak and Parasara Sridhar Duggirala. Simulation-equivalent reachability of large linear systems with inputs. In *International Conference on Computer Aided Verification*, pages 401–420. Springer, 2017.
- [DCJ⁺19] Souradeep Dutta, Xin Chen, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. Sherlock - A tool for verification of neural network feedback systems: demo abstract. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control, HSCC*, pages 262–263, 2019.
- [GPT21] Dario Guidotti, Luca Pulina, and Armando Tacchella. pyn-ever: A framework for learning and verification of neural networks. In *Automated Technology for Verification and Analysis - 19th International Symposium, ATVA*, pages 357–363, 2021.
- [GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR (Poster)*, 2015.
- [HKR⁺20] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xiping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.
- [KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [KHI⁺19] Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljic, David L. Dill, Mykel J. Kochenderfer, and Clark W. Barrett. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification - 31st International Conference, CAV*, pages 443–452, 2019.
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [LM17] Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. *CoRR*, abs/1706.07351, 2017.
- [LNPT18a] Francesco Leofante, Nina Narodytska, Luca Pulina, and Armando Tacchella. Automated Verification of Neural Networks: Advances, Challenges and Perspectives. *arXiv e-prints*, page arXiv:1805.09938, May 2018.
- [LNPT18b] Francesco Leofante, Nina Narodytska, Luca Pulina, and Armando Tacchella. Automated verification of neural networks: Advances, challenges and perspectives. *CoRR*, abs/1805.09938, 2018.
- [NKR⁺18] Nina Narodytska, Shiva Prasad Kasiviswanathan, Leonid Ryzhyk, Mooly Sagiv, and Toby Walsh. Verifying properties of binarized deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 6615–6624, 2018.
- [SZS⁺14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014.
- [TLM⁺19] Hoang-Dung Tran, Diago Manzananas Lopez, Patrick Musau, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, and Taylor T Johnson. Star-based reachability analysis of deep neural networks. In *International Symposium on Formal Methods*, pages 670–686. Springer, 2019.
- [WPW⁺18] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6369–6379, 2018.
- [WWR⁺18] Min Wu, Matthew Wicker, Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. A game-based approximate verification of deep neural networks with provable guarantees. *CoRR*, abs/1807.03571, 2018.
- [Zhe19] Yu Zheng. Computing bounding polytopes of a compact set and related problems in n-dimensional space. *Computer-Aided Design*, 109:22–32, 2019.

A DSL-based modeling approach for energy harvesting IoT/WSN

Lelio Campanile
Mauro Iacono
Fiammetta Marulli
Dip. di Matematica e Fisica
Università degli Studi della
Campania "L. Vanvitelli"
viale Lincoln 5
81100, Caserta, Italy

Marco Gribaudo
Dip. di Elettronica,
Informatica e Bioingegneria
Politecnico di Milano
via Ponzio 51
20133, Milano, Italy

Michele Mastroianni
Dip. di Informatica
Università degli Studi di Salerno
Via Giovanni Paolo II, 132
84084 Fisciano, Italy

KEYWORDS

Performance evaluation; multiformalism modeling; energy management; wireless sensor networks; simulation; ns-3; Internet of Things; edge computing

ABSTRACT

The diffusion of intelligent services and the push for the integration of computing systems and services in the environment in which they operate require a constant sensing activity and the acquisition of different information from the environment and the users. Health monitoring, domotics, Industry 4.0 and environmental challenges leverage the availability of cost-effective sensing solutions that allow both the creation of knowledge bases and the automatic process of them, be it with algorithmic approaches or artificial intelligence solutions. The foundation of these solutions is given by the Internet of Things (IoT), and the substanding Wireless Sensor Networks (WSN) technology stack. Of course, design approaches are needed that enable defining efficient and effective sensing infrastructures, including energy related aspects.

In this paper we present a Domain Specific Language for the design of energy aware WSN IoT solutions, that allows domain experts to define sensor network models that may be then analyzed by simulation-based or analytic techniques to evaluate the effect of task allocation and offloading and energy harvesting and utilization in the network. The language has been designed to leverage the SIMTHESys modeling framework and its multiformalism modeling evaluation features.

I. INTRODUCTION

The consolidation of IoT/WSN devices as a commodity enables a widespread adoption of sensing nodes that are capable of autonomous computing and may constitute intelligent sensor networks. IoT/WSN technologies are largely used in small size installations, such in the case of home automation applications, as well as in large settlements, such in the case of smart cities; in controlled environments, such as in industrial plants within Industry 4.0 projects, and in the wild, e.g. for environmental monitoring in forests, seas, or farmlands;

in predictable situations, such as surveillance systems, and where is little information before deployment, such as in emergency response and management; in static configurations, such as in museums or campuses, and in highly dynamic and critical scenarios, such as in battlefield and tactical military operations. In the most advanced applications, IoT/WSN-based solutions are enabled to be deployed in any scenario by leveraging energy management techniques and node computing capabilities, that, at the state of the art, are non negligible with a limited cost per node and allow advanced features, such as support to fog and edge computing solutions and dynamic reconfiguration. The possibility of adopting energy harvesting, e.g. exploiting vibrations, tidal waves, wind, electromagnetic radiation, enables the design of large scale distributed reactive systems with advanced sensing capabilities that can be deployed in locations that do not provide energy grids and can survive and perform their mission for long periods without attendance or maintenance, possibly exploiting reconfiguration and task migration strategies to keep operating in degraded conditions until a minimal resource amount, in terms of energy and nodes, is available.

The design process of systems with these features must be supported by proper tools in order to ensure the compliance with non-functional specifications related to energy, coverage, resilience, survivability. Design issues include the application level, the architecture of each single node class, the definition of sensing strategies and devices, the organization of the network, the definition of the most fit networking technology and protocols, the energy management features, the reconfiguration policies, including task scheduling and offloading solutions. They belong to different domains and impact different aspects, often correlated and in conflict with each other. Even the description of models should consider abstractions that are derived from the application domain, in order to allow modelers to focus on the scenarios and have an holistic approach to the characteristics of the components of the system: for this purpose, defining a DSL (Domain Specific Language) may be a profitable choice, provided that mod-

els specified by this language can be then analyzed or processed by means of proper tools.

This work is based on the general methodology presented in [10] and proposes a possible implementation for it, that is fit to partially fulfill the goals of the ePassion research project. At the moment, it supports the definition and the evaluation of single scenarios. In fact, in this paper we present a DSL-based modeling approach that allows the description of IoT/WSN systems with energy harvesting capabilities and task offloading features, to support the evaluation of fog and edge enabled applications. The language is designed in the framework of the SIMTHESys [3] modeling approach, to exploit its solver generation features and its multiformalism capabilities. We introduce a suitable DSL and show how it can be used to describe and analyze a scenario.

After this introduction, the paper is organized as follows: Section II analyzes the WSN energy aspects to be modeled; Section III presents the general modeling approach; Section IV presents a simple application to demonstrate the approach; Section V points at some relevant literature; conclusions close the paper.

II. IoT/WSN CHARACTERISTICS AND ENERGY RELATED ISSUES

A IoT/WSN node is a low-cost embedded computer that is capable of communicating with other nodes and with a server in different network configurations by means of standard wireless and Internet technologies. The main responsibility of a node is the operation of a number of sensors, the nature of which depends on the mission the application should accomplish. While nodes may be used in scenarios in which they can be plugged into the power grid, in this paper we are interested in nodes that are powered by batteries and can harvest energy from the environment. A node has a complete software stack and might run applications in real time mode: in this paper we will only consider at the operating system level what is related to managing the problem of task allocation, execution and offloading, and at application level the tasks intended as workload to be managed at network level. Consequently, with respect to the analysis presented in [10], in this paper we consider a reduced architecture for each node with a special accent on software tasks, sensors, interaction mechanisms with the environment and task allocation strategies.

The way in which nodes use energy depends on many factors. A baseline is given by the energy needed to keep a node in stand-by mode, in the case in which its duty cycle (that is, the organization of time slots in which it is actively operating (by sensing, transmitting, computing) or waiting for the next activation event) and architecture allow a low-energy state; another contribution is needed to run application tasks; moreover, sensing and related preprocessing operations require energy as well. A significant contribution is given by network transmission, both for exchanging payload and for network management: this contribution depends on

the kind of protocols used, the chosen hardware, the frequency of reconfigurations, the interconnection services provided to other nodes, the position of the node in the network, natural obstacles, interferences and noise, frequency of data exchanges according to the application, the existence of an infrastructure, the availability and reliability of nodes, the scale of the network and the average distances between nodes. In some applications, nodes also may be equipped with motors to be able to change position, or to contrast external actions that interfere with their positioning (e.g., see [9]).

The design of energy performances of IoT/WSN systems should target the energy balance of harvesting nodes, the general energy management policy of the network, the survivability of the network: consequently, different models are needed at different design stages, depending on the level of detail, and possibly for different targets (e.g. an analytical modeling technique for a coarse grain or a global behavioral model, or a hardware-in-the-loop technique applied to a prototype for a very detailed understanding of the system). As interactions and dynamics depend on the evolution of the components, a good level of prediction quality may be achieved, once the design is sufficiently detailed, by using event-based simulation to explore alternative scenarios.

III. MODELING APPROACH

In our framework, on one side, we analyzed a number of actual scenarios from commercial or custom installations, on the other we analyzed the general domain to identify, on a metamodeling basis, the entities that characterize performance models for the domain and their behaviors (including mutual interactions), according to the SIMTHESys approach.

In general, as discussed in [10], modeling a node implies modeling the behavior of its hardware configuration, its software and scheduling, its communication related activities and the utilization and harvesting recharge processes of its power source. The environment impacts on the schedule, as well as the network configuration and organization. On these premises, we identified a set of *elements*, with their *properties* and *behaviors*, a minimal set of which is shown in Fig. 1 and Fig. 2 that is suitable to describe the most of the analyzed scenarios. Properties are attributes of an element that define its current state, while behaviors describe its dynamics and change its state. As the goal of these models is to evaluate the energy consumption according to the configuration and the course of events of a scenario, behaviors in these models are essentially related to task scheduling and migrations (in their effects on the energy available in a node and workload configuration), task activation in reaction to sensing events (in the same perspective) and energy harvesting phenomena (that raise the available energy), resulting in a state update related to the energy balance in the system, so they will not be described nor analyzed for the sake of simplicity.

The chosen modeling structure exploits hierarchical

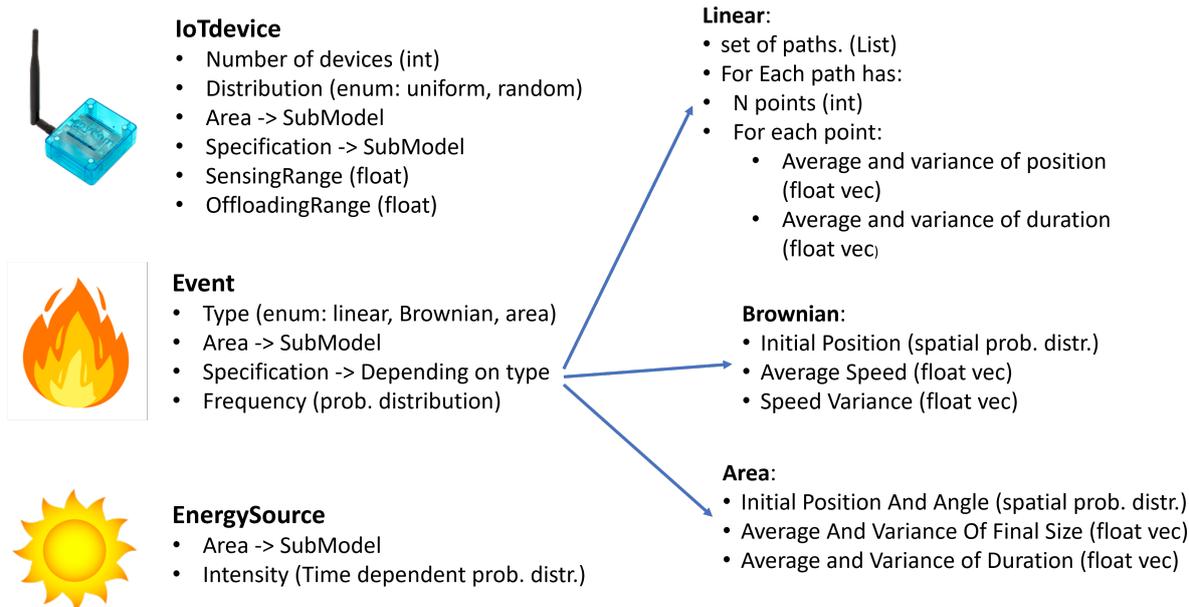


Fig. 1. The elements of the DSL and their properties

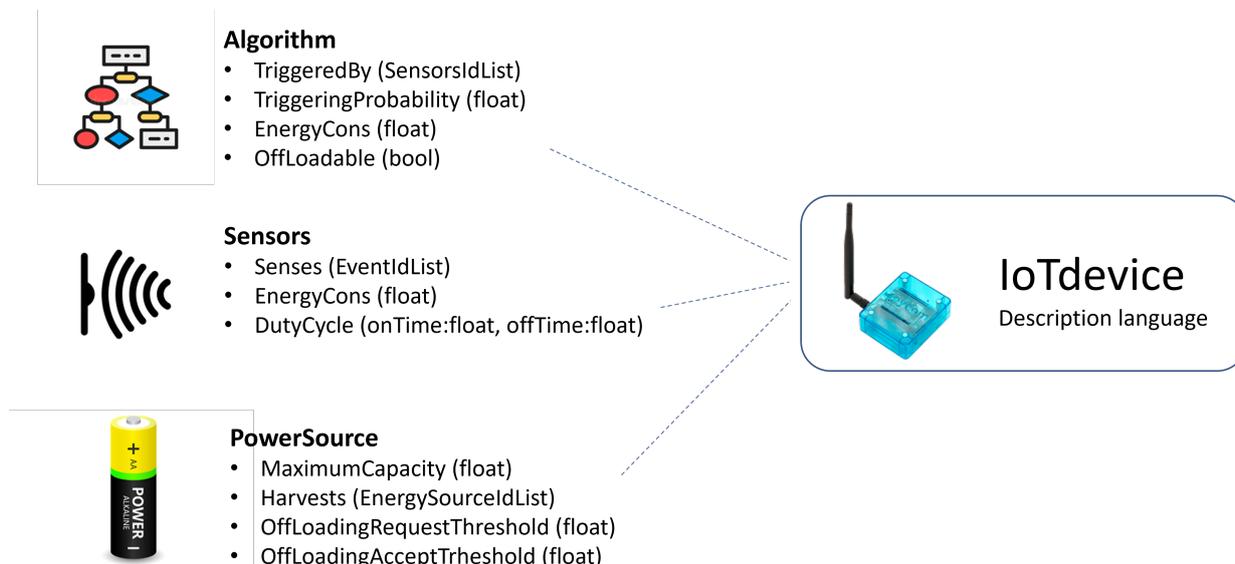


Fig. 2. The elements of the DSL describing the components of an IoTdevice and their properties

modeling, in order to simplify the modeling procedure and to better map to the domain the approach. In fact, elements such as *IoTdevice* and *Event* include submodels, that also add flexibility.

A. Model elements

Element *IoTdevice* describes a generic IoT node that can be equipped with diverse sensors and is capable of executing software. Its properties are: *NumberOfDevices*, that takes into account the number of instances of such a node which compose a scenario; *Distribution*, that provides the spatial distribution of nodes of this kind in the scenario, by means of a probability distribu-

tion¹; *SensingRange*, the radius of the area, centered on each node of this kind, in which a sensor is catching external events; *OffloadingRange*, the radius of the area, centered on each node of this kind, in which a sensor can ask a neighbour node for task offloading when the available energy is below a defined threshold; *Area*, a submodel describing the covered area; *Specification*, a submodel describing the configuration, in terms of component *Sensor*, *Algorithm* and *PowerSource* elements.

Element *Event* describes an event class that may solicit *Sensor* elements, inducing task executions on *IoTdevice* and related state changes. Its properties are: *Type*, that describes where in the scenario events are

¹In the full formalism, it is also possible to provide all positions, but the consequences are not in the scope of this work.

generated by means of a probability distribution² and the area in which the event may manifest; *Frequency*, that defines when the events manifest with a probability distribution; *Specification*, that provides additional information on the logic with which the positions in which events manifest are described; *Area*, a submodel describing the covered area. In this version of the formalism, if *Type* is Linear, *Specification* describes a set of trajectories consisting of N points, which they pass by, the average and variance around each point and the average and variance of the duration of the event around each point; if it is Brownian, the initial point of trajectories, the average speeds of a reference trajectory and their variances are provided; if it is Area, events are generated within an area and an initial position and angle and the average and variance of the final size of the covered area are provided, together with average and variance of the duration of the events generation.

Element *EnergySource* describes a type of possible source of energy which can be harvested. Properties are: *Intensity*, which is described by a time dependent probability distribution to match the oscillations of the source during the timespan of the evaluation; *Area*, a submodel describing the covered area.

Specification submodels are composed of three kinds of elements: *Algorithm* describes a workload, with properties *TriggeredBy*, a list of on-board sensors which schedule its execution, *TriggeringProbability*, the probability that, as a consequence of a sensor triggering, the workload is actually scheduled, *EnergyConsumed*, the energy amount required for a run, and *OffLoadable*, defining if this workload can be offloaded in case of need or it cannot (for privacy reasons, or because tightly connected to the given *IoTdevice* instance); *Sensor*, defining one of the on-board sensors, with properties *Senses*, providing the list of events it reacts to, *EnergyConsumed*, the energy amount required for a sensor operation, and *DutyCycle*, specifying the time duration of its active part and inactive part in its activity cycle; and *PowerSource*, describing one of the on-board energy storage units with properties *MaximumCapacity*, *Harvests*, which provides the list of *EnergySource* elements that provide it energy, *OffLoadingRequestThreshold* and *OffLoadingAcceptThreshold*, that respectively define the energy level under which the node may ask neighbors for offloading or over which it may accept offloadings from neighbors.

B. Example

Fig. 3 shows a simple scenario example.

In this example, there is a single type of *IoTdevice*, with 100 instances that are uniformly distributed in *Area Area₃*. Each of the instances has a sensing area with a radius of 50 meters and can request for offloads, or can be requested for offloads by, other instances

²In the full formalism, it is also possible to provide all positions, but the consequences are not in the scope of this work; here, due to the nature of the case study, that focuses on catching movements of different subjects in an area for surveillance and early alert, events are generated by means of trajectories.

within a radius of 200 meters. Each instance schedules and runs two workloads of two different *Algorithm* kinds, namely A_1 and A_2 , both activated by an instance of the R_1 *Sensor*. A_1 is triggered by R_1 with probability 0.1, consumes 2 mW and cannot be offloaded; A_2 is triggered by R_1 with probability 0.02, consumes 20 mW and can be offloaded. Each *IoTdevice* instance has a single *PowerSource* B_1 , with a maximum capacity of 5000 A, requesting offloading with less than 750 A available and accepting offloading with more than 2500 A available, and capable of harvesting the only *EnergySource* H_1 in the model, that radiates over *Area Area₄* with the intensity described by the formula in figure.

Two types of *Event* sources are present in the example. The first one is of Brownian type, acting in *Area₁*, in which events are exponentially distributed with rate 0.1 events per hour (details are omitted); the second one is of Linear type, with the same event rate and a trajectory description that is omitted, but graphically represented in the figure as different paths passing by a number of points in *Area₂*.

Area elements are described as polygons as in the figure.

IV. EVALUATING AN EXAMPLE SCENARIO

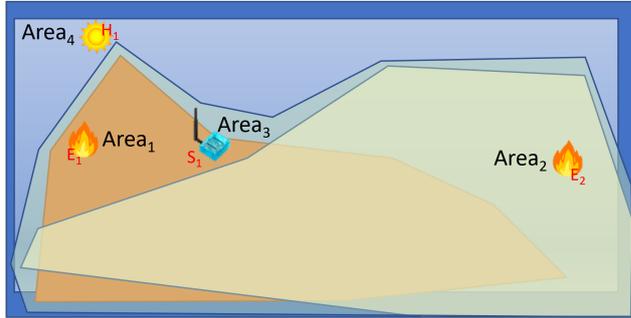
To demonstrate the approach, we evaluated the example presented in Section III-B in a scenario in which different levels of energy harvesting are considered. The system has been simulated by starting from a description in the proposed DSL and generating a simulation model by means of the SIMTHESys modeling framework, that relies on the technique described in [5], integrated in a SIMTHESys solver. Fig. 4 shows an instance in which the sensors have been randomly placed according to their number, and to the area defined by the model. Figure shows also six paths leading to events generation: three follow the linear model, and three the Brownian approach. During the motion among the considered paths, mobile entities can produce events on the monitored area, according to two different levels of severity, as shown in Fig. 5. In particular, the image shows where events are missed: the considered path and the position of the sensors, combined with the area, creates a great hole in the left side of the region. A sensor relocation, placing more devices in the area, would improve the reliability of the system, significantly reducing the number of losses.

Figures 6, 7 and 8 show the usage of batteries respectively with no harvesting, an average of 0.2 mAh obtained by harvesting and an average of 0.4 mAh. In particular, interpolating the energy decrease, it can be seen that the two harvesting can increase the lifetime of the system respectively of 66% and 400%, showing the importance of this feature: even a minimal increase in harvesting power can effectively increase the lifetime of the system. Some sensors, however, due to their position where events are more frequent, and where there are no neighbors with which they might share

Example

- 1 type of sensor
 - 1 type of energy
 - 2 types of event
- IoT devices S_1 have:
- 2 algorithms
 - 1 sensor
 - 1 energy source

- H_1 : Area: $Area_4$
- Intensity: $\delta(x-3(1+\cos(2\pi t/24))) W$



- Type: **Brownian**
- Area: $Area_1$
- Specification ->
 - Initial Position: ...
 - Average Speed: ...
 - Speed Variance: ...
- Frequency: $EXP<0.1 h^{-1}>$

- Type: **linear**
- Area: $Area_2$
- Specification ->
- Frequency: $EXP<0.1 h^{-1}>$

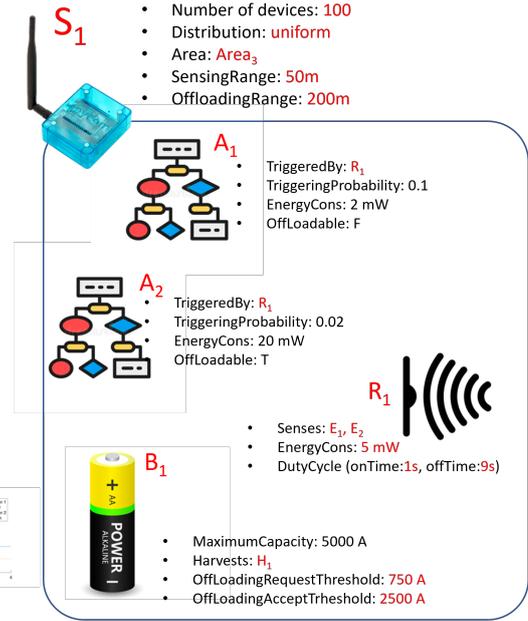


Fig. 3. An example of scenario

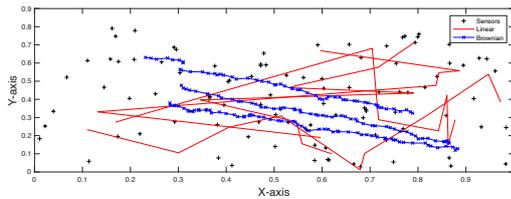


Fig. 4. Generated trajectories

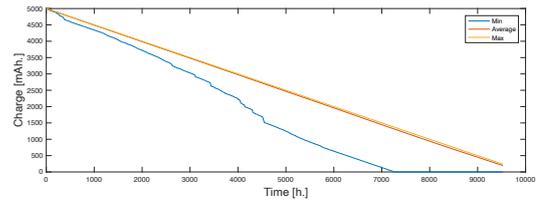


Fig. 6. No harvesting

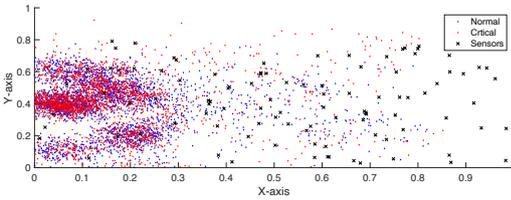


Fig. 5. Generated events

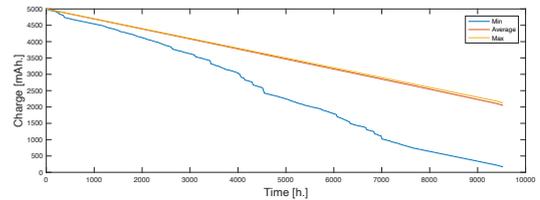


Fig. 7. Harvesting, average 0.2 mAh

their load, are more stressed than others, and tend to deplete their battery much faster than others. Again, a better placement of the devices should allow to solve this issue: the use of the proposed method in an automatic search procedure could help precisely determining the locations in which the addition of few sensors can increase the coverage and the total lifetime of the system.

V. RELATED WORK

Energy management in WSN is a topic extensively covered in literature, in which the different aspects of the subject are analyzed, regardless it remains one of the most important challenge [16]. In [26], [17] and [4], a general introduction is provided and an introduction about the main elements to define a modeling

framework is offered. All devices that are part of a WSN/IoT consume a certain amount of energy to perform their functionality [2]. This causes the batteries in WSN devices to drain quickly and require frequent battery replacement. In a distributed network, changing batteries can also be risky, costly and laborious, and impossible in some scenarios. Thus, energy harvesting is the only viable option to provide unlimited energy resources to such low-power devices in the IoT [20].

Harvesting techniques are widely used in CPS (Cyber-Physical System), because of their nature, that implies integrating, controlling and monitoring physical devices using sensors and actuators through WSN/IoT systems [15]. Various harvesting techniques are used in CPS applications to guarantee a continuous energy to devices [8]. Lastly, a review of energy sources in CPS

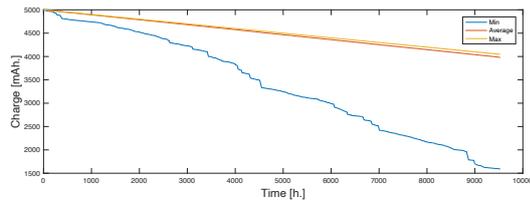


Fig. 8. Harvesting, average 0.4 mAh

is presented in [12].

In [13] and [19] an overview about power management methods is available. In [25] and [23] energy saving in protocols and network management is examined. Different energy optimization techniques have been applied to WSN, including neural networks [21][14], reinforcement learning [22], game theory [7], fuzzy logic [6] and genetic algorithms [11].

The definition of the characteristics of an IoT system may be extremely difficult to obtain in the modeling and design phases; a modeling approach is described in [18] and [24]. A more focused review on DSL methods may be found in [1].

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented a DSL for the description of energy harvesting IoT/WSN, designed to allow technicians to evaluate setups in which sensors are deployed in large areas with no access to the electrical grid, with the purpose of implementing surveillance applications. We leveraged our previous results to obtain the generation of complex scenarios, including events or visitors, from a user-friendly description. Future works include the release of a complete solution, including assisted validation tools to manage the parameterization of the models to match the characteristics of off-the-shelf sensors and nodes and the implementation of further behaviors and event categories, and the design of the needed SIMTHESys components to implement more complex offloading and energy management strategies from literature.

VII. ACKNOWLEDGEMENTS

This work has been partially funded by the internal competitive funding program “VALERE: VANviteLli pEr la RicErca” of Università degli Studi della Campania “Luigi Vanvitelli”, and is part of the research activity realized within the project PON “Ricerca e Innovazione” 2014-2020, action IV.6 “Contratti di ricerca su tematiche Green”.

REFERENCES

- [1] A. Abouzahra, A. Sabraoui, and K. Afdel. Model Composition in Model Driven Engineering: A systematic literature review. *Information and Software Technology*, 125:106316, 2020.
- [2] R. Arshad, S. Zahoor, M. Shah, A. Wahid, and H. Yu. Green iot: An investigation on energy saving practices for 2020 and beyond. *IEEE Access*, 5:15667–15681, 2017.
- [3] E. Barbierato, M. Gribaudo, and M. Iacono. Modeling hybrid systems in SIMTHESys. *Electronic Notes in Theoretical Computer Science*, 327:5 – 25, 2016.

- [4] E. Bitar, E. Baeyens, and K. Poolla. *Energy management in wireless sensor networks*. 2012.
- [5] L. Campanile, M. Gribaudo, M. Iacono, and M. Mastroianni. Hybrid simulation of energy management in iot edge computing surveillance systems. In P. Ballarini, H. Castel, I. Dimitriou, M. Iacono, T. Phung-Duc, and J. Walraevens, editors, *Performance Engineering and Stochastic Modeling - 17th European Workshop, EPEW 2021, and 26th International Conference, ASMTA 2021, Virtual Event, December 9-10 and December 13-14, 2021, Proceedings*, volume 13104 of *Lecture Notes in Computer Science*, pages 345–359. Springer, 2021.
- [6] L. Campanile, M. Iacono, F. Marulli, M. Mastroianni, and N. Mazzocca. Toward a fuzzy-based approach for computational load offloading of iot devices. *J. Univers. Comput. Sci.*, 26(11):1455–1474, 2020.
- [7] E. Campos-Nanez, A. Garcia, and C. Li. A game-theoretic approach to efficient power management in sensor networks. *Operations Research*, 56(3):552–561, 2008.
- [8] P. Castiglione, O. Simeone, E. Erkip, and T. Zemen. Energy-harvesting for source-channel coding in cyber-physical systems. In *2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE, 2011.
- [9] M. D’Arienzo, M. Iacono, S. Marrone, and R. Nardone. Petri net based evaluation of energy consumption in wireless sensor nodes. *J. High Speed Networks*, 19(4):339–358, 2013.
- [10] L. De Arcangelis, M. Iacono, and E. Lippiello. Towards a multiparadigm approach to model energy management in WSN for IoT based edge computing applications. *Communications of the ECMS*, 34(1):1–7, 2020.
- [11] K. Ferentinos and T. Tsiligiridis. Adaptive design optimization of wireless sensor networks using genetic algorithms. *Computer Networks*, 51(4):1031–1051, 2007.
- [12] G. Honan, N. Gekakis, M. Hassanaliheragh, A. Nadeau, G. Sharma, and T. Soyata. Energy harvesting and buffering for cyber physical systems: A review. *Cyber-Physical Systems: A Computational Perspective*, pages 191–217, 2015.
- [13] J. Khan, H. Qureshi, A. Iqbal, and C. Lacatus. Energy management in wireless sensor networks: A survey. *Computers and Electrical Engineering*, 41(C):159–176, 2015.
- [14] U. Kulkarni, D. Kulkarni, and H. Kenchannavar. Neural network based energy conservation for wireless sensor network. pages 1312–1316, 2018.
- [15] C. K. Lee, C. L. Yeung, and M. N. Cheng. Research on iot based cyber physical system for industrial big data analytics. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1855–1859. IEEE, 2015.
- [16] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. *IEEE Internet of Things Journal*, 4(5):1125–1142, 2017.
- [17] R. Mini and A. Loureiro. Energy-efficient design of wireless sensor networks based on finite energy budget. *Computer Communications*, 35(14):1736–1748, 2012.
- [18] P. Murali, A. Challa, M. Kasyap, and C. Hota. A generalized energy consumption model for wireless sensor networks. pages 210–213, 2010.
- [19] E. Popovici, M. Magno, and S. Marinkovic. Power management techniques for wireless sensor networks: A review. pages 194–198, 2013.
- [20] F. K. Shaikh and S. Zeadally. Energy harvesting in wireless sensor networks: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 55:1041–1054, 2016.
- [21] Y. Shen and B. Guo. Dynamic power management based on wavelet neural network in wireless sensor networks. pages 431–436, 2007.
- [22] P. Sridhar, T. Nanayakkara, A. Madni, and M. Jamshidi. Dynamic power management of an embedded sensor network based on actor-critic reinforcement based learning. pages 76–81, 2007.
- [23] S. Tyagi and N. Kumar. A systematic review on clustering and routing techniques based upon leach protocol for wireless sensor networks. *Journal of Network and Computer Applications*, 36(2):623–645, 2013.
- [24] Q. Wang and W. Yang. Energy consumption model for power management in wireless sensor networks. pages 142–151, 2007.
- [25] X.-S. Yi, P.-J. Jiang, X.-W. Wang, and S.-C. Zhang. Survey

of energy-saving protocols in wireless sensor networks. pages 208–211, 2011.

- [26] B. Zhang, R. Simon, and H. Aydin. Energy management for time-critical energy harvesting wireless sensor networks. *Lecture Notes in Computer Science*, 6366:236–251, 2010.

AUTHOR BIOGRAPHIES



LELIO CAMPANILE is a tenured teacher of Computer Science and a research associate at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy, where he has been a technician, a network administrator and an expert for many local and regional projects, and is a member of the Data and Computer Science group. He holds a M. Sc. Degree

in Computer Science and a Ph.D.. His email is lelio.campanile@unicampania.it.



MARCO GRIBAUDO is an Associate Professor in Computer Science at Politecnico di Milano, Italy. He works in the performance evaluation group. His current research interests are multi-formalism modeling, queueing networks, mean-field analysis and spatial models. The main applications to which the previous methodologies are applied comes from cloud computing, multi-core architectures and wireless sensor networks.



MAURO IACONO is an Associate Professor in Computing Systems at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy, where he leads the Computer Science section of the Data and Computer Science research group. He received the Ph.D. in Electrical Engineering from Seconda Università degli Studi di Napoli. His research activity is mainly centred on the field of performance modeling of complex computer-based systems, with special attention for multiformalism modeling techniques. His email is mauro.iacono@unicampania.it. For more information: <http://www.mauroiacono.com>.



FIAMMETTA MARULLI is an Assistant Professor in Computing Sys-

tems at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy. She works in the Data and Computer Science research group. Her research interests lie in Cognitive Computing and Artificial Intelligence methodologies applied to Deep Neural Networks design for Natural Language Processing (NLP), Data Analytics and Cyber-Physical Systems Security (CPSS) applications.



MICHELE MASTROIANNI is Assistant Professor in Computer Science at Dipartimento di Informatica of Università degli Studi di Salerno, Italy, and is also a research associate at Dipartimento di Matematica e Fisica of Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy (where he has been teaching and Data Protection Officer, Network Manager, project leader and expert for many local, regional and national technical projects), with the Data and Computer Science research group. He holds a M. Sc. degree in Electrical Engineering and a Ph.D. degree in Management Engineering. His email is mmastroianni@unisa.it.

SOME ERGODICITY AND TRUNCATION BOUNDS FOR A SMALL SCALE MARKOVIAN SUPERCOMPUTER MODEL

Rostislav Razumchik
Federal Research Center “Computer Science and Control”
of the Russian Academy of Sciences
44-2 Vavilova Str., Moscow 119333, Russian Federation,
Peoples Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation
Email: rrazumchik@ipiran.ru, razumchik-rv@rudn.ru

Alexander Rumyantsev
Institute of Applied Mathematical Research,
Karelian Research Centre of RAS
11 Pushkinskaya Str.
Petrozavodsk 185910, Russian Federation
Email: ar0@krc.karelia.ru

KEYWORDS

supercomputer model, multiserver job model, transient analysis, speed of convergence to steady state

ABSTRACT

In this paper we address the transient analysis of a markovian two-server supercomputer model where customers are served by a random number of servers simultaneously. The Markov process, which described the model’s evolution, is of quasi–birth–death type. It is shown that, at least under low load conditions, the logarithmic norm method can be used to obtain ergodicity bounds for the model. This allows one to solve both the stability detection problem (i.e. determine when the computations of the time–dependent performance measures can be terminated) and the truncation problem (i.e. locate the level at which the infinite system of Kolmogorov forward equations must be truncated in order to guarantee certain accuracy). An illustrative numerical example is provided.

INTRODUCTION

Most of the standard computing systems are now parallel. Ranging from multicore battery-powered devices to large-scale datacenters and supercomputers, all these are capable of processing compute load in parallel way. As a response to this trend, multiserver queueing models are actively studied in recent decades.

From a software perspective, parallel computing technologies are used to empower the software engineers with relevant hardware capabilities. An essential feature present both in a supercomputer and on, say, laptop is the possibility to run a specific code simultaneously on a number of cores/servers which results in overall computation time reduction. However, from the modeling perspective, such a model, called simultaneous service multiserver system (a.k.a. multiserver job model, cluster model, hereinafter referred as *supercomputer model*) is known to be hard to analyze (Harchol-Balter, 2021).

The distinctive feature of the supercomputer model is simultaneous occupation and simultaneous release of a (random) number of servers by a customer in a rigid

way (Filippopoulos and Karatza, 2007) (in contrast to classical single-server customers), causing the workload process to be non-work-conserving (Rumyantsev and Morozov, 2017). While supercomputer model is used to analyze the stability and performance characteristics of supercomputers (Morozov and Rumyantsev, 2011; Rumyantsev and Morozov, 2017), these models are also applicable to the study of social service systems (Brill and Green, 1984; Kim, 1979).

Stability (Rumyantsev and Morozov, 2017; Rumyantsev, 2020) and performance characteristics of supercomputer model are difficult to obtain even for small scale instances (Filippopoulos and Karatza, 2007; Chakravarthy and Karatza, 2013), and therefore a significant number of problems associated with such systems are open (Harchol-Balter, 2021, 2022). Surprisingly, among the most interesting problems related to the supercomputer model recently announced in (Harchol-Balter, 2022), performance analysis in transient regime is not enumerated. Yet time-dependent characteristics are rather important for energy consumption analysis, which goes in the context of energy efficiency studies focused on supercomputer model’s energy-performance tradeoff, see e.g. (Rumyantsev et al., 2021). The present paper is a step in this important direction.

Here consideration is given to time–dependent performance characteristics (e.g. average number of customers at the time instant t) of the *small scale* supercomputer model in *markovian* case. While in general this problem can be considered from the aspects of computation time, accuracy, complexity, storage etc., we establish *upper bounds* on the transient performance under light load, which constitutes the main contribution of the paper. Below we give the motivation for the importance of this result.

In transient regime, basic performance characteristics can be obtained as the solution the infinite system (see (7)) of ordinary differential equations (ODE). As such, solution techniques are closure approximations (Taaffe and Ong, 1987; Clark, 1981; Massey and Pender, 2013), uniformization (Van Dijk et al., 2018) as well as various differential equation solvers (Arns et al., 2010), to name a few. However, to improve the effi-

ciency of any solution technique, two questions need to be addressed. The first question concerns *stability detection*: time-dependent performance computation (from ODE) can be terminated once the model has reached stability regime. The second question deals with the *truncation*: the infinite system of ODEs needs to be truncated at a finite level before numerical technique can be applied. Both questions are addressed in the present paper using the well-known *logarithmic norm* method, see e.g. (Zeifman et al., 2021, Section 2). Compared to the previous studies, here we give one more evidence that the method can be applied to the analysis of the so-called quasi-birth-death (QBD) processes (with finite number of phases). Even though the particular case considered in the present paper can be treated using other methods from the literature (see, for example, (Burak and Korytkowski, 2020)), comparison of the methods, not being the goal of the paper, was not undertaken.

For the sake of brevity, further attention is paid only to the case of homogeneous servers. Although the generalization to the non-homogeneous case is possible (by following the lines drawn in the present paper), the generalization to n -server case is not straightforward and will be considered elsewhere.

The paper is organized as follows. In the next section, the detailed problem statement is given. Section 3 reviews the necessary theory, which is used in the Sections 4-6 to obtain the solutions. Some results of the numerical experiments and the main conclusions of the research are briefly summarized in Sections 7-8.

Notation

In what follows by $\|\cdot\|$ we denote the l_1 -norm, i.e. if \vec{x} is an $(l+1)$ -dimensional column vector, then $\|\vec{x}\| = \sum_{k=0}^l |x_k|$. The choice of operator norms will be the one induced by the l_1 -norm on column vectors i.e. $\|\mathbb{H}\| = \sup_{0 \leq j \leq l} \sum_{i=0}^l |h_{ij}|$ for a linear operator (matrix) \mathbb{H} . The vectors throughout the paper are regarded as column vectors (dimensions are clear from the context), $\vec{1}^T$ — row vector of 1's with T denoting the matrix transpose, \mathbb{I} — identity matrix.

THE MODEL AND THE PROBLEM STATEMENT

The schematic representation of the model considered in this paper is best given by the Fig. 1 in (Filippopoulos and Karatza, 2007). For the sake of completeness we reproduce it below.

The system consists of two identical servers and a queue of infinite capacity which follows the FIFO scheduling discipline. Customers of two classes arrive to the system according to the Poisson flow at joint rate $\lambda = \lambda p_1 + \lambda p_2$, where p_i is the probability of class- i customer arrival, $i = 1, 2$, and $p_1 + p_2 = 1$. Class- i customer requires i servers simultaneously for the same service time exponentially distributed with the rate μ (independent of customer class). In case there are insufficient resources (servers) available, the head-of-queue

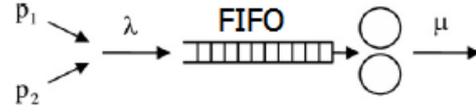


Figure 1: Two identical servers process customers from the infinite-capacity queue on the FIFO basis. For each customer i servers (with the probability p_i) are required to start service

customer prevents subsequent ones (if any) from entering service. We assume the most interesting case $p_1 \in (0, 1)$, since at the endpoints the model degenerates either to the classic $M/M/1$, or to classic $M/M/2$ queue already studied using the logarithmic norm method, see (Zeifman et al., 2019b). Thus we intentionally do not consider the boundary cases $p_1 = 0$ and $p_1 = 1$.

It is assumed that the customer class becomes known upon customer's arrival and remains unchanged during customer's sojourn time in the system. However, in the model only two oldest (in the order of arrival) customer classes are tracked. This is correct, since other customers (if any) may have generic class which indeed becomes known only upon arrival to the *head of the queue*, for more discussion of this issue see (Rumyantsev and Morozov, 2017). As such, we adopt from (Rumyantsev and Morozov, 2017) the following three-dimensional continuous-time Markov process describing the system,

$$\{S(t) = (N(t), X_1(t), X_2(t)), t \geq 0\}, \quad (1)$$

where $N(t) \geq 0$ is the number of customers in the system, and $X_i(t)$ is the class of i th oldest customer in the system at time $t \geq 0$, if any. The state space \mathcal{X} of the model consists of subsets (levels)

$$\mathcal{X} = \{0\} \cup \mathcal{X}_1 \cup \mathcal{X}_2 \dots,$$

where $\{0\}$ denotes an empty system, and the set \mathcal{X}_n , $n \geq 1$ corresponds to all possible states with n customers in the system. Thus \mathcal{X}_1 consists of two states $\{(1, 1), (1, 2)\}$ (we do not include the empty component); \mathcal{X}_n consists of four triplets (n, x_1, x_2) , where n is the total number of customers in the system, and $x_1, x_2 \in \{1, 2\}$ are the classes of oldest, second oldest customers, respectively.

As such, the process (1) is the irreducible QBD process with infinitesimal generator \mathbb{Q} of the form

$$\mathbb{Q} = \begin{pmatrix} N_0 & L_0 & 0 & 0 & \dots \\ M_0 & N & L & 0 & \dots \\ 0 & M & N & L & \dots \\ 0 & 0 & M & N & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where the blocks are defined as

$$\mathbb{N}_0 = \begin{pmatrix} -\lambda & \lambda p_1 & \lambda p_2 \\ \mu & -\lambda - \mu & 0 \\ \mu & 0 & -\lambda - \mu \end{pmatrix}, \mathbb{M}_0 = \begin{pmatrix} 0 & 2\mu & 0 \\ 0 & 0 & \mu \\ 0 & \mu & 0 \\ 0 & 0 & \mu \end{pmatrix},$$

$$\mathbb{L}_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \lambda p_1 & \lambda p_2 & 0 & 0 \\ 0 & 0 & \lambda p_1 & \lambda p_2 \end{pmatrix}, \mathbb{L} = \lambda \mathbb{L}_0,$$

$$\mathbb{N} = \begin{pmatrix} -2\mu - \lambda & 0 & 0 & 0 \\ 0 & -\mu - \lambda & 0 & 0 \\ 0 & 0 & -\mu - \lambda & 0 \\ 0 & 0 & 0 & -\mu - \lambda \end{pmatrix},$$

$$\mathbb{M} = \begin{pmatrix} 2\mu p_1 & 2\mu p_2 & 0 & 0 \\ 0 & 0 & \mu p_1 & \mu p_2 \\ \mu p_1 & \mu p_2 & 0 & 0 \\ 0 & 0 & \mu p_1 & \mu p_2 \end{pmatrix}.$$

Denote the time-dependent probability distribution of the Markov process (1) by $p_{n,i,j}(t)$ i.e.

$$\begin{aligned} p_{n,i,j}(t) &= \mathbb{P}\{N(t) = n, X_1(t) = i, X_2(t) = j\}, \quad n \geq 2, \\ p_{1,i}(t) &= \mathbb{P}\{N(t) = 1, X_1(t) = i\}, \\ p_0(t) &= \mathbb{P}\{N(t) = 0\}. \end{aligned}$$

Let

$$\vec{p}_n(t)^T = (p_{n,1,1}(t), p_{n,1,2}(t), p_{n,2,1}(t), p_{n,2,2}(t)), \quad n \geq 2,$$

$$\vec{p}_1(t)^T = (p_{1,1}(t), p_{1,2}(t)),$$

and

$$\vec{p}(t)^T = (p_0(t), \vec{p}_1(t)^T, \vec{p}_2(t)^T, \dots).$$

It is known (Brill and Green, 1984), that when the inequality $\lambda < 2\mu/(2 - p_1^2)$ holds, the model is stable and thus $\vec{p}(t) \rightarrow \vec{p}$ (element-wise) as $t \rightarrow \infty$, where the vector \vec{p} can be found from the system of global balance equations $\mathbb{Q}^T \vec{p} = \vec{0}$, $\vec{p}^T \vec{1} = 1$. The exact procedures to compute the entries of \vec{p} are already available in the literature (see, for example, (Filippopoulos and Karatzas, 2007)). In what follows we are basically interested in the two questions:

(i) given $\epsilon > 0$, find t^* such that

$$\|\vec{p}(t) - \vec{p}\| < \epsilon \text{ for } t > t^*; \quad (2)$$

(ii) given $\epsilon > 0$, find a positive integer N^* such that $\|\vec{p}(t) - \vec{p}^*(t)\| < \epsilon$, where $\vec{p}^*(t)$ denotes the time-dependent probability distribution vector of the super-computer model with the queue of finite capacity N^* (this results in changes for the corresponding generator matrix \mathbb{Q}^* given explicitly in (18)).

AUXILIARY RESULTS

In order to construct the upper bounds for $\|\vec{p}(t) - \vec{p}\|$ (and, as will be seen, for $\|\vec{p}(t) - \vec{p}^*(t)\|$ as well) we will use the notion of the logarithmic norm of locally integrable operator functions and (known) estimates for the differential equations. Consider an ODE system¹

$$\frac{d}{dt} \vec{y}(t) = \mathbb{H}(t) \vec{y}(t), \quad t \geq 0, \quad (3)$$

where the entries of the matrix $\mathbb{H}(t) = (h_{ij}(t))_{i,j=1}^\infty$ are locally integrable on $[0, \infty)$ and $\mathbb{H}(t)$ is bounded in the sense that $\|\mathbb{H}(t)\|$ is finite for any fixed t . Then

$$\frac{d}{dt} \|\vec{y}(t)\| \leq \gamma(\mathbb{H}(t)) \|\vec{y}(t)\|, \quad (4)$$

where

$$\gamma(\mathbb{H}(t)) = \sup_i \left\{ h_{ii}(t) + \sum_{j \neq i} |h_{ji}(t)| \right\}. \quad (5)$$

is called the logarithmic norm of $\mathbb{H}(t)$. Thus from (4) one gets the following upper bound²:

$$\|\vec{y}(t)\| \leq e^{\int_0^t \gamma(\mathbb{H}(u)) du} \|\vec{y}(0)\|. \quad (6)$$

STABILITY DETECTION

Given any proper initial condition $\vec{p}(0)$, the Kolmogorov forward equations for the time-dependent distribution $\vec{p}(t)$ of (1) can be written as

$$\frac{d}{dt} \vec{p}(t) = \mathbb{Q}^T \vec{p}(t), \quad t \geq 0, \quad (7)$$

with the normalization condition

$$\vec{p}(t)^T \vec{1} = 1. \quad (8)$$

It is straightforward to check that the logarithmic norm $\gamma(\mathbb{Q}^T)$ is always positive. Thus the right part of (6) grows with t and is useless in solving both (i) and (ii). As such, a specific transformation is needed which we describe below.

Fix a positive constant, say c , and a non-decreasing sequence of positive numbers $\{d_i, i \geq 1\}$ with $d_1 = 1$. Let $D_i = \prod_{n=1}^i d_n$. Introduce two infinite matrices, say \mathbb{D} and \mathbb{C} , having the form:

$$\mathbb{D} = \text{diag}(D_1, D_2, D_2, \underbrace{D_3, D_3, D_3, D_3}_{4 \text{ entries}}, \dots, \underbrace{D_n, D_n, D_n, D_n}_{4 \text{ entries}}, \dots),$$

¹The definitions and results, which are stated without any details below, can be fully recovered from, for example, (Zeifman, 1995, Appendix).

²It is worth mentioning, that for the bound (6) to hold, it is not necessary for $H(t)$ to be bounded to all $t \geq 0$. In such a case the right part of (5) is the generalization of the logarithmic norm (see (Zeifman et al., 2019a)).

$$\mathbb{C} = \begin{pmatrix} c & c & c & c & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Now consider two stochastic vectors, $\vec{p}^*(t)$ and $\vec{p}^{**}(t)$, solving (7). It follows from (8) that the solution of (7) is not affected by performing row operations on \mathbb{Q}^T and, in particular, adding or subtracting some constant (componentwise) from a single row in the matrix \mathbb{Q}^T . As such, we get the following easily verifiable identity:

$$\frac{d}{dt} \mathbb{D}(\vec{p}^*(t) - \vec{p}^{**}(t)) = \underbrace{\mathbb{D}(\mathbb{Q}^T - \mathbb{C})}_{=\mathbb{A}} \mathbb{D}^{-1} \mathbb{D}(\vec{p}^*(t) - \vec{p}^{**}(t)).$$

Assume now that such \mathbb{D} (i.e. the sequence $\{d_i, i \geq 1\}$) and \mathbb{C} exist that the logarithmic norm of the matrix \mathbb{A} is negative³, i.e.

$$\gamma(\mathbb{A}) < 0. \quad (9)$$

Then, since $d_i \geq 1$, it follows from (6) that

$$\|\vec{p}^*(t) - \vec{p}^{**}(t)\| \leq e^{\gamma(\mathbb{A})t} \|\mathbb{D}(\vec{p}^*(0) - \vec{p}^{**}(0))\|. \quad (10)$$

By plugging into the previous inequality \vec{p} instead of $\vec{p}^{**}(t)$ and solving (2) for t , one gets the solution of (i):

$$t^* = \frac{1}{\gamma(\mathbb{A})} \ln \left(\frac{\epsilon}{\|\mathbb{D}(\vec{p}(0) - \vec{p})\|} \right). \quad (11)$$

It remains to establish the conditions of existence of \mathbb{D} and \mathbb{C} that guarantee $\gamma(\mathbb{A}) < 0$, which is done in the next section.

Note that the condition (9) allows one to obtain useful insights into the model. For example, assuming that $\inf_{i \geq 1} (i^{-1} D_{i+1}) > 0$, for the average number $EN(t)$ of customers at instant t we have:

$$EN(t) = \sum_{n=1}^{\infty} n \vec{p}_n(t)^T \vec{1} \leq \left[\inf_{i \geq 1} \frac{D_{i+1}}{i} \right]^{-1} \|\mathbb{D} \vec{p}(t)\|. \quad (12)$$

By left-multiplying (7) with \mathbb{D} , one obtains the upper bound for $\|\mathbb{D} \vec{p}(t)\|$:

$$\|\mathbb{D} \vec{p}(t)\| \leq e^{\gamma(\mathbb{A})t} \|\mathbb{D} \vec{p}(0)\| + \frac{c}{\gamma(\mathbb{A})} (e^{\gamma(\mathbb{A})t} - 1). \quad (13)$$

Once the initial condition $\vec{p}(0)$ is fixed, by plugging (13) into (12) one immediately gets the upper bounds for the average number of customers in the model at instant t and for its steady state value. Since the Little's law holds for the considered model, one has at once also the upper bound for the average response time.

³Note, that even when $\gamma(\mathbb{A}) < 0$, the matrix \mathbb{A} may have negative row elements off the main diagonal.

LOGARITHMIC NORM

Let us proceed establish the conditions for (9) to hold good. Put $c = \mu$ and $d_i = d > 1$ for all $i \geq 2$. Then the right part of (5) (after plugging \mathbb{A} instead of $\mathbb{H}(t)$) becomes equal to

$$\lambda(d-1) - \mu + \mu \frac{d+1}{d^2} = \frac{f(d)}{d^2}. \quad (14)$$

It can be shown, that if $f(d) < 0$, then the range of possible value of the customer's arrival rate λ is limited to the interval

$$\lambda \in \left(0, \mu \cdot \min \left(\frac{1 + \sqrt{5}}{2}, \frac{2}{2 - p_1^2} \right) \right). \quad (15)$$

But even when (15) holds, it may happen⁴ that $f(d) > 0$. It is well-known that $f(d) = 0$ has three distinct real roots if its discriminant, say Δ , is positive. Assume⁵ $\Delta > 0$. Descartes rule of signs shows that two of the three roots of $f(d) = 0$ are positive; denote them δ_1 and δ_2 ($\delta_1 < \delta_2$). Since $f(0) > 0$ and $f(\infty) > 0$, then $f(d) < 0$ always for $d \in (\delta_1, \delta_2)$. Since $f(1) > 0$ and $f'(1) < 0$, then $(\delta_1, \delta_2) \subset (1, \infty)$. Moreover, Sturm's theorem applied to the polynomial $f(d)$ shows that $(\delta_1, \delta_2) \subset (1, 1 + \frac{\mu}{\lambda})$. Now, in order to find the proper value of $\gamma(\mathbb{A})$, it remains to choose $d \in (\delta_1, \delta_2)$, which minimizes (14) i.e.

$$\gamma(\mathbb{A}) = \min_{d \in (\delta_1, \delta_2)} \left\{ \lambda(d-1) - \mu + \mu \frac{d+1}{d^2} \right\}. \quad (16)$$

For any $p_1 \in (0, 1)$ it can be checked numerically, that the assumption $\Delta > 0$ holds whenever

$$\lambda \in \left(0, \alpha(p_1) \cdot \mu \cdot \min \left(\frac{1 + \sqrt{5}}{2}, \frac{2}{2 - p_1^2} \right) \right), \quad (17)$$

where α is the curve with the box markers in the Fig. 2 (upper curve).

The approximation for the function α allows one to estimate the fraction of the stability interval, in which it is possible⁶ to make the logarithmic norm $\gamma(\mathbb{A})$ negative: it varies from $\approx 29\%$ (when $p_1 \approx 0$) to $\approx 15\%$ (when $p_1 \approx 1$).

TRUNCATION BOUNDS

Fix a positive integer N^* and assume that the total number of customers in the model can never be greater than N^* i.e. those customers which find the queue full are considered as lost. It is convenient to describe the new model dynamics by the QBD, say $\{S^*(t), t \geq 0\}$, which is identical to (1), except for the fact that the new generator, say

⁴Because the cubic equation $f(d) = 0$ (in one variable d) does not have three real roots.

⁵If $\Delta = 0$ all roots of $f(d) = 0$ are real, with one root repeated. Due to the Descartes rule of signs the latter is positive. Thus, since $f(0) > 0$ and $f(\infty) > 0$, $f(d)$ can never be negative for $d \in (0, \infty)$.

⁶When $c = \mu$ and $d_i = d > 1$ for all $i \geq 2$.

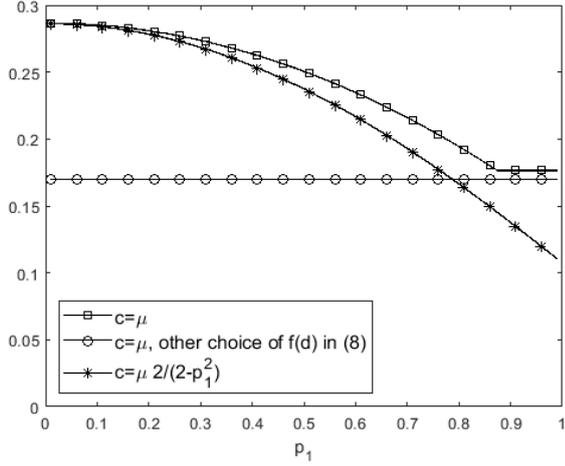


Figure 2: Dependencies of peak allowed values of α on the value of p_1 .

\mathbb{Q}^* , has the form

$$\mathbb{Q}^* = \begin{pmatrix} N_0 & L_0 & 0 & 0 & 0 & \dots \\ M_0 & N & L & 0 & 0 & \dots \\ 0 & M & N & L & 0 & \dots \\ 0 & 0 & M & -\text{diag}(M\vec{1}) & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (18)$$

Denote the time-dependent probability vector of the new model states with $\vec{p}^*(t)$; by analogy with $\vec{p}(t)$, we define it as

$$\vec{p}^*(t)^\top = (p_0^*(t), \vec{p}_1^*(t)^\top, \vec{p}_2^*(t)^\top, \dots, \vec{p}_{N^*}^*(t)^\top, 0, \dots).$$

Let us assume that the initial state probability distributions of both QBDs $\{S(t), t \geq 0\}$ and $\{S^*(t), t \geq 0\}$ are identical. Then, by applying the well-known truncation technique for birth-and-death processes (see, for example, (Zeifman et al., 2014; Satin et al., 2017)), one has that

$$\begin{aligned} \|\mathbb{D}(\vec{p}(t) - \vec{p}^*(t))\| &\leq (\|\mathbb{M} - \mathbb{N}\| + \lambda d_{N^*+1}) D_{N^*} \times \\ &\times \frac{e^{\gamma(\mathbb{A})t} - 1}{\gamma(\mathbb{A})} \sup_{u \in [0, t]} p_{N^*}^*(u). \end{aligned} \quad (19)$$

Once $\vec{p}^*(0)$ (equal to $\vec{p}(0)$) is chosen, the right-most term on the right-hand side of (19) can be computed from the (practically) finite system of ODEs $\frac{d}{dt} \vec{p}^*(t) = (\mathbb{Q}^*)^\top \vec{p}^*(t)$, with the conventional numerical methods. Clearly, since $d_i \geq 1$, $\|\vec{p}(t) - \vec{p}^*(t)\| \leq \|\mathbb{D}(\vec{p}(t) - \vec{p}^*(t))\|$. Thus (provided that the appropriate values of d_i and c are known for the given values of λ , p_1 and μ) one can answer the question (ii) from (19) by applying simple exhaustive search.

In order to highlight the usefulness of such results like (19), note that the value of N^* , which solves (ii), does not necessarily guarantee, for example, that the average number $EN^*(t) = \sum_{n=1}^{\infty} n \vec{p}^*(t)^\top \vec{1}$ of customers in

the model with the finite queue is close to $EN(t)$. But since

$$|EN(t) - EN^*(t)| \leq \frac{1}{\inf_{i \geq 1} \frac{d_{i+1}}{i}} \|\mathbb{D}(\vec{p}(t) - \vec{p}^*(t))\|,$$

one can again use (19) and exhaustive search to detect the value of N^* , which makes $EN^*(t)$ as close to $EN(t)$ as required.

NUMERICAL EXPERIMENT

In order to illustrate the findings of the previous sections, consider the following simple example. Fix the service rate $\mu = 1$ and the arrival rate $\lambda = 0.25$. Let $p_1 = 0.55$ i.e. almost half of the customers require two servers. Under these conditions the model is stable and the load is equal to ≈ 0.21 . From the Fig. 2 it can be seen that for $p_1 = 0.55$ and $\lambda = 0.25$ it is possible to choose c and $d > 1$ such that the logarithmic norm $\gamma(\mathbb{A})$ is negative. Clearly, $c = \mu$. Computations from (16) yield $d = 2.65$ and $\gamma(\mathbb{A}) = -0.0677$. Assuming that initially the model is empty, one can apply (11) to compute the instant t^* beyond which the system can be considered stable with the error less than $\epsilon = 10^{-2}$. This value ($t^* \approx 73.5$) is depicted as the vertical dashed line in the Fig. 3 alongside with the time-dependent probabilities $p_2(t)$, $p_3(t)$ and $p_4(t)$.

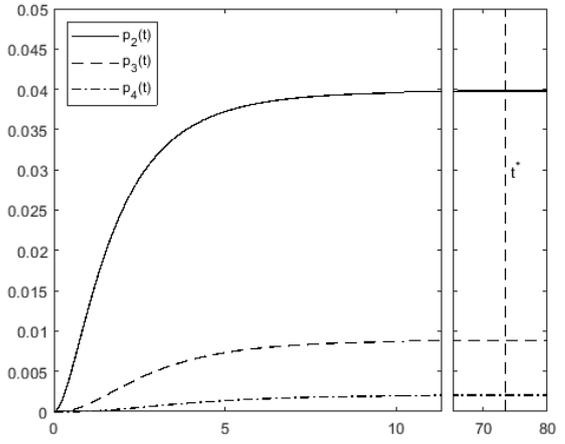


Figure 3: Time-dependent probabilities $p_2(t)$, $p_3(t)$ and $p_4(t)$ as the function of time t

The probabilities depicted in the Fig. 3 are the solutions of the systems (7), which was truncated at some high level, chosen arbitrarily. If now one uses (19) to detect the proper value of N^* (with the same ϵ), then one finds that $N^* = 16$. In the Fig. 4 one can see the behaviour of the average number of customers in the model at instant t and its upper bound according to (12). Using the steady state value of the upper bound (it is equal to ≈ 5.54) one immediately obtains from the Little's law that the customer's average sojourn time in the system is below ≈ 22.16 . As it is commented below, this bound can be significantly improved by choosing other values of c and d ; for example, for $d = 2.14$ and $c = 0.437$ we have $\gamma(\mathbb{A}) \approx -0.1522$ and $\lambda^{-1}EN \approx 5.37$.

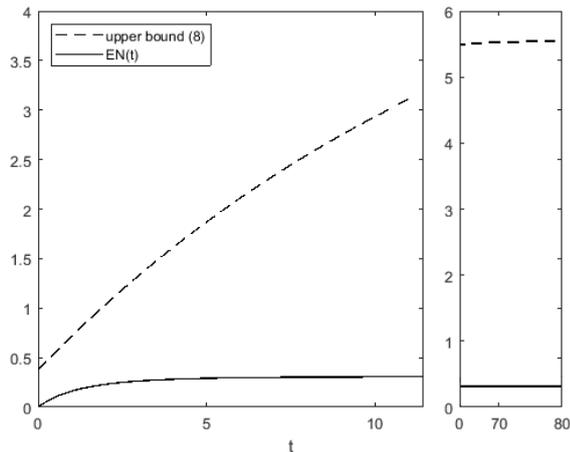


Figure 4: Average number $EN(t)$ of customers in the model at instant t and its upper bound according to (12)

SUMMARY

Even though the two questions (i) and (ii) considered above concerned certain upper bounds, the logarithmic norm method can also be used to obtain inequalities similar to (6) but reversed. This naturally leads to the solution of some questions, which involve lower bounds (see (Zeifman et al., 2018b)); for example, find an estimate of the instant $t_* > 0$ such that the model cannot become stable earlier than at $t = t_*$.

All the analysis in the paper was carried out under the assumption that the model parameters are time-independent. The most promising aspect of the adopted method is that it allows generalizations (at least) towards the time-varying arrival rate. This can clearly be seen from (6), which permits the generator to be time-dependent. Yet the conditions under which the logarithmic norm $\gamma(\mathbb{A}(t))$ is negative, are to be found. This seems to be possible at least in those settings, when the arrival rate function is bounded (such are known for a long time, (Calzarossa and Serazzi, 1985)).

One of the drawbacks of the obtained solutions to the questions (i) and (ii) is that the obtained upper bounds are not sharp. Specifically, with respect to the value t^* (see (10)–(11)), it means that the model becomes stable earlier than at instant $t = t^*$. This is clearly seen from the figures in the previous section. According to the general theory (see, for example, (Zeifman et al., 2018a; Satin et al., 2020)) this effect is due to the fact that the matrix \mathbb{A} has negative row elements off the main diagonal. In our experiments we were unable to detect the values of d_i , $i \geq 2$, which would fix the issue.

Probably the most serious defect in the solutions is that they are applicable not in the whole stability region of the model: with $c = \mu$ and $d_i = d$, $i \geq 2$, low system's load (below 0.3 when $p_1 \approx 0$ and below 0.15 when $p_1 \approx 1$) is the only feasible region (see (17) and Fig. 2). Yet there do exist several ways to transform it. One can manipulate $f(d)$ in (14), or vary the values of c and d_i , $i \geq 2$. For example, by noting that $\frac{f(d)}{d^2} < \lambda(d-1) - \mu + \frac{2\mu}{d}$ one can repeat the derivations of Section 5 and obtain that

the feasible region is $\lambda \in (0, 0.1635\mu)$ (see the line with the circle markers in the Fig. 2). It is narrower than (17): below 0.17 when $p_1 \approx 0$ and below 0.09 when $p_1 \approx 1$. If one keeps $f(d)$ in (14) and puts $c = \mu \frac{2}{2-p_1}$ instead of $c = \mu$ in \mathbb{A} (keeping $d_i = d > 1$ for all $i \geq 2$ unchanged), then the feasible region is transformed but again becomes narrower than the original (see the line with the asterisk markers in the Fig. 2). Our preliminary analysis shows that the good choice for c seems to be $c = \mu \cdot g(d, p_1)$, where a function g is defined in the semi-infinite strip $(d, p_1) \in (1, \infty) \times (0, 1)$ and is everywhere positive and non-increasing in d . Computations show that when the feasible region is made wider (by manipulating c), the value of $\gamma(\mathbb{A})$ becomes smaller (invoking too pessimistic upper bounds). Thus, if the described steps are followed, a trade-off must exist between the applicability and the desired accuracy. Filling the gaps here, as well as the choice of d_i , are the directions of further research.

Acknowledgements The research has been prepared with the support of Russian Science Foundation according to the research project No. 21-71-10135.

REFERENCES

- Arns, M., P. Buchholz, and A. Panchenko. 2010. On the numerical analysis of inhomogeneous continuous-time Markov chains. *INFORMS Journal on Computing*. Vol. 22. No. 3. Pp. 416–432.
- Brill, P. H., and L. Green. 1984. Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Management Science*. Vol. 30. No. 1. Pp. 51–68. DOI: 10.1287/mnsc.30.1.51
- Burak, M. R., and P. Korytkowski. 2020. Inhomogeneous CTMC birth-and-death models solved by uniformization with steady-state detection. *ACM Transactions on Modeling and Computer Simulation*. Vol. 30. No. 3. Pp. 1–18.
- Calzarossa, M., and G. Serazzi. 1985. A characterization of the variation in time of workload arrival patterns. *IEEE Trans. Comput.* Vol. C-34. No. 2. Pp. 156–162.
- Chakravarthy, S.R., and H.D. Karatza. 2013. Two-server parallel system with pure space sharing and markovian arrivals. *Computers & Operations Research*. Vol. 40. No. 1. Pp. 510–519.
- Clark, G. M. 1981. Use of Polya distributions in approximate solutions to nonstationary $M/M/s$ queues. *Communications of the ACM*. Vol. 24. No. 4. Pp. 206–217.
- Filippopoulos, D., and H. Karatza. 2007. An $M/M/2$ parallel system model with pure space sharing among rigid jobs. *Mathematical and Computer Modelling*. Vol. 45. Pp. 491–530.
- Harchol-Balter, M. 2021. Open problems in queueing theory inspired by datacenter computing. *Queueing Systems*. 2021. Vol. 97. Pp. 3–37.
- M. Harchol-Balter. 2022. The multiserver job queueing model. *Queueing Systems*. DOI: 10.1007/s11134-022-09762-x.

- Kim, S. 1979. *M/M/s* queueing system where customers demand multiple server use. PhD Thesis. Southern Methodist University.
- Massey, W. and J. Pender. 2013. Gaussian skewness approximation for dynamic rate multiserver queues with abandonment. *Queueing Systems*. Vol. 75. No. 2. Pp. 243–27.
- Morozov, E. V., and A. S. Rumyantsev. 2011. Multi-server models to analyze high performance cluster. *Transactions of Karelian Research Centre of the Russian Academy of Sciences*. No. 5. Pp. 75–85. (in Russian)
- Rumyantsev, A., and E. Morozov. 2017. Stability criterion of a multiserver model with simultaneous service. *Ann. Oper. Res.* Vol. 252. No. 1. Pp. 29–39.
- Rumyantsev A. 2020. Stability of Multiclass Multi-server Models with Automata-type Phase Transitions. *Proceedings of the Second International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY 2020)*. pp 213225. URL: <http://ceur-ws.org/Vol1-2792/#paper16>
- Rumyantsev A. et al. 2021. A Three-Level Modelling Approach for Asynchronous Speed Scaling in High-Performance Data Centres. *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. Association for Computing Machinery, New York, NY, USA, pp 417423. DOI: 10.1145/3447555.3466580
- Satin, Y., K. Kiseleva, S. Shorgin, V. Korolev, and A. Zeifman. 2017. Two-sided truncations for the $M_i/M_i/S$ queueing model. *Proceedings of the 31st European Conference on Modelling and Simulation*. Pp. 635–641.
- Satin, Ya., A. Zeifman, and G. Shilova. 2020. On approaches to constructing limiting regimes for some queueing models. *Informatika i ee Primeneniya — Inform. Appl.* Vol. 15. No. 2. Pp. 19–24.
- Taaffe, M. R., and K. L. Ong. 1987. Approximating nonstationary $Ph(t)/M(t)/s/c$ queueing systems. *Annals of Operations Research*. Vol. 8. No. 1. Pp. 103–116.
- Van Dijk, N. M., S. P. J. van Brummelen, and R. J. Boucherie. 2018. Uniformization: Basics, extensions and applications. *Performance evaluation*. Vol. 118. Pp. 8–32.
- Zeifman, A., Y. Satin, I. Kovalev, R. Razumchik, and V. Korolev. 2021. Facilitating numerical solutions of inhomogeneous continuous time Markov chains using ergodicity bounds obtained with logarithmic norm method. *Mathematics*. Vol. 9. No. 1. Art. ID 42. 20 p.
- Zeifman, A.I. 1995. Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. *Stoch. Proc. Appl.* Vol. 59. Pp. 157–173.
- Zeifman, A., Y. Satin, K. Kiseleva, and V. Korolev. 2019. On the rate of convergence for a characteristic of multidimensional birth–death process. *Mathematics*. Vol. 7. Iss. 5. Art. ID 477. 10 p.
- Zeifman, A., Y. Satin, V. Korolev, and S. Shorgin. 2014. On truncations for weakly ergodic inhomogeneous birth and death processes. *Int. J. Appl. Math. Comput. Sci.* Vol. 24. Pp. 503–518.
- Zeifman, A., A. Sipin, V. Korolev, G. Shilova, K. Kiseleva, A. Korotysheva, and Y. Satin. 2018. On sharp bounds on the rate of convergence for finite continuous–time markovian queueing models. *LNCS*. Vol. 10672. Pp. 20–28.
- Zeifman, A. I., V. Y. Korolev, Y. A. Satin, and K. M. Kiseleva. 2018. Lower bounds for the rate of convergence for continuous–time inhomogeneous Markov chains with a finite state space. *Statistics & Probability Letters*. Vol. 137. Pp. 84–90.
- Zeifman, A., Y. Satin, K. Kiseleva, V. Korolev, and T. Panfilova. 2019. On limiting characteristics for a non-stationary two-processor heterogeneous system. *Applied Mathematics and Computation*. Vol. 351(C). Pp. 48–65.

AUTHOR BIOGRAPHIES

ROSTISLAV RAZUMCHIK received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as the leading research fellow at the Institute of Informatics Problems of the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS). 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

RUMYANTSEV ALEXANDER received his PhD from Petrozavodsk State University. He is now a researcher in the Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences. His research interests include Stochastic Processes, Queueing Systems, High-Performance and Distributed Computing, Multi-Core and Many-Core Systems. His email address is ar0@krc.karelia.ru.

EFFECT OF IMPURITIES ON STABILITY OF THE SKYRMION PHASE IN A FRUSTRATED HEISENBERG ANTIFERROMAGNET

Mariia Mohyl'na and Milan Žukovič
Institute of Physics
Faculty of Science
Pavol Jozef Šafárik University in Košice
041 54 Košice, Slovakia
Email: mariia.mohyl'na@upjs.sk

KEYWORDS

Heisenberg antiferromagnet; Geometrical frustration; Skyrmion lattice; Nonmagnetic impurities; Hybrid Monte Carlo

ABSTRACT

We employ a hybrid Monte Carlo simulation implemented on GPU to study the effect of nonmagnetic impurities in a frustrated Heisenberg antiferromagnetic (AFM) model on a triangular lattice with Dzyaloshinskii-Moriya interaction in the presence of the external magnetic field. We focus on the skyrmion lattice phase (SkX), which in the pure model is known to be stabilized in a quite wide temperature-field window. We aim to confront the effect of impurities on the SkX phase in the present frustrated AFM model with that in the nonfrustrated ferromagnetic counterpart as well as to consider more realistic conditions in the proposed experimental realizations of the present model. We show, that up to a fairly large concentration of the impurities, $p \approx 35\%$, the SkX phase can survive albeit in somewhat distorted form. Distortion of the SkX phase due to formation of bimerons, reported in the ferromagnetic model, was not observed in the present case.

INTRODUCTION

Magnetic skyrmions, topologically nontrivial twisted magnetic spin configurations, have recently attracted a lot of attention due to the wide variety of properties, which make them promising candidates for the new generation of memory storage devices (Zhang et al. 2015b), logic gates (Zhang et al. 2015a), microwave detectors (Finocchio et al. 2015) and others. Similar to the well-known for decades one-dimensional topological objects - domain walls - and two-dimensional magnetic vortexes, skyrmions carry a certain property called topological charge or topological number, which sets them apart from the topologically trivial spin textures like ferromagnetic (FM) or antiferromagnetic (AFM) order and distinguishes one topological object from another. The idea of the existence of topologically-nontrivial structures in magnetic materials was actively theoretically developed in the end of the previous century (Belavin and Polyakov 1975; Bogdanov and Yablonskii 1989; Roessler et al. 2006), but it was not until 2009 when the first direct experimental proof of the presence of the hexagonal skyrmion crystal phase in a bulk ferromagnet was obtained in 2009 by Mühlbauer et al. (2009). After that the search for new skyrmion-hosting materials began in order to identify most promising materials for technological implementa-

tion. They were found in Hall ferromagnets, ferromagnetic monolayers, multilayers and ferrimagnets.

Stability of skyrmions is guaranteed by topological charge as it is an invariant. However, in real materials it is not absolute and thermal fluctuations can bring the system over the finite energy barrier separating one spin configuration from another. The reason for the stabilization of the skyrmion lattice (SkX) state is usually the presence of Dzyaloshinskii-Moriya interaction (DMI) (Dzyaloshinsky 1958; Moriya 1960), which breaks the inversion symmetry. Nevertheless, there are some other mechanisms, that can lead to the formation of skyrmions, among which of particular interest are frustrated interactions. It was recently demonstrated by Okubo et al. (2012), that such interactions are capable of stabilizing both skyrmion and antiskyrmion crystals on any lattices of the trigonal symmetry with next-nearest interactions. Although skyrmions were first encountered in FM materials, recently the focus shifted to the alternatives (Barker and Tretiakov 2016; Bessarab et al. 2019), that were proven to be capable of hosting the SkX phase. It was shown by Rosales et al. (2015) and Osorio et al. (2017), that SkX can be stabilized in the classical Heisenberg AFM on triangular lattice with moderate DMI in a quite wide temperature-field window due to the combined effect of the frustration and the DMI. Further studies demonstrated the possibility of the SkX phase stabilization even at very small values of DMI (Mohyl'na and Žukovič 2020; Mohyl'na et al. 2021).

The presence of nonmagnetic impurities (spin vacancies) is a common feature in magnetic solids. In the case of frustrated spin systems with a ground-state degeneracy the problem of collective impurity behaviour can become rather nontrivial due to a possible "order by quenched disorder" effect with a profound impact on the phase diagram. In particular, for the classical Heisenberg AFM on a triangular lattice in an external magnetic field but without DMI it has been shown that competition between thermal fluctuations and nonmagnetic impurities leads to a complicated temperature-field phase diagram with the emergence of a conical state at low temperatures (Maryasin and Zhitomirsky 2013). Considerable effect of nonmagnetic impurities has also been observed in the nonfrustrated FM Heisenberg model on a square lattice in the field with DMI (Silva et al. 2014), which displays the SkX phase. In particular, it was found that even very tiny concentrations of the vacancies induce the formation of bimerons in both helical (HL) and SkX states. In the considered system they show up as elongated configurations similar to a skyrmion with its disk-shaped central core

shared by two half disks separated by a rectangular stripe domain. The presence of bimerons is found to cause deformation of both the HL and SkX states. While in the former case it occurs due to their appearance between the vacancies, thus breaking stripe-domain structures, in the latter case bimerons make the skyrmion positions in the skyrmion lattice change in a nontrivial way and decrease their overall number. The nonmagnetic impurities thus distort both the skyrmion configuration and the skyrmion lattice.

Monte Carlo (MC) simulations are among the most powerful tools in studying the behaviour of the magnetic materials at finite temperatures and, in particular, identifying the most promising candidates for the skyrmion-hosting environment. It is very important, though, to emulate the real materials used in experiments, as closely as possible. Although some studies on the Heisenberg AFM with DMI already attempted to introduce some more realistic features, such as the presence of a single-ion anisotropy, (Fang et al. 2021; Mohylina and Žukovič 2022) and quantum effects (Liu et al. 2020), the effect of nonmagnetic impurities, very common in real materials, remains to be investigated. In this work we study the influence of nonmagnetic impurities on the stability of the SkX phase in the frustrated classical Heisenberg AFM on a triangular lattice in the presence of the DMI by means of the hybrid Monte Carlo simulations. Since the problem is computationally very demanding and allows for parallelization, the simulations are implemented on a highly parallelized architecture of GPU using CUDA programming language.

MODEL AND METHOD

We investigate the classical Heisenberg AFM on a triangular lattice with the following Hamiltonian

$$\mathcal{H} = -J \sum_{\langle i,j \rangle} \vec{S}_i \cdot \vec{S}_j + \sum_{\langle i,j \rangle} \vec{D}_{ij} \cdot [\vec{S}_i \times \vec{S}_j] - h \sum_i S_i^z, \quad (1)$$

where \vec{S}_i is a classical unit-length Heisenberg spin at the i th site, $J < 0$ is the AFM exchange coupling constant, h is the external magnetic field applied perpendicular to the lattice plane (along the z direction) and $\langle i,j \rangle$ denotes the sum over nearest-neighbour spins. \vec{D}_{ij} is the DMI vector whose orientation is defined by the crystal symmetries. In this study it is chosen to point along the radius-vector $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$ between two neighbouring sites, i.e., $\vec{D}_{ij} = D \frac{\vec{r}_{ij}}{|\vec{r}_{ij}|}$ (Fig. 1), which results into the formation of the Bloch-type skyrmions. The magnitude of the parameter D defines the strength of the DMI. The presence of nonmagnetic impurities is simulated by randomly replacing a certain percentage p of spins on the lattice with vacancies. In the following we set $J = -1$ to fix the energy scale and absorb the Boltzmann constant in temperature by setting its value to $k_B = 1$.

In order to identify the presence of the SkX phase we use the skyrmion chirality, a discretization of a continuum topological charge (Berg and Lüscher 1981), which reflects the number and the nature of topological objects present in the system. The topological charge of a single skyrmion is ± 1 for the core magnetization $\pm |\vec{S}|$ (Zhang et al. 2020). The skyrmion chirality κ and the corresponding susceptibility χ_κ are defined as follows:

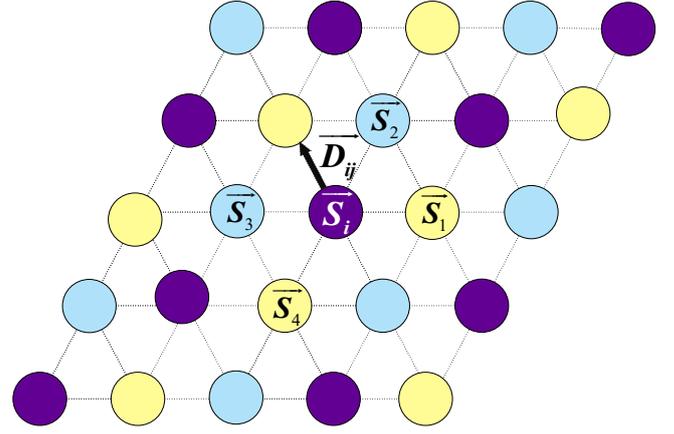


Fig. 1. Three-sublattice decomposition of the triangular lattice shown by different colors. \vec{S}_i is the central spin, $\vec{S}_1, \dots, \vec{S}_4$ are the spins involved in calculation of the local chirality, and \vec{D}_{ij} represents the Dzyaloshinskii-Moriya vector.

$$\kappa = \frac{\langle K \rangle}{N} = \frac{1}{8\pi N} \left\langle \sum_i (\kappa_i^{12} + \kappa_i^{34}) \right\rangle, \quad (2)$$

$$\chi_\kappa = \frac{\langle K^2 \rangle - \langle K \rangle^2}{NT}, \quad (3)$$

where $\kappa_i^{ab} = \vec{S}_i \cdot [\vec{S}_a \times \vec{S}_b]$ is the chirality of a triangular plaquette of three neighbouring spins (area of the triangle spanned by those spins) and $\langle \dots \rangle$ denotes the thermal average. The chirality is calculated for the whole lattice and the summation runs through all the spins with $\{\vec{S}_a, \vec{S}_b\}$ corresponding to $\{\vec{S}_1, \vec{S}_2\}$ and $\{\vec{S}_3, \vec{S}_4\}$ in Fig. 1. Spins are taken in counter-clockwise fashion to keep the sign in accordance with the rules in (Berg and Lüscher 1981). For construction of the phase diagram it is useful to calculate some other basic thermodynamic quantities, such as the magnetization m , the magnetic susceptibility χ_m , and the specific heat c , as follows:

$$m = \frac{\langle M \rangle}{N} = \frac{1}{N} \left\langle \sum_i S_i^z \right\rangle, \quad (4)$$

$$\chi_m = \frac{\langle M^2 \rangle - \langle M \rangle^2}{NT}, \quad (5)$$

$$c = \frac{\langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2}{NT^2}. \quad (6)$$

In order to identify the presence of the skyrmion phase and construct the phase diagrams we implement the hybrid Monte Carlo (HMC), which combines the standard Metropolis algorithm with the over-relaxation (OR) method (Creutz 1987). The OR method is a deterministic energy preserving perturbation method, which leads to the faster relaxation of the system due to faster decorrelation. To perform configurational averaging we run independent simulations on 50 replicas with different configurations of randomly distributed nonmagnetic impurities for each value of the impurity concentration p . In simulations we use $2 \cdot 10^6$ MC sweeps, half of which are used for the equilibration. The lattice size in all the simulations is $L = 48$ and periodic boundary conditions are implemented. Due to high

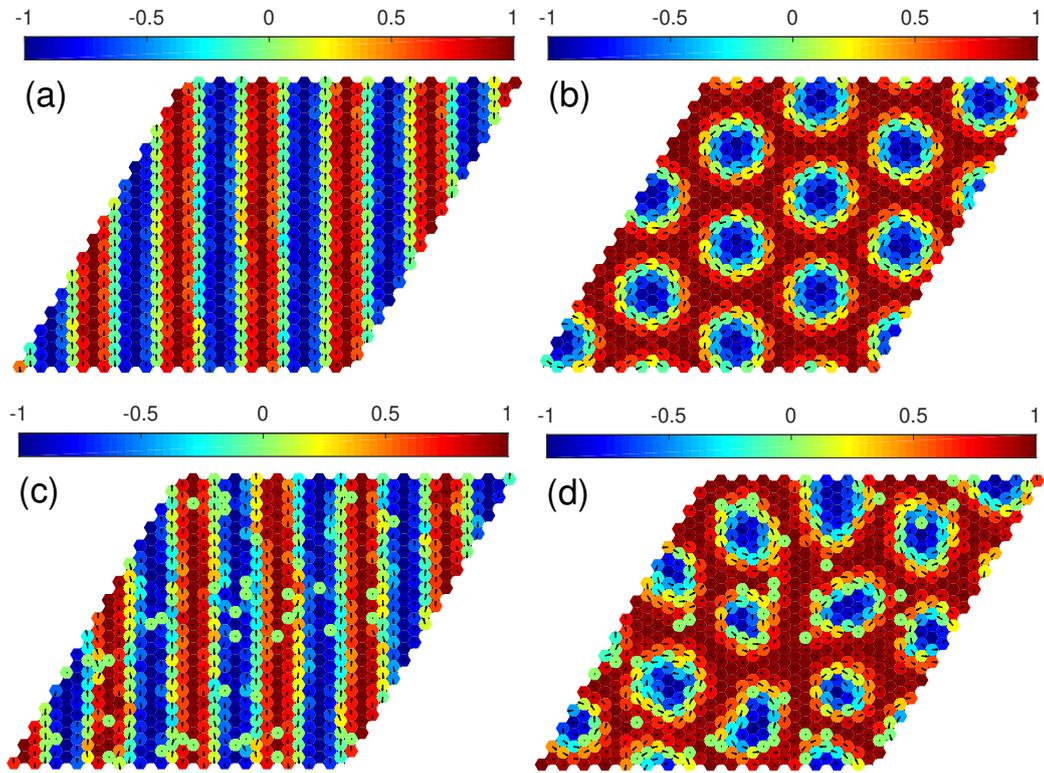


Fig. 2. (a,c) HL states at $h = 1.6$ and (b,d) SkX states at $h = 2.4$ for (a,b) pure system and (c,d) with $p = 5\%$ of vacancies. The remaining parameters are $T = 0.01$, $D = 0.5$, and $L = 48$. The green circles in (c,d) represent the vacancy locations. The snapshots are taken for one of the replicas.

computational demands the simulations are carried out on General Purpose Graphical Processing Units (GPGPU) using CUDA, which allows for the massive parallelization of the calculations.

RESULTS

Pure model

The frustrated Heisenberg AFM with DMI has been previously intensively studied in a wide parameter space (Rosales et al. 2015; Mohylina and Žukovič 2020; Mohylina et al. 2021). The phase diagram of a pure isotropic model with a moderate DMI ($0.2 < D < 1$) consists of three distinct phases: the helical (HL) phase with chiral stripes rotating in the $x - y$ plane; the skyrmion lattice (SkX) phase, which consists of three interpenetrating skyrmion lattices on each of the sublattices and the V-like (VL) phase (Mohylina et al. 2021). The typical snapshots of the HL and the SkX phases on one of the sublattices in the case of no impurities are depicted in Fig. 2 (a) and (b), respectively. The colours represent the values of the z -component: the red spins are those pointing along the external magnetic field and the blue ones pointing opposite to it. The arrows show the projections of the spins to the $x - y$ plane. The representative phase diagram for the model without impurities with $D = 0.5$ is shown in Fig. 3 in black circles. The SkX phase occupies a relatively big part of the $T - h$ plane and emerges at the fields around $h \approx 0.24$ for the lowest temperature, which is reflected in a sharp increase of the chirality signaling the first-order phase transition (Rosales et al. 2015), as also shown in Fig. 4(a). The increase of temperature results in some distortion of the skyrmions' profile and conse-

quently it leads to smoothing of the chirality and the change of the transition type to the second-order one (Mohylina et al. 2021).

Effect of impurities

To simulate the presence of nonmagnetic impurities in our system we randomly replaced $p\%$ of the sites with vacancies and studied their effect on the HL and SkX states. In Fig. 2 we present the respective states in the pure system (a,b) and with $p = 5\%$ of nonmagnetic impurities (c,d) on one of the three interpenetrating sublattices in the HL phase (a,c) and the SkX phase (b,d). The snapshots for one of the running replicas are shown. We can observe that, compared to the pure systems, the presence of the vacancies naturally results in some distortion of the respective phases. Nevertheless, there are no signs of the formation of bimerons, as it was the case in the FM model (Silva et al. 2014). We believe that the present AFM system is more resilient against creation of bimerons than the FM one due to the fact that both HL and SkX textures are formed on each of the three interpenetrating sublattices. Therefore, for example the stripe-domain structures in the HL state are more robust and it is more difficult for nonmagnetic impurities to break them into pieces (bimerons) than in the FM system. In the SkX phase, one can notice that, besides the skyrmion distortion, the presence of vacancies also reduces their number. It is interesting to study their mutual interaction. It is known that isolated skyrmions in isotropic 2D Heisenberg magnets are attracted to the spin vacancies (Subbaraman et al. 1998; Pereira and Pires 2003), which allows an impurity to be at a skyrmion center. On the other hand, skyrmions in a skyrmion crystal do not generally appear centered at vacancies. In such

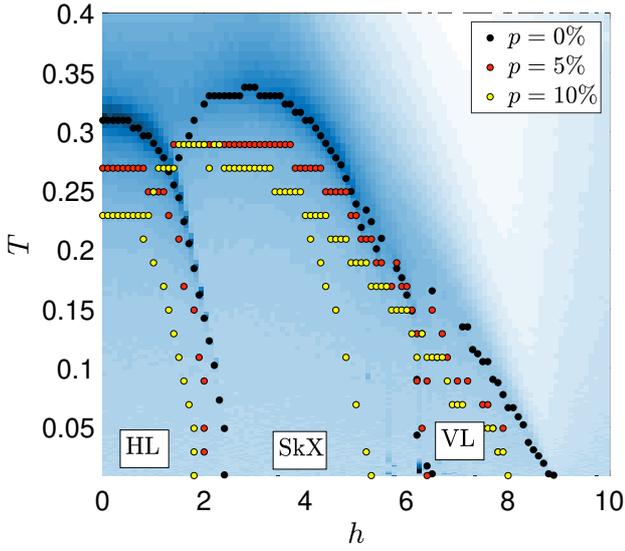


Fig. 3. Phase diagrams of the Heisenberg AFM model in $T-h$ plane with $D = 0.5$ and different p . The black, red and yellow circles represent the cases with $p = 0\%$, 5% and 10% , respectively.

case it is energetically more favorable if configurations involve the whole crystal instead of isolated skyrmions and, therefore, the skyrmion lattice is formed in such a way that the spin vacancies become, in general, localized between skyrmions. This phenomenon can be also observed in the present case (see Fig. 2(d)) when the skyrmions have a tendency to rearrange in such way so that the vacancies stay at their outskirts.

In Fig. 4 we plot field and temperature dependencies of the chirality for selected values of the concentration of non-magnetic vacancies p . The presence of impurities leads to distortion of the skyrmion profiles and consequently reduction of the magnitude of the chirality. At the same time it smears out its abrupt change, particularly at the HL-SkX phase transition, with smaller but finite values within the higher-field part of the HL phase (see Fig. 4(a)). The effect of the chirality reduction with the increasing concentration of the nonmagnetic impurities is also demonstrated in Fig. 4(b), in which the chirality is plotted versus temperature for a wide range of the concentration p . Similar effect of distortion of the skyrmion profiles and reduction of the chirality is also produced by thermal fluctuations. Our rough estimate of the critical threshold of the magnetic dilution below which no skyrmions can survive even at the lowest temperatures is $p_c \approx 35\%$. This value is in a good correspondence with the value estimated for the FM model on the square lattice (Silva et al. 2014).

The phase diagrams for $p = 5\%$ (red circles) and $p = 10\%$ (yellow circles) are presented in Fig. 3. Their overall topology remains unchanged as compared to the case of the pure model and consists of the HL phase in the low-field region, the VL phase in the high-field region and the SkX phase sandwiched between them. However, a tendency of the SkX phase to shrink and smear with the increasing number of impurities can be noticed. Nevertheless, the area of the SkX phase still occupies a big part of a phase diagram even for $p = 10\%$.

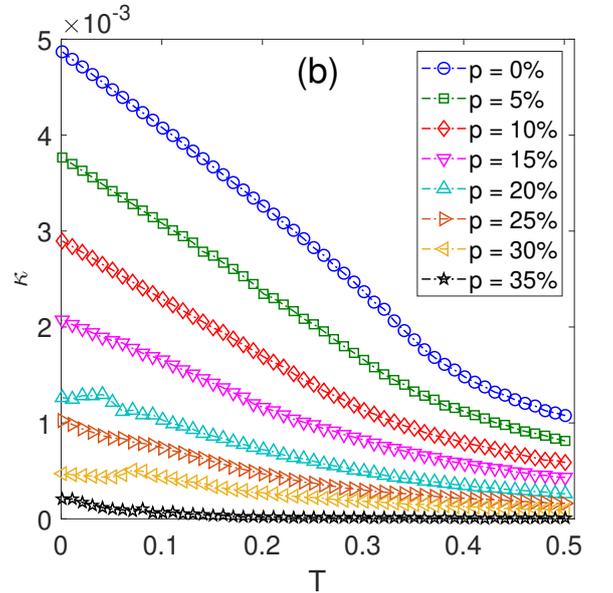
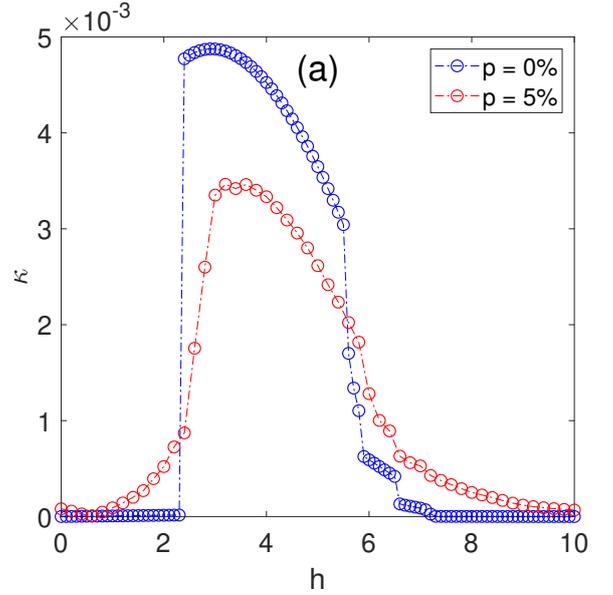


Fig. 4. (a) Field dependence of the chirality at $T = 0.01$ and (b) temperature dependence of the chirality at $h = 3.4$, for selected values of the concentration of nonmagnetic vacancies p .

CONCLUSIONS

In this study we investigated the effect of the presence of nonmagnetic impurities in the frustrated Heisenberg AFM model with the DMI, hosting the SkX phase. The purpose was to confront the effects of impurities on the SkX phase in the present frustrated AFM model with that in the non-frustrated FM model (Silva et al. 2014) and also to take into account a more realistic situation in the possible experimental realizations of the present model (Fang et al. 2021). We focused on the influence of impurities on the shape, size and position of individual skyrmions on the lattice as well as overall stability of the SkX phase. Our findings suggest, that both the HL and the SkX phases in the present frustrated AFM system are more robust to the distortion, induced by those impurities, than in the nonfrustrated FM counterpart, where the formation of bimerons occurs already at low con-

centrations of $p \approx 1\%$. In our case the skyrmion lattice is formed in such a way, that the vacancies become localized in-between the skyrmions. The effect caused by the presence of impurities is similar to the one caused by thermal fluctuations: the skyrmions shape is distorted and the chirality is both reduced and smeared out, in particular at the HL-SkX phase boundary. This tendency becomes more pronounced with the increasing percentage of the impurities. We estimate that the threshold below which no SkX phase can be stabilized is $p_c \approx 35\%$.

ACKNOWLEDGMENTS

This work was supported by the grants of the Slovak Research and Development Agency (Grant No. APVV-20-0150) and the Scientific Grant Agency of Ministry of Education of Slovak Republic (Grant No. 1/0531/19). Part of computations was held on the basis of the HybriLIT heterogeneous computing platform LIT, JINR (Adam et al. 2018).

REFERENCES

- Adam, G., Bashashin, M., Belyakov, D., Kirakosyan, M., Matveev, M., Podgainy, D., Sapozhnikova, T., Streltsova, O., Torosyan, S., Vala, M., et al. (2018). It-ecosystem of the hybridit heterogeneous platform for high-performance computing and training of it-specialists. In *English, in CEUR Workshop Proceedings, V. Korenkov, A. Nechaevskiy, T. Zaikina, and E. Mazhitova, Eds*, volume 2267, pages 638–644.
- Barker, J. and Tretiakov, O. A. (2016). Static and dynamical properties of antiferromagnetic skyrmions in the presence of applied current and temperature. *Physical review letters*, 116(14):147203.
- Belavin, A. and Polyakov, A. (1975). Metastable states of two-dimensional isotropic ferromagnets. *JETP lett*, 22(10):245–248.
- Berg, B. and Lüscher, M. (1981). Definition and statistical distributions of a topological number in the lattice σ -model. *Nuclear Physics B*, 190(2):412–424.
- Bessarab, P., Yudin, D., Gulevich, D., Wadley, P., Titov, M., and Tretiakov, O. A. (2019). Stability and lifetime of antiferromagnetic skyrmions. *Physical Review B*, 99(14):140411.
- Bogdanov, A. and Yablonskii, D. (1989). Thermodynamically stable “vortices” in magnetically ordered crystals. the mixed state of magnets. *Zh. Eksp. Teor. Fiz.*, 95(1):182.
- Creutz, M. (1987). Overrelaxation and monte carlo simulation. *Physical Review D*, 36(2):515.
- Dzyaloshinsky, I. (1958). A thermodynamic theory of “weak” ferromagnetism of antiferromagnetics. *Journal of physics and chemistry of solids*, 4(4):241–255.
- Fang, W., Raeliarijaona, A., Chang, P.-H., Kovalev, A. A., and Belashchenko, K. D. (2021). Spirals and skyrmions in antiferromagnetic triangular lattices. *Physical Review Materials*, 5(5):054401.
- Finocchio, G., Ricci, M., Tomasello, R., Giordano, A., Lanuzza, M., Puliafito, V., Burrascano, P., Azzerton, B., and Carpentieri, M. (2015). Skyrmion based microwave detectors and harvesting. *Applied Physics Letters*, 107(26):262401.
- Liu, Z., dos Santos Dias, M., and Lounis, S. (2020). Theoretical investigation of antiferromagnetic skyrmions in a triangular monolayer. *Journal of Physics: Condensed Matter*, 32(42):425801.
- Maryasin, V. and Zhitomirsky, M. (2013). Triangular antiferromagnet with nonmagnetic impurities. *Physical review letters*, 111(24):247201.
- Mohylna, M., Buša Jr, J., and Žukovič, M. (2021). Formation and growth of skyrmion crystal phase in a frustrated heisenberg antiferromagnet with dzyaloshinskii-moriya interaction. *Journal of Magnetism and Magnetic Materials*, 527:167755.
- Mohylna, M. and Žukovič, M. (2020). Emergence of a skyrmion phase in a frustrated heisenberg antiferromagnet with dzyaloshinskii-moriya interaction. *Acta Physica Polonica, A.*, 137(5).
- Mohylna, M. and Žukovič, M. (2022). Stability of skyrmion crystal phase in antiferromagnetic triangular lattice with dmi and single-ion anisotropy. *Journal of Magnetism and Magnetic Materials*, 546:168840.
- Moriya, T. (1960). Anisotropic superexchange interaction and weak ferromagnetism. *Physical review*, 120(1):91.
- Mühlbauer, S., Binz, B., Jonietz, F., Pfleiderer, C., Rosch, A., Neubauer, A., Georgii, R., and Böni, P. (2009). Skyrmion lattice in a chiral magnet. *Science*, 323(5916):915–919.
- Okubo, T., Chung, S., and Kawamura, H. (2012). Multiple-q states and the skyrmion lattice of the triangular-lattice heisenberg antiferromagnet under magnetic fields. *Physical review letters*, 108(1):017206.
- Osorio, S. A., Rosales, H. D., Sturla, M. B., and Cabra, D. C. (2017). Composite spin crystal phase in antiferromagnetic chiral magnets. *Physical Review B*, 96(2):024404.
- Pereira, A. and Pires, A. (2003). Solitons in the presence of a static spin vacancy on a 2d heisenberg antiferromagnet. *Journal of magnetism and magnetic materials*, 257(2-3):290–295.
- Roessler, U. K., Bogdanov, A., and Pfleiderer, C. (2006). Spontaneous skyrmion ground states in magnetic metals. *Nature*, 442(7104):797–801.
- Rosales, H. D., Cabra, D. C., and Pujol, P. (2015). Three-sublattice skyrmion crystal in the antiferromagnetic triangular lattice. *Physical Review B*, 92(21):214439.
- Silva, R., Secchin, L., Moura-Melo, W., Pereira, A., and Stamps, R. (2014). Emergence of skyrmion lattices and bimerons in chiral magnetic thin films with nonmagnetic impurities. *Physical Review B*, 89(5):054434.
- Subbaraman, K., Zaspel, C., and Drumheller, J. E. (1998). Impurity-pinned solitons in the two-dimensional antiferromagnet detected by electron paramagnetic resonance. *Physical review letters*, 80(10):2201.
- Zhang, X., Ezawa, M., and Zhou, Y. (2015a). Magnetic skyrmion logic gates: conversion, duplication and merging of skyrmions. *Scientific reports*, 5(1):1–8.
- Zhang, X., Zhao, G., Fangohr, H., Liu, J. P., Xia, W., Xia, J., and Morvan, F. (2015b). Skyrmion-skyrmion and skyrmion-edge repulsions in skyrmion-based racetrack memory. *Scientific reports*, 5(1):1–6.
- Zhang, X., Zhou, Y., Song, K. M., Park, T.-E., Xia, J., Ezawa, M., Liu, X., Zhao, W., Zhao, G., and Woo, S. (2020). Skyrmion-electronics: writing, deleting, reading and processing magnetic skyrmions toward spintronic applications. *Journal of Physics: Condensed Matter*, 32(14):143001.

MARIIA MOHYLNA obtained her PhD degree in theoretical physics from Pavol Jozef Šafárik University in Košice, Slovakia in 2021. She continued to pursue her research in the field of numerical modeling of the magnetic systems and currently is employed by Pavol Jozef Šafárik University in Košice as a postdoc researcher. Her e-mail address is: mariia.mohylna@upjs.sk

MILAN ŽUKOVIČ was born in Svidník, Slovakia and obtained his PhD degree in applied physics from Kyushu University, Japan in 2000. He pursued his research in the field of theoretical condensed matter physics at Kyushu University for two more years, before he assumed a position in automotive industry within Yazaki Corporation. In 2006-2008, he was involved in the research in modeling of spatial random fields as a Marie-Curie postdoc at Technical University of Crete, Greece. Since 2009 he has been with the Institute of Physics, Pavol Jozef Šafárik University in Košice, Slovakia, currently as a Professor of Physics. His e-mail address is: milan.zukovic@upjs.sk and his Webpage can be found at: <https://ufv.science.upjs.sk/zukovic/>.

MASSIVE DEGENERACY AND ANOMALOUS THERMODYNAMICS IN A HIGHLY FRUSTRATED ISING MODEL ON HONEYCOMB LATTICE

Milan Žukovič

Department of Theoretical Physics and Astrophysics, Institute of Physics

Faculty of Science

Pavol Jozef Šafárik University in Košice

Park Angelinum 9, 041 54 Košice, Slovak Republic

Email: milan.zukovic@upjs.sk

KEYWORDS

Frustrated spin model; Macroscopic degeneracy; Metropolis algorithm; Replica exchange Monte Carlo

ABSTRACT

We numerically study a prototypical frustrated Ising model on a honeycomb lattice with competing nearest- and second-nearest-neighbor antiferromagnetic interactions, J_1 and J_2 , for the highly frustrated case of $J_1 = J_2$. We employ both standard (SMC) and a more sophisticated replica exchange (REMC) Monte Carlo simulations in effort to demonstrate the difficulties of using the former and advantages of using the latter approaches. The ground state of the system is highly degenerate and consists of frozen superantiferromagnetic (SAF) domains, separated by zero-energy domain walls (ZEDW), thus showing no conventional long-range ordering. We demonstrate that such states are difficult to access by using SMC due to complex multimodal energy landscape, characteristic for such systems. On the other hand, the REMC approach turns out to be more efficient in exploring it and thus reaching also states not accessible to SMC. At finite temperatures, the system shows a peculiar behavior with multiple anomalies in thermodynamic functions, which can be attributed to the transitions between several states with the SAF-like ordering characterized by different types of ZEDW.

INTRODUCTION

In the last decades frustrated spin models have been a hot topic in condensed matter physics (Diep and Koibuchi 2020). Frustration can be generated by the lattice geometry and/or by the competition between different kinds of interaction. It often results in highly degenerate ground states with new induced symmetries, which may give rise to unexpected and exotic behaviors at finite temperatures. In spite of intensive investigations, many properties of frustrated systems are still not very well understood. Their numerical simulations have turned out to be challenging and standard approaches, such as the Metropolis Monte Carlo (MC) simulation typically used in the study of classical spin systems, encountered many difficulties or completely failed. Typical problems result from the presence of complex multimodal free energy landscapes with local minima that are separated by entropic barriers. Conventional MC approaches suffer from long equilibration times due to the suppressed tunneling through these barriers or get trapped in metastable states. Overcoming large energy barriers at strong first-order phase transitions may be another problem.

Ising models on bipartite lattices with further-neighbor antiferromagnetic (AF) interactions are paradigmatic examples of the spin systems frustrated due to the interaction competition. In the most studied two-dimensional Ising antiferromagnet on a square lattice the competition between the nearest-neighbor interaction J_1 (ferromagnetic or antiferromagnetic) and the increasing AF second-nearest-neighbor interaction J_2 leads to a gradual increase of frustration, which is reflected in reduction of the critical temperature down to zero at $R \equiv J_2/|J_1| = -1/2$ (Landau 1980; Grynberg and Tanatar 1992; Morán-López et al. 1993; Moran-Lopez et al. 1994). A possible change of the transition to the first order near $R = -1/2$ has also been reported (Jin et al. 2013; Bobák et al. 2015). For $R < -1/2$ the system shows a phase transition to a peculiar striped or superantiferromagnetic (SAF) state with a lot of controversy regarding its character. A series of earlier studies predicted a second-order transition with non-universal critical exponents for any $R < -1/2$ (see Malakis et al. (2006) and references within), however, some more recent approaches favored a first-order transition for $R^* < R < -1/2$ and a second-order one only for $R < R^*$ (Morán-López et al. 1993; Moran-Lopez et al. 1994; dos Anjos et al. 2008; Kalz et al. 2008, 2011; Jin et al. 2012; Bobák et al. 2015). The value of R^* was in different studies estimated to range between -1.1 (Morán-López et al. 1993) and -0.67 (Jin et al. 2012).

Much less attention has been devoted to the frustrated $J_1 - J_2$ model on other bipartite lattices, albeit, their critical behavior might be quite different from the square lattice one. In the present study we focus on the $J_1 - J_2$ Ising model on a honeycomb lattice, in which the small coordination number together with frustrated interactions has been shown to have interesting effects in the Heisenberg model with a magnetically disordered region and a spin-liquid phase (Zhang and Lamas 2013; Cabra et al. 2011). This frustrated model is also interesting from the experimental point of view, since it can be applied to some real materials (Matsuda et al. 2010; Tsirlin et al. 2010). The ground state of the $J_1 - J_2$ Ising model was investigated in some earlier papers (Houtappel 1950; Kudō and Katsura 1976; Katsura et al. 1986) but its critical behavior was studied only recently by the effective field theory (EFT) (Bobák et al. 2016), the MC simulation (Žukovič et al. 2020; Žukovič 2021), cluster mean-field (CMF) (Schmidt and Godoy 2021) and machine learning (ML) (Corte et al. 2021; Acevedo et al. 2021) approaches. For smaller degree of frustration, namely within the interval $-1/4 < R < 0$, the critical behavior resembles that for the square lattice. In particular, with the increasing frustration

due to increasing $|J_2|$ the transition temperature gradually decreases down to zero at $R = -1/4$. A possible crossover to the first-order behavior was suggested by the EFT (Bobák et al. 2016) but not confirmed by either MC (Žukovič 2021) or CMF (Schmidt and Godoy 2021).

In the present study we focus on the highly frustrated region of $R < -1/4$, for which both the EFT and CMF analytical approximations predicted no phase transition but the numerical approaches indicated some kind of phase transition to a highly degenerate low-temperature phase with no long-range ordering (Žukovič et al. 2020; Corte et al. 2021; Acevedo et al. 2021). In our previous study, focused on the less frustrated case of $-1/4 < R < 0$, we demonstrated that with the increasing frustration ($R \rightarrow -1/4$ and $T \rightarrow 0$) the system tends to freeze in metastable domain states separated by large energy barriers with extremely sluggish dynamics. In such cases the standard MC simulations turn out to be either very inefficient or completely failing in reaching the stable thermal equilibrium state and that the tunneling through the energy barriers can be considerably improved by using the replica exchange MC approach (Žukovič 2021). Therefore, to better handle the above mentioned problems, here we employ the replica exchange MC method and confront the obtained results with those from the standard MC simulation.

MODEL AND METHODS

Model

We study the Ising model on the honeycomb lattice described by the Hamiltonian

$$\mathcal{H} = -J_1 \sum_{\langle i,j \rangle} s_i s_j - J_2 \sum_{\langle i,k \rangle} s_i s_k, \quad (1)$$

where $s_i = \pm 1$ is the Ising spin variable at the i th site and the summations $\langle i,j \rangle$ and $\langle i,k \rangle$ run over all nearest and second-nearest spin pairs, respectively. The competition between the nearest- and second-nearest neighbor AF interactions ($J_1 < 0$ and $J_2 < 0$) leads to frustration. In the following we restrict our considerations to the case of $J_1 = J_2$, i.e., $R = -1$.

Methods

Standard Monte Carlo (SMC): We consider the system sizes $L \times L$, with $L = 12 - 72$, and apply periodic boundary conditions. We note that the presence of high frustration restricts our considerations to relatively small values of L since the above mentioned difficulties escalate very quickly as the system size increases. In the SMC simulations we employ a vectorized (checkerboard) single-spin flip Metropolis algorithm. For thermal averaging of the calculated quantities at each value of the temperature T we perform from 10^5 up to 10^6 MC sweeps (MCS) after discarding the initial burn-in period corresponding to twenty percent of those numbers for thermalization. Due to the low-temperature freezing of the system in metastable states in our simulations we use both cooling and heating protocols. Namely, the simulations are initialized from random (SAF) state at high (low) temperatures and then proceed towards lower (higher) temperatures with the step $k_B \Delta T / |J_1| = -0.01$ (0.01). The

next simulation starts from the final configuration obtained at the previous temperature to keep the system close to the equilibrium during the entire simulation.

Replica exchange Carlo (REMC): In effort to alleviate the difficulties of the standard MC approach we apply the REMC or parallel tempering method (Hukushima and Nemoto 1996), which is designed to more efficiently overcome energy barriers by a random walk in temperature space and allows better exploration of complex energy landscapes. Using the REMC method, we run simulations at $N_T = 120$ temperatures (replicas) in parallel and propose 10^5 swaps between replicas followed by 100 MCS over the complete lattice. The swap acceptance rate between neighboring replicas is $P(\beta_i \leftrightarrow \beta_{i+1}) = \min\{1, \exp(\Delta\beta \Delta\mathcal{H})\}$, where $\Delta\beta = \beta_{i+1} - \beta_i$ and $\Delta\mathcal{H} = \mathcal{H}_{i+1} - \mathcal{H}_i$ are differences between the neighboring inverse temperatures, $\beta_i = |J_1|/k_B T_i$, and the energies of the corresponding configurations, respectively. The temperatures T_i are chosen with the focus on the simulation bottlenecks in the vicinity of the expected transition temperatures to ensure sufficient acceptance rates (approximately flat) within the whole temperature interval.

The calculated quantities are restricted to the internal energy per spin $e \equiv E/N|J_1| = \langle \mathcal{H} \rangle / N|J_1|$, where $N = L^2$ is the number of spins, and the specific heat per spin obtained as

$$C/N = \frac{\langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2}{NT^2}, \quad (2)$$

where $\langle \dots \rangle$ denotes the thermal average. The peaks in the latter quantity are useful for capturing various anomalies and/or phase transitions.

RESULTS

In the regime of $R < -1/4$ the ground state (GS) corresponds to a spin arrangement with the energy $E_{SAF}/N|J_1| = -1/2 + R$ (Kudō and Katsura 1976; Katsura et al. 1986; Bobák et al. 2016). Such energy is realized in any spin configuration with the following local arrangement: out of the three nearest-neighbor bonds two are antiferromagnetic and one ferromagnetic and out of the six second-nearest-neighbor bonds four are antiferromagnetic and two ferromagnetic. However, there are many lattice spin configurations (see Fig. 1 for two examples) fulfilling these constraints. This leads to macroscopic degeneracy and the absence of any long-range ordering in the system.

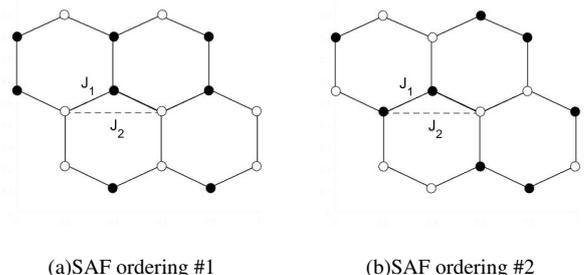
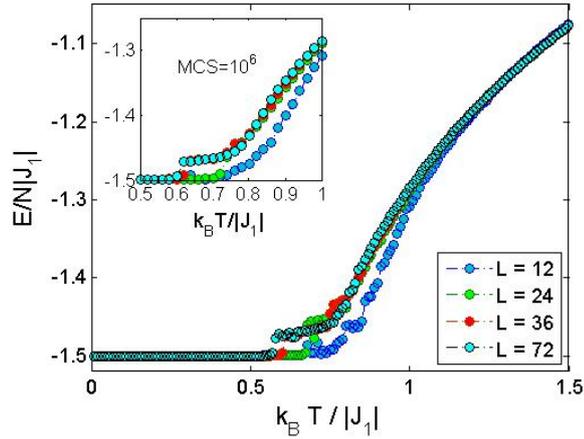
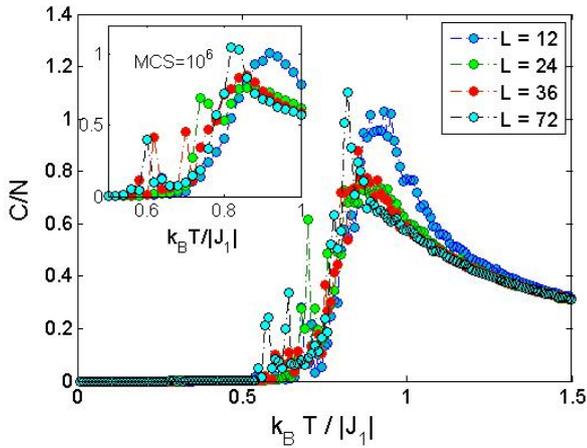


Fig. 1. Ground state for $R < -1/4$: Two examples of possible SAF ordering. The empty and filled circles represent spin-up and spin-down states, and the solid and dashed lines represent the interactions J_1 and J_2 .



(a) Internal energy (SMC)



(b) Specific heat (SMC)

Fig. 2. Temperature dependencies of (a) the internal energy and (b) the specific heat for different lattice sizes L and 10^5 MCS, obtained from SMC. The insets show the results for 10^6 MCS.

Fig. 2 shows temperature variations of the internal energy and the specific heat obtained from the shorter runs (10^5 MCS) using SMC. For the smallest lattice size the internal energy curve shows at $k_B T/|J_1| \approx 0.9$ an anomaly characteristic for a continuous phase transition reflected in a specific heat peak (Fig. 2(b)). However, for larger L another anomaly appears at lower temperatures as a discontinuous decrease in the internal energy and the specific heat showing a spike-like peak at $k_B T/|J_1| \approx 0.6$, suggesting a first-order phase transition. The insets demonstrate that these features also persist in much longer runs (10^6 MCS). Just below this jump of the internal energy thermal fluctuations are strongly suppressed and the system freezes in a state corresponding to the expected GS energy of $E_{SAF}/N|J_1| = -3/2$.

To better understand this unconventional behavior we performed several independent SMC runs for the largest lattice size, $L = 72$, following the temperature-decreasing as well as temperature-increasing processes. The former are initialized at high temperatures by random configurations and the latter start at low temperatures from the SAF states. We can see a relatively large variability in the calculated quantities corresponding to the individual runs. An example showing the temperature variations of the energy for two runs of each

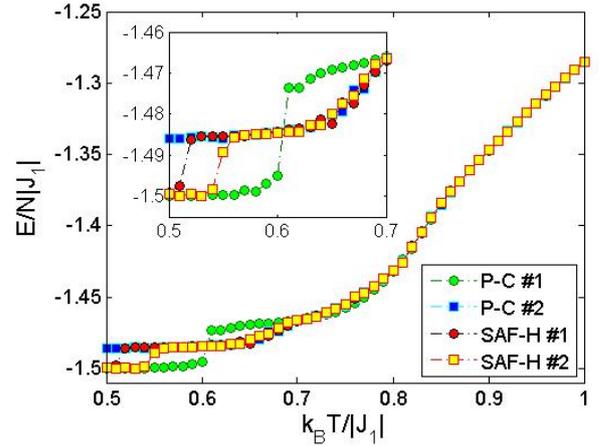


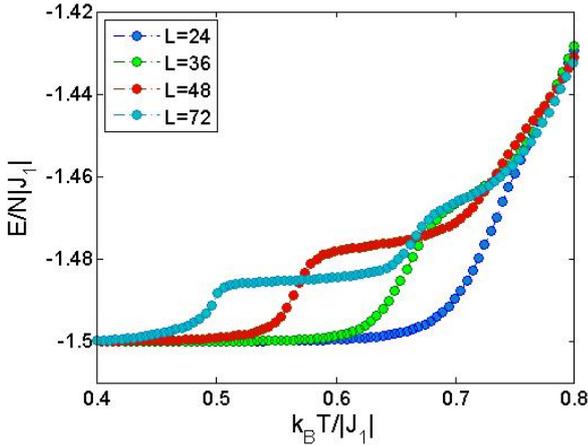
Fig. 3. Temperature dependencies of the internal energy for $L = 72$ and 10^6 MCS, obtained from SMC for two independent replicas cooled from paramagnetic phase (P-C #1,2) and two others heated from SAF phase (SAF-H #1,2).

kind are presented in Fig. 3. One can see that in the cooling process one simulation (green circles) lead to the low-temperature states with the energies close to the GS value, while the other one (blue squares), after some decrease below $k_B T/|J_1| \approx 0.7$ got stuck in a metastable state with the energy larger than the GS one. On the other hand, in the heating process all the runs start from the (SAF) configurations corresponding to the GS energy $E_{SAF}/N|J_1| = -3/2$. At some point of the heating the energies show a discontinuous increase but the temperatures at which this jump occurs in different runs are different (see red circles and yellow squares). At higher temperatures there are some additional smaller anomalies in the energy curves. This peculiar behavior signals the presence of metastable states corresponding to local minima in the rugged energy landscape.

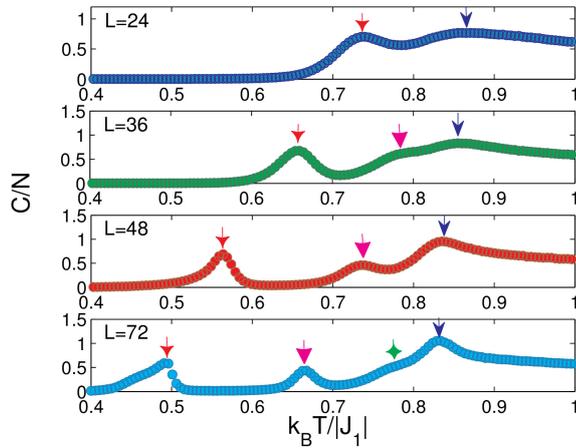
To secure thermal equilibrium states we subsequently applied REMC simulations. The internal energy and the specific heat temperature variations for different lattice sizes are presented in Fig. 4. Compared to the similar dependencies obtained from SMC, shown in Fig. 2, the REMC curves are smoother and all the energy curves reached the values close to the GS value at all temperatures below $k_B T/|J_1| \approx 0.4$. Nevertheless, the anomalies observed in the SMC simulations are reproduced also in the REMC runs. In particular interesting is their lattice size dependence. As L increases the number of anomalies in both the energy and specific heat curves also increases, starting from two for $L = 24$ (or one for $L = 12$ not shown here) up to four for $L = 74$. The new ones keep forming at gradually lower temperatures creating wave-like dependencies.

Furthermore, the character of the lower-temperature anomalies resembles that observed at the first-order phase transitions. This is demonstrated in Fig. 5(a) by showing in the insets a bimodal character of the energy histograms at the temperatures corresponding to two low-temperature anomalies for the largest system size. On the other hand, the highest-temperature anomaly in the energy curve remains smooth but the corresponding specific heat peak increases with the lattice size like at a phase transition (see Fig. 5(b)).

Finally, let us inspect the character of the phases in the



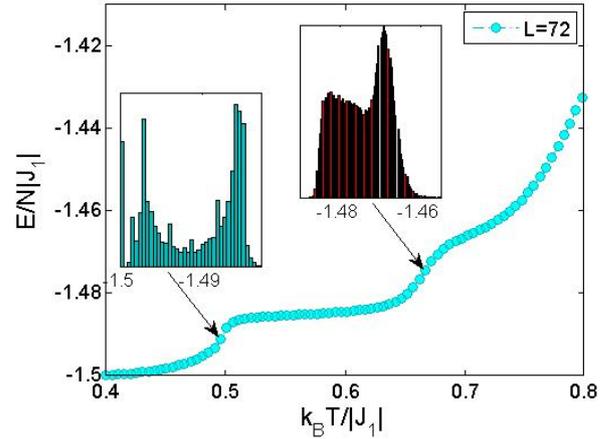
(a) Internal energy (REMC)



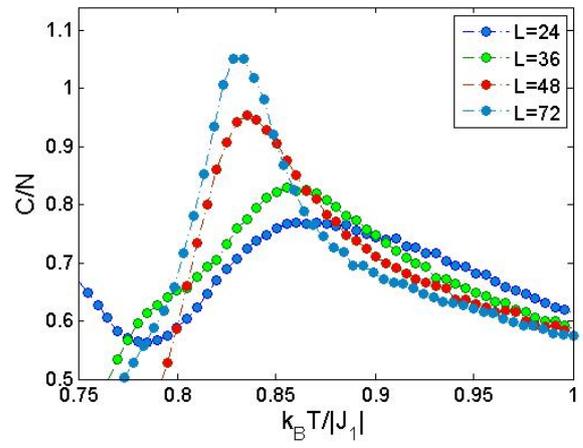
(b) Specific heat (REMC)

Fig. 4. Temperature dependencies of (a) the internal energy and (b) the specific heat for different lattice sizes L , obtained from REMC. Different arrows in (b) show the positions of the respective anomalies.

temperature ranges between different anomalies. In Fig. 6 we show some typical spin configurations at different temperatures from REMC simulations for $L = 72$. The empty and filled circles represent spin-up and spin-down states, respectively. At the lowest temperature, $k_B T/|J_1| = 0.4$, one can notice that there is no long-range ordering that would span the entire lattice. Instead, the lattice is split into several domains with different types of SAF ordering inside each domain. The respective domains form horizontal bands of different widths and are separated from each other by zero-energy domain walls (ZEDW) spanning the entire lattice in the horizontal direction. At higher temperature, $k_B T/|J_1| = 0.57$, which falls within the temperature range between the two low-temperature anomalies, the spin texture is somewhat different. The lattice is still split into several domains with SAF ordering, which are separated by ZEDW, however, the domain wall direction is now diagonal. One can also notice some isolated defects, which deform the boundary shape and increase the domain wall energy (empty and filled circles with magenta and cyan edges). We believe that these defects result from the increased thermal fluctuations which partially disrupt the perfect SAF-like ordering and slightly increase the energy. At still higher temper-



(a) Internal energy (REMC)



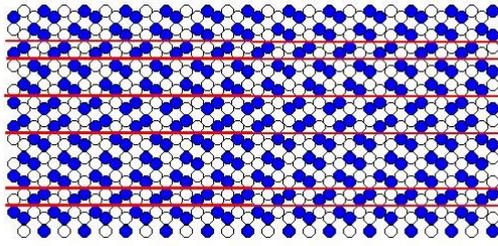
(b) Specific heat (REMC)

Fig. 5. Temperature dependencies of (a) the internal energy at lower temperatures and $L = 72$ and (b) the specific heat at higher temperatures and different L . The insets in (a) show the energy histograms at the temperatures corresponding to the anomalies.

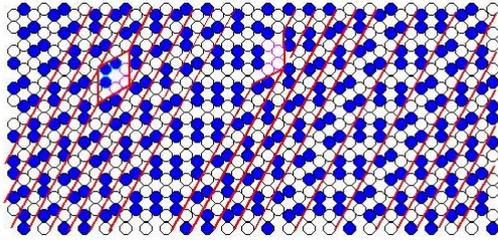
ature, $k_B T/|J_1| = 0.72$, above the second low-temperature anomaly, there is still SAF-type of ordering fragmented in multiple domains with ZEDW. Here, however, the domains are smaller and do not necessarily span the entire system. Consequently, there is a mixture of shorter horizontal and vertical domain walls as illustrated in Fig. 6(c) by highlighting just a few of them. At the elevated temperature, naturally, there is also higher occurrence of defects due to thermal fluctuations.

CONCLUSIONS

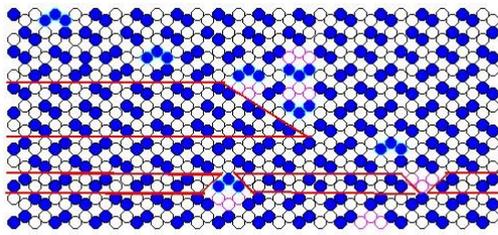
We studied the frustrated Ising model on a honeycomb lattice with competing nearest- and second-nearest-neighbor AF interactions, J_1 and J_2 , in the highly frustrated region of $R \equiv J_2/|J_1| < -1/4$. We employed both SMC and REMC simulations in effort to demonstrate the difficulties of using the former and advantages of using the latter approaches in a numerical study of such a paradigmatic example of frustrated spin systems. We showed that the ground state of the system is highly degenerate and consists of frozen SAF-like domains, separated by zero-energy domain walls (ZEDW), with no conventional magnetic long-



(a) $k_B T / |J_1| = 0.4$



(b) $k_B T / |J_1| = 0.57$



(c) $k_B T / |J_1| = 0.72$

Fig. 6. Typical spin snapshots at different temperatures from REMC simulations for $L = 72$. The empty (filled) circles represent spins up (down) and (a) the horizontal, (b) the diagonal and (c) the horizontal and diagonal red lines mark SAF domain walls. The empty and filled circles with magenta and cyan edges in (b) and (c) mark defects in perfect SAF-like ordering.

range ordering. We also demonstrated that such a state might be difficult to access by using the SMC, which might get stuck in metastable states. On the other hand, the REMC approach can more efficiently explore complex energy landscapes, characteristic for such systems, and thus reach also states not accessible to the SMC.

At finite temperatures, the system displayed a peculiar behavior with multiple anomalies in thermodynamic functions. The anomalies can be attributed to the transitions between several states with the SAF-like ordering, characterized by different types of ZEDW. The low-temperature anomalies bear some features of the first-order phase transitions, nevertheless, their number and positions show strong finite-size effects. Therefore, in order to verify whether the observed anomalies are just finite-size artifacts or at least some of them signal true phase transitions, a careful finite-size scaling analysis involving much larger system sizes to inspect the approach to the thermodynamic limit would be desirable. This goal might be achievable by implementing the REMC approach on the highly parallel GPU architecture, since the GPU-accelerated MC simulations of spin systems have been reported to achieve speedups up to three orders of magnitude compared to the standard CPU implementations (Weigel 2012).

ACKNOWLEDGMENTS

This work was supported by the grants of the Slovak Research and Development Agency (Grant No. APVV-18-0197) and the Scientific Grant Agency of Ministry of Education of Slovak Republic (Grant No. 1/0531/19).

REFERENCES

- Acevedo, S., Arlego, M., and Lamas, C. A. (2021). Phase diagram study of a two-dimensional frustrated antiferromagnet via unsupervised machine learning. *Physical Review B*, 103(13):134422.
- Bobák, A., Lučivjanský, T., Borovský, M., and Žukovič, M. (2015). Phase transitions in a frustrated ising antiferromagnet on a square lattice. *Physical Review E*, 91(3):032145.
- Bobák, A., Lučivjanský, T., Žukovič, M., Borovský, M., and Balcerzak, T. (2016). Tricritical behaviour of the frustrated ising antiferromagnet on the honeycomb lattice. *Physics Letters A*, 380(34):2693–2697.
- Cabra, D. C., Lamas, C. A., and Rosales, H. D. (2011). Quantum disordered phase on the frustrated honeycomb lattice. *Physical Review B*, 83(9):094506.
- Corte, I., Acevedo, S., Arlego, M., and Lamas, C. (2021). Exploring neural network training strategies to determine phase transitions in frustrated magnetic models. *Computational Materials Science*, 198:110702.
- Diep, H. T. and Koibuchi, H. (2020). Frustrated magnetic thin films: Spin waves and skyrmions. In *Frustrated Spin Systems: 3rd Edition*, pages 631–719. World Scientific.
- dos Anjos, R. A., Viana, J. R., and de Sousa, J. R. (2008). Phase diagram of the ising antiferromagnet with nearest-neighbor and next-nearest-neighbor interactions on a square lattice. *Physics Letters A*, 372(8):1180–1184.
- Grynberg, M. D. and Tanatar, B. (1992). Square ising model with second-neighbor interactions and the ising chain in a transverse field. *Physical Review B*, 45(6):2876.
- Houtappel, R. M. F. (1950). Order-disorder in hexagonal lattices. *Physica*, 16(5):425–455.
- Hukushima, K. and Nemoto, K. (1996). Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608.
- Jin, S., Sen, A., Guo, W., and Sandvik, A. W. (2013). Phase transitions in the frustrated ising model on the square lattice. *Physical Review B*, 87(14):144406.
- Jin, S., Sen, A., and Sandvik, A. W. (2012). Ashkin-teller criticality and pseudo-first-order behavior in a frustrated ising model on the square lattice. *Physical Review Letters*, 108(4):045702.
- Kalz, A., Honecker, A., Fuchs, S., and Pruschke, T. (2008). Phase diagram of the ising square lattice with competing interactions. *The European Physical Journal B*, 65(4):533–537.
- Kalz, A., Honecker, A., and Moliner, M. (2011). Analysis of the phase transition for the ising model on the frustrated square lattice. *Physical Review B*, 84(17):174407.
- Katsura, S., Ide, T., and Morita, T. (1986). The ground states of the classical heisenberg and planar models on the triangular and plane hexagonal lattices. *Journal of statistical physics*, 42(3):381–404.
- Kudō, T. and Katsura, S. (1976). A method of determining the orderings of the ising model with several neighbor interactions under the magnetic field and applications to hexagonal lattices. *Progress of Theoretical Physics*, 56(2):435–449.
- Landau, D. (1980). Phase transitions in the ising square lattice with next-nearest-neighbor interactions. *Physical Review B*, 21(3):1285.
- Malakis, A., Kalozoumis, P., and Tyraskis, N. (2006). Monte carlo studies of the square ising model with next-nearest-neighbor interactions. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1):63–67.
- Matsuda, M., Azuma, M., Tokunaga, M., Shimakawa, Y., and Kumada, N. (2010). Disordered ground state and magnetic field-induced long-range order in an $s = 3/2$ antiferromagnetic honeycomb lattice compound $\text{Bi}_3\text{Mn}_4\text{O}_{12}$ (no 3). *Physical review letters*, 105(18):187201.
- Morán-López, J., Aguilera-Granja, F., and Sanchez, J. (1993). First-order phase transitions in the ising square lattice with first-and second-neighbor interactions. *Physical Review B*, 48(5):3519.
- Moran-Lopez, J., Aguilera-Granja, F., and Sanchez, J. (1994). Phase transitions in ising square antiferromagnets with first-and second-neighbour interactions. *Journal of Physics: Condensed Matter*, 6(45):9759.
- Schmidt, M. and Godoy, P. (2021). Phase transitions in the ising antifer-

- romagnet on the frustrated honeycomb lattice. *Journal of Magnetism and Magnetic Materials*, 537:168151.
- Tsirlin, A. A., Janson, O., and Rosner, H. (2010). β -cu 2 v 2 o 7: A spin-1 2 honeycomb lattice system. *Physical Review B*, 82(14):144416.
- Weigel, M. (2012). Performance potential for simulating spin models on gpu. *Journal of Computational Physics*, 231(8):3064–3082.
- Zhang, H. and Lamas, C. A. (2013). Exotic disordered phases in the quantum j 1-j 2 model on the honeycomb lattice. *Physical Review B*, 87(2):024415.
- Žukovič, M. (2021). Critical properties of the frustrated ising model on a honeycomb lattice: A monte carlo study. *Physics Letters A*, 404:127405.
- Žukovič, M., Borovský, M., Bobák, A., Balcerzak, T., and Szałowski, K. (2020). Spin-glass-like ordering in a frustrated j 1-j 2 ising antiferromagnet on a honeycomb lattice. *Acta Physica Polonica, A.*, 137(5).

MILAN ŽUKOVIČ was born in Svidník, Slovakia and obtained his PhD degree in applied physics from Kyushu University, Japan in 2000. He pursued his research in the field of theoretical condensed matter physics at Kyushu University for two more years, before he assumed a position in automotive industry within Yazaki Corporation. In 2006-2008, he was involved in the research in modeling of spatial random fields as a Marie-Curie postdoc at Technical University of Crete, Greece. Since 2009 he has been with the Institute of Physics, Pavol Jozef Šafárik University in Košice, Slovakia, currently as a Professor of Physics. His e-mail address is: milan.zukovic@upjs.sk and his Webpage can be found at: <https://ufv.science.upjs.sk/zukovic/>.

AUTHOR INDEX

- 5 Abbass, Hussein
 226 Alaliyat, Saleh Abdel-Afou
 226 Alfaro, Claudia Viviana Lopez
 236, 245 Amundsen, Andreas
 226 Angelaki, Stavroula
 245 Aspen, Dina
 58 Bagirova, Anna
 36 Bandinelli, Romeo
 90 Banditvilai, Somsri
 150 Basova, Olga A.
 226 Besenecker, Ute
 36 Bindi, Bianca
 210 Bobis, Daniel
 64 Boros, Eszter
 29 Boukadi, Khouloud
 167 Burmistrov, Vladimir
 317 Campanile, Lelio
 262 Chaganova, Olga
 114 Claus, Thorsten
 189 Csoban, Attila
 226, 236 da Silva Torres, Ricardo
 245
 254 Daragmeh, Sarah M.
 310 Demarchi, Stefano
 135 Denkova, Zapryana
 135 Denkova-Kostova, Rositsa
 99, 107 El-Mihoub, Tarek A.
 36 Fani, Virginia
 29 Fattouch, Najla
 70 Felfoeldi-Szuecs, Nora
 13, 143 Fottner, Johannes
 20 Gaspar, Henrique M.
 285, 296 Gaza, Lukasz
 135 Goranov, Bogdan
 203 Graeff, Jozsef
 317 Gribaudo, Marco
 150, 167 Grigoryev, Anton
 262, 278
 296 Grzonka, Daniel
 310 Guidotti, Dario
 7 Gupta, Manish
 128 Haag, Stefan
 121 Haas, Florian
 159 Haghbayan, Hashem
 45 Hajlasz, Maria
 226, 254 Hameed, Ibrahim A.
 226 Hassan, Muhammad Umair
 114 Herrmann, Frank
 317 Iacono, Mauro
 159 Immonen, Eero
 291 Imre, Kayhan M.
 219 Jaeger, Miriam
 285 Jakobik, Agnieszka
 174 Jemai, Ahlem
 159 Karami, Masoomah
 254 Karlsen, Anniken Th.
 135 Kostov, Georgi
 51 Kovacs, Erzsebet
 70 Kralik, Balazs
 219 Kurrle, Marius
 29 Lahmar, Imen Ben
 245 Leplat, Leo
 271 Levis, Alexander H.
 9 Lie, Knut-Andreas
 226, 236 Major, Pierre
 317 Marulli, Fiammetta
 317 Mastroianni, Michele
 210 Mate, Tamas
 45 Mielczarek, Bozena
 236 Mikalsen, Egil Tennfjord
 20 Misund, Armand
 331 Mohylina, Mariia
 189 Molnar, Jakab
 121 Morelli, Frank
 150 Nikolaev, Dmitry P.
 99, 107 Nolle, Lars
 174
 36 Nunziatini, Andrea
 167 Panfilova, Ekaterina
 203 Pasty, Laszlo
 310 Pitto, Andrea
 159 Plosila, Juha
 324 Razumchik, Rostislav
 324 Rummyantsev, Alexander
 219 Sauer, Alexander
 121 Schaeffer, Frank

219 Schlereth, Andreas
181 Selmair, Maximilian
159 Shahsavari, Sajad
78 Shmarova, Irina
135 Shopska, Vesela
58, 78 Shubat, Oksana
143 Siciliano, Giulia
128 Simon, Carlo
271 Sliva, Amy
296 Sosnicki, Adrian
236 Stadsnes, Pernille
174 Stahl, Frederic
226 Styve, Arne
303 Suchacka, Grazyna
51 Szanyi-Nagy, Sara
64 Sztano, Gabor
310 Tacchella, Armando
203 Tamas, Kornel
278 Teplyakov, Lev
114 Terbrack, Hajo
99, 107 Tholen, Christoph
70 Varadi, Kata
84 Varga, Erzsebet Terez
51 Vaskoevi, Agnes
13 Wegerich, Benjamin
196 Wotzka, Daria
13 Wuennenberg, Maximilian
219 Yesilyurt, Ozan
143 Yu, Yue
128 Zakfeld, Lara
278 Zhdanovskiy, Vyacheslav
107, 174 Zielinski, Oliver
331, 336 Zukovic, Milan
189, 210 Zwierczyk, Peter T.

