

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 978-3-937436-68-5 (Print)

ISSN 2522-2414 (Print)

ISBN 978-3-937436-69-2 (CD-ROM)

ISSN 2522-2422 (Online)

ISSN 2522-2430 (CD-ROM)

**Cover pictures front and
and back side**

© pictures taken by Matthias Friel

Printed by

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

Communications of the ECMS

Volume 34, Issue 1, June 2020

Proceedings of the 34th International ECMS Conference on Modelling and Simulation ECMS 2020

June 2020, United Kingdom

Edited by:

Mike Steglich
Christian Mueller
Gaby Neumann
Mathias Walther

Organized by:

ECMS - European Council for Modelling and Simulation
Technical University of Applied Sciences TH Wildau

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 2020 ORGANIZATION

Conference Chair

Mike Steglich

Technical University of Applied Sciences
TH Wildau, Germany

Conference Co-Chair

Christian Mueller

Technical University of Applied Sciences
TH Wildau, Germany

Programme Chair

Gaby Neumann

Technical University of Applied Sciences
TH Wildau, Germany

Programme Co-Chair

Mathias Walther

Technical University of Applied Sciences
TH Wildau, Germany

Editors in Chief

Lars Nolle

Jade University of Applied Sciences, Germany

Evtim Peytchev

Nottingham Trent University, United Kingdom

Managing Editor

Martina-Maria Seidel

St. Ingbert, Germany

Editorial Advisory Board

Andrzej Bargiela

Nottingham, United Kingdom

Zuzana Kominkova Oplatkova

Tomas Bata University in Zlin, Czech Republic

Khalid Al-Begain

Kuwait College of Science and Technology, Kuwait

Editorial Board

Andrzej Bargiela	United Kingdom	Mohamed Gaber	United Kingdom
Khalid Al-Begain	Kuwait	Zuzana Kominkova	Czech Republic
Lars Nolle	Germany	Roman Senkerik	Czech Republic
Evtim Peytchev	United Kingdom	Agnes Vidovics-Dancs	Hungary
Eugene Kerckhoffs	Netherlands	Jiri Vojtesek	Czech Republic
Kata Varadi	Hungary	Frantisek Gazdos	Czech Republic
Peter T. Zwierczyk	Hungary	Janos P. Radics	Hungary
Frank Herrmann	Germany	Joanna Kolodziej	Poland
Michael Manitz	Germany	Mauro Iacono	Italy
Marco Trost	Germany	Agneszka Jakobik	Poland
Edward J. Williams	USA	Rostislav Razumchik	Russia
Romeo Bandinelli	Italy	Christoph Tholen	Germany
Jens Werner	Germany	Ricardo da Silva Torres	Norway
Frederic T. Stahl	Germany	Henrique M. Gaspar	Norway
Marwan Hassani	Netherlands	Marco Gribaudo	Italy
Lelio Campanile	Italy		

INTERNATIONAL PROGRAMME COMMITTEE

Industrial Process Modelling and Simulation

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

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

Simulation of Intelligent Systems

Track Chair: **Zuzana Kominkova Oplatkova**
Tomas Bata University of Zlin, Czech Republic

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

Finance, Economics and Social Science

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

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

Modelling, Simulation and Control of Technological Processes

Track Chair: **Jiri Vojtesek**
Tomas Bata University in Zlin, Czech Republic

Co-Chair: **Frantisek Gazdos**
Tomas Bata University in Zlin, Czech Republic

Open and Collaborative Models and Simulation Methods

Track Chair: **Henrique M. Gaspar**
Norwegian University of Science and Technology, Norway

Co-Chair: **Ricardo da Silva Torres**
Norwegian University of Science and Technology, Norway

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

Finite – Discrete – Element Simulation

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

Co-Chair: **Janos P. Radics**

Budapest University of Technology and Economics, Hungary

Machine Learning for Big Data

Track Chair: **Frederic Theodor Stahl**
DFKI German Research Center for Artificial Intelligence, Germany

Co-Chairs:

Mohamed Gaber

Birmingham City University, United Kingdom

Marwan Hassani

Eindhoven University of Technology, Netherlands

Modeling and Simulation for Performance Evaluation of Computer-based Systems

Track Chair: **Mauro Iacono**
University of Campania "Luigi Vanvitelli", Italy

Co-Chairs for DIS:

Agnieszka Jakobik

Cracow University of Technology, Poland

Lelio Campanile

University of Campania "Luigi Vanvitelli", Italy

Chair for PSTAT: **Rostislav Razumchik**

Institute of Informatics Problems, FRC CSC RAS, Russia

IPC Members

Eric Afful-Dadzie, University of Ghana, Ghana

Walailak Atthirawong, King Mongkut's Inst. of Technology Ladkrabang, Thailand

Anna Bagirova, Ural Federal University, Russia

Enrico Barbierato, Catholic University of Sacred Heart, Brescia, Italy

Thomas Beltrame, Federal University of Sao Carlos, Brazil

Denes Bencsik, Robert Bosch Kft., Hungary

Daniel Beres, Metropolitan University of Budapest, Hungary

Simona Bernardi, University of Zaragoza, Spain

Sandor Bozoki, Corvinus University of Budapest, Hungary

Aleksander Byrski, AGH University of Science and Technology, Poland

Damian Fernandez Cerero, University of Sevilla, Spain

Petr Chalupa, Tomas Bata University in Zlin, Czech Republic

Donald Davendra, Central Washington University, USA

Antinisca Di Marco, University of L'Aquila , Italy

Sergei Dudin, Belarusian State University, Belarus

Frantisek Dusek, University of Pardubice, Czech Republic

Virginia Fani, University of Florence, Italy

Thomas Farrenkopf, TH Mittelhessen Friedberg, Germany

Nora Felfoeldi-Szucs, John von Neumann University, Hungary

Alexandre Ferreira, State University of Campinas, Brazil

Massimo Ficco, University of Campania "Luigi Vanvitelli", Italy

Szabolcs Fischer, Szechenyi Istvan University, Hungary

Icaro Aragao Fonseca, Norwegian University of Science and Technology, Norway

Ingo Frank, OTH Regensburg, Germany

Ildiko Gelanyi, Corvinus University of Budapest, Hungary

Horacio Gonzalez-Velez, National College of Ireland, Ireland

Marco Gribaudo, Polytechnic University of Milan, Italy

Alexander Grusho, Institute of Informatics Problems, FRC CSC RAS, Russia
Daniel Grzonka, Cracow University of Technology, Poland
Michael Guckert, TH Mittelhessen Friedberg, Germany
Stefan Haag, HS Worms University of Applied Sciences, Germany
Mahmood Hammoodi, University of Babylon, Iraq
Benjamin Hoffmann, TH Mittelhessen Friedberg, Germany
Daniel Honc, University of Pardubice, Czech Republic
Teruaki Ito, University of Tokushima, Japan
Michal Janosek, University of Ostrava, Czech Republic
Bogumil Kaminski, Warsaw School of Economics, Poland
Stelios Kapetanakis, University of Brighton, United Kingdom
Petia Koprinkova-Hristova, Bulgarian Academy of Sciences, Bulgaria
Victor Korolev, Moscow State University, Russia
Andreasz Kosztopulosz, University of Szeged, Hungary
Martin Kotyrba, University of Ostrava, Czech Republic
Marek Kubalcik, Tomas Bata University in Zlin, Czech Republic
Frederick Lange, Maschinenfabrik Reinhausen GmbH, Regensburg, Germany
Alexander H. Levis, George Mason University, USA
Thomas Lienert, Technical University of Munich, Germany
Luis Gustavo Lorgus Decker, State University of Campinas, Brazil
Oleg Lukashenko, Karelian Research Centre RAS, Russia
Lubomir Macku, Tomas Bata University in Zlin, Czech Republic
Andrea Marin, University of Venice, Italy
Stefano Marrone, University of Campania "Luigi Vanvitelli", Italy
Giovanna Martinez-Arellano, University of Nottingham, United Kingdom
Michele Mastroianni, University of Campania "Luigi Vanvitelli", Italy
Radek Matusu, Tomas Bata University in Zlin, Czech Republic
Nicolas Meseth, HS Osnabrück University of Applied Sciences, Germany
Lusine Meykhanadzhyan, Financial University under the Government of the Russian Federation, Russia

Thiago Gabriel Monteiro, Norwegian University of Science and Technology, Norway
Frank Morelli, University of Applied Sciences in Pforzheim, Germany
Daniel Mota, Norwegian University of Science and Technology, Norway
Christian Müller, University of Applied Sciences TH Wildau, Germany
Andras Oliver Nemeth, Corvinus University of Budapest, Hungary
Jakub Novak, Tomas Bata University in Zlin, Czech Republic
Laszlo Oroszvary, Knorr-Bremse Research, Hungary
Francesco Palmieri, University of Salerno, Italy
Libor Pekar, Tomas Bata University in Zlin, Czech Republic
Allan Pinto, State University of Campinas, Brazil
Pawel Piskur, Polish Naval Academy, Poland
Michal Pluhacek, Tomas Bata University in Zlin, Czech Republic
Navya Prakash, DFKI GmbH Oldenburg, Germany
Daniel Prata Vieira, University of Sao Paulo, Brazil
Michal Przybylski, Polish Naval Academy, Poland
Peter Rausch, Nuremberg Institute of Technology, Germany
Simone Righi, University College London, United Kingdom
David Romero, Monterrey Institute of Technology, Mexico
Faruk Savasci, Kronos AG, Germany
Veronica Scuotto, University of Turin, Italy
Maximilian Selmair, BMW-Group Munich, Germany
Oleg Shestakov, Moscow State University, Russia
Oksana Shubat, Ural Federal University, Russia
Markus Siegle, Bundeswehr University Munich, Germany
Carlo Simon, HS Worms University of Applied Sciences, Germany
Janos Simonovics, Budapest University of Technology and Economics, Hungary
Tuanjai Somboonwiwat, King Mongkut's Inst. of Technology Ladkrabang, Thailand
Grazyna Suchacka, Opole University, Poland
Grzegorz Szafranski, University of Lodz, Poland

Janos Szaz, Corvinus University of Budapest, Hungary

Melinda Szodorai, Corvinus University of Budapest & Keler CCP, Hungary

Armando Tacchella, University of Genova, Italy

Enrico Teich, Technical University Dresden, Germany

Hajo Terbrack, Technical University Dresden, Germany

Nikolai Ushakov, Norwegian University of Science and Technology, Norway

Adam Viktorin, Tomas Bata University in Zlin, Czech Republic

Eva Volna, University of Ostrava, Czech Republic

Thananya Wasusri, King Mongkut's Inst. of Technology Ladkrabang, Thailand

Victor Zakharov, Institute of Informatics Problems, FRC CSC RAS, Russia

Alexander Zeifman, Vologda State University, Russia

PREFACE

The European Conference on Modelling and Simulation ECMS 2020 is dedicated to help define the state of the art in different fields involved in creating, defining and building innovative simulation systems, simulation and modelling tools and techniques, and novel applications for modelling and simulation. For this year, the Technical University of Applied Sciences Wildau was asked to organise and host the 34th annual conference of the European Council for Modelling and Simulation international conference. But the Corona pandemic changed our plans as for many other conferences and events. In order to protect the health of all participants, the organisers as well as the lecturers and students at the Technical University of Applied Sciences, we were forced to cancel this conference with a heavy heart.

The Technical University of Applied Sciences Wildau, founded in 1991, is a modern and compact campus with a direct rail connection to Berlin, aspiring academics enjoy an ideal environment for their studies in a range of disciplines covering the natural sciences, engineering, economics, law, business administration and management in several Bachelor and Master programmes. As a university of applied sciences, it is designated to provide scientific innovation and development which makes Wildau a desirable partner for innovative small and medium-sized businesses, as well as large international companies. One of the reasons to take over hosting the ECMS 2020 was to strengthen the international collaborations of our university which actively promotes international exchanges for researchers and students. It is for these reasons that the university cooperates with more than 70 partner universities around the world.

After the decision that the conference cannot take place, the question arose how the scientific exchange can still be carried out. The board of the European Council for Modelling and Simulation decided and offered to publish the conference contributions as an annual edition of the book series “Communications of the ECMS”. We are convinced that this book gives all interested researchers and practitioners a good impression about the newest developments in the field of simulation, modelling and operations research. It also shows, particularly in Corona times, the importance of sciences for solving real problems in our society.

We would like to thank all participants for submitting the papers, the editors for reviewing the articles, all members of the local organising team for all the preparations that have unfortunately become redundant due to the cancellation. Furthermore, we would like to thank Martina-Maria Seidel for keeping the conference alive and also Lars Nolle, the president of the ECMS for his trust in our organisation team.

And finally, we are pleased to announce that the ECMS2021 will take place in Wildau from June 8th to 11th, 2021. We are looking forward to welcoming you next year in Wildau!

Mike Steglich, Gaby Neumann, Christian Müller and Mathias Walther

Wildau, April 2020

TABLE OF CONTENTS

Simulation of Intelligent Systems

A Model For Ground Transportation Systems Simulation At Airports Under Centralized Control

Farid Saifutdinov, Jurijs Tolujevs5

Simulation Of Underwater Color Images Using Banded Spectral Model

*Denis A. Shepelev, Valentina P. Bozhkova, Egor I. Ershov,
Dmitry P. Nikolaev* 11

A Conceptual Model Of An IOT-Based Smart And Sustainable Solid Waste Management System: A Case Study Of A Norwegian Municipality

Wajeeha Nasar, Anniken Th. Karlsen, Ibrahim A. Hameed 19

Optimal Receiver Configuration Of Short-Baseline Localisation Systems Using Particle Swarm Optimisation

*Christoph Tholen, Tarek El-Mihoub, Lars Nolle, Oliver Ralle,
Robin Rofallski*.....25

Using The CMA Evolution Strategy For Locating Submarine Groundwater Discharge

Tarek A. El-Mihoub, Christoph Tholen, Lars Nolle.....32

Industrial Process Modelling and Simulation

Numerical Simulation Of Condensing Ammonia In Plate Heat Exchangers Using CFD

*Alexander Dietrich, Mario Nowitzki, Ron van de Sand,
Joerg Reiff-Stephan*41

Discrete Event Simulation – Model Of A Call Center In SIMUL8 Software

Martina Kuncova, Jan Fabry, Anna Marie Klimova48

Balancing Assembly Line In The Footwear Industry Using Simulation: A Case Study

Virginia Fani, Bianca Bindi, Romeo Bandinelli56

Finance and Economics and Social Science

The European Stability Mechanism And Sovereign Bond Yields: An Analysis In Light Of New Debates

Eszter Boros, Gabor Sztano65

Modelling The Relationship Between Demographic Structures Of The Russian Population

Anna Bagirova, Oksana Shubat.....73

Russian Grandparenting: Demographic And Statistical Modelling Experience

Oksana Shubat, Anna Bagirova.....78

What Is The Best Way To Help?

Central Bank Strategies And The Interbank Market

Gabor Kuerthy, Agnes Vidovics-Dancs, Janos Szaz, Peter Juhasz84

Clustering EU Countries Based On Death Probabilities

Kolos Csaba Agoston, Agnes Vaskoevi.....91

Circular Economy:

A Coloured Petri Net Based Discrete Event Simulation Model

*Marco Gribaudo, Marco Pironti, Paola Pisano, Daniele Manini,
Veronica Scutto*97

Forecasting Residential Electricity Consumption Based On Urbanization And Income Projections

Emilia Nemeth-Durko, Peter Juhasz, Fanni Dudas..... 104

Sales Forecasting And Newsboy Model Techniques

Integrated For Merchandise Planning And Business Risk Optimization

Tomasz Brzeczek 111

Income Inequality In Hungary

Ildiko Gelanyi, Andras Oliver Nemeth, Erzsebet Terez Varga 116

The Necessary Size Of The Skin-In-The-Game To Stay In The Game

*Kira Muratov-Szabo, Andrea Prepuk, Melinda Szodorai,
Kata Varadi* 122

Compensation Scheme With Shapley Value

For Multi-Country Kidney Exchange Programmes

*Peter Biro, Marton Gyetvai, Xenia Klimentova, Joao Pedro Pedroso,
William Pettersson, Ana Viana*..... 129

Modelling, Simulation and Control of Technological Processes

Retrofit Optimization Of Battery Air Cooling By CFD And Machine Learning <i>Eero Immonen, Janne Sovela, Samuli Ranta, Kirill Murashko, Paula Immonen</i>	139
Analytical Approaches For Determining The Effects Of Wort Extract On The Specific Growth Rate Of The Yeast Population <i>Georgi Kostov, Rositsa Denkova-Kostova, Vesela Shopska, Bogdan Goranov, Kristina Ivanova</i>	146
Kinetics Of Microbial Inactivation Of Human Pathogens By Biological Factors <i>Georgi Kostov, Rositsa Denkova-Kostova, Vesela Shopska, Zapryana Denkova, Bogdan Goranov, Desislava Teneva</i>	153
Automatic Production Of Patient Adapted Orthopaedic Braces Using 3D - Modelling Technology <i>Paul Steffen Kleppe, Webjoern Rekdalsbakken</i>	161
Navigation System For Landing A Swarm Of Autonomous Drones On A Movable Surface <i>Anam Tahir, Jari Boeling, Mohammad-Hashem Haghbayan, Juha Plosila</i>	168

Machine Learning for Big Data

A Novel Oversampling Technique To Handle Imbalanced Datasets <i>Ayat Mahmoud, Ayman El-Kilany, Farid Ali, Sherif Mazen</i>	177
Modelling Interleaved Activities Using Language Models <i>Eoin Rogers, Robert J. Ross, John D. Kelleher</i>	183
Predicting Business Process Bottlenecks In Online Events Streams Under Concept Drifts <i>Yorick Spenrath, Marwan Hassani</i>	190
Estimating Relationships In Multi-Dimensional Data Sets By Means Of Asymmetric Fuzzy Regression <i>Raphael A. Krauthann, Tobias Kruse, Hinnerk Jannis Mueller, Michael Stumpf, Peter Rausch</i>	197

Open and Collaborative Models and Simulation Methods

Fundamentals Of Digital Twins Applied To A Plastic Toy Boat And A Ship Scale Model

Icaro A. Fonseca, Henrique M. Gaspar.....207

Simulation Of The Conceptual Design Of Offshore Salt Caves For CO₂ Storage

Daniel Prata Vieira, Kazuo Nishimoto, Felipe Ferrari de Oliveira, Henrique M. Gaspar.....214

A Model For Forecasting Mental Fatigue In Maritime Operations

Thiago G. Monteiro, Henrique M. Gaspar, Houxiang Zhang, Charlotte Skourup.....221

INTEGRA: An Open Tool To Support Graph-Based Change Pattern Analyses In Simulated Football Matches

Nicolo Oreste Pinciroli Vago, Yuri Lavinias, Daniele Rodrigues, Felipe Moura, Sergio Cunha, Claus Aranha, Ricardo da Silva Torres.....228

Enabling Python Driven Co-Simulation Models With PythonFMU

Lars Ivar Hatledal, Houxiang Zhang, Frederic Collonval.....235

Finite – Discrete - Element Simulation

Lightweight Industrial Trailer By Using Composite Material - A New Concept Design

Federico Ceresoli, Andrea Buffoli.....243

Failure Analysis Of A Custom-Made Acetabular Cage With Finite Element Method

Martin O. Doczi, Robert Szoedy, Peter T. Zwierczyk.....250

A New Variable For Characterising Irregular Element Geometries In Experiments And DEM Simulations

Katalin Bagi, Akos Orosz.....256

Analysis Of The Stress State Of A Railway Sleeper Using Coupled FEM-DEM Simulation

Akos Orosz, Peter T. Zwierczyk.....261

A VCCT Approach Of Crack Propagation In Railway Wheels

Tamas Mate, Peter T. Zwierczyk.....266

Simulation and Optimization

Global Stability Of Fractional Positive Nonlinear Feedback Systems With Interval State Matrices

Tadeusz Kaczorek275

Scenario-Based Simultaneous Investment, Financing And Operational Planning

Mike Steglich280

Influence Of Company Sizes In Adapted Master Production Scheduling For Improving Human Working Conditions

Marco Trost, Thorsten Claus, Frank Herrmann287

Simulatable Reference Models To Transform Enterprises For The Digital Age – A Case Study –

Carlo Simon, Stefan Haag294

Simulation-Based Evaluation Of Reservation Mechanisms For The Time Window Routing Method

Thomas Lienert, Florian Wenzler, Johannes Fottner301

Mathematical Simulation Of Adjacent-Coupling Ammonia Absorptive Reactor

Wenchan Qi, Rene Banares-Alcantara308

Implementation Of The Optimizer Of SOA System Deployment Architecture

Adrian P. Wozniak315

Efficient Task Prioritisation For Autonomous Transport Systems

Maximilian Selmair, Vincent Pankratz, Klaus-Juergen Meier322

An Approach To Creating A Simple Digital Twin For Optimizing A Small Electric Concept Vehicle Drivetrain

Tamas Doka, Peter Horak328

Deviation In Energy Consumption On Aggregate Production Planning Level In Industrial Practice

Hajo Terbrack, Thorsten Claus, Frank Herrmann334

Modeling and Simulation for Performance Evaluation of Computer-based Systems

Modelling and Simulation of Data Intensive Systems - Special Session -

Computing Resilience Of Interconnected Systems By Piecewise Linear Lyapunov Functions

Alberto Tacchella, Armando Tacchella345

Towards Artificial Neural Network Hashing With Strange Attractors Usage

Jacek Tchorzewski, Agnieszka Jakobik.....354

Towards A Multiparadigm Approach To Model Energy Management In WSN For IoT Based Edge Computing Applications

Lucilla De Arcangelis, Mauro Iacono, Eugenio Lippiello361

3D-Stacked Memory For Shared-Memory Multithreaded Workloads

Sourav Bhattacharya, Horacio Gonzalez-Velez.....368

AWS EC2 Spot Instances For Mission Critical Services

Jerry Danysz, Victor del Rosal, Horacio Gonzalez-Velez376

A Simulation Study On A WSN For Emergency Management

*Lelio Campanile, Mauro Iacono, Fiammetta Marulli,
Michele Mastroianni*.....384

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

Probability Model Of Concepts Recovery In Small Sample Learning

*Alexander A. Grusho, Nick A. Grusho, Michael I. Zabezhailo,
Elena E. Timonina, Vladislav V. Kulchenkov*393

A Simple Dispatching Policy For Minimizing Mean Response Time In Non-Observable Queues With SRPT Policy Operating In Parallel

Mikhail Konovalov, Rostislav Razumchik.....398

Method For Bounding The Rate Of Convergence For One Class Of Finite-Capacity Markovian Time-Dependent Queues With Batch Arrivals When Empty

*Anastasiya Kryukova, Viktoriya Oshushkova, Alexander Zeifman,
Rostislav Razumchik*.....403

Author Index	407
---------------------------	-----

ECMS 2020

SCIENTIFIC PROGRAM

Simulation of Intelligent Systems

A MODEL FOR GROUND TRANSPORTATION SYSTEMS SIMULATION AT AIRPORTS UNDER CENTRALIZED CONTROL

Farid Saifutdinov and Jurijs Tolujevs
Transport and Telecommunication Institute
Lomonosova str. 1, LV-1019 Riga, Latvia
f.saifutdinov@mail.ru

KEYWORDS

Airport, Ground Transportation System, Simulation, Centralized Control.

ABSTRACT

This paper formulates requirements for a simulation model designed to study the movement of all types of vehicles in the ground space of the aerodrome on condition that a centralized control system provides continuous automatic control of their movement and sends commands of control to prevent dangerous situations. Examples of object movement models in 2D space are considered. The advantages of using a grid-based space for analyzing the interaction of objects are shown. It is concluded that commercial packages of processes modeling with discrete events or packages of transport flows modeling do not provide the conditions for implementing all the formulated requirements to the model. There are formulated principles of building models that can be implemented using universal programming languages. An example of developing a test model using the VBA language in MS Excel is described.

INTRODUCTION

All the main functions of organizing and controlling vehicles movement at the airport are currently performed by a person working as an air traffic controller. His main task is to ensure safety by preventing aircraft collision with each other, ground transport or other objects. Also, the vehicles themselves (aircraft, various types of cars, buses, tow trucks, etc.) are controlled today by people who perform the functions of pilots or drivers. Very often, the coordination of participants' actions in the transport process is made only by voice information exchange through the radio communication. It is obvious that the ideal level of safety of transport processes at airports cannot be achieved while maintaining a large share of human participation in the planning and implementation of the individual vehicle's movement. The main way to solve this problem is to increase the level of automation of traffic control processes in the ground space of the airport.

The document entitled "Vision 2050", published by International Air Transport Association (IATA 2011), notes that in the future, traffic control in the airfield zone will be transferred to automated systems, as a

result of which air and ground traffic controllers will primarily perform the role of operators and monitor the functioning of such systems. Polish specialists (Augustyn and Znojek 2015) describe the airport process functioning of the future, accompanying the path of a particular aircraft from the moment of its landing to the parking stand and from the parking stand to lining up position before takeoff. Centralized objects control on the territory of the airfield starts with obtaining accurate data about each object location in real time. The control system issues commands that set the speed and direction of movement of all moving objects, and this takes into account not only the values of physical parameters of objects (location, geometric dimensions, speed and movement direction), but also data such as the type of object, the final destination of the route, the function and the priority level.

An implementation of automatic centralized control, of course, does not cancel the use of decentralized (local) vehicle control methods. These methods include both conventional manual control, which is provided by drivers or pilots, and the most modern technologies for automatic control of unmanned vehicles. Decentralized control methods are based on the construction of the space dynamic model, which directly surrounds a specific vehicle. Both data from various types of sensors located onboard the vehicle, and data obtained during communication with other participants of transportation process, are used as source data for such a model. A lot of publications are devoted to the research of decentralized control methods (see, for example, (Le-Anh and de Koster 2006) and (Craighead et al. 2007)), but they are not a matter of consideration in this article.

In (Augustyn and Znojek 2015), the concept of dividing the airfield area by zones, the occupancy state of which should be taken into account by the automatic dispatcher, is briefly mentioned. In turn, the zones where vehicles are moving can be divided into cells that are sequentially occupied or vacated during the movement of a particular object. The idea of allocating controlled zones is described in detail (Tabares and Mora-Camino 2017). Fig. 1 shows the GSE (Ground Support Equipment) zones and directions in the vicinity of the aircraft parking area.

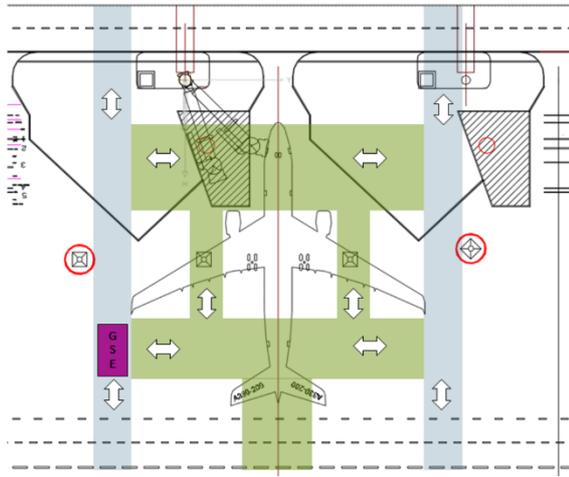


Figure 1: Proposed zoning and directions for GSE automated solution (Tabares and Mora-Camino 2017)

It is known that simulation is the most efficient method of studying transport processes at airports (Alomar et al. 2017). In order to study the process of mobile object control, a model that includes the following components must be developed:

- a model of objects movement on the selected paths and the model of the free movement on the grounds;
- a model for determining the location coordinates, speed, and direction of objects movement;
- a model of an automatic dispatcher that generates commands to control the movement of objects.

This work is devoted to the analysis of alternative methods for developing simulation models, for which the above-mentioned properties are an attribute. The result of an analysis that has been done, is the decision to develop special methods of conceptual and computer modeling that allow to solve set tasks of researching transport processes at airports in a centralized control environment.

A PROCESS MODELING PARADIGM CHOOSING

The main feature of the object movement model is the need to take into account their relative positions in 2D space. In this regard the real shape and size of each object should be taken into account, since one of the main tasks of the air traffic controller is to prevent collisions between both moving objects with each other and moving objects with obstacles. A straightforward way to solve this problem is to represent the shape of a moving or stationary object in the form of a polygon (Fig. 2). Then it becomes necessary to check the intersection condition of all polygons that display moving objects with their neighboring objects or obstacles nearby. The above-mentioned check in the model has to be carried out with a time interval of about 0.1 seconds, since during this time at a speed of 30-40 km/h the object can travel a distance of about 1 meter. The problem of checking the intersection condition of

two polygons can be solved using geometric modeling methods, but this problem will have to be solved 600 times per each minute of the model time, which with dozens of interacting objects will lead to a noticeable slowdown of the simulation program.

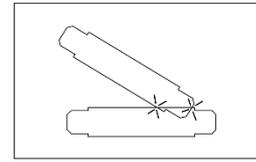


Figure 2: Objects collision detection represented by polygons

The second feature of the object movement model is the necessity to take into account accelerations and related changes in the objects speed movement. It is clear that the processes of acceleration and deceleration of aircraft will differ from those observed in ground vehicles. It is important to note that such processes can occur both at the initiative of drivers and in response to the commands of the automatic dispatcher. Fig. 3 shows the results of speed measuring of 10 aircraft on one particular taxiway during for 180 seconds (Mazur and Schreckenberg 2018). It is seen that the speeds vary over a wide range, for example, from 6 to 10 m/s.

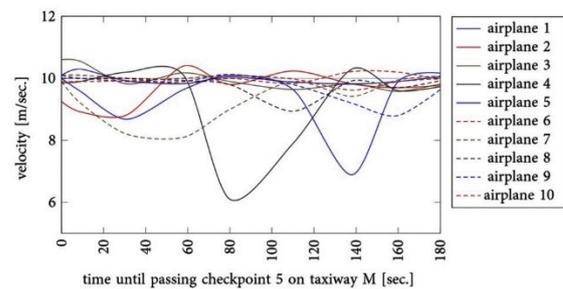


Figure 3: Speed-time plot of 10 airplanes taxiing on taxiway (Mazur and Schreckenberg 2018)

The two features of the object movement model that have already been mentioned above make it impossible to use the discrete event simulation paradigm. The necessity of application the time step “delta T” into a model is also related to the fact that a real automatic dispatcher also has to receive information about the current state of all traffic participants several times per second. The “delta T” method is used in commercial traffic flow simulation software packages, such as Vissim (PTV Vision). But in such models, the movement of objects is carried out mainly along the dedicated lanes, i.e. in mathematical terms, the models are one-dimensional, since they only take into account the distance between the cars following by each other. Models based on cellular automata are also one-dimensional (Mazur and Schreckenberg 2018). In traffic flow models, the principles of local control are applied, when a new speed value is determined for each car in the “delta T” step, depending on the speed of the car in

front and the distance to it. In such models there is no aim to collect data about individual objects and centralized control of these objects. Since standard methods of transport flow modeling do not allow to solve the set task of modeling in 2D space, there appears a necessity to consider other areas of modeling application for studying the interaction of controlled moving objects.

EXAMPLES OF MOVING OBJECTS MODELING IN 2D SPACE

Since the beginning of the 90s, transport systems based on Automated Guided Vehicles (AGVs) have been used in production and logistics. Most often, such systems are modeled under the assumption that individual vehicles move along fixed sections of the route that collectively make up the transport network. The Central control system usually sets only a driving route to each vehicle and then does not control its position during the trip. Conflict situations that may occur between vehicles are resolved by means of local control systems that are installed on the vehicles themselves. A typical example of a model created using the Plant Simulation package based on the Discrete Event Simulation paradigm can be found in (Selmair et al. 2019).

This work is interested in simulations of the systems with multiple free navigating AGVs. A serious research is the master thesis (de Groot 1997), in which moving objects and obstacles are described using polygons, and the “delta T” principle is used to display the dynamics of the process. To predict a collision of objects the geometric model mentioned above is used (see Fig. 2). The simulation model is implemented using the Python programming language. An example of a scenario simulation involving 10 AGVs is shown in Fig. 4.

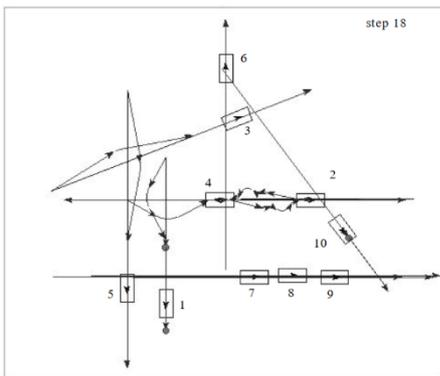


Figure 4: A simulation test with 10 AGVs (de Groot 1997)

A simulation model of the system with free navigating AGVs can also be found in the paper (Berman et al. 2003). It mainly focuses on local management of distinct AGVs in order to prevent AGVs from colliding with each other as well as with obstacles (Fig. 5). The simulation model is implemented using the Visual C++ programming language.

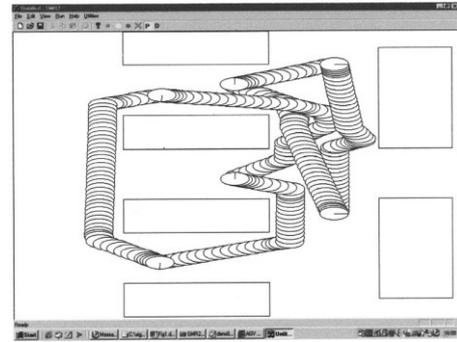


Figure 5: A simulation test with 6 AGVs (Berman et al. 2003)

APPLICATION OF 2D GRID SPACE

All the models described above used a continuous 2D space, in which the position of each point is described by a pair of coordinates (x, y) . It is in such space, in particular, a rather difficult task of identifying the intersection of graphic images of objects, described in polygons, was solved. This space additionally can be further divided into cells with coordinates $[j, i]$, where i and j are the row and column numbers, respectively. If dimensions for a cell are assigned for example, 2x2 meters, so the following coordinates conversion is trivial: $(10.7; 15.3) \rightarrow [6; 8]$. Since for each point with coordinates (x, y) , the coordinates of the cell in which it is located are known, many tasks of controlling the objects movement can be reduced for checking whether the object points belong to the corresponding cells or not. For example, two polygons are not in the intersection state if they do not have any sharing single cells. The accuracy of calculating the real distances between the points depends on the size of cell d . In Fig. 6 it is shown that two points p_1 and p_2 will be located in different cells both in the case of $\{p_1^1, p_2^1\}$, and in the case of $\{p_1^2, p_2^2\}$. Since the distance between the centers of the cells is $d\sqrt{2}$, the error in interpreting the position of the points in both cases will be equal to this value. For example, if $d = 2$ meters, the maximum error will be 2.83 meters.

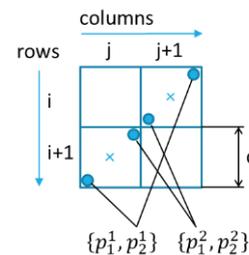


Figure 6: The limits in points location of neighboring cells

Since the accuracy of technical systems for determining the location of moving objects rarely exceeds this value, such accuracy can be considered sufficient, for example, when modeling the movement of vehicles on the airfield

area. In case of higher accuracy of determining the coordinates of objects, the value of d could be reduced, for example, to 1 meter.

The article (Szlapczynski 2006) describes the application of 2D grid space, which the author calls the term “raster grids”. The model shows the process of maneuvering ships in the water port area, where there is a risk of collision between each other or with port facilities. It should be noted that the concept of grid-based space is widely implemented in modeling processes that occur in certain geographical areas. Most often, these processes belong to class of a very slow ones, for example, the processes of urban areas development. There are models that were created to research also relatively fast processes, such as floods or fires. The article (Mazzoleni et al. 2006) reports on the development of a universal software package with which ecological models such as animal distribution, fire propagation, seed dispersal, and different models for simulation of vegetation dynamics have been created. The main part of the package has been implemented in Visual Basic (VB6) using ActiveX and COM components. The article (Taillandier et al. 2016) discusses how to improve the accuracy of raster models by adding objects represented by vector models to them.

SELECTED METHOD OF MODEL IMPLEMENTATION

The simulation examples discussed above, which have at least some properties that are similar to the properties of the set task of modeling transport processes on the airfield, indicate that such problems can be solved only with the help of universal programming languages. Commercial process simulation packages with Discrete Event Simulation packages or traffic flow simulation packages do not provide the conditions for implementing all the stated requirements for the model. The principled decision related to object movement processes modeling with an orientation on the application of universal programming languages are described below.

1. Moving objects and obstacles are located in a 2D space in which it uses both continuous coordinates (x, y) and cell coordinates $[i, j]$.
2. Each moving or stationary object is described with the usage of a convex polygon, the points of which can be located in several cells of a discrete space.
3. One of the points of a moving object is declared as a reference point. The current position of the object is determined by the coordinates of the reference point (x_{ref}, y_{ref}) and the angle of rotation α relative to the northbound angle.
4. It is agreed that there are eight discrete directions of moving objects orientations, i. e. $\alpha \in \{N, NE, E, SE, S, SW, W, NW\}$.
5. For each class of moving objects, eight graphical models are created that display the location of the occupied cells by the specified coordinates (x_{ref}, y_{ref}) and the rotation angle α (Fig. 7).

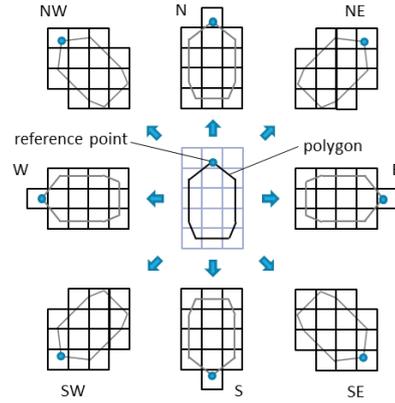


Figure 7: Eight graphic models of a moving object

6. The actual trajectory of the R_{real} object is replaced by a sequence of cells according to Bresenham's line algorithm. The simulated R_{sim} path passes through the center points of the cells that make up this path (Fig. 8). At any given time, the p_{ref} reference point can be located at any point in the R_{sim} path, depending on the distance passed.

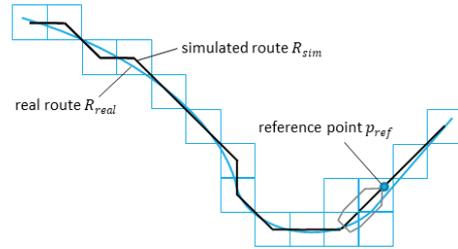


Figure 8: The replacement of real route with a sequence of cells

The first experiments, where the principles described above have been implemented, were conducted in MS Excel using macros written in the VBA programming language. The dimensions of the Excel table cells that display the airfield plan are selected so that they form a grid-based space with the cell size $d = 2$ meters. Fixed routes for aircraft and ground vehicles, as well as work areas and obstacles, are coded by assigning a color to the corresponding cells. Graphic models of moving objects are created in the form of ShapeRange. The objects positions on the screen are set using the Left, Top, and Rotation properties of these objects. Continuous animation is provided by interrupting the execution of a VBA program using the DoEvents function. In Fig. 9 there is a fragment of the simulation model, which shows the paths of ground vehicles and aircraft.

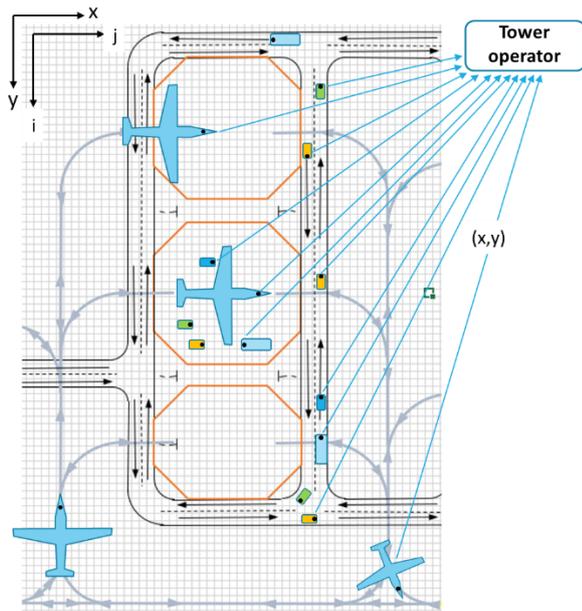


Figure 9: A simulation model fragment using grid-based space

The model generates a data stream from moving objects, which is recorded in the form of a protocol (see Fig. 10). This protocol displays events that can occur at any time. The main task of each event is to transmit the moving object actual position by the radio channel of local coordinates (x, y) . In this way, the data flow from the supposed real equipment for determining the location of moving objects is simulated. Based on this data, the speeds, accelerations, and directions of movement of objects are calculated.

Event protocol			
Real-time stamp	Object ID	X-coordinate	Y-coordinate
10:45:12.84	ob_2	323.5	156.5
10:45:12.91	ob_1	247.2	140.7
10:45:13.03	ob_2	323.5	147.2
10:45:13.12	ob_3	323.5	170.6
10:45:13.84	ob_1	256.4	140.7

State protocol								
Discrete time	ob_1				ob_2			
	X-coord.	Y-coord.	i-cell	j-cell	X-coord.	Y-coord.	i-cell	j-cell
10:45:12.8	246.2	140.7	71	124	323.5	156.5	79	162
10:45:13.0	248.2	140.7	71	125	323.5	154.5	78	162
10:45:13.2	250.2	140.7	71	126	323.5	152.5	77	162
10:45:13.4	252.2	140.7	71	127	323.5	150.5	76	162
10:45:13.6	254.2	140.7	71	128	323.5	148.5	75	162
10:45:13.8	256.2	140.7	71	129	323.5	146.5	74	162

Figure 10: A creation of a state protocol based on event protocol data

By processing the event protocol data, a secondary protocol is created with the name "State protocol". This Protocol uses a discrete time with a "delta T" step, which is selected taking into account the speed of automatic system of situation analysis in the transport

system of the aerodrome. In Fig. 10 an example where the time was equal to 0.2s is shown. The state protocol, along with the coordinates (x, y) , at the same time contains the coordinates (i, j) that show the position of the object reference point in grid-based space.

HOW TO USE THE GTSS PROGRAM

The program developed by the authors was named GTSS (Ground Traffic Scenario Simulation). This program is not intended for statistical modeling of aircraft and ground vehicles mass flows, but for modeling specific, precisely described scenarios in order to test the feasibility of centralized control algorithms. The GTSS program is a universal tool for solving the tasks described above, since it can be applied to research the transport process at any airport. To prepare the program for researching a new object it is necessary to do the following:

- on the "Aerodrome" Excel sheet, create a network of square cells by setting, for example, the following parameter values: ColumnWidth=0.5, RowHeight=5;
- set the display scale of the airfield plan, for example, when the side of the cell is 1 or 2 meters;
- display the airfield plan graphically, for example, as shown in Fig. 9;
- for each fixed path of object moving, mark all cells, which are a part of the path, with a color and number (see Fig. 8);
- create classes of movable objects as Shape objects and create eight "cell" models for each class, as shown in Fig. 7.

The rest of the model source data is set by filling in specially prepared tables. The tables of the "Scenario" type, which describe the processes of specific mobile objects appearing at specified points in the transport network of the airfield, are of particular importance.

CONCLUSIONS AND FUTURE WORK

The chosen method of modeling, which uses discretization both space and time, makes it possible to significantly simplify the solution of problems of analyzing the relative positions of objects in 2D space. The developed model of object movement in 2D space is the basis for implementing the model of an automatic dispatcher that generates commands for controlling the movement of objects. The data in the state protocol is an information base for various algorithms of analyzing the current situation and predicting the development of this situation for the next minutes and seconds. In particular, as a result of this analysis, the situations that require the intervention of the dispatcher should be identified. Examples of operational solutions for an automatic system or dispatcher are the following:

- to stop the movement of one or more objects if they are in danger of colliding with each other or with an obstacle;
- to prohibit all vehicles other than one or more precisely defined moving through a specific

intersection in order to provide their arriving at destination as soon as possible;

- to report new routes to distinct vehicles if congestions occur at specific nodes in the transport network.

This study is particularly at the stage of developing models of automatic decision-making for operational control of the ground vehicles and aircraft movement on the airfield.

REFERENCES

- Alomar I., Tolujevs J., Medvedevs A. 2017. "Simulation of Ground Vehicles Movement on the Aerodrome". In *Procedia Engineering*, Vol. 178, 340-348.
- Augustyn S., Znojek B. 2015. "The new vision in design of airport". *Scientific Research & Education in the Air Force – AFASES*, Vol. 2, 369-372.
- Berman S., Edan Y., Jamshidi M. 2003. "Navigation of Decentralized Autonomous Automatic Guided Vehicles in Material Handling". *IEEE Transactions on Robotics and Automation*, Vol. 19, no. 4, 743-749.
- Craighead J., Murphy R., Burke J. et al. 2007. "A survey of commercial & open source unmanned vehicle simulators". In *IEEE International Conference on robotics and automation (ICRA)*, 852-857.
- de Groot R. M. 1997. *Dynamic traffic control of free navigating automatic guided vehicles*. Master thesis University of Twente, Computer Sciences, SPA.
- International Air Transport Association 2011. *Vision 2050*, Singapore, February 12, 2011.
- Le-Anh T., de Koster R. 2006. "A Review of Design and Control of Automated Guided Vehicle Systems". *European Journal of Operational Research*, Vol. 171, Issue 1, 1-23.
- Mazur F., Schreckenber M. 2018. "Simulation and Optimization of Ground Traffic on Airports using Cellular Automata". *Collective Dynamics*, 3, A14, 1-22.
- Mazzoleni S., Giannino F., Mulligan M. et al. 2006. "A new raster-based spatial modelling system: 5D environment. Summit on Environmental Modelling and Software". In *Proceedings of the iEMSs*, Vol. 1, Burlington, USA.
- Selmair M., Hauers S., Gustafsson-Ende L. 2019. "Scheduling Charging Operations of Autonomous AGVs in Automotive In-house Logistics". In *Simulation in Production and Logistics*, Wissenschaftliche Scripten, Auerbach, 315-324.
- Szlapczynski R. 2006. "A new method of ship routing on raster grids, with turn penalties and collision avoidance". *The Journal of Navigation*, Vol. 59, 27-42.
- Tabares D. A., Mora-Camino F. 2017. "Aircraft ground handling: Analysis for automation". In *17th AIAA Aviation Technology, Integration, and Operations Conference*, Denver, United States. AIAA.
- Taillandier P., Banos A., Drogoul A. et al. 2016. "Simulating Urban Growth with Raster and Vector Models: A Case Study for the City of Can Tho, Vietnam". In *Autonomous Agents and Multiagent Systems AAMAS 2016*, Lecture Notes in Artificial Intelligence, Springer, 154-171.

AUTHOR BIOGRAPHIES



FARID SAIFUTDINOV graduated from the Riga Aeronautical Institute in 2016 with the qualification of Manager of Transportation Enterprise and Professional Master's degree in business management. He has been working in the aviation sphere in the field of air navigation at airports in Kazakhstan since 2009. He has been working as an air traffic controller, air traffic controller instructor, senior controller and supervisor. He is currently a PhD student at the Transport and Telecommunication Institute, Riga, where he studies and implements new technologies for controlling transport processes at airports. His email address is: f.saifutdinov@mail.ru



JURIJS TOLUJEVS is a professor of Mathematical Methods and Modelling at the Transport and Telecommunication Institute (Riga, Latvia). Besides, he is a project manager at the Fraunhofer Institute for Factory Operation and Automation IFF in Magdeburg, Germany. He received a doctoral degree in automation engineering from the Riga Technical University. He also received a habil. degree in computer science from Otto von Guericke University Magdeburg. His research interests include the simulation-based analysis of production and logistics systems, protocol-based methods for analyzing processes in real and simulated system. He is an active member in the ASIM, the German organization of simulation. His email address is: jurijs1949@gmail.com.

SIMULATION OF UNDERWATER COLOR IMAGES USING BANDED SPECTRAL MODEL

Denis A. Shepelev
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia;
Moscow Institute of Physics and Technology
Institutskiy per. 9, Dolgoprudny 141701, Russia
E-mail: shepelev@iitp.ru

Egor I. Ershov
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: ershov@iitp.ru

Valentina P. Bozhkova
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19, Moscow, 127051, Russia
E-mail: bgk@iitp.ru

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

Underwater photography; Dataset augmentation; Color rendering; Spectral color models; Underwater image enhancement

ABSTRACT

This paper considers methods of color underwater images simulation based on terrestrial images. These methods avoid the expensive process of collecting real data to develop algorithms for improving underwater images. To model light attenuation underwater, existing works often use channel-wise approximation of the Beer–Lambert–Bouguer law. The accuracy of such approximation significantly depends not only on the camera sensitivities but also on its color calibration parameters. In this paper, we propose a method for underwater images simulation, based on the banded spectral model, which is devoid of the last drawback. In the proposed approach, the spectrum of the registered radiance is estimated from the camera response and attenuation modeling is performed in the spectral space. The average reproduction angular error in simulation based on the banded spectral model is 20-30% less than in channel-wise approximation.

INTRODUCTION

Recently, interest in creating algorithms for the analysis and improvement of underwater images has grown markedly. The latter are required, for example, during automatic inspection of underwater objects (Foresti 2001), prevention of drownings in public swimming pools (Lavest et al. 2002), research and visualization of underwater archaeological artifacts (Kahanov and Royal 2001; Mangeruga et al. 2018; Skarlatos et al. 2016), and so on. This area of study has relevance in widespread practice. It is worth mentioning that there has been a massive transition of mobile phone manufacturers to the IP68 standard (National Electric

Manufacturers Association 2004) with increased requirements for the degree of protection of electronic devices from water.

When developing any image processing algorithms, including underwater ones, formal criteria are required that evaluate the quality of their results. This is necessary to rank the algorithms and to optimize the algorithms according to their parameters and to determine whether the problem was solved with the proper quality. In image enhancement tasks, these formal criteria are sometimes based on full-scale psycho-physical experiments (Gracheva et al. 2020), but this approach is too time-consuming in the development of a new method when thousands of comparisons of different versions of the algorithm are required. Therefore, it is desirable that the quality criteria are formally computable. With this, it becomes possible to automatically compare versions of the algorithm.

As a rule, a statistic (for example, the average value) of a certain image comparison metric serves as a formal criterion for the quality of the algorithm’s performance as a whole. This metric is calculated on a set of images in which an ideal target image is specified for each input image, and the result of the algorithm is compared with this ideal image (see figure 1). In addition, a good set of such data should be very diverse. The question is, how can one create a large and diverse set of images with corresponding ideals for the task of improving underwater images?

It is clear that the direct collection of such data is difficult. One may notice that as the complexity of data gathering increases, both the final size and diversity of the dataset decrease. For example, for the task of document recognition 500 video sequences were collected (Arlazarov et al. 2019), for the color segmentation and color constancy tasks in the controlled conditions – 432 images (Smagina et al. 2020), and for the underwater image improvement problem in the controlled conditions – 82 images (Duarte et al. 2016). At the moment,

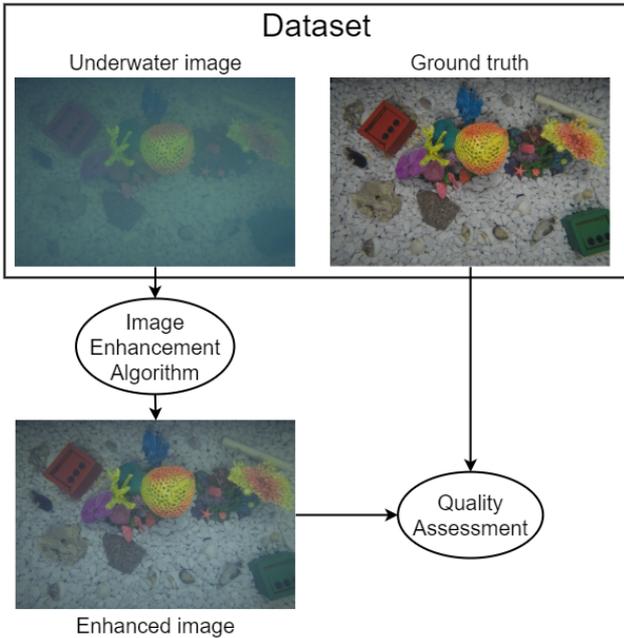


Figure 1: Quality assessment scheme for image enhancement algorithm. Images were taken from (Duarte et al. 2016) and were modified for illustrative purposes.

a simulation is used to obtain a set of images equipped with ideals.

Thorough simulation of image formation (rendering) implies a formal description of all the objects in the scene, their color parameters, light field, etc., as well as the calculation of multiple reflections and light scattering. It is extremely complex and time-consuming computationally (Aranha 2005; Boffety et al. 2012). At the same time, rendering models may turn out to be oversimplified, and the simulation results being far from photorealistic.

The combined approach is more widely used – simulation by converting images taken in the air. In this approach, one of the inputs of the simulator is a full-scale image obtained under certain conditions, and the result is an underwater image that models other registration conditions. The fundamental limitation of this method should be noted: the typical semantic content of ground scenes differs from underwater ones.

These methods can be divided into three groups: methods that use direct simulation (Anwar et al. 2018; Chang et al. 2019; Hu et al. 2018; Li et al. 2016; Peng and Cosman 2017; Schechner and Karpel 2004, 2005; Zhao et al. 2015; Uplavikar et al. 2019; Ding et al. 2019; Yu et al. 2018; Li et al. 2017; Aranha 2005; Boffety et al. 2012), neural network based methods (Fabbri et al. 2018; Li et al. 2018) and methods using a combination of direct simulation and neural networks (Li et al. 2017). First, let us consider the second group. Methods from this group are based on generative-adversarial neural networks which are trained to transfer image style from “ideal” ones (for example, terrestrial images, as in (Li et al. 2018)) to the target “distorted” underwater ones. Their statement of the simulation problem in reality is a style transfer problem since no correct

answer is given for the input image, but the goal is to transform it so that it belongs to some target class of images without any guarantee the correct colors of the objects will be obtained in the result.

The methods of the first and the third groups use a particular physical underwater image formation model, and the quality of their work depends on the accuracy of this model. The usage of neural networks, or lack thereof, have no effect on this fact. Any of these methods can be used and are used to train neural network algorithms to improve the quality of underwater images (Anwar et al. 2018; Hu et al. 2018; Uplavikar et al. 2019; Ding et al. 2019; Yu et al. 2018; Li et al. 2017). Thus, the improvement of augmentation models leads to an improvement of the accuracy of underwater image enhancement methods based on neural networks.

In this paper, we study the problem of the effect of light attenuation simulation underwater based on terrestrial images. The attenuation coefficient in water is hundreds of times higher than in air which leads to a decrease in observed brightness of the object when increasing its submerging depth or its distance to the camera. Moreover, light attenuation in water increases with wavelength which results in color distortions in underwater images. To simulate the attenuation of light underwater, the works presented in the literature use a channel-wise approximation of the Beer–Lambert–Bouguer law. In this paper, we show the incorrectness of this approach and propose a method for the attenuation effect simulation using spectral models (Nikolaev and Nikolayev 2005; Nikolaev et al. 2006; Nikolaev and Nikolayev 2007). Spectral models allow us to estimate the radiation from the camera response, imposing model restrictions on the spectrum. Our proposed approach allows us to achieve a more accurate simulation of the effect of light attenuation underwater compared to the approach based on channel-by-channel approximation which is demonstrated by numerical experiments.

CLASSICAL MODEL OF UNDERWATER IMAGE FORMATION

The authors of the first underwater image formation models (Jaffe 1990; McGlamery 1980) suggested to consider separately the illumination of the sensor created by radiation reflected from the scene objects, and illumination which is not directly related to the observed objects. This additional illumination, provided by the light scattered by the water, was called “backscattered light” F_{bs} .

Considering the radiation that came from objects, they identified two main components in it. The first is light reflected from the object (the authors called it “direct light”) – it gradually weakens in the water and creates illumination F_d on the sensor. The rate of attenuation of direct light, travelling in the water media, depends on the wavelength. This is associated with color distortion amplifying with the distance to the object. Another component, called “forward-scattered light” F_{fs} , represents a small part of the scattered di-

rect light. The article (Schechner and Karpel 2005) shows that the contribution of F_{fs} to contrast degradation is much smaller than F_{bs} , so the component F_{fs} can be omitted in simulation, that we do in this paper.

Thus, the final signal is represented as the sum of two components that are formed independently, namely:

$$F(\vec{x}, \lambda) = F_d(\vec{x}, \lambda) + F_{bs}(\vec{x}, \lambda), \quad (1)$$

where \vec{x} are the coordinates in the image corresponding to the direction in which light comes from the scene; λ is the wavelength. To model the final image, the expression (1) must be integrated which results in the final image being the sum of two images:

$$\begin{aligned} \vec{f}(\vec{x}) &= \vec{f}_d(\vec{x}) + \vec{f}_{bs}(\vec{x}), \\ \vec{f}_d(\vec{x}) &= \int F_d(\vec{x}, \lambda) \vec{\chi}(\lambda) d\lambda, \\ \vec{f}_{bs}(\vec{x}) &= \int F_{bs}(\vec{x}, \lambda) \vec{\chi}(\lambda) d\lambda, \end{aligned} \quad (2)$$

where $\vec{\chi}(\lambda)$ – sensor sensitivity curves. Hereinafter the integration range is limited by visible spectrum.

According to (2), each component can be modeled independently, so the issues of their modeling can be considered separately from each other. In this paper, using the direct component as an example, we consider the problems of underwater color images simulation based on terrestrial images, and how to solve them using spectral models.

DIRECT LIGHT MODELING

The direct light component is described by the Beer–Lambert–Bouguer law:

$$\begin{aligned} F_d(\vec{x}, \lambda) &= C(\vec{x}, \lambda) T(\rho(\vec{x}), \lambda), \\ T(\rho(\vec{x}), \lambda) &= e^{-\beta(\lambda) \rho(\vec{x})}, \end{aligned} \quad (3)$$

where $\rho(\vec{x})$ is the distance between the optical center of the camera and the object point, projected to a point \vec{x} of the image; $C(\vec{x}, \lambda)$ is radiance of the object; $\beta(\lambda)$ is the attenuation coefficient; and $T(\rho(\vec{x}), \lambda)$ is called the transmission map of the water media.

From equations (2) and (3), it follows that the direct light can be modeled as follows:

$$\begin{aligned} \vec{f}_d(\vec{x}) &= \int C(\vec{x}, \lambda) T(\rho(\vec{x}), \lambda) \vec{\chi}(\lambda) d\lambda = \\ &= \int C(\vec{x}, \lambda) e^{-\beta(\lambda) \rho(\vec{x})} \vec{\chi}(\lambda) d\lambda. \end{aligned} \quad (4)$$

Thus, for modeling the direct light, besides distances to points in the scene $\rho(\vec{x})$, the attenuation coefficient of water $\beta(\lambda)$, and camera sensitivity curves $\vec{\chi}(\lambda)$, one also need to know the spectrum of $C(\vec{x}, \lambda)$ in the visible range.

In the case of modeling based on terrestrial images, instead of $C(\vec{x}, \lambda)$, only the corresponding sensor responses $\vec{c}(\vec{x}) = \int C(\vec{x}, \lambda) \vec{\chi}(\lambda) d\lambda$ are known, which does not allow direct use of equation (4). This leads to an ill-posed problem of estimating the spectrum based on the camera response.

In the works (Anwar et al. 2018; Chang et al. 2019; Hu et al. 2018; Li et al. 2016; Peng and Cosman 2017; Schechner and Karpel 2004, 2005; Zhao et al. 2015; Uplavikar et al. 2019; Ding et al. 2019; Yu et al. 2018; Li et al. 2017) when modeling, the researchers act radically by approximating the expression (3) channel-by-channel which can be written as follows:

$$\vec{f}_d(\vec{x}) = \vec{c}(\vec{x}) \otimes e^{-\vec{\beta} \rho(\vec{x})} \quad (5)$$

where $\vec{c}(\vec{x})$ – sensor response at \vec{x} ; $\vec{\beta}$ – the vector of attenuation; \otimes – termwise multiplication; $e^{\vec{a}}$ – termwise exponentiation.

From our point of view, this approach is an oversimplification. Even if the approximation of weak dependence of physical effects on wavelength is adequate, which allows us to switch to scalar expressions for different spectral ranges, we cannot consider that the camera measures radiance in spaced spectral ranges – the camera spectral sensitivity curves for different channels significantly overlap (Akkaynak and Treibitz 2018).

Also, in the context of equation (5), the important question is in which color space the modeling should be carried out, in the original color space of the camera, or in some linearly related to it, for example, in the CIE XYZ space, to which the original color space is usually converted by some linear transformation. So, if in a certain colour space the result of the modeling using equation (5) match the ground truth, obtained according to equation (4), in other linearly connected color space the error is guaranteed. Therefore for channel-wise simulation, it is reasonable to consider which color space channel-wise approximation is correct. This was not considered in the previously mentioned works.

In such cases, one should use spectral models that have already been studied in detail in the field of technical color vision related to the ensuring color constancy (Maloney 1986; Finlayson et al. 1993; Nikolaev and Nikolayev 2005; Nikolaev et al. 2006; Nikolaev and Nikolayev 2007; Chong et al. 2007; Gusamutdinova et al. 2017) and solve the previously mentioned problems of channel-wise approximation. According to this approach, it is necessary to convert the responses of the camera to the spectra of the incident radiation where the simulation operations are consistent with the model of image formation. In addition, the total simulation error is measured in the initial space of the responses in order to assess the quality of the approach.

Using an approach based on spectral models, we are able to achieve more accurate simulation compared to the approach based on channel-wise model when applied to direct light simulation.

SIMULATION BASED ON CHANNEL-WISE MODEL

To simulate according to equation (5), one needs to know the attenuation vector $\vec{\beta}$. The attenuation vectors $\vec{\beta}$ used in works (Anwar et al. 2018; Hu et al. 2018; Li et al. 2016; Peng and Cosman 2017; Schechner and Karpel 2004, 2005; Zhao et al. 2015) were not selected

for the reasons of simulation accuracy, so they cannot be used when comparing with spectral methods. In order to compare the channel-wise and spectral approaches, we consider the following channel-wise model that does not depend on the specific choice of the attenuation vector $\vec{\beta}$:

$$\begin{aligned} \vec{f}_d^{CW}(\vec{x}) &= \vec{c}(\vec{x}) \otimes \vec{t}(\rho(\vec{x})), \\ \vec{c}(\vec{x}) &= \int C(\vec{x}, \lambda) \vec{\chi}(\lambda) d\lambda, \\ \vec{t}(\rho) &= \int e^{-\beta(\lambda)\rho} \vec{\chi}(\lambda) d\lambda. \end{aligned} \quad (6)$$

Further in the paper, simulation using model (6) will be referred to as CW based simulation.

When the ranges of sensitivity curves intersect in modern cameras, this approximation is not correct. Unless, the restrictions are imposed on the functions $C(\vec{x}, \lambda)$ and $T(\rho, \lambda)$.

SIMULATION BASED ON SPECTRAL MODELS

In order to directly use equation (4), it is necessary to estimate the spectrum of $\hat{C}(\lambda, \vec{x})$ by $\vec{c}(\vec{x})$. The following spectral models are already known from the literature: banded spectral, Gaussian, von Mises (Nikolaev and Nikolayev 2005; Nikolaev et al. 2006; Nikolaev and Nikolayev 2007). The use of spectral models for simulating underwater images based on terrestrial images makes this approach not only consistent with the classic underwater imaging models but also more accurate than the channel-wise approach which we demonstrate using the banded spectral model.

Banded spectral model

The key constraint of the banded spectral model (BSM) is statement that any spectral function $C(\lambda)$ can be expressed as follows:

$$C(\lambda) = \vec{c}_{BSM}^T \vec{\delta}(\lambda), \quad (7)$$

where \vec{c}_{BSM} – is some constant vector, and $\vec{\delta}(\lambda)$ is a vector function of the wavelength so that each of its components are:

$$\delta_i(\lambda) = \begin{cases} 1, & \text{if } \lambda \in \Delta_i \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

and $\{\Delta_i\}$ are such that $i \neq j : \Delta_i \cap \Delta_j = \emptyset$. Functions $\delta_i(\lambda)$ form a basis in the space of spectral functions, and vector \vec{c}_{BSM} represents the coordinates of the spectral function $C(\lambda)$ in the basis.

It can be shown that the transition from the sensor response space to the parameter space of the banded spectral model is a linear transformation:

$$\vec{c} = L \vec{c}_{BSM}, \quad (9)$$

where \vec{c} – sensor response; \vec{c}_{BSM} – BSM coordinates; L – matrix, with each element equals:

$$L_{ij} = \int_{\Delta_j} \chi_i(\lambda) d\lambda. \quad (10)$$

Assuming that L is invertible we get:

$$\vec{c}_{BSM} = L^{-1} \vec{c}, \quad (11)$$

Note that in the case of color space linear conversion, the matrix L will be changed as follows:

$$L' = M L, \quad (12)$$

where M is a color space linear transformation matrix. From equations (11) and (12) it follows that parameters \vec{c}_{BSM} are invariants with respect to the color space linear conversion.

For a sensor that has an integral for each sensitivity curve equal to 1, you can get the following properties:

- For all i and j : $0 \leq L_{ij} \leq 1$.
- For every row i : $\sum_j L_{ij} \leq 1$.
- If $\forall i : \Omega_i \subset \cup_j \Delta_j$ where $\Omega_i = \{\lambda \in \mathbb{R}_+ : \chi_i(\lambda) \neq 0\}$, then

$$\vec{e} = L \vec{e}, \quad (13)$$

where $\vec{e} = (1 \ 1 \ 1)^T$.

The latter property shows that a unit equi-energy spectrum that corresponds to an all-ones vector in the camera space in the banded spectral representation will also correspond to a unit equi-energy spectrum. Without loss of generality, we will assume that the integral for each sensitivity curve is equal to 1.

The coordinates of the spectral function $C(\lambda)$ in the spectral parameter space are as follows:

$$\vec{c}_{BSM} = L^{-1} \int C(\lambda) \vec{\chi}(\lambda) d\lambda. \quad (14)$$

Multiplication and addition of functions that satisfy the limitations of the banded spectral model corresponds to the termwise addition and multiplication of their parameters.

BSM based simulation

Modeling direct light based on BSM can be expressed as follows:

$$\begin{aligned} \vec{f}_d^{BSM}(\vec{x}) &= L \left(L^{-1} \vec{c}(\vec{x}) \otimes L^{-1} \vec{t}(\rho(\vec{x})) \right), \\ \vec{c}(\vec{x}) &= \int C(\vec{x}, \lambda) \vec{\chi}(\lambda) d\lambda, \\ \vec{t}(\rho) &= \int e^{-\beta(\lambda)\rho} \vec{\chi}(\lambda) d\lambda. \end{aligned} \quad (15)$$

Further in the paper, simulation using model (15) will be referred to as BSM based simulation.

As can be seen from equation (10), the L matrix is determined not only by the sensitivity curves $\vec{\chi}(\lambda)$ but also by the intervals of the basic functions $\{\Delta_i\}$, which must be pre-selected. Next, we propose a procedure for finding the optimal parameters of the spectral model for simulating direct light.

Optimal parameters determination of the BSM

To find the optimal matrix for BSM, we used the following algorithm. BSM matrix L depends on the

sensor sensitivities $\vec{\chi}(\lambda)$ and $\{\Delta_i\}$ intervals. Let us assume that $\{\Delta_i\} = \{[350, \lambda_l], [\lambda_l, \lambda_r], [\lambda_r, 800]\}$, where 350 and 800 are chosen as the typical boundaries of the cameras visible range. Varying λ_l and λ_r one can obtain various BSM matrices $L(\lambda_l, \lambda_r)$. For the given sensitivities we need to find λ_l and λ_r of matrix $L(\lambda_l, \lambda_r)$ which will achieve better simulating results.

To find the best parameters we solve the following optimization problem. For the given sets of spectra $\{C_n(\lambda)\}_{n=1}^N$, attenuation coefficients $\{\beta_m(\lambda)\}_{m=1}^M$ and ranges $\{\rho_k\}_{k=1}^K$, we find λ'_l and λ'_r that minimize the following error function:

$$\lambda'_l, \lambda'_r = \arg \min_{\lambda_l, \lambda_r} E(\lambda_l, \lambda_r)$$

$$E(\lambda_l, \lambda_r) = \sum_{n,m,k=1}^{N,M,K} \varphi\left(\vec{f}_{nmk}, \vec{f}_{nmk}^{BSM}(\lambda_l, \lambda_r)\right), \quad (16)$$

where \vec{f}_{nmk} – ground truth response of the sensor $\vec{\chi}(\lambda)$ obtained by simulating light $C_n(\lambda)$ traveled a distance ρ_k through a water with attenuation coefficient $\beta_m(\lambda)$ using equation (4); $\vec{f}_{nmk}^{BSM}(\lambda_l, \lambda_r)$ – response of the sensor $\vec{\chi}(\lambda)$ obtained by simulating light $C_n(\lambda)$ traveled a distance ρ_k through a water with attenuation coefficient $\beta_m(\lambda)$ according to equation (15) using BSM matrix $L(\lambda_l, \lambda_r)$; $\varphi(\cdot, \cdot)$ – reproduction angular error (Finlayson et al. 2017). We solve this optimization problem by exhaustive search assuming $\lambda_l < \lambda_r$.

EVALUATION

To compare the quality of the described simulating algorithms, we conducted the following numerical experiments. For attenuation coefficients we used the vertical attenuation coefficients of 10 types of Jerlov’s waters (Jerlov 1968) which are illustrated in figure 3. For each type of water, equation (4) was used to simulate the ideal direct light of underwater images of the color target Digital ColorChecker SG (Babelcolor 2005) at various distances from 1 to 10 meters.

Next, for each type of water, using equations (6) and (15), we simulated images of the direct light of the color target for the previously specified distances using the CW based and BSM based algorithms, respectively. The parameters of the BSM were pre-optimized for the selected sensor for all distances, attenuation coefficients, and reflectances of the color target patches used as a set of radiations.

Further, for the obtained pairs of ideal images and images simulated by one of the algorithms, the average values of the angular reproduction error (Finlayson et al. 2017) for all patches, distances, and attenuation coefficients were calculated; in addition, for each type of water, the average reproduction errors for all patches and distances were calculated separately (see the table 1).

Numerical experiments were performed for two cameras: Nikon D90 and Canon 500D, the sensitivity curves of which were taken from (Jiang et al. 2013). Sensitivities and calculated optimal BSM parameters of the cameras are shown in figure 2.

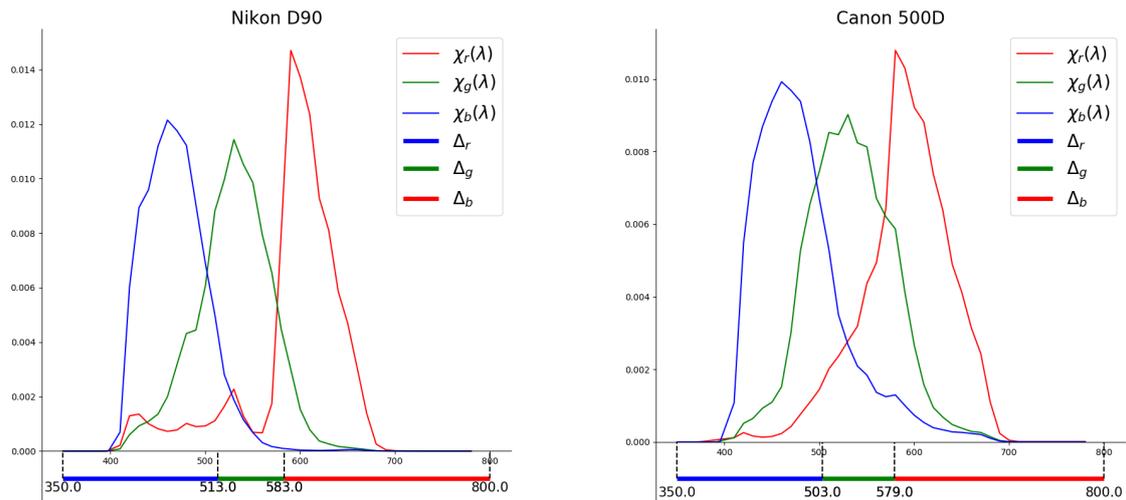
Table 1: Errors of channel-wise simulation using equation (6) (columns “CW”) and simulation based on the banded spectral model using equation (15) (columns “BSM”) for two cameras: Nikon D90 and Canon 500D. The “All” line contains average reproduction errors (Finlayson et al. 2017) for all water types, for all Digital ColorChecker SG color target patches, and for all distances from 1 to 10 meters. In the other lines, the average reproduction errors are shown for each type of water separately. For each camera, the lowest average errors in the line are shown in bold.

	Nikon D90		Canon 500D	
	CW	BSM	CW	BSM
All	3.10	2.21	3.20	2.44
I	2.91	1.12	2.70	1.62
IA	2.93	1.12	2.71	1.64
IB	2.96	1.13	2.73	1.66
II	2.97	1.13	2.72	1.72
III	2.92	1.24	2.63	1.99
1C	2.75	1.91	2.48	2.76
3C	2.42	2.29	2.41	2.91
5C	1.99	2.65	2.49	2.80
7C	3.10	3.12	3.98	2.78
9C	6.03	6.43	7.10	4.49

According to the table 1, the proposed method when compared to the channel-wise one shows better results. So, for all oceanic water types (see lines “I”, “IA”, “IB”, “II”, “IIP”), average reproduction angular errors of the BSM based algorithm (columns “BSM”) are less than the corresponding errors of the CW algorithm (columns “CW”). For Nikon D90 on the “7C” water type, the difference between algorithms errors are negligible (0.02 degrees), and for the “9C” water type, the errors for both are quite large (more than 6 degrees). I.e. for the Nikon D90 camera, the proposed method significantly worse on the water type “5C”, and for the Canon 500D camera – on the water types “1C”, “3C”, “5C”. The water types in the lower half of the table are coastal, where a significant drop in brightness is observed at distances around 3m, which isn’t taken into account by reproduction angular error. We believe that in such cases, it is reasonable to use metrics that take into account errors not only in chromaticity but also in brightness. This can significantly change the estimated accuracy of algorithms for these water types. In any case, for Nikon D90 and Canon 500D, the reproduction angular errors of the BSM based algorithm averaged for all water types, are less than ones of the CW based algorithm by 29% and 24%, respectively (see row “All”).

CONCLUSION

It is shown that in modern works when simulating the direct light component of underwater images based on terrestrial images, the question of correctness and accuracy of the channel-wise approximation of model (4) is omitted. This paper shows that using spectral models, it is possible to solve these problems. In particular, us-



(a) Nikon D90 BSM parameters: $\lambda_l = 513$, $\lambda_r = 583$.

(b) Canon 500D BSM parameters: $\lambda_l = 503$, $\lambda_r = 579$.

Figure 2: Sensitivities and calculated banded spectral model parameters for Canon 500D and Nikon D90 camera.

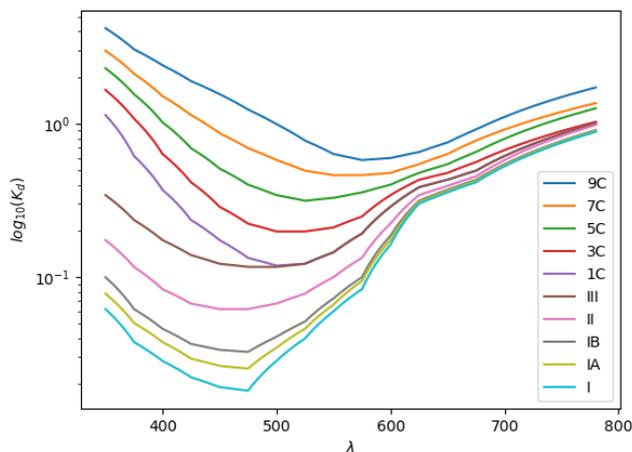


Figure 3: Downwelling attenuation coefficients $K_d(\lambda)$ of 10 Jerlov Water Types (Jerlov 1968), used as attenuation coefficients in the numerical experiments.

ing the example of the proposed simulation algorithm based on banded spectral model (15), we were able to show that the average accuracy of the color reproduction of the proposed method is higher than that of the algorithm based on channel-wise model (6).

In the future, we plan to develop algorithms for simulating underwater images based on terrestrial images using spectral models and study properties of these algorithms using numerical experiments on a larger dataset.

REFERENCES

Akkaynak, D. and T. Treibitz. 2018. “A revised underwater image formation model.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6723–6732.

Anwar, S.; C. Li; and F. Porikli. 2018. “Deep underwater image enhancement.” *arXiv preprint arXiv:1807.03528*.

Aranha, M. 2005. “Realistic underwater visualisation.” *Computer Graphics Group, University of Bristol*.

Arlazarov, V.V.; K.B. Bulatov; T.S. Chernov; and V.L. Arlazarov. 2019. “Midv-500: A dataset for identity document analysis and recognition on mobile devices in video stream.” *Computer Optics*, 43(5):818–824. doi: 10.18287/2412-6179-2019-43-5-818-824. Cited By 0.

Babelcolor. 2005. “SG color chart reflectances.” http://www.babelcolor.com/index_htm_files/Digital%20ColorChecker%20SG.txt.

Boffety, M.; F. Galland; and A.-G. Allais. 2012. “Color image simulation for underwater optics.” *Applied Optics*, 51(23):5633–5642.

Chang, H.; C. Cheng; and C. Sung. 2019. “Single underwater image restoration based on depth estimation and transmission compensation.” *IEEE Journal of Oceanic Engineering*, 44(4):1130–1149. ISSN 2373-7786. doi: 10.1109/JOE.2018.2865045.

Chong, H.Y.; S.J. Gortler; and T. Zickler. 2007. “The von kries hypothesis and a basis for color constancy.” In *2007 IEEE 11th International Conference on Computer Vision*, 1–8.

Ding, X.; Y. Wang; Y. Yan; Z. Liang; Z. Mi; and X. Fu. 2019. “Jointly adversarial network to wavelength compensation and dehazing of underwater images.”

Duarte, A.; F. Codevilla; J.O. Gaya; and S.S.C. Botelho. 2016. “A dataset to evaluate underwater image restoration methods.” In *OCEANS 2016-Shanghai*, 1–6.

Fabbri, C.; M.J. Islam; and J. Sattar. 2018. “Enhancing underwater imagery using generative adversarial networks.” In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 7159–7165.

Finlayson, G.D.; M.S. Drew; and B.V. Funt. 1993. “Color constancy: enhancing von kries adaption via sensor transformations.” In *Human Vision, Visual Processing, and Digital Display IV*, volume 1913, 473–484.

Finlayson, G.D.; R. Zakizadeh; and A. Gijsenij. 2017. “The reproduction angular error for evaluating the performance of illuminant estimation algorithms.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1482–1488. ISSN 1939-3539. doi: 10.1109/TPAMI.2016.2582171.

Foresti, G.L. 2001. “Visual inspection of sea bottom structures by an autonomous underwater vehicle.” *IEEE*

- Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(5):691–705.
- Gracheva, M.A.; V.P. Bozhkova; A.A. Kazakova; I.P. Nikolaev; and G.I. Rozhkova. 2020. “Subjective assessment of the quality of static and video images from mobile phones.” In Jianhong Zhou Wolfgang Osten, Dmitry Nikolaev, editor, *ICMV 2019*, volume 11433 (SPIE, 2020). DOI: 10.1117/12.2559367.
- Gusamutdinova, N.; E. Ershov; S. Gladilin; and D. Nikolaev. 2017. “Verification of applicability two multiplicative closed spectral models for multiple reflection effect description.” In *2016 International Conference on Robotics and Machine Vision*, volume 10253, page 1025305.
- Hu, Y.; K. Wang; X. Zhao; H. Wang; and Y. Li. 2018. “Underwater image restoration based on convolutional neural network.” In *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, 296–311.
- Jaffe, J.S. 1990. “Computer modeling and the design of optimal underwater imaging systems.” *IEEE Journal of Oceanic Engineering*, 15(2):101–111.
- Jerlov, N. 1968. “Irradiance optical classification.” *Optical Oceanography*, 118–120.
- Jiang, J.; D. Liu; J. Gu; and S. Süssstrunk. 2013. “What is the space of spectral sensitivity functions for digital color cameras?” In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 168–179.
- Kahanov, Y. and J.G. Royal. 2001. “Analysis of hull remains of the dor d vessel, tantura lagoon, israel.” *The International journal of nautical archaeology*, 30(2):257–265.
- Lavest, J.-M.; F. Guichard; and C. Rousseau. 2002. “Multi-view reconstruction combining underwater and air sensors.” In *Proceedings. International Conference on Image Processing*, volume 3, 813–816.
- Li, C.; J. Guo; and C. Guo. 2018. “Emerging from water: Underwater image color correction based on weakly supervised color transfer.” *IEEE Signal Processing Letters*, 25(3):323–327.
- Li, C.-Y.; J.-C. Guo; R.-M. Cong; Y.-W. Pang; and B. Wang. 2016. “Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior.” *IEEE Transactions on Image Processing*, 25(12):5664–5677.
- Li, J.; K.A. Skinner; R.M. Eustice; and M. Johnson-Roberson. 2017. “Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images.” *IEEE Robotics and Automation letters*, 3(1):387–394.
- Maloney, L.T. 1986. “Evaluation of linear models of surface spectral reflectance with small numbers of parameters.” *Journal of the Optical Society of America A*, 3(10):1673–1683.
- Mangeruga, M.; M. Cozza; and F. Bruno. 2018. “Evaluation of underwater image enhancement algorithms under different environmental conditions.” *Journal of Marine Science and Engineering*, 6(1):10.
- McGlamery, B.L. 1980. “A computer model for underwater camera systems.” In *Ocean Optics VI*, volume 208, 221–232.
- National Electrical Manufacturers Association. 2004. “Ansi/iec 60529–2004, degrees of protection provided by enclosures (ip code).” Technical report.
- Nikolaev, D.P. and P.P. Nikolayev. 2005. “Comparative analysis of gaussian and linear spectral models for colour constancy.” In *Proceedings of 19th European Conference on Modelling and Simulation*, 300–305.
- Nikolaev, D.P. and P.P. Nikolayev. 2007. “On spectral models and colour constancy clues.” In *Proceedings of 21st European Conference on Modelling and Simulation*, 318–323.
- Nikolaev, D.P.; P.P. Nikolayev; and V.P. Bozhkova. 2006. “Efficiency comparison of analytical gaussian and linear spectral models in the same colour constancy framework.” *International Journal of Simulation–Systems,(IJSSST, Special Issue on: Vision and Visualization)*, 21–36.
- Peng, Y.-T. and P.C. Cosman. 2017. “Underwater image restoration based on image blurriness and light absorption.” *IEEE transactions on image processing*, 26(4):1579–1594.
- Schechner, Y.Y. and N. Karpel. 2004. “Clear underwater vision.” In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, 536–543.
- Schechner, Y.Y. and N. Karpel. 2005. “Recovery of underwater visibility and structure by polarization analysis.” *IEEE Journal of oceanic engineering*, 30(3):570–587.
- Skarlatos, D.; P. Agraftotis; T. Balogh; F. Bruno; F. Castro; B. D. Petriaggi; S. Demesticha; A. Doulamis; P. Drap; and A. Georgopoulos. 2016. “Project imareculture: advanced vr, immersive serious games and augmented reality as tools to raise awareness and access to european underwater cultural heritage.” In *Euro-Mediterranean Conference*, 805–813.
- Smagina, A.; E. Ershov; and A. Grigoryev. 2020. “Multiple light source dataset for colour research.” In Jianhong Zhou Wolfgang Osten, Dmitry Nikolaev, editor, *ICMV 2019*, volume 11433 (SPIE, 2020). DOI: 10.1117/12.2559491.
- Uplavikar, P.; Z. Wu; and Z. Wang. 2019. “All-in-one underwater image enhancement using domain-adversarial learning.”
- Yu, X.; Y. Qu; and M. Hong. 2018. “Underwater-gan: Underwater image restoration via conditional generative adversarial network.” In *International Conference on Pattern Recognition*, 66–75.
- Zhao, X.; T. Jin; and S. Qu. 2015. “Deriving inherent optical properties from background color and underwater image enhancement.” *Ocean Engineering*, 94:163–172.

AUTHOR BIOGRAPHIES



DENIS A. SHEPELEV was born in Tambov, Russia. He studied applied physics and mathematics, and obtained his Master degree in 2016 from Moscow Institute of Physics and Technology. Since 2015 he works as a researcher at RAS Institute for Information Transmission Problems. In 2016 he started his Ph.D. in Moscow Institute of Physics and Technology.

His research activities focus on the areas of computer vision and robotics. His e-mail address is shepelev@iitp.ru.



VALENTINA P. BOZHKOVA obtained her master's degree in physics in 1966 in Moscow State University, the Ph.D. degree in biology in 1972, and the degree of Doctor of Sciences (in biology) in 1987. She works at the Institute for Information Transmission Problems, Russian Academy of Sciences since 1975.

Her primary research subjects are biological processes modelling and colour vision. Her e-mail address is bgk@iitp.ru.



EGOR I. ERSHOV was born in Moscow, USSR. He studied engineering science and mathematics, obtained his master's degree in 2014 at Moscow Institute of Physics and Technology and Ph.D. degree in 2019 at Institute for Information Transmission Problems. He is working in Vision Systems Lab at the Institute for Information Transmission Problems RAS since 2014.

His research activities focus on the areas of color computer vision, fast Hough and Radon transforms, visual odometry. His e-mail address is ershov@iitp.ru.



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 dimonstr@iitp.ru.

A CONCEPTUAL MODEL OF AN IOT-BASED SMART AND SUSTAINABLE SOLID WASTE MANAGEMENT SYSTEM: A CASE STUDY OF A NORWEGIAN MUNICIPALITY

Wajeeha Nasar¹, Anniken Th. Karlsen² and Ibrahim A. Hameed³

Department of ICT and Natural Sciences

Faculty of Information Technology and Electrical Engineering

Norwegian University of Science and Technology

Email: wajeehan@stud.ntnu.no¹, anniken.t.karlsen@ntnu.no², ibib@ntnu.no³

KEYWORDS

Conceptual Model, Internet Of Things, Smart And Sustainable, Infrared Sensors, Ultrasonic Sensors, RFID Sensors, Iots Based Smart Bins, Smart Bin App, Data Transfer Technologies, LoRaWAN

ABSTRACT

The core processes of waste management have been changed during the last few decades. Through advanced technologies, sensors, cameras, actuators, IoT controls, data driven and data transfer technologies, the old and insufficient processes for waste management can be replaced. In this paper, we propose a conceptual model for an IoT-based smart and sustainable waste management system for a Norwegian municipality. The model illustrates all the aspects needed to develop a smart IoT-based waste management system. A Norwegian municipality constituted our case study. Our conceptual model proposed here, provides a design solution with data analysis in such a way that it can easily be adopted by the current infrastructure and practices of the municipality. Finally, features of a prototype system are suggested.

I. INTRODUCTION

With the advent of recent advancements of smart devices, the abstraction of connecting everyday objects via the existing networks has become highly favorable. The Internet of things (IoTs) is an arrangement of web related items that can accumulate and exchange data. A result of the evolution of conventional networks that link billions of connected devices together defines a world where almost anything can connect and interact in a smarter fashion than before (Silva, et al. 2018). Technological advancements in ubiquitous computing (UC), wireless sensor networks (WSN), and machine-to-machine (M2M) communication have further strengthened the notion of IoT (Vaisali, et al. 2017). Moreover, linked devices share their information and access authorized information of other devices to support contextual decision making. As a result of these developments, new business areas and opportunities have originated, summarized into various smart city and smart factory concepts. Due to the dramatic urbanization all over the world, the continuous developments into smart cities

have become the main concern in the past few decades. Information and communication technology (ICT) have made cities efficient in several aspects. However, incorporating only ICT does not fully interpret the smart city concept. In general terms, a smart city is an urban environment that utilizes ICT and other related technologies to improve performance efficiency of regular city operations and quality of service provided to urban citizens (Kamm, et al., 2020). IoTs link various areas/operations of a smart city into holistic entities. In Figure 1, the concept of an IoT enabled smart and sustainable city is illustrated. In a smart and sustainable city, all the aspects of a society are connected through shared IoT clouds. This enables the use of the new opportunities offered by IoT platform, thereby empowering us to set a sustainable footprint to the world. Smart waste management is one fundamental concern in smart and sustainable city development (ITU 2019). According to Periathamby et al. (2014) the global population will increase into 9 billion in 2050. In addition to that, the increased level of urbanization will lead to a massive pressure on the current infrastructures and practices of municipalities. This led us towards investigating good practice of solid waste handling. Current practices for a waste management system includes waste collection, waste sorting, waste recycling and its transportation as in Figure 4. They can often be improved by reengineering. A vital concept in this circumstance is Key performance indicators (KPIs) for solid waste, categorized by the EU report agenda (ITU 2019). This is a conceptual framework that achieved the KPIs described in Figure 2.

Our paper is structured as follows: The related literature for smart and sustainable solid waste management systems based on IoTs technologies are discussed in section II. The conceptual models are described in detail in section III. Current practices and relative techniques to develop a smart and sustainable solid waste management system are proposed. In the discussion section, we also provide future recommendations.

II. LITERATURE REVIEW

A solid waste material hierarchy, Figure 3, can be described as follows: The material should be prevented as much as possible and if it can't then it goes for reuse.

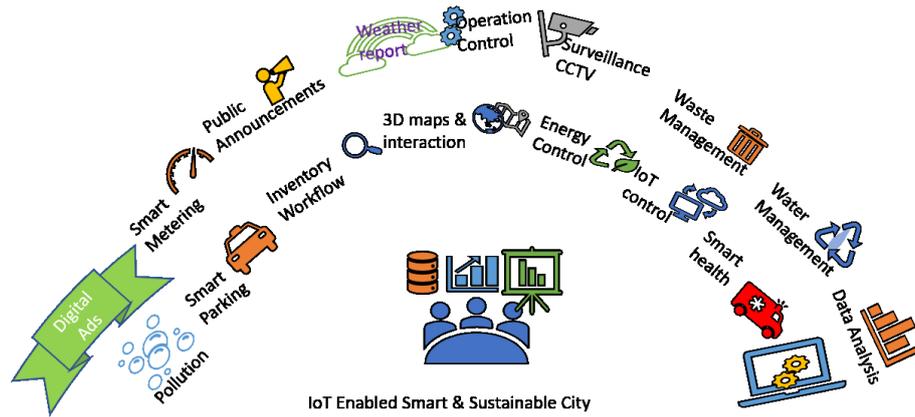


Figure 1: A Conceptual Model of an IoT Enabled Smart And Sustainable City

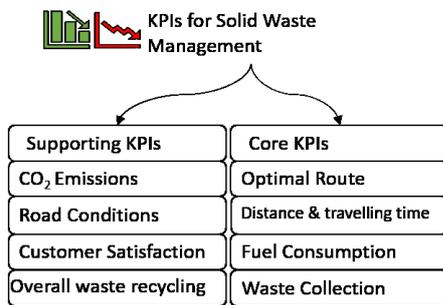


Figure 2: Smart And Sustainable Solid Waste Management KPIs

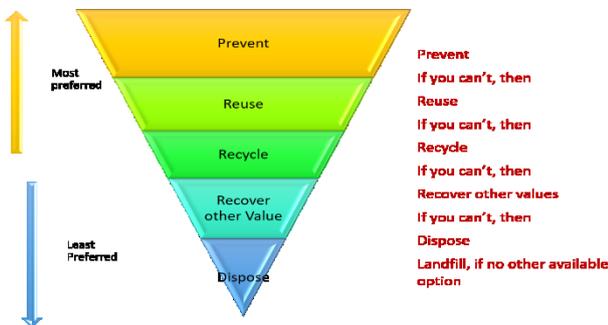


Figure 3: Solid Waste Material Hierarchy

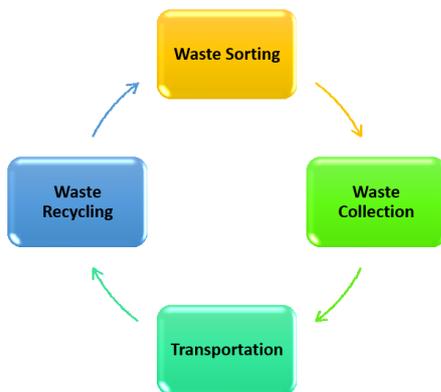


Figure 4: Smart And Sustainable Solid Waste Management System

If reuse isn't possible, then it goes for recycling. When recycling is not an option, a possibility might be to recover other values. If that isn't possible disposing is the last option.

The disposal and reuse procedure associated with each waste material is different. Additionally, there are several solid waste management projects executed worldwide in rural and urban areas. The solutions proposed in these projects are based on different techniques and data transfer technologies (Kamm, et al. 2020).

One important issue in waste management is transportation. The transportation of waste includes collection from waste bins and the transportation to various disposal sites as illustrated in Figure 4. For waste collection and transportation several methods have been proposed in literature. For example, in an article by Mingaleva et al. (2019) waste management in green and smart cities are discussed. Further on, current practices and further actions towards sustainable cities are described. In Patel et al. (2019) dry and wet dustbins are segregated, and different sensors and Wifi module for waste collection are used for their proposed model. Dugdhe et al. (2016) propose a method for waste collection scheduling for truck drivers, using mathematics to calculate the shortest route between filled-up bins and bins producing harmful gases.

With all the proposed IoT techniques described in literature, we are now able to solve many obstacles associated with waste management systems. Still there are many issues that need to be solved pertaining to reliability, scaling, bandwidth, security and power consumption.

III. CONCEPTUAL MODEL

A conceptual model can be defined as a simplified representation of a system used to describe its main physical features and principal processes (Helmig 1997; Tatomir et al. 2018). In what follows, we propose a conceptual model of an IoT-based smart and sustainable solid waste management system. The model provides a

Norwegian municipality with a blueprint of a potential system design that can easily be adopted into the municipality's current infrastructure and practices. We assume that the model might also be relevant for others as an initial template for building smarter waste management systems.

Current practices

The current practices and infrastructure of the Norwegian municipality's solid waste management system relates to different types of bins (for example, standard volume waste bins, underground waste bins and sensor based underground waste bins) for all types of waste (household waste, paper, plastic, glass, metal and food waste), mounted and scattered around the municipality. In Nasar et al. (2020) current practices for the Norwegian municipality under study are described.

Both 2G and 3G communication techniques are used for data transfer and data analysis (Nasar et al. 2020). Experience indicates that battery life of sensors devices is a major problem that should be resolved. Additionally, the product portfolio of service providers in the municipality differ from each other. Pertaining to the ultrasonic sensors used to sense the fill-up volume of waste in the bins, most of them communicate through GSM technology.

As regards data transmission via GSM, this also faces some challenges in relation to high power consumption and dependency of the network provider. In our proposed conceptual model, all these facts such as data transfer technologies, sensors, planning to future development etc are considered to enable the construction of a smart and sustainable solid waste management system based on IoT technologies.

Towards an IoT-based smart and sustainable solid waste management system

Our conceptual model is presented in Figure 5. Several stakeholders or actors, for example management companies, truck drivers and citizens, are in this improved system, connected to a database so that they can make choices and corresponding actions as regards the functioning of the waste management system. Additionally, truck drivers are equipped with a display screen with GPS and GIS information for waste collection from the bins scattered in the city.

As regards the citizens, they are connected to and can interact with the smart waste bin application via their mobile phones. The waste management companies get

direct access to the database in the same manner. That the stakeholders communicate through the same platform makes communication smarter by being easier and more efficient.

A challenge in current practice is that the waste truck drivers follow a traditional way for waste collection whereby optimization has not been an issue. After deploying the proposed conceptual model focusing on IoT use, this problem can be solved by making active use of the possibilities provided by sensors combined with the use of KPIs and optimization algorithms to achieve reduced cost, distance and time. An evident outcome of an implementation of the conceptual model will be reduced CO₂ emission in the atmosphere as shown in Figure 2.

Sensor-based bins

In Figure 6, our proposal for how an IoT-based smart waste bin system can be designed, is illustrated.

In current practices, sensors-based bins are subject to many problems, due to different service providers using different platforms for data transfer and data handling etc. Another experienced problem is low battery life. The proposed IoT-based smart waste bin solution attach ultrasonic sensors, RFID sensors, Infrared sensors, and solar batteries with the cloud for data capturing and data transmission process. In the IoT-cloud, machine learning techniques are used to predict the fill-up volume of the waste bin. Optimization algorithms are used to find the truck drivers' optimal routes by taking total waste cost management into consideration.

The waste management cost for N number of bins can be described as:

$$W_{total} = \sum_{i=1}^N W_i^c + \sum_{i=1}^N W_i^t + \sum_{i=1}^N W_i^p + \sum_{i=1}^N W_i^d \quad (1)$$

where W_i^c is the collection cost, W_i^t is the transportation cost, W_i^p is the processing cost, W_i^d is the disposal cost for k number of unused or produced waste material after processing and W_i is the constant cost that depends on other parameters such as accident, maintenance of collection center, transfer station and trucks.

The profit gain by the N number of sources is:

$$W_{management_cost}(profit) = \sum_{i=1}^p R_i - W_i \quad (2)$$

Where R_i is achieved by the recyclable materials, sales of compost products and electricity sales.

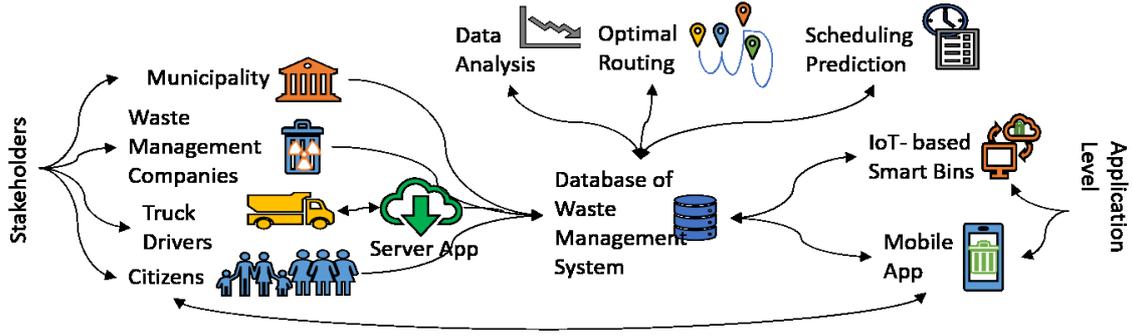


Figure 5: A Conceptual Model of a Smart and Sustainable Solid Waste Management System

In general, the objective of a waste management system is to develop mechanisms that will increase the overall profit $W_{management_cost}(profit)$ associated with the system by reducing the $\sum_{i=1}^N W_i^c$, $\sum_{i=1}^N W_i^t$ and increase $W_{management_cost}(profit)$.

A transportation problem for a waste management system in such cases is referred as a combinatorial optimization problem. The main objective for such problems is to reduce cost, distance, travelling time and fuel consumption. Besides that, to build a prediction model for smart waste collection, we must consider both road and weather conditions as these will impact on fuel consumption.

In our proposed conceptual model, ultrasonic sensors are embedded to sense the present waste volume in the waste bins. Infrared sensors are implemented to sense the type of waste material. These will be helpful to achieve correct waste sorting and correspondingly to increase the waste management company's revenue. Implementation of RFID sensors will be used to identify where bins are placed. As the performance and lifetime of sensors depends on batteries, we suggest the use of solar batteries. Whilst using solar batteries have the potential of solving the experienced problems associated with present battery life. Developments within these types of batteries provide enough energy for all system actions in a very sustainable manner. Further on, it is very important to produce low power consuming hardware and to schedule the sleep mode for the proposed waste management system. The sleep mode, as in several electronics' devices, are embedded to enable energy savings when a sensor-based bin is not measuring, processing or sending data.

In order to provide independent telecommunication for enabling smart and sustainable city initiatives, an IoT network will be installed. The term IoT-WAN (also known as Low Power Wide Area Network-LPWAN) contains a variety of technologies such as LoRa, Sigfox or NB-IoT (Kamm, et al. 2020).

Features of a Smart Waste Bin Application

User End

In the proposed conceptual model, the application level is divided into two parts. One part constitutes IoT-based

smart waste bins, and the other one contains a smart waste bin mobile application. In current practice, the waste management company offers a mobile app for customers with limited possibilities. Features of an improved application, Figure 8, can be:

- *Waste volume status* – The app will notify about the volume of waste in the customer's bin which will help the customer to decide whether to go out for throwing the trash or not.
- *Scheduled route* – The app will notify about the scheduled trips of waste trucks, so the citizens can put their waste bins outside.
- *Parking area status* – In current infrastructure, a vast area is dedicated specifically for waste trucks to collect waste. Instead of this practice, in the prototype app customers are notified about the waste trucks' schedule, thereby orienting the customer of when certain areas can be used for regular parking.
- *Customer credit* – In the present infrastructure there are different waste bins provided to the customers for different types of waste. These different types of waste are to be treated differently. For example, food waste needs to be emptied more frequently than other waste types. Similarly, the household waste is not considered as being recycling material and therefore are usually to be sent to disposal sites. Paper, plastic and other waste types are commonly sent for recycling in various recycling plants. To motivate customers to act properly pertaining to sorting waste, our prototype embeds the idea of giving away customer credits in the form of bonus points or some appreciating messages.

Smart App Server

Based on the data transfer technologies, a smart waste management application can be built (Kamm, et al. 2020). Various Python libraries can be used to build the platform shown in Figure 9. The implementation can be divided into three parts:

1. A smart bin App based on sensors which will detect waste material and sense the fill up volume of bins. Through the obtained data, the data analysis can be done as shown in Figure 7.

In this Figure, the fill-up volume before and after emptying the waste bins is shown. Sensor data from two types of waste, i.e., paper and household waste, are subject for data analysis.

- The second part is data connectivity between the sensor based smart app, a decoder and an application manager handler through an IoT-WAN infrastructure which is suggested to build the app.
- The last part is a smart waste application server.

IV. DISCUSSIONS & FUTURE RECOMMENDATIONS

In this paper a conceptual model is proposed as a blueprint for a smart and sustainable waste management system based on IoT technologies. We investigated a Norwegian municipality as a case study and based on our knowledge of current practice and the needs and wishes of the municipality we propose a design solution.

In the future it is our hope that this conceptual model can be implemented by the municipality using IoT technologies, sensors, cameras, actuators and Python for software development.

By implementing some of our suggestions we hope the municipality will end up with an IoT-based waste management system providing optimal routes and schedules for truck drivers via prediction models. For this different machine learning and artificial intelligence techniques can be used. To solve optimization problems, we can use techniques such as multiple travelling salesman problem (MTSP), constraint vehicle routing problem (CVRP) and multi objective optimization (MOO). These are all issues and possibilities to be investigated in future work.

In Nasar et al. (2020), the waste collection and transportation problem is addressed. The multi-objective traveling salesman problem (MOP-TSP) is used to find the optimal shortest possible route with multiple constraints such as minimum traveling time and distance as shown in Figure 10. The obtained results are compared with current practices and it is concluded that the proposed method is 34% cost and time saving. The solid waste KPIs i.e. stated in Figure 2 have been achieved with this proposed solution.

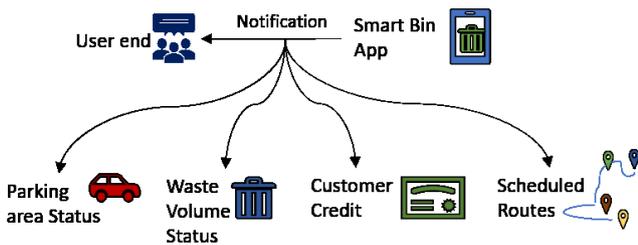


Figure 6: Smart Waste Bin Application at User end

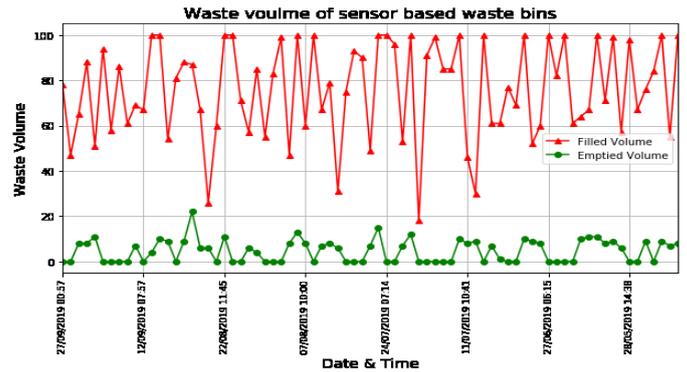


Figure 7: Sensor Measurement Of Waste Volume In Waste Bins

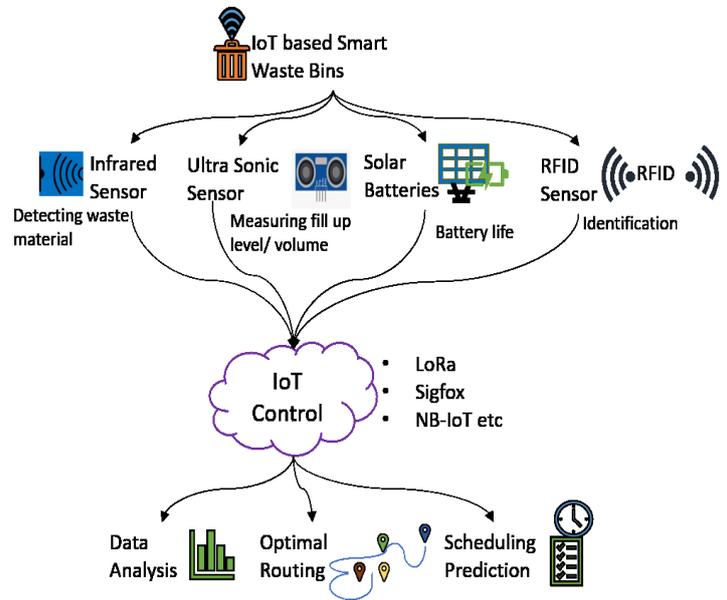


Figure 8: IoT-Based Smart Waste Bins

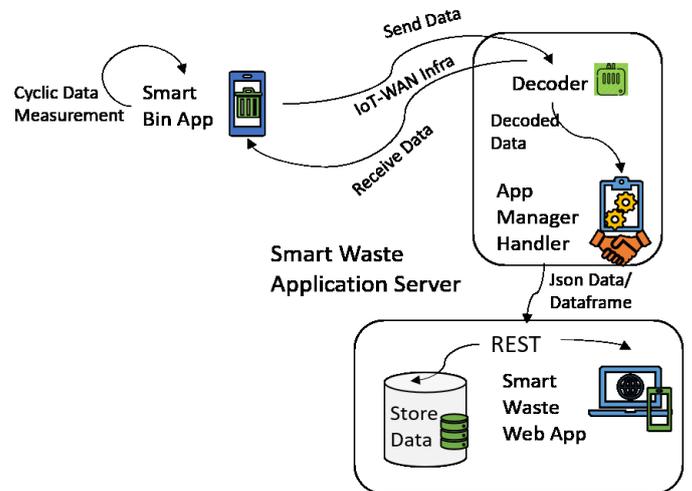


Figure 9: Smart Waste Bin Application Server

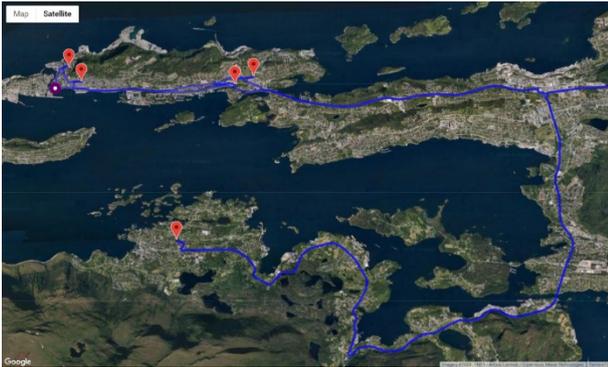


Figure 10: MOP-TSP optimal route finding with minimum distance and minimum time (Nasar et al. 2020)

REFERENCES

- Dugdhe, S., Shelar, P., Jire, S. and Apte, A. (2016): "Efficient waste collection system," *International Conference on Internet of Things and Applications (IOTA)*, pp. 143-147.
- Helmig, R. (1997): "Multiphase Flow and Transport Processes in the Subsurface: A Contribution to the Modeling of Hydrosystems, 1st edn.", *Springer, Berlin Heidelberg New York*.
- ITU (2019): [url: http://handle.itu.int/11.1002/1000/12627](http://handle.itu.int/11.1002/1000/12627)
- Kamm, M., Gau, M., Schneider, J., & Vom, B. J. (2020): "Smart Waste Collection Processes-A Case Study about Smart Device Implementation". *Paper presented at the Hawaii International Conference on System Sciences, Hawaii. (VHB 3: C)*
- Mingaleva, Z., Vukovic, N., Volkova, I. and Salimova, T. (2019): "Waste Management in Green and Smart Cities: A Case Study of Russia." *Sustainability*. 12. 10.3390/su12010094.
- Nasar, W., Karlsen, A. Th., Hameed, I. A. and Dwivedi, S. (2020) "An Optimized IoT-based Waste Collection and Transportation Solution: A Case Study of a Norwegian Municipality", *Submitted in 3rd International Conference on Intelligent Technologies and Applications, INTAP 2020*.
- Nirde, K, Mulay, P. S. and Chaskar, U. M. (2017): "IoT based solid waste management system for smart city," *International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai*, pp. 666-669.
- Patel, D., Kulkarni, A. and Sharma, H. (2019): "Smart Dustbins for Smart Cities". *International Journal of Trend in Scientific Research and Development*. Volume-3. 1828-1831. 10.31142/ijtsrd22993.
- Periathamby, A. and Tanaka M. (2014): "Municipal Solid Waste Management in Asia and the Pacific Islands: Challenges and Strategic Solutions". *Springer Singapore*. 10.1007/978-981-4451-73-4.
- Silva B.N., Khan M. and Han K. (2018): "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges" *in smart cities, Sustainable Cities and Society*, Volume 38, Pages 697-713, ISSN 2210-6707.
- Tatomir, et al. (2018): "Conceptual model development using a generic Features, Events, and Processes (FEP) database for assessing the potential impact of hydraulic fracturing on groundwater aquifers". *Advances in Geosciences*. 45: 185–192. doi:10.5194/adgeo-45-185-2018.
- Vaisali, Bhargavi G. and Kumar K.S., (2017): "Smart solid waste management system by IOT". *International Journal of Mechanical Engineering and Technology*. 8. 841-846.

AUTHOR BIOGRAPHIES



Wajeeda Nasar is lecturer at the Department of ICT and Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway. She is master's student of simulation and Visualization from The Norwegian University of Science and Technology (NTNU), Norway. She has a MSc degree in Electrical Engineering from Institute of space technology, Pakistan. Her current research interest includes Optimization, Routing problem, and Artificial Intelligence.



Anniken T. Karlsen is Head of Department at the Department of ICT and Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway. 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 in several sectors, including maritime, marine, offshore, food, consultant, health and banking. Her main interests are digital transformation from a holistic business perspective.



Ibrahim A. Hameed is 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. He is also program coordinator of the department's international master program in simulation and visualization. Among others, 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. He is author of more than 120 journal and conference articles. His current research interest includes Artificial Intelligence, Machine Learning, Optimization, and Robotics.

OPTIMAL RECEIVER CONFIGURATION OF SHORT-BASELINE LOCALISATION SYSTEMS USING PARTICLE SWARM OPTIMISATION

Christoph Tholen, Tarek El-Mihoub,
Lars Nolle, Oliver Ralle
Autonomous Systems Research Group
Department of Engineering Science
Jade University of Applied Sciences
Email: {christoph.tholen | tarek.el-mihoub |
lars.nolle }@jade-hs.de
oliver.ralle@student.jade-hs.de

Robin Rofallski
Institute for Applied Photogrammetry and
Geoinformatics
Jade University
of Applied Sciences
Oldenburg, Germany
robin.rofallski@jade-hs.de

KEYWORDS

SBL, AUV, Optimisation, Short-Baseline-Localisation, Underwater Acoustic Localisation, Autonomous-Underwater-Vehicles

ABSTRACT

This work investigates the localisation error of a short-baseline system used for the localisation of submerged underwater vehicles. In a first step, different possible error influences are identified and their numerical simulation are described. In a second step, this simulation is used to determine the optimal position of the acoustic receivers of the system using Particle Swarm Optimisation and Monte-Carlo simulations. The positioning error of the optimised receiver arrangement is 6.64 % smaller than the error of a standard arrangement.

INTRODUCTION

The use of autonomous underwater vehicles (AUVs) requires a method for robust, reliable and accurate determination of an AUV's position. Autonomous vehicle usually rely on the availability of global navigation satellite systems (GNSS) in order to estimate their global position. However, the electromagnetic signals of such satellite systems cannot be received by submerged AUVs (Wu, et al., 2019). Different possible methods for localisation of submerged AUVs were proposed in the past. These methods can be classified into inertial or dead reckoning, acoustic methods and geophysical methods (Paull, et al., 2014). This paper focusses on the accuracy of an acoustic based localisation method called short-baseline localisation (SBL) (Paull, et al., 2014). The system under investigation is an off-the-shelf low-cost positioning system (Water Linked AS, 2020). In this work, different error sources, decreasing the accuracy of an acoustic based localisation system, are investigated. In order to increase the localisation accuracy of the system, the optimal receiver configuration is determined using Monte Carlo simulations and Particle Swarm Optimisation.

SHORT-BASELINE-LOCALISATION

Acoustic based localisation methods, like SBL, use the time of flight principle of acoustic waves (Paull, et al., 2014). The system usually consists of a couple of transducers, mounted on a mothership or a jetty and a single receiver mounted on an AUV (Figure 1). The transducers are emitting acoustic signals. The time of flight (TOF), i.e. the time between the emission and the reception of the signals, is measured by the receiver mounted on the AUV. Subsequently, the TOF can be used to calculate the position of the AUV.

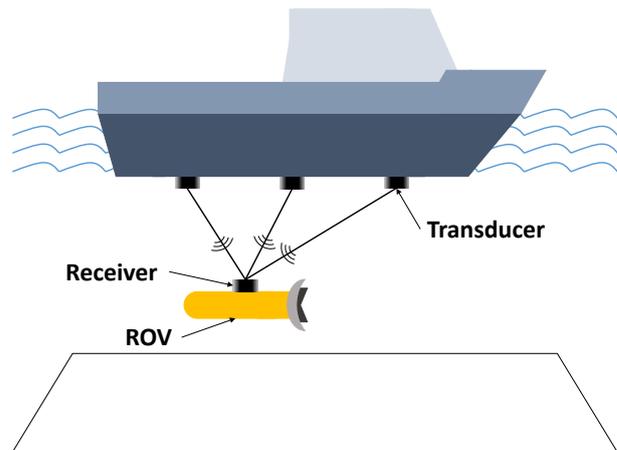


Figure 1: Principle of Short-Baseline localisation methods modified after (Wu, et al., 2019)

However, the Water Linked system used in this research uses a single transducer, or locator, mounted on the AUV and four receivers mounted on a mothership or a jetty (Water Linked AS, 2020). The locator emits an acoustic signal, which is received by the four receivers. The time between the emission and the reception of the signal is measured individually by the four receivers. The principle of the Water Linked SBL system is shown in Figure 2.

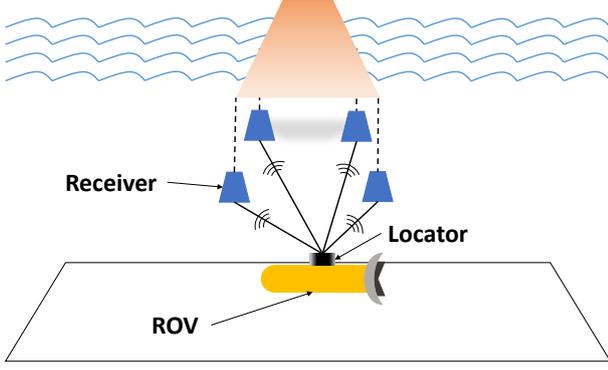


Figure 2: Water Linked SBL System

To localise an AUV using an SBL system usually a spherical-based algorithm is used (Turetta, et al., 2014). It uses the distances between the AUV and the SBL receivers to estimate the AUV's position. The distances d_i are determined by measuring the time of flight t_i for each receiver. Given the speed of sound v , the distances d_i for n receivers are calculated as follows:

$$d_i = t_i \cdot v \quad (1)$$

Where:

- d_i : Distance between receiver i and AUV,
- t_i : Time of flight measured by receiver i ,
- v : Speed of sound.

In a Cartesian coordinate system, the Euclidean distances d_i can be decomposed into the x , y , and z components as follows:

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = d_i^2 \quad (2)$$

Where:

- $[x, y, z]$: Coordinates of the AUV,
- $[x_i, y_i, z_i]$: Coordinates of the receiver i .

The z -coordinate of the AUV usually is determined using an on-board pressure sensor (Turetta, et al., 2014). Thus, the 3D problem can be reduced to a 2D problem in the x - y -plane with the planar distances r_i between the AUV and the receiver i :

$$r_i^2 = d_i^2 - (z - z_i)^2 \quad (3)$$

From equations (2) and (3) follows:

$$(x - x_i)^2 + (y - y_i)^2 = r_i^2 \quad (4)$$

Equation (4) leads to the linear relationship with four receivers and the position vector $X = [x, y]^T$:

$$A \cdot X = R - D \quad (5)$$

With:

$$A = \begin{bmatrix} x_1 - x_2 & y_1 - y_2 \\ x_2 - x_3 & y_2 - y_3 \\ x_3 - x_4 & y_3 - y_4 \\ x_4 - x_1 & y_4 - y_1 \end{bmatrix},$$

$$R = \frac{1}{2} \begin{bmatrix} r_2^2 - r_1^2 \\ r_3^2 - r_2^2 \\ r_4^2 - r_3^2 \\ r_1^2 - r_4^2 \end{bmatrix},$$

$$D = \frac{1}{2} \begin{bmatrix} (x_2^2 + y_2^2) - (x_1^2 + y_1^2) \\ (x_3^2 + y_3^2) - (x_2^2 + y_2^2) \\ (x_4^2 + y_4^2) - (x_3^2 + y_3^2) \\ (x_1^2 + y_1^2) - (x_4^2 + y_4^2) \end{bmatrix}.$$

By using the least squares method, the position of the AUV can be estimated as follows (Turetta, et al., 2014):

$$\hat{X} = (A^T \cdot A)^{-1} \cdot A^T \cdot (R - D) \quad (6)$$

With:

$$\hat{X} = [\hat{x}, \hat{y}]^T.$$

In the absence of measurement errors, the described spherical-based algorithm guarantees an accurate estimation of the actual AUV position. However, in a real world application, measurement errors affect the accuracy of the SBL system. The accuracy of an acoustic positioning system can be determined using experiments (Almeida, et al., 2016) or numerical simulations (Turetta, et al., 2014). In the next chapter, different error sources and their impact on the localisation process are described.

ERROR FORMULATION

From equations (1) to (3), the following potential sources of error are identified: determination of the receivers' positions, accuracy of the measurements of TOF, determination of the speed of sound and accurate calculation of the actual depth of the AUV. These potential sources of error are discussed in more detail below.

Receiver Position

Usually, the position of the receivers is obtained using different measurement methods, like GPS (Almeida, et al., 2016), tape measurements or using a total station. The accuracy of the measurements depend on the chosen method. It ranges from millimetre to decimetre accuracy. During operation, the receivers might be affected by currents and waves, resulting in periodical drift of the receivers.

In this work tape measuring is simulated. Therefore, both effects are modelled using a truncated Gaussian distribution with the maximum value of $x_{max} = 0.02$ m, the true locator position as mean value and a standard deviation of $\sigma_{xy} = 0.05$ m for the x and y position of the receivers. The error of the z position is modelled using a Gaussian distribution with the true position as mean value and a standard deviation of $\sigma_z = 0.05$ m.

Time of Flight

The TOF measurements are affected if the clocks used in the locator and in the receiver are not synchronised

(Paull, et al., 2014). Also, a quantisation error is introduced by the signal-processing unit (Turetta, et al., 2014).

The Water Linked system, used in this work, utilises GPS time to avoid these synchronisation issues (Water Linked AS, 2020). Based on the specifications of the system used, the TOF error is simulated using a Gaussian distribution with the true time of flight as mean value and a standard deviation of $\sigma_{TOF} = 10^{-6}$ s.

Speed of Sound

The speed of sound in seawater depends on temperature, salinity and pressure of the seawater (UNESCO, 1983). In the water column, temperature and salinity vary over time (Tholen, et al., 2020). Hence, the speed of sound also varies. For the accurate determination of the speed of sound, the path d_i of the sound wave from the locator to receiver i has to be taken into account:

$$v_i = \frac{1}{d_i} \int_0^{d_i} v(S(d), T(d), P(d)) dd \quad (7)$$

Where:

- d_i : Path from transducer to the receiver i ,
- $S(d)$: Salinity as function of the path,
- $T(d)$: Temperature as function of the path,
- $P(d)$: Pressure as function of the path.

The calculation of the speed of sound using (7) requires detailed knowledge about the temperature and salinity distribution within the area under investigation. However, for this work, the knowledge is actually not available. Therefore, the speed of sound is calculated using the UNESCO formula (UNESCO, 1983) with a fixed value for the temperature, the salinity and the pressure. The values of the environmental parameters are randomly selected from a Gaussian distribution. The mean values and the standard deviation are calculated from data recorded during a test dive in a harbour in Fremantle, Western Australia. The mean values and the standard deviations of the environmental parameters are summarised in Table 1.

Table 1: Environmental Parameters simulated

Parameter	\bar{x}	σ
Temperature	22.54 °C	0.054 °C
Salinity	23.35 PSU	0.675 PSU
Pressure	1086.1 mbar	28.41 mbar

AUV Depth

Usually, the actual depth of an AUV is calculated using an on-board pressure sensor. The pressure measured by the sensor is the sum of the barometric pressure P_{amb} and the hydrostatic pressure P_{hyd} caused by the water column above the sensor. Due to the slow change rate and the small variation of the barometric pressure, compared to the hydrostatic pressure, changes of the barometric pressure are neglected in this paper.

The hydrodynamic pressure is affected by the density of the seawater $\rho_{seawater}$, the gravity acceleration g and, the height of the water column z . The density of seawater is a function of temperature, salinity and pressure (UNESCO, 1983). The pressure dependency can be neglected if the maximum depth is less than 100 m (Nayar, et al., 2016). The temperature and salinity vary within the depth. The pressure at depth z can be calculated as follows:

$$P = g \cdot \int_0^z \rho(S(z), T(z)) dz \quad (8)$$

Where:

- g : gravity acceleration,
- $S(z)$: Salinity as function of the depth z ,
- $T(z)$: Temperature as function of the depth z ,
- $\rho()$: Density as function of salinity and temperature,
- z : Depth.

The calculation of the pressure using (8) requires detailed knowledge about the temperature and salinity distribution. However, for this work, this information is not available. Therefore, the speed of sound is calculated using a Gaussian distribution of the temperature and salinity (Table 1). In the simulations presented in this work, the pressure is calculated as follows:

$$P = g \cdot \rho(\bar{T}, \bar{S}) \cdot z \quad (9)$$

Where:

- g : gravity acceleration,
- $\rho(\bar{T}, \bar{S})$: Density as function of average temperature and salinity,
- z : Depth.

The calculated pressure value is measured by the on-board pressure sensor of the ROV. The pressure reading of the sensor is affected by measurement errors. This measurement error is modelled using a Gaussian distribution with the actual value of P as mean value and a standard deviation of $\sigma_{Pressure\ Sensor} = 40\ Pa$. This pressure is used to calculate the estimated depth \hat{z} of the ROV. In addition, the ROV measures the temperature and the salinity in order to estimate the speed of sound and the density. The measurement errors are modelled using a Gaussian distribution with the real values as mean values and standard deviations based on the sensor specifications. The standard deviation of the temperature sensor was set to $\sigma_{Temperature\ Sensor} = 0.05\ ^\circ C$ and the standard deviation of the salinity sensor was set to $\sigma_{Salinity\ Sensor} = 0.84\ PSU$.

SIMULATIONS

As shown in the previous section, different parameters, like the speed of sound or the movement of the receivers, have an influence on the performance of an acoustic based localisation system. In addition, the chosen baseline, i.e. the distance between the receivers, has an influence on the performance of the SBL system (Paull,

et al., 2014). The error of an acoustic based localisation system depends on the distance between the transducer and the receiver (Turetta, et al., 2014) and, if the baseline is not symmetric, the position of the ROV with respect to the position of the baseline. Hence, $n=100$ positions within the search radius of 100 m were selected randomly. The points are selected once and used for all simulations, in order to allow a fair comparison between the different solutions.

All error sources described above are modelled using normal distributed random numbers. Hence, the localisation error depends on the random number generation. Therefore, for each chosen ROV position $m=100$ position evaluations are carried out following the algorithm presented in Figure 3. Here, variables marked with an asterisk represent true values, whereas variables marked with a tilde represent error affected values.

```

function compute estimated position:
  calculate  $T^*(\bar{T}, \sigma_T)$  %real temperature|
  calculate  $S^*(\bar{S}, \sigma_S)$  %real salinity
  calculate  $v^*(T^*, S^*)$  %real speed of sound
  calculate  $\rho^*(T^*, S^*)$  %real density
  calculate  $t_i^*(x_i^*, x', v^*)$  %real time of flight (ToF)
  measure  $\tilde{t}_i(t_i^*, \sigma_t)$  %ToF error affected (EA)
  calculate  $P^*(\rho^*, z)$  %real pressure at depth z (Eq. 9)
  measure  $\tilde{P}(P^*, \sigma_P)$  %pressure (EA)
  measure  $\tilde{T}(T^*, \sigma_T)$  %temperature (EA)
  measure  $\tilde{S}(S^*, \sigma_S)$  %salinity (EA)
  calculate  $\tilde{\rho}(\tilde{T}, \tilde{S})$  %density (EA)
  calculate  $\tilde{z}(\tilde{P}, \tilde{\rho})$  %depth of the ROV (EA)
  calculate  $\tilde{v}(\tilde{T}, \tilde{S})$  %speed of sound (EA)
  calculate  $\tilde{d}_i(\tilde{t}_i, \tilde{v})$  %distance ROV and receivers (EA)
  calculate  $\tilde{x}_i(x_i^*, \sigma_x, \max_x)$  %position of the receivers (EA)
  calculate  $\tilde{r}_i(\tilde{d}_i, \tilde{x}_i, \tilde{z})$  %planar radius (EA) (Eq.3)
  calculate  $\hat{x}(\tilde{r}_i, \tilde{x}_i)$  %estimated position (Eq. 5&6)
return  $\hat{x}$ 

```

Figure 3: Pseudocode to compute the estimated position

The localisation error of an estimated position is calculated as the Euclidean distance between the estimated position \hat{x} and the real position x of the ROV as follows:

$$\epsilon_j = \sqrt{(x - \hat{x}_j)^2 + (y - \hat{y}_j)^2 + (z - \hat{z}_j)^2} \quad (10)$$

Where:

- ϵ_j : Error of the evaluation j ,
- x, y, z : Real position of the ROV,
- $\hat{x}, \hat{y}, \hat{z}$: Estimated position of the ROV.

Equation 10 gives the error of a single position estimation. In order to evaluate the performance of a chosen locator configuration, a fitness function is needed. The chosen fitness function should consider the errors of all ROV positions and all evaluations equally. Therefore, the fitness value of a chosen receiver configuration is calculated as follows:

$$f = \frac{\sum_{i=1}^n \left(\frac{\sum_{j=1}^m \epsilon_j}{m} \right)}{n} \quad (11)$$

Where:

- f : Fitness value of the receiver configuration,
- ϵ_j : Error of the evaluation j ,
- n : Number of target positions,
- m : Number of evaluations.

Previous published work recommends to setup the baseline as long as possible (Bingham, 2009). Furthermore, during simulations, the receivers are usually positioned on a rectangular shaped baseline, in order to simplify the spherical-based algorithm (Turetta, et al., 2014). According to the length of the locator cables and the requirements of a rectangular shaped baseline, a common baseline, using the Waterlinked SBL system, is shown in Figure 4. It can be observed from the figure that the locators are positioned at different depths. This shall improve the performance of the localisation algorithm (Water Linked AS, 2020). The localisation error of this common receiver configuration is determined to be $f_{common} = 1.5386$ m. The error of the optimised locator configuration should be less than the error of this standard configuration.

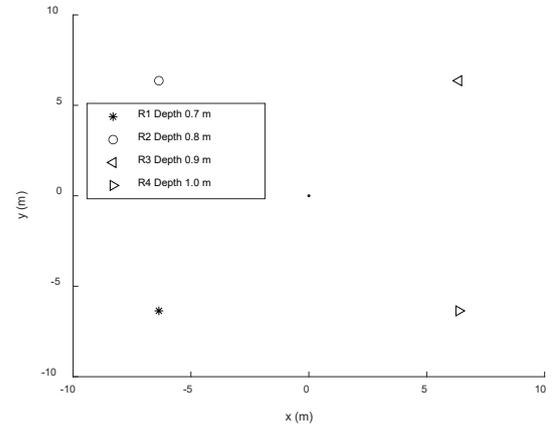


Figure 4: Common Baseline Configuration for the Waterlinked SBL

Since the length of the receiver cables is limited, the possible positions of the receivers are restricted. In addition, the cables should not be placed along the shortest possible routes, as there is a risk that propellers could damage the cables. The cables are placed in the x - y plane and then lowered to depth z . All receivers are connected to the main box of the SBL system, positioned at the origin of the coordinate system. Therefore, the cable length needed is calculated as follows:

$$d_i = \sqrt{x_i^2 + y_i^2} + z_i \quad (12)$$

Where:

- $[x_i, y_i, z_i]$: Coordinates of the receiver i .

All calculated cable length must be less than 10 m. Otherwise the cable length of the available receivers is not sufficient and the chosen receiver configuration cannot be used. In this case the fitness of this receiver

configuration will not be evaluated by using the simulation (Figure 3). Instead, a penalty value, bigger than the typical fitness values, is set as fitness value (Equation 13). This penalty strategy is commonly known as “death penalty” (Coello, 1999).

$$Penalty = \begin{cases} true; & \max(d_i) > 10 \text{ m} \\ false; & \max(d_i) \leq 10 \text{ m} \end{cases} \quad (13)$$

Where:

d_i : Euclidean distance between the receiver i and the origin of the SBL system.

PARTICLE SWARM OPTIMISATION

In this work, the optimal locator positions for the Water Linked SBL system are determined using particle swarm optimisation (PSO).

PSO is modelled on the behaviour of collaborative real world entities (particles), for example fish schools or flocks of birds, which work together to achieve a common goal (Kennedy & Eberhart, 1995). Each individual of the swarm searches for itself. However, the other swarm members also influence the search behaviour of each individual.

In the beginning of a search, each particle of the swarm starts at a random position and a randomly chosen velocity for each direction of the n -dimensional search space. Then, the particles move through the search space with an adjustable velocity. The velocity of a particle is based on its current fitness value, the best solution found so far by the particle (cognitive knowledge) and the best solution found so far by the whole swarm (social knowledge) (14):

$$\vec{v}_{i+1} = \vec{v}_i \omega + r_1 c_1 (\vec{p}_b - \vec{p}_i) + r_2 c_2 (\vec{g}_b - \vec{p}_i) \quad (14)$$

Where:

\vec{v}_{i+1} : new velocity of a particle,
 \vec{v}_i : current velocity of a particle,
 ω : inertia weight,
 c_1 : cognitive scaling factor,
 c_2 : social scaling factor,
 r_1, r_2 : random number from range [0,1],
 \vec{p}_i : current position of a particle,
 \vec{p}_b : best known position of a particle,
 \vec{g}_b : best known position of the swarm.

After calculating the new velocity of the particle, the new position \vec{p}_{i+1} can be calculated as follows:

$$\vec{p}_{i+1} = \vec{p}_i + \vec{v}_{i+1} \Delta t \quad (15)$$

Where:

\vec{p}_{i+1} : new position of a particle,
 \vec{p}_i : current position of a particle,
 \vec{v}_{i+1} : new velocity of a particle,
 Δt : time step (one unit).

In (15), Δt , which always has the constant value of one unit, is multiplied to the velocity vector \vec{v}_{i+1} in order to receive consistency in the physical units (Nolle, 2015). In this research the control parameter values for all

experiments were chosen as follows (Eberhardt & Shi, 2000):

$$\begin{aligned} \omega &= 0.1, \\ c_1 &= 1.49, \\ c_2 &= 1.49. \end{aligned}$$

Figure 5 shows pseudocode of the optimisation framework described afore.

```

Init PSO
for each Iteration do:
  for each Particle do:
    choose Position of Locators
    calculate cable length (Eq. 12)
    check Penalty (Eq. 13)
    if Penalty is true do:
      set fitness value to 10
    else do:
      for each target-position do:
        for each evaluation do:
          compute estimated position (Fig. 3)
          calculate error (Eq. 10)
        end for
      end for
      calculate fitness value (Eq. 11)
    end if
    update  $p_b$  and  $g_b$ 
    update  $\vec{v}$  and  $\vec{p}$ 
  end for
end for

```

Figure 5: Pseudocode for Optimisation of Locator Position

RESULTS

Seven experiments were carried out, which took approximately 30 hours on a high-end workstation. Figure 6 shows the development of the fitness values of the seven optimisation runs over time, i.e. iterations. It can be observed from the figure that, except from one experiment, PSO was able to minimise the fitness of the given problem.

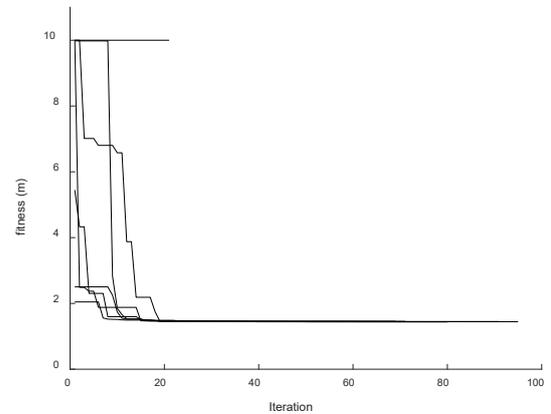


Figure 6: Fitness over Time for seven experiments

The g_b value of the different optimisation runs is given in Table 2. The mean fitness, except for run two, is $\bar{f} =$

1.4470 m and the standard deviation of the fitness, except run two is $\sigma_f = 0.0072$ m.

Table 2: g_{best} Values of the Optimisation runs

Run	g_{best}
1	1.4364
2	10
3	1.4540
4	1.4431
5	1.4496
6	1.4550
7	1.4441

The best receiver configuration was found in optimisation run one. The optimal position of the receivers is summarised in Table 3.

Table 3: Optimal Position of the Receivers

Receiver	x (m)	y (m)	z (m)
1	6.29	- 5.53	1.30
2	3.52	2.11	0.01
3	- 5.30	- 7.61	0.16
4	- 1.82	- 3.07	2.78

The optimal position of the four receivers is shown in Figure 7. It can be obtained from the figure, that, against the assumption, in the optimal configuration the receivers are not positioned in a rectangular shape.

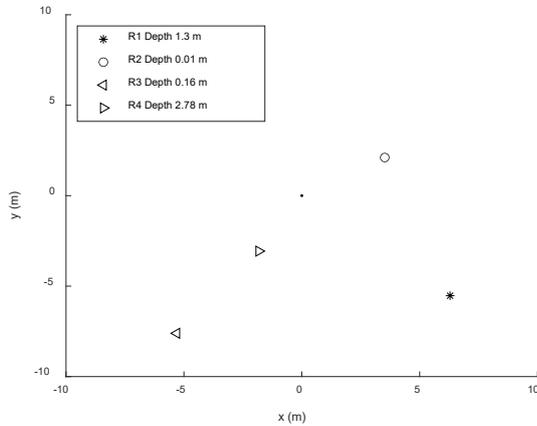


Figure 7: Optimised Receiver configuration

DISCUSSION

The mean localisation error of all receiver configurations, except run two, found by the PSO is 5.95 % smaller than the error of the original receiver configuration. The best solution is 6.64 % better than the original configuration. The optimal solution for the receiver configuration seems not to be a symmetric configuration (Figure 7). Figure 8 shows the receiver configuration of all successful optimisation runs. It can be observed from the figure that all solutions found by the PSO are not symmetric configurations.

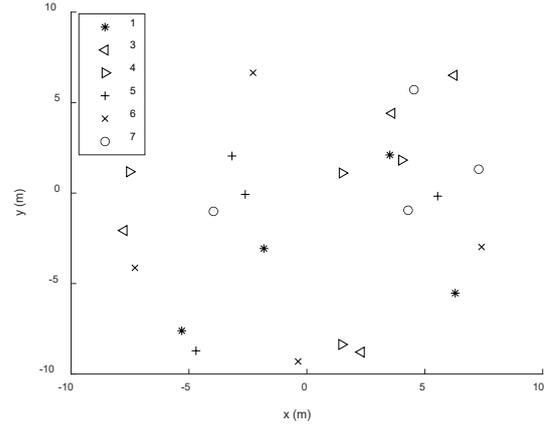


Figure 8: Receiver Configuration of all Successful Optimisation Runs

It can be observed from Figure 8 that in all configurations, except configuration 6, the receivers form a T-shaped baseline. Configuration six has the worst fitness value, compared to the other optimised solutions. The chosen penalty function does not allow the PSO to exploit any information from invalid receiver configurations. Therefore, if the whole population is outside the allowed area, the PSO is sometimes not able to find a suitable solution. Potentially, a penalty function, which takes the degree of violation into account when calculating the penalty value, might improve the performance of the optimisation.

CONCLUSION AND FUTURE WORK

The focus of this paper was an investigation of the localisation error using an SBL system. For this purpose, in a first step, different possible error influences are identified and a possible numerical simulation of the influence on the SBL was discussed. In the second step, the described simulation was used to determine the optimal position of the receivers of the system using PSO and Monte-Carlo simulations. For comparison, a standard configuration of the receivers, arranged in a symmetrical rectangular shaped baseline, was also simulated. The positioning error of the optimised receiver arrangement is 6.64 % smaller than the error of the standard arrangement. The accuracy specified by the manufacturer is one percent of the range, i.e. at a range of 100 m the accuracy is 1 m (Water Linked AS, 2020). The localisation error of the simulation is in the same order of magnitude as the proposed accuracy of the real SBL system used.

In future work the theoretical localisation error, calculated in this work, will be compared to real measured localisation errors. Furthermore, a cellular automaton (Tholen, et al., 2019) will be used in order to simulate the spatial and temporal changes in the speed of sound distribution and for the calculation of the hydrostatical pressure. In addition, other possible error sources, for example multi path propagation, will be

modelled and their impact on the localisation error will be analysed.

REFERENCES

- Almeida, R., Melo, J. & Cruz, N., 2016. Characterization of measurement errors in a LBL positioning system. *IEEE OCEANS 2016 - Shanghai*, pp. 1-6.
- Bingham, B., 2009. Predicting the Navigation Performance of Underwater Vehicles. *The 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 261-266.
- Coello, C. A. C., 1999. *A survey of Constraint Handling Techniques used with Evolutionary Algorithms*, Avanzada: Lania-RI-99-04, Laboratorio Nacional de Informática Avanzada.
- Eberhardt, R. & Shi, Y., 2000. Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization. *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00*, pp. 84-88.
- Kennedy, J. & Eberhart, R., 1995. Particle swarm optimization. *IEEE International Conference on: Neural Networks, Vol. 4*, pp. 1942-1948.
- Nayar, K., Sharqawy, M. & Banchik L. D. Lienhard, J., 2016. Thermophysical properties of seawater: A review and new correlations that include pressure dependence. *Desalination (390)*, pp. 1-24.
- Nolle, L., 2015. On a search strategy for collaborating autonomous underwater vehicles. *Proceedings of Mendel 2015, 21st International Conference on Soft Computing*, pp. 159-164.
- Parsopoulos, K. E. & Vrahatis, M. N., 2005. Unified Particle Swarm Optimization for Solving Constrained Engineering Optimization Problems. *Advances in Natural Computation. ICNC 2005. Lecture Notes in Computer Science, vol 3612*, pp. 582-591.
- Paull, L., Saeedi, S., Seto, M. & Li, H., 2014. "AUV Navigation and Localization: A Review,". *IEEE Journal of Oceanic Engineering, vol. 39, no. 1*, pp. 131-149.
- Tholen, C. et al., 2019. Automated Tuning Of A Cellular Automata Using Parallel Asynchronous Particle Swarm Optimisation. *Proceedings of the 33rd International ECMS Conference on Modelling and Simulation*, pp. 30-36.
- Tholen, C., El-Mihoub, T., Nolle, L. & Zielinski, O., 2018. On the robustness of self-adaptive Levy-flight. *OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO), Kobe*, pp. 1-5.
- Tholen, C. & Nolle, L., 2017. Parameter Search for a small swarm of AUVs using Particle Swarm Optimisation. *Artificial Intelligence XXXIV. SGAI 2017. Lecture Notes in Computer Science, vol 10630*, pp. 384-396.
- Tholen, C. et al., 2020. On the localisation of artificial submarine groundwater discharge sites using a low-cost multi-sensor-platform. *2020 OCEANS Singapore (to appear)*.
- Turetta, A. et al., 2014. Analysis of the Accuracy of a LBL-based Underwater Localization Procedure. *IEEE Oceans - St. John's (Newfoundland)*, pp. 1-7.
- UNESCO, 1983. Algorithms for computation of fundamental properties of seawater. *Unesco technical papers in marine sciences (44)*.
- Water Linked AS, 2020. *Water Linked AS Underwater GPS*. [Online] Available at: <https://waterlinked.com/underwater-gps/>
- Wu, Y. et al., 2019. Survey of underwater robot positioning navigation. *Applied Ocean Research (90)*, pp. 101845,.

AUTHOR BIOGRAPHIES

CHRISTOPH THOLEN graduated from the Jade University of Applied Science in Wilhelmshaven, Germany, with a Master degree in Mechanical Engineering in 2015. Since 2016 he is a research fellow at the Jade University of Applied Science in a joint project of 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.

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 and autonomous underwater vehicles.

LARS NOLLE graduated from the University of Applied Science and Arts in Hanover, Germany, with a degree in Computer Science and Electronics. He obtained a PgD in Software and Systems Security and an MSc in Software Engineering from the University of Oxford as well as an MSc in Computing and a PhD in Applied Computational Intelligence from The Open University. He worked in the software industry before joining The Open University as a Research Fellow. He later became a Senior Lecturer in Computing at Nottingham Trent University and is now a Professor of Applied Computer Science at Jade University of Applied Sciences. His main research interests are computational optimisation methods for real-world scientific and engineering applications.

ROBIN ROFALLSKI graduated from Jade University of Applied Sciences in Oldenburg, Germany with a Master degree in Geodesy and Geoinformatics in 2016. Since 2016, he holds a research fellow position at the Institute of Applied Photogrammetry and Geoinformatics at Jade University. His research interests are in underwater photogrammetry, camera calibration and dynamic metrology.

OLIVER RALLE graduated from the Jade University of Applied Science with a Bachelor degree in Mechanical Engineering in 2019. Currently he is studying his Master degree in Mechanical Engineering at Jade University of Applied Sciences.

USING THE CMA EVOLUTION STRATEGY FOR LOCATING SUBMARINE GROUNDWATER DISCHARGE

Tarek A. El-Mihoub, Christoph Tholen and Lars Nolle
Department of Engineering Science
Jade University of Applied Sciences
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
Email: {tarek.el-mihoub | christoph.tholen | lars.nolle}@jade-hs.de

KEYWORDS

Covariance matrix adaptation evolution strategy, Autonomous underwater vehicles, Submarine groundwater discharge, Population-based search.

ABSTRACT

For effective localisation of a search target by a swarm of Autonomous Underwater Vehicles (AUVs), a suitable cooperative search strategy should be utilised. Various aspects of the search task should be taken into account when selecting a search strategy. The nature of the search environment, the search target and the search agents should be considered. The Covariance Matrix Adaption Evolution Strategy (CMA-ES) is a well-known search strategy that proves its success in solving different continuous optimisation problems. This paper investigates utilising the CMA-ES to locate a Submarine Groundwater Discharge (SGD) using the temperature of water as a tracer. The impact of introducing some of the constraints, which are imposed by the search task, on the CMA-ES performance are studied. The influence of the number of the AUVs and their energy capacities on the search performance is investigated. The effect of the resolution of the temperature sensors together with the localisation and the navigation problems on the search behaviour are explored. The results show that these constraints have varying degrees of impact on the performance of the search strategy.

INTRODUCTION

Technological advances in Autonomous Underwater Vehicles (AUVs) have opened the doors to explore and access areas previously considered inaccessible (Bhat & Stenius, 2018). Different types of AUVs have been developed and used in different applications (Paull, et al., 2014).

Searching is an important class of AUVs applications. AUVs can be used, for example, to detect mines, locate groundwater discharge sources, search for harmful dumped waste and lost ship containers (Zielinski, et al., 2009).

A swarm of AUVs can be used to explore a predefined search area to locate mobile or stationary targets (Nolle, 2015). A huge number of search algorithms has been successfully applied to solve real-world problems. However, selecting a suitable search algorithm to guide

an AUV towards a point of interest is not an easy task. Different aspects should be considered when selecting or developing a cooperative search strategy for AUVs. Sensors' quality, energy constraints, localisation errors, navigation capabilities and communication quality are among the factors that influence the performance of a swarm of AUVs (Tholen, et al., 2017).

A cooperative search strategy can be used to guide a group of AUVs, as search agents, towards the most promising region. This search strategy should have the capability to analyse the search information gathered by the AUVs to suggest the best path to the target. Different population-based search algorithms have the capability of efficient utilisation of search information to locate a global optimum. However, using AUVs as search agents can affect the behaviour of these algorithms.

The efficiency of a population-based search algorithm depends on its ability to utilise the shared search experience to capture a global view of the search problem. Building a global view of the search problem depends on the population size, which is determined by the number of the available AUVs. The captured global view also depends on the quality of the shared information. The shared information includes the location information and the target information (i.e. the tracer information of the target).

The quality of the search information depends on the quality of the sensors that collect the location information and the tracer information. The quality of the sensor information depends on the accuracy and the resolution of the sensor. It also depends on the sampling rate and the response time together with the speed of the AUV. To acquire search information with a specific quality, sensors can impose constraints on the acquiring rate of search information and on the speed of the AUVs.

AUVs, as real-time search agents, impose other constraints on the search algorithm. There are restrictions in terms of their physical movement and their search range. To evaluate a search decision for exploring a search region, an AUV should move to that region and collect the requested search information. Such an evaluation can take some time depending on the speed of the AUV and the distance to that region. In addition, the search information can be changed before collecting the requested information. This change can be due to the dynamic nature of both the search algorithm and the phenomenon. Furthermore, the decision for a detailed

exploration of the current search region should be taken immediately and should not be delayed for collecting more search information. This delay can waste energy of the AUVs by revisiting some locations more than once. It can further restrict the ability of the AUVs and the search algorithm to explore the whole search space due to the limited energy capacities of the AUVs.

Efficient utilisation of the energy of an AUV is essential for search algorithms to locate a target. The energy consumption can be minimised through avoiding exploring unpromising search regions. It can also be reduced by decreasing sharp changes in the AUVs directions. A search algorithm that creates smooth search paths for AUVs can help in extending the search range of the AUVs.

AUVs as search agents when navigate to explore the search space are prone to localisation and navigation errors (Paull, et al., 2014). There are different ways to alleviate localisation and navigation problems. However, an effective search strategy should consider these errors. Special attention should be paid for localisation errors. The localisation errors can influence the search behaviour through effecting the accuracy of the search information. In addition, the aim of the search process is to define the exact location of the global optimum with an acceptable accuracy.

Evolution Strategies (ESs) (Schwefel, 1981) are black box population-based optimisation techniques. They have been studied for decades, leading to the many variants. The Matrix Adaption Evolution Strategy (CMA-ES) (Hansen & Ostermeier, 1996) is a well-known variant of ESs. It was originally designed for small population sizes and has been successfully applied to a considerable number of real world continuous domain problems (van Rijn, et al., 2017).

In this paper, the possibilities of utilising the CMA-ES as a search strategy for a swarm of AUVs to locate a Submarine Groundwater Discharge (SGD) using the temperature of water as a tracer is investigated. The robustness of the performance of the CMA-ES against some constrains of the task of locating SGDs is evaluated. These constrains include the number of AUVs and their energy capacity. They also include the resolution of the sensors. Furthermore, the impact of the localisation and the navigation problems on the algorithm behaviour is studied.

This paper is organised as follows. A very short introduction into SGDs is given in the second section. The CMA-ES algorithm, as described in (Hansen, 2006), is reviewed in the third section. The paper concludes by presenting and discussing the simulations' results.

SUBMARINE GROUNDWATER DISCHARGE

Submarine Groundwater Discharge (SGD) is the flow of water across the sea floor. Groundwater discharge may be pure groundwater entering the sea from a coastal aquifer, or it may be recirculated seawater, or some combination of the two (Burnett, et al., 2006).

SGDs connect the land and ocean in the global water cycle (Taniguchi, et al., 2019). They can have a significant influence on the costal environment (Taniguchi, et al., 2019). They are important sources of nutrients, dissolved inorganic carbon or trace metals to coastal waters. This continuous loading of nutrients and trace metals alters the water quality and may lead to environmental degradation of coastal regions (Prakash, et al., 2018).

Due to their impact on coastal regions, locating SGDs and tracking their dispersal is an important as well as challenging task. Natural tracers can be used to locate and quantify SGDs. Natural tracers, other than temperature, include nutrients, radioisotopes, salinity, and trace elements such as silica, barium, methane and others (Ray & Dogan, 2016). By measuring the changes in the natural tracers, SGDs can be located and quantified.

The contrasts between groundwater and sea surface temperatures can be used to locate SGDs. Detecting such contrast in temperature can be done using simple temperature sensors. The temperature difference can also modify the colour and the transparency of seawater. These changes in colour and transparency enable identifying SGDs form aerial photographs or satellite image. Temperature as tracer can be used to identify shallow SGDs and SGDs with high flow rates. However, it might not be suitable for detecting deep SGDs or SGDs with low discharge flow due to the high heat capacity of the seawater (Kelly, et al., 2013).

Different search strategies have been used to guide a swarm of AUVs to locate an SGD using the temperature as a tracer (El-Mihoub, et al., 2019; Tholen, et al., 2018; Tholen, et al., 2017). The reported results of applying these strategies show that the search task's constrains influence their performances. To gain insight into the relations between these constrains and the search performance, an investigation in applying a variant of the CMA-ES algorithm as a search strategy for a swarm of AUVs is conducted.

THE CMA-EVOLUTION STRATEGY

The CMA-ES algorithm optimises a fitness function $f: x \in \mathbb{R}^n \rightarrow f(x) \in \mathbb{R}$ by sampling a population of λ solutions (individuals) from a multi-variate normal distribution. It selects the best μ solutions (parents) out of the λ individuals to adaptively estimate the local covariance matrix of the objective function.

At generation g , the CMA-ES samples λ individuals according to

$$x_k^{g+1} \sim \mathcal{N} \left(m^{(g)}, (\sigma^{(g)})^2 C^{(g)} \right) \sim m^{(g)} + \sigma^{(g)} \mathcal{N} \left(0, C^{(g)} \right), \quad (1)$$

$$k = 1, \dots, \lambda$$

where

\sim denotes the same distribution on the left and right side.

$\mathcal{N}(0, C^{(g)})$ is the multi-variate normal distribution with zero mean and a covariance of $C^{(g)}$.

x_k^{g+1} is k -th offspring of generation $g + 1$.

$m^{(g)}$ is the mean value of the distribution at generation g .

$\sigma^{(g)}$ is the overall standard deviation, step size, at generation g .

$C^{(g)}$ is the covariance matrix at generation g .

$\lambda \geq 2$, is the population size or the sample size.

These λ individuals are evaluated and ranked. The mean $m^{(g+1)}$ of the distribution is updated and set to the weighted sum of the best μ individuals.

$$m^{(g+1)} = \sum_{i=1}^{\mu} w_i x_{i:\lambda}^{(g+1)} \quad (2)$$

where

$w_i > 0$ for $i = 1, \dots, \mu$

$\sum_{i=1}^{\mu} w_i = 1$,

$x_{i:\lambda}^{(g+1)}$ is the i -th best ranked individual out of λ individuals of $x^{(g+1)}$.

The CMA-ES updates the covariance matrix through considering the evolution path, p_c . The evolution path is the search path the strategy takes over a number of generation steps. It can be expressed as a sum of consecutive steps of the mean. The evolution path can be constructed through exponential smoothing. Starting with $p_c^{(0)} = 0$, $p_c^{(g+1)}$ can be calculated.

$$p_c^{(g+1)} = (1 - c_c)p_c^{(g)} + \sqrt{c_c(2 - c_c)\mu_{eff}} \frac{m^{(g+1)} - m^{(g)}}{\sigma^{(g)}} \quad (3)$$

where

$p_c^{(g)}$ is the evolution path at generation g .

$c_c \leq 1$, is the learning rate for cumulation update of the covariance matrix.

$\mu_{eff} = \frac{1}{\sum_{i=1}^{\mu} w_i^2}$, is the variance effective selection mass.

The evolution path is used for updating the covariance matrix. The rank-one and rank- μ updates of the CMA can be combined in a single formula with μ_{cov} to determine their relative weighting.

$$C^{(g+1)} = (1 - c_{cov})C^{(g)} + \frac{c_{cov}}{\mu_{cov}} U_{rank-one} + c_{cov} \left(1 - \frac{1}{\mu_{cov}}\right) U_{rank-\mu} \quad (4)$$

where

$U_{rank-one} = p_c^{(g+1)} p_c^{(g+1)T}$, is the rank-one update of the covariance matrix

$U_{rank-\mu} = \sum_{i=1}^{\mu} w_i \left(\frac{x_{i:\lambda}^{(g+1)} - m^{(g)}}{\sigma^{(g)}}\right) \left(\frac{x_{i:\lambda}^{(g+1)} - m^{(g)}}{\sigma^{(g)}}\right)^T$, is the rank- μ update of the covariance matrix.

$\mu_{cov} \geq 1$, is a parameter for weighting between rank-one and rank- μ update.

$c_{cov} \approx \min(\mu_{cov}, \mu_{eff}, n^2)/n^2$, is the learning rate for the covariance matrix update.

The CMA-ES uses a step size control for better estimation of the step size, σ^g . The cumulative path length control adapts the step size utilising the concept of evolution path. It compares the length of the evolution path with the expected length under random selection, i.e. $E \|\mathcal{N}(0, I)\|$. The cumulative path length control increases the step size if the evolution path is longer than expected, which indicates single steps are pointing to similar directions. On the other hand, the step size is

decreased, if the evolution path is shorter than expected, which indicates single steps cancel each other.

To make the length of evolution path independent of its direction for estimating the step size a conjugate evolution path p_σ can be calculated.

$$p_\sigma^{(g+1)} = (1 - c_\sigma)p_\sigma^{(g)} + \sqrt{c_\sigma(2 - c_\sigma)\mu_{eff}} C^{(g)-\frac{1}{2}} \frac{m^{(g+1)} - m^{(g)}}{\sigma^{(g)}} \quad (5)$$

where

c_σ is the learning rate for the step size update.

The conjugate evolution path is used to update the step size, σ^{g+1} based on the current step size.

$$\sigma^{g+1} = \sigma^g \times \exp\left(\frac{c_\sigma}{d_\sigma} \left(\frac{\|p_\sigma\|}{E \|\mathcal{N}(0, I)\|} - 1\right)\right). \quad (6)$$

where

d_σ is a damping parameter, which scales the change magnitude of the step size.

In the CMA-ES algorithm, there is a number of strategy parameters that control, separately, change rates of the mean m^g , the covariance matrix, C^g , and the step size, σ^g . However, default values for these strategy parameters have been defined and are applicable to a range of optimisation problems (Hansen, 2006).

The success of CMA-ES algorithms is due to estimating the parameters of the mutation distribution based on a set of selected steps not on a set of points (Hansen & Ostermeier, 1996). The use of cumulative path length control also enables adapting nearly optimal step sizes (Hansen, 2006). This improves convergence speed and global search capabilities at the same time.

SIMULATION AND RESULTS

A set of experiments was conducted using CMA-ES to locate the SGD, which has the highest discharge rate in an intermedia sized area with two SGDs. The aim of these experiments is to investigate the effect of some constrains, which are imposed by the search task, on the CM-ES behaviour. The strategy parameters of the CMA-ES algorithm were set to the default values as defined in (Hansen, 2006). The experiments were conducted with the assumption that AUVs can travel to any point in the search space in zero time and without any communication problems.

The problem of locating SGDs in a marine environment with the dimensions of 400m x 400m is simulated by a two-dimensional search space. In this space, two SGDs are located randomly. A Gaussian shape with maximum temperature at its centre is used to represent an SGD.

The average temperature of the water is set to 30 °C. The temperature of the centre of the SGD with the highest rate is set to 24 °C and the temperature of the centre of the second SGD was set to 26 °C (Akawwi, 2006). The radiuses of the basin of SGDs are selected randomly in the range from 10 to 20 m. The plume areas with these radiuses can be produced by SGDs with flow rates in the range from 0.00433 m³/s to 0.01524 m³/s (Kelly, et al., 2013).

The performance criterion in the simulations is the accuracy of locating the global SGD. The SGD with the

centre of highest temperature is the global SGD. The distance between the best-found location by an algorithm and the exact location of the global SGD's centre was defined as the error of that algorithm.

Each experiment was repeated for 100 times using the mentioned above search environment. The experiments' results were used to extract the cumulative distribution of the errors of each algorithm in each experiment. This cumulative distribution is used to estimate the probability of an algorithm to locate the global optimum with less than or equal to a specific error value.

Each experiment was conducted with a different number of AUVs, to study the combined effect of each constrain and the population size on the performance. The number of AUVs was set to λ_{min} , $2.5\lambda_{min}$, $5\lambda_{min}$ and $10\lambda_{min}$. λ_{min} is the default value for the population size strategy parameter. $\lambda_{min} = 4 + \lceil 3 \ln(n) \rceil$, where n is the dimension of the optimisation problem (Hansen, 2006).

Energy Capacity

The aim of the first set of experiments is to study the influence of the energy capacity of the AUV on the performance of the CMA-ES.

In literature, the number of function evaluations is usually used as a termination criterion when evaluating the performance of optimisation algorithms. However, for the task of locating SGDs using AUVs, energy consumption is a more realistic termination criterion. The cost of function evaluations, which is sensing the water's temperature, can be ignored. The goal of the swarm of AUVs is to locate the target before consuming their energy. With the goal of studying the effect of the number of AUVs and their energy capacity on the performance of the CMA-ES algorithm, the stopping criteria for search is consuming the energy stored on the AUVs.

The energy capacity of the AUVs can be translated in terms of meters travelled by the AUV. The energy

capacity of the AUV used in this research is sufficient for travelling a distance of 4,500 meter with an average speed of 1 m/s (Tholen, et al., 2018). This capacity was used as a reference in these experiments. The experiments were conducted for AUVs with 2250 meter, 4500 meter, 9000 meter, and unlimited energy capacities. The cost of changing the direction by an angle of 180° was assumed to be equivalent to travelling 4 meters (El-Mihoub, et al., 2019).

The experiments were conducted with the assumption that the AUVs estimate their location without errors, navigate with zero navigation error, and are using ideal sensors.

Figure 1 shows the performance of the CMA-ES algorithm with different number of AUVs and with different energy capacities. The figure shows that even with a population size of 60 AUVs and without any constrains, the CMA-ES is only able to find the global optimum with an accuracy of less than 5 meter in about 70% of the experiments. It also demonstrates as expected that the decrease in the energy capacities of the AUVs degrade the algorithm performance. It also shows that as the capacity decreases, the difference in the performance between the algorithms with 60, 30 and 15 AUVs decreases. In other words, there is a minimum of energy requirements for effective cooperation regardless of the number of cooperating AUVs. The figure shows that increasing the energy capacity and increasing the number of AUVs does not guarantee locating the global optimum even with an off the shelf state-of-the-art search algorithm. Another set of experiments has been done using the genetic algorithm optimisation tool of MATLAB to solve this problem. The results, which are not shown here, demonstrate that the probability of locating the global optimum with an accuracy of less than 5 meter is less than 0.6.

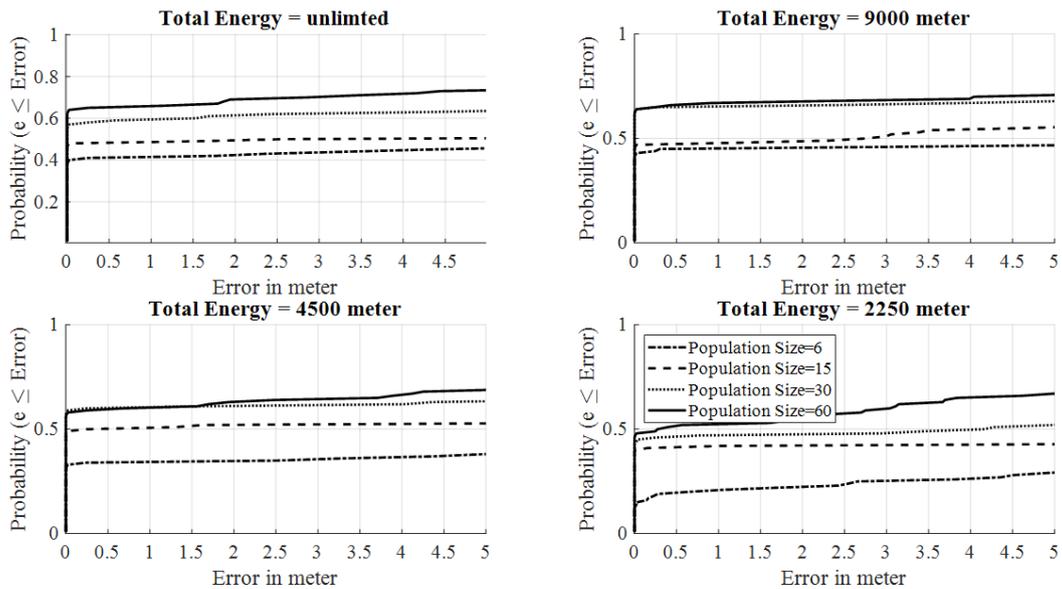


Figure 1: The change in the performance for different population size and different energy capacities

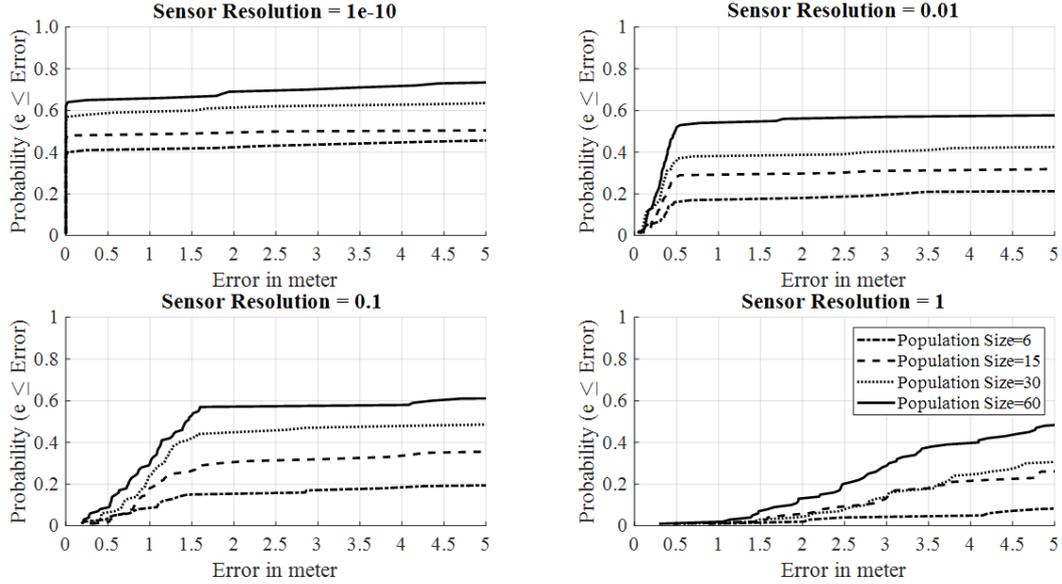


Figure 2: The Influence of the Sensor Resolution on the Performance

Sensor Resolution

Search algorithms evaluate the quality of a search region for further detailed exploration based on the quality of the samples of the region. The quality of these samples depends on the feature used to trace the target and the quality of the sensor for quantifying this feature. For locating SGDs in this paper, temperature is used as a tracer. The relation between the location of an SGD and the current location depends to some extent on the difference in the temperature. This relation is the objective function of the search algorithm. It is the only mean for differentiating the quality of the sampled locations.

The search algorithm uses the readings of the temperature sensor at selected locations as the objective value of these locations. The actual objective function used by the search algorithm depends on the details of the sensor. Sensors can modify the original relation between the SGD location and the temperature. Sensors can introduce some noise in mapping locations to temperature depending on their accuracy. Furthermore, the response time of the sensor can introduce errors in this mapping. Instead of optimising the original, response time can lead to optimising $f: x \in \mathbb{R}^n \rightarrow f(x - \Delta_r) \in \mathbb{R}$, Δ_r is the accumulated effect of the response time. The sensor resolution, which defines the smallest measurement a sensor can reliably indicate, can transform the objective function into a staircase function.

In this section, the influence of the sensor resolution on the performance of the CMA-ES is evaluated. The experiments were conducted for sensors with resolutions of 1e-10 (an ideal resolution), 0.01, 0.1 and 1.0 °C. The sensors are assumed to have no accuracy errors and zero response time.

The experiments were conducted with the assumption that the AUVs estimate their location without errors, navigate with zero navigation error, and have unlimited energy capacity.

Figure 2 shows the results of these experiments. The figures show that the sensors resolution influences both the accuracy and the probability of locating the global optimum. The effect of changing the sensor resolution on the performance is more significant than that of the energy capacity. The graphs show that the algorithm with 60 AUVs is able to locate the global optimum with an acceptable accuracy with resolutions up to 0.1 °C. The algorithm shows poor performance with a resolution of 1°C. The results of these experiments are expected due to the impact of the resolution on the ability of the algorithm to differentiate between sampled locations.

Localisation Errors

To assess the impact of the localisation errors on the CMA-ES performance, another set of experiments was conducted. Gaussian probability distributions with standard deviations of {0.0, 0.1, 0.3, 1.0} were used to model the localisation errors. These distributions can produce errors in the range roughly from -3σ to 3σ . For AUVs with zero localisation error, a distribution with $\sigma = 0$ is assumed. $\sigma = 0.1$ is assumed for AUVs, that rely on GPS for localisation. $\sigma = 0.3$ and $\sigma = 1.0$ is assumed for AUVs, that use underwater localisation techniques. The experiments were carried out with the assumption that the AUVs navigate with zero navigation error, have unlimited energy capacity and ideal temperature sensors.

The results of the experiments, as shown in Figure 3, demonstrate that the change in the localisation errors affect the accuracy of locating the global optimum. The plots show that these errors do not misguide the search but can decrease its ability to find the exact location of the global optimum. It is worth mentioning that the effect on the accuracy is related to the range of localisation errors. This error can be due to the error in reporting the exact location of the global optimum.

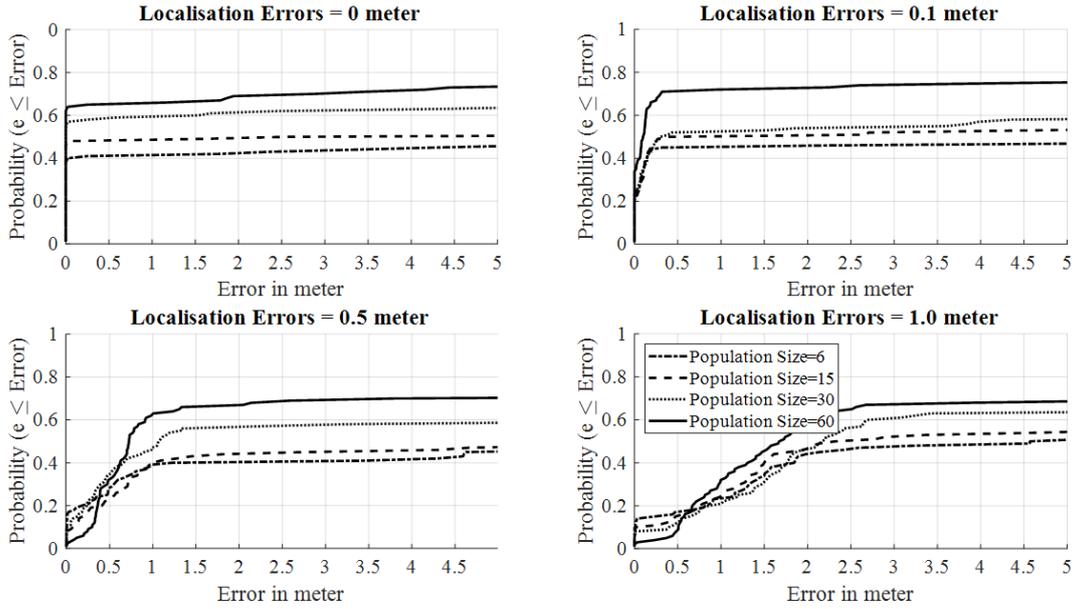


Figure 3: The Impact of Localisation Errors on the Performance

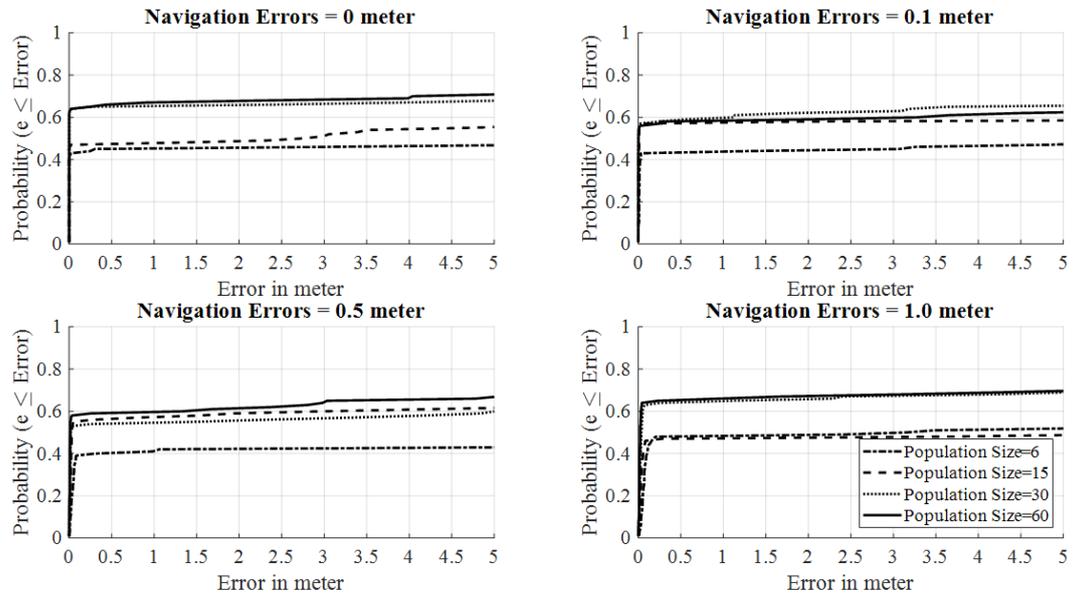


Figure 4: The Navigation Problem and the Performance

On the other hand, the localisation errors have a small impact on the probability of locating the global optimum. The localisation errors in the experiments do not accumulate over time. The localisation process at each sampled location sets upper limits on the localisation error values. The localisation errors can introduce errors in mapping the locations to their quality. However, the CMA-ES algorithm is resilient against these errors.

Navigation Errors

The last set of experiments were carried out to evaluate the effect of the navigation error on the CMA-ES performance. Gaussian distributions with standard deviation values as those defined for localisation errors are used to model the navigation errors.

These experiments were executed with the assumption that the AUVs can estimate their locations without any error, have unlimited energy capacity and have ideal temperature sensors.

The results of these experiments, depicted in Figure 4, show that the navigation errors have the least significant impact on the performance compared with other constrains. The navigation errors alone can lead the search to sample points other than the target points. The distance between the sampled points and the target points depends on the navigation errors. However, navigation errors do not cause any errors in evaluating the sampled points. It does not introduce any errors in mapping the location to the objective function. Since the CMA-ES is a stochastic algorithm, non-accumulative navigation

errors does not introduce a significant effect on its search behaviour. They can affect slightly the accuracy of locating an SGD as they can guide the search to a location near the exact location in the final stages of the search, as shown in graphs of small population sizes. Moreover, navigation errors can improve the diversity of the population and improve the search results, as shown in Figure 4 for navigation errors with a standard deviation of 0.1.

CONCLUSION AND FUTURE WORK

Selecting a suitable cooperative search algorithm for a swarm of AUVs necessitate studying its robustness against constrains, which are imposed by the search task. To shed light on the impact of these constrains on the algorithm performance, the performance of CMA-ES as a state-of-the-art algorithm against some of these constrains was evaluated. The performance of the algorithm was studied in locating the global SGD using the water's temperature as a tracer.

The experiments show that using an off the shelf state-of-the-art search algorithm cannot guarantee solving the problem even with a large number of AUVs and with unlimited energy capacities. The experiments also illustrate that using the CMA-ES with a population size, which is recommended in literature (Hansen, 2006), produces a poor performance in locating the global SGD in a search space with two SGDs.

The experiments also demonstrate that the resolution of the sensor has a significant influence on the search behaviour. Sensors with high resolutions empower the discrimination ability of the algorithm between similar solutions. On the other hand, sensors with low resolutions transform the objective function to a kind of staircase function and degrade the algorithm ability of discrimination between sampled solutions.

The localisation errors also have a considerable impact on the performance. The localisation errors can lead to associate the quality of sampled solutions to other solutions. This can lead to errors in evaluating the quality of the sampled solution and can lead the search towards non-optimal solutions. The simulations show that, in most cases, the localisation errors do not prevent the search from locating the global optimum. Meanwhile, they can degrade the algorithm accuracy in reporting the location of the global optimum.

The next step in this research is to investigate possible ways to improve the CMA-ES algorithm performance in locating SGDs. It will include studying the impact of the physical movement of the AUVs and the communication constrains on the algorithm performance. Utilising the $(1,\lambda)$ -ES with mirrored sampling and sequential selection (Auger, et al., 2011) as a search algorithm for a single AUV search will be also investigated.

REFERENCES

- Akawwi, E. J., 2006. Locating Zones and Quantify the Submarine Groundwater Discharge into the Eastern Shores of the Dead Sea-Jordan.
- Auger, A., Brockhoff, D. & Hansen, N., 2011. Analyzing the Impact of Mirrored Sampling and Sequential Selection in Elitist Evolution Strategies. Schwarzenberg, Austria, s.n., pp. 127-138.
- Bhat, S. & Stenius, I., 2018. Hydrobotics: A Review of Trends, Challenges and Opportunities for Efficient and Agile Underactuated AUVs. Porto, Portugal, IEEE, pp. 1-8
- 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. *Science of The Total Environment* 367 (2-3), pp. 498-543.
- El-Mihoub, T., Tholen, C. & Nolle, L., 2019. Informed search patterns for alleviating the impact of the localisation problem. Caserta, Italy, pp.37-42
- 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. & Ostermeier, A., 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. s.l., s.n., p. 312–317.
- Kelly, J., Glenn, C. & Lucey, P., 2013. High-resolution aerial infrared mapping of groundwater discharge to the coastal ocean. *Limnology and Oceanography Methods*, pp. 262-277.
- Nolle, L., 2015. On a search strategy for collaborating autonomous underwater vehicles. Brno, s.n., pp. 159-164.
- Paull, L., Saedi, S., Seto, M. & Li, H., 2014. AUV Navigation and Localization: A Review. *IEEE JOURNAL OF OCEANIC ENGINEERING*, pp. 131-149.
- Prakash, R., Srinivasamoorthy, K., Gopinath, S. & Saravanan, K., 2018. Measurement of submarine groundwater discharge using diverse methods in Coleroon Estuary, Tamil Nadu, India. *Applied Water Science*, p. 8:13.
- Ray, L. & Dogan, A., 2016. Contemporary Methods for Quantifying Submarine Groundwater Discharge to Coastal Areas. In: *Emerging Issues in Groundwater Resources*. s.l.:Springer, Cham, pp. 327-364.
- Schwefel, H., 1981. *Numerical Optimization of Computer Models*. s.l.:John Wiley & Sons.
- Taniguchi, M. et al., 2019. Submarine Groundwater Discharge: Updates on Its Measurement Techniques, Geophysical Drivers, Magnitudes, and Effects. *Frontiers in Environmental Science*, Volume 7, pp. 1-26.
- Tholen, C., El-Mihoub, T., Nolle, L. & Zielinski, O., 2018. *On the robustness of self-adaptive Levy-flight*. Kobe, IEEE, pp. 1-5.
- Tholen, C., Nolle, L. & Werner, J., 2017. *On the Influence of Localisation and Communication Error on the Behaviour of a Swarm of Autonomous Underwater Vehicles*. Recent Advances in Soft Computing. MENDEL 2017, pp 68-79.
- van Rijn, S., Wang, H., van Stein, B. & Bäck, T., 2017. *Algorithm configuration data mining for CMA evolution strategies*. Berlin, Germany, ACM, pp. 737-744 .
- Zielinski, O. et al., 2009. Detecting marine hazardous substances and organisms: sensors for pollutants, toxins, and pathogens. *Ocean Science*, pp. 329-349

Industrial Process Modelling and Simulation

NUMERICAL SIMULATION OF CONDENSING AMMONIA IN PLATE HEAT EXCHANGERS USING CFD

Alexander Dietrich, Mario Nowitzki, Ron van de Sand and Prof. Dr.-Ing Jörg Reiff-Stephan
Faculty of Engineering and Natural Sciences
Technical University of Applied Sciences Wildau
15745, Wildau, Germany
E-mail: dietrich@th-wildau.de

KEYWORDS

Condensation, CFD, plate heat exchanger, refrigeration

ABSTRACT

For reasons of environmental compatibility and sustainability, the demand for more efficient heat exchangers is growing. The design of heat exchangers is based on experimentally determined correlations. For the use of heat exchangers under flow conditions that have not been researched yet, complex experimental investigations are necessary or a safety factor is applied. A third alternative is the investigation using Computational Fluid Dynamics (CFD). For this, however, it is necessary to generate an exact numerical model. Therefore, this study deals with the generation of a numerical model to simulate a condensation process of a widely used natural refrigerant, namely ammonia, within a plate heat exchanger (PHE). Based on the results of the pressure drop and the heat flow, the simulation's plausibility is checked with a validation test case. Thus, a statement about the accuracy of the CFD simulation is given.

INTRODUCTION

The consideration of condensation processes forms an integral part in the development of high-efficiency heat exchanger, such as heat pipes, nuclear industry reaction towers or refrigeration cycles (Huang et al. 2012) and is often the focus of attention in terms of reliability and efficiency aspects. Two-phase cooling schemes can deliver orders of magnitude enhancement in heat transfer coefficient compared to their single-phase counterparts. For that reason, condensation within heat exchangers is increasingly addressed across the literature in recent years (Kim and Mudawar 2013).

Especially with regard to vapor compression refrigeration systems (VCRS), predicting the condensation of the refrigerant within the condenser is crucial for the design of the overall system. However, as vapor and liquid flows occur simultaneously, the accurate estimation of the refrigerant flow characteristic, as well as the heat transfer, is more complex compared to the single-phase flow (Bhramara et al. 2009). Besides, condensers often consist of plate heat exchangers, which are generally more thermally efficient compared to their shell-and-tube counterparts (Huang et al. 2012). Due to the greater complexity of the channel geometry (Wang et al. 2007), which promotes a high degree of turbulence of the flow (Huang et al. 2012), the heat transfer estimation

is additionally complex. Consequently, the accuracy of predicting the two-phase heat transfer in such non-circular channels by use of empirical correlations is often insufficient and may require experiments. Another approach is to use the Computational Fluid Dynamics to model the two-phase flow and heat phenomena in a PHE. In the field of fluid mechanics, such a model is most commonly developed using CFD, which can be applied to simulate the condensation process within a PHE. This allows an improved design of the VCRS and enables the theoretical evaluation of the system efficiency. Therefore, this work presents an approach towards the development of a CFD based numerical model for predicting the condensation process of ammonia in PHE. Furthermore, it describes the simulation and discusses its result based on a validation test case.

RELATED WORKS

Approximation of pressure drop and heat transfer during condensation of refrigerants in plate heat exchangers has gained attraction of researchers and many approaches are well described across the literature. Most authors found appropriate models though experimental investigations, which are widely applied for the development of VCRS. The empirical correlations describe the heat transfer and pressure drop of different refrigerants for different ranges of validity. For example, correlations for R134a were formulated by Yan et al. (Yan et al. 1999) and Zhang et al. (Zhang et al. 2019). Other authors, in turn, describe formulations for other refrigerants, such as R410a (Kuo et al. 2005) or hydrocarbons (Thonon and Bontemps 2002). García-Cascales et al. gives a large overview of correlations for single-phase and two-phase heat transfer (García-Cascales et al. 2007). Similar papers from Park and Hrnjak (Park and Hrnjak 2008) and Numrich and Müller (Numrich and Müller 2013) give general predictions for the condensation heat transfer and pressure drop in thin tubes.

Yet, a full understanding of the prediction of ammonia during condensation in plate heat exchanger is still lacking. Khan et al. (Khan et al. 2012) and Djordjevic and Kabelac (Djordjevic and Kabelac 2008) have already dealt with the formulation of correlation for the evaporation heat transfer and pressure drop of ammonia in plate heat exchangers. However, these correlations cannot be applied to condensation, since evaporation and condensation heat transfer are different (Soler 1996).

A problem with the simulation of two-phase flows is that locally different morphologies of the phase interface can

occur (Höhne and Vallée 2010). For that reason, an Algebraic Interfacial Area Density model (AIAD) was implemented based on the mixture model for the application in CFD (Höhne and Vallée 2010). In several studies, the results of the simulation with experimental data are compared (Höhne and Vallée 2010, Höhne and Lucas 2011). The developed model allows switching between correlations for the respective predominant morphology in the flow. Although the prediction of heat transfer and pressure drop of condensing refrigerant within PHEs has been addressed throughout the literature, the numerical approach still represents a challenge as the flow-describing conservation equations are not known/correct for many phenomena (e.g. multi-phase flow, turbulence, combustion) (Ferziger and Peric 2008). However, through the application of CFD, complex structures can be modeled and the condensation process could be observed within a wide range of applications.

In the two-phase model, the interfacial area density is one of the most important parameters (Wu et al. 1997). In industrial practice, however, the Sauter mean diameter is used to determine the interfacial area density, which is usually estimated through empirical correlations (Castellano et al. 2018). Another method for predicting the particle size distribution is through population balance equations PBE. PBE's are used to describe how the disperse phase develops as a population of entities in a continuous phase (Ramkrishna 2000). For an accurate description, however, a large number of population classes are required, which increases the computational effort (Lo and Zhang 2009). A simpler representation of the particle size distribution is obtained by a combination of breakup and coalescence models. The modification and development of new models for the interfacial area density is the topic of numerous publications (Yao and Morel 2004, Ishii and Kim 2001).

VALIDATION TEST CASE

The validation test case result data were taken from (Alfa Laval 2011) and are based on an ammonia water-glycol countercurrent plate heat exchanger condenser (Type Alfa Laval AlfaNova 76-80H). It is derived from the study carried out in (van de Sand et al. 2019) of which the collected data is used throughout this paper to estimate the numerical model deviation to the measurements.

Generally, the core of a plate heat exchanger is the stack of corrugated metal plates. In two separate circuits, refrigerant ammonia and coolant water-glycol are alternately passed through the plates. It is understood that the pressure drop, as well as the heat transfer, are significantly affected by the geometry and characteristics of the plate shapes (Focke et al. 1985). As illustrated in Figure 1, the chevron pattern embossings of two adjacent plates are arranged in opposite directions. Each plate is characterized by a chevron angle β , a wall thickness s , a corrugated pitch Λ , an amplitude b , a plate length L_P and a plate width W_P . The diameter of the ports which

provide access to the flow passages on either side (inlet/outlet) is given by D_P .

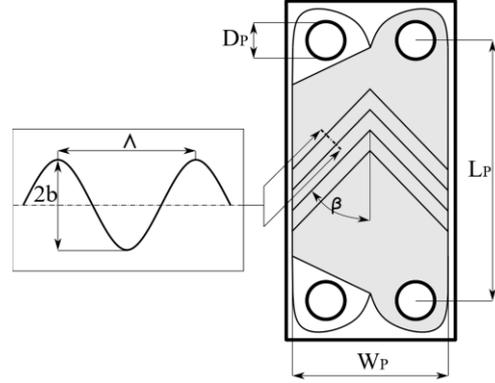


Figure 1: Corrugated Plate

To calculate the heat transfer and pressure drop for that validation test case, dimensionless number correlations for plate heat exchangers described by Martin (Martin 1996) are used. With a dimensional analysis, it can be shown that the Nusselt number Nu , the Reynolds number Re and the Prandtl number Pr are crucial parameters for the heat transfer and pressure drop (Stephan et al. 2019). In order to determine the dimensionless numbers, Martin (Martin 1996) shows, that the characteristic length can be replaced by the following hydraulic diameter, see equation 1.

$$d_h = \frac{4b}{\Phi} \quad (1)$$

Where Φ is the dimensionless parameter for the area enlargement factor and the result of the increased heat transfer surface due to the corrugated embossings. The dimensionless area enlargement factor definition can be seen in equation 2.

$$\Phi = \frac{1}{6} \cdot \left(1 + \sqrt{1 + X} + 4 \cdot \sqrt{1 + \frac{X^2}{2}} \right) \quad (2)$$

Where X is the dimensionless corrugation parameter, which can be determined from the heat exchanger plate corrugated pitch Λ and amplitude b , see equation 3 and Figure 1.

$$X = 2 \cdot \pi \cdot \frac{b}{\Lambda} \quad (3)$$

NUMERICAL MODEL

Grid generation

In this study, the CFD analysis is conducted based on a 3D model representing the continuum (channel) between two corrugated plates of the PHE. For the analysis of heat transfer and two-phase flow in a plate heat exchanger, it is possible to reduce the computational effort by simulating a single channel instead of the entire PHE. In

order to perform the simulation, the software StarCCM+ is used in this study. The inlet and outlet ports are extended so that the inlet and outlet boundary conditions have less effect on the flow conditions in the heat transfer area inside the channel.

The 3D model is meshed with polyhedral elements. In the channel, there are thin layer elements on the walls, in order to resolve the temperature boundary layer. In the segment where the corrugated embossings cross (main heat transfer area) the mesh is locally refined to better resolve the higher turbulence and thus heat transfer rates. Table 1 shows the mesh parameters. The mesh quality meets the recommendations of Lloyd and Espanoles (Lloyd and Espanoles 2002). In advance, the numerical model was tested on a tube condenser. A mesh-independent study was carried out for a tube condenser. Finally, the simulation results of the tube condenser were validated with an appropriate correlation from Stephan et al. (Stephan et al. 2019). The result of this latter previously investigation (mesh independence study) is a deviation of the numerical model of -49.9% for the pressure drop and -53% for the heat flow. Accordingly, a discretization error for the PHE simulation of approximately -2% for the pressure drop and approximately -10% for the heat transfer can be expected.

Table 1: Plate heat exchanger mesh parameters

Total layer thickness	0.4mm
Number of layers	3
Prism layer stretching	1.3
Base size	1.7mm
Min. face size	0.102mm
Cell element	Polyeder

Flow Model

To take into consideration the separation of the two-phase flow inside the heat exchanger channel, the gravity force must be included in the simulation model. The plate heat exchanger, which is examined in this study, is positioned vertical, so that the condensate can collect at the bottom of the heat exchanger and can be removed to the outlet nearby the bottom.

In order to model the two-phase flow within the PHE, the Eulerian Multiphase flow model is further applied, in which both the liquid and gas phase are considered as separate continua (Parekh and Rzehak 2018). For each of the two phases, separate conservation equations for mass and momentum are solved. The Algebraic Interface Area Density Model describes the interaction between the phases (Ishii and Hibiki 2010).

Due to the small temperature change close to the dew point, the simulation can be simplified. The gas phase is assumed as an ideal gas and the liquid phase with constant density. In the present study, the two-equation turbulence model $k-\varepsilon$ is used to represent the turbulent properties of the flow.

Algebraic Interface Area Density Model

The Algebraic Interface Area Density Model (AIAD) was developed to capture different flow patterns in two-phase flow. Ishii and Hibiki (Ishii and Hibiki 2010) are giving an overview of different opportunities to mathematically describe two-phase flows. Depending on the gas/liquid volume fraction, three regimes are introduced of which each regime is provided with its own correlations and coefficients for the exchange of moments. Hence it is possible to switch between the correlations depending on the respective flow pattern locally. The three regimes are:

- Disperse bubbles in continuous liquid phase (bubbly regime)
- Liquid droplets in continuous gas phase (droplet regime)
- Separated Flow (free surface regime)

A blending function makes it possible to switch between regimes. This enables the localization and detection of the different regimes.

A regime is defined by its volume fraction. The bubbly regime is used for vapor volume fractions smaller 30 % and the droplet regime for vapor volume fraction greater than 70%. If the volume fraction is between 30 and 70%, the free surface regime is used. Figure 2 shows the relation between regimes and blending function.

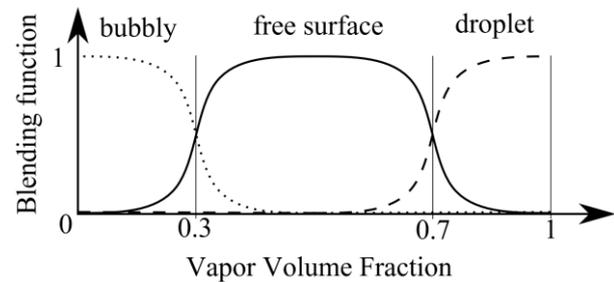


Figure 2: blending function for multiple flow regimes according to Porombka and Höhne (Porombka and Höhne 2015)

Generally, the momentum transport between phases must be modeled. The sum of the drag force and non-drag forces is the interfacial momentum transfer. The most crucial force for interfacial momentum transport is the drag force (Porombka and Höhne 2015).

Drag Force:

The drag force is the resistance on disperse particles in a continuous phase caused by different relative velocities to each other (Ishii and Hibiki 2010). The drag force is calculated as follows:

$$F_D = C_D \cdot A_{\text{proj}} \cdot |U_{\text{rel}}|^2 \cdot \frac{\rho}{2} \quad (4)$$

where U_{rel} is the relative velocity between the phases, ρ the density of the continuum phase and A_{proj} the projected area of a particle. C_D is the drag force coefficient. In eq.

4 C_D and A_{proj} are unknown values that must be calculated. Thus, for the droplet and bubble regime, the correlation for C_D according to (Schiller and Naumann 1933) is used, which is suitable for flows consisting of small spherical bubbles/droplets. For the free surface regime, an adapted method of Štrubelj and Tiselj (Štrubelj Tiselj 2011) is used to calculate the interface drag coefficient.

$$A_{\alpha\beta} = \frac{A_{proj}}{V_{sphere}} \quad (5)$$

The Interfacial Area Density $A_{\alpha\beta}$, as shown in eq. 5, specifies the ratio of interfacial area A_{proj} per volume V_{sphere} . α and β represent the bubbly and droplet regime. For both the bubbly and droplet regime the spherical particle approach and for the free surface regime the mixture approach is used, see (Nowitzki 2020).

Surface Tension:

The surface tension is added to computation to ensure a correct fluid particle wall interaction. According to Hyvärinen et al., the surface tension for pure ammonia at 25°C is 0.021 mN/m (Hyvärinen et al. 2005).

Non-Drag Forces:

In addition to the Drag Force, which points in the opposite direction of the velocity vector, there are other forces perpendicular to the direction of flow affecting the momentum transfer. Ishii and Hibiki divide these forces into lift force, wall lubrication force and turbulent dispersion force (Ishii and Hibiki 2010). According to Legendre and Magnaudet, the lift force describes the lift/shear force on particles generally moving in a rotational flow (Legendre and Magnaudet 1998). In this study, the Tomiyama model (Tomiyama et al. 2002) is used for the bubbly regime and a constant coefficient of 0.25 is used for the droplet regime, following (Lance and Bataille 1991). Other Non-Drag forces than the lift force are neglected (Méndez et al. 2005).

Boundary conditions

Inlet condition:

In this study, according to the measurements in (Alfa Laval 2011) the ammonia gas mass flow value per single channel can be assumed as 0.003 kg/s.

Outlet condition:

The ammonia leaves the plate channel at a pressure of 1.16 MPa. The outlet values for temperature and the volume fraction of each phase that needs to be chosen are determined by a function, that is calculating the average value on a plane just before the outlet port.

Wall condition:

All heat transferring surfaces are assumed as a convective heat source to model the heat transfer on the coolant side. Heat transfer surfaces are all surfaces where the channel volume is touched by the plates. The remaining surfaces are assumed adiabatic, such as the wall of the inlet/outlet

extensions or the sides of the plate channel. The temperature of the coolant rises logarithmically (Cartaxo and Fernandes 2011). For that reason, the average temperature of the coolant (water-glycol) is 24.4 °C and is calculated from the logarithmic mean of the temperatures given in (Alfa Laval 2011) at the inlet of 22 °C and at the outlet of 27 °C in the coolant circuit. The heat transfer coefficient of the coolant channel α_{WG} is calculated based on the correlation of Focke et al. (Focke et al. 1985) for water as:

$$\alpha_{WG} = 0.44 \cdot Re^{0.64} \cdot Pr^{0.5} \cdot \frac{\lambda_{WG}}{d_h} \quad (6)$$

where Pr is the Prandtl number, Re the Reynolds number, λ_{WG} is the thermal conductivity and d_h the hydraulic diameter, see equation 1. The Prandtl number consists of material properties of water-glycol. The Reynolds number is calculated as follows in equation 7:

$$Re_{WG} = \frac{U_{WG} \cdot d_h \cdot \rho_{WG}}{\mu_{WG}} \quad (7)$$

where ρ_{WG} is the density and μ_{WG} the dynamic viscosity of the coolant. The calculation of U_{WG} can be found in (Martin 1996).

Nusselt number

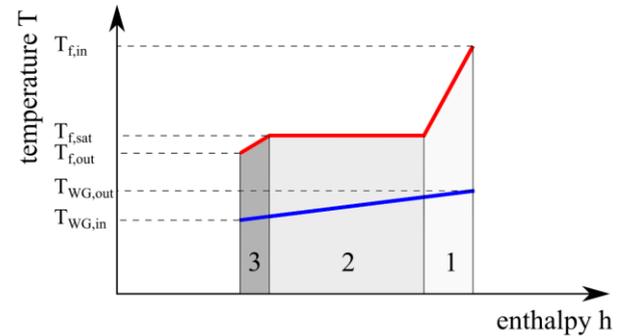


Figure 3: temperature profile of different heat transfer regions

It is necessary to specify the Nusselt number in order to simulate the heat transfer. As illustrated in Figure 3, the red graph represents the temperature drop of the refrigerant side and the blue graph the rise in temperature of the coolant side. $T_{f,in}$ and $T_{f,out}$ represent the inlet and outlet temperatures of the refrigerant circuit, $T_{WG,in}$ and $T_{WG,out}$ the inlet and outlet temperatures of the coolant circuit and $T_{f,sat}$ is the saturation temperature of the refrigerant. The whole heat transfer process is divided into three regions, as shown in Figure 3. Region 1: vapor desuperheating, region 2: saturation condensation, and region 3: liquid subcooling. Since the heat transfer rates vary a lot in these later different regions, it is necessary to treat them separately. Table 2 shows the used sources to determine the Nusselt number for each regime.

Table 2: Used Nusselt number for different regimes

Region	Nusselt number
1	Benchmark following (Hell 1992)
2	Determining according to calculation model from (Zhang et al. 2019)
3	Determining according to (Martin 1996)

Modeling the interaction length scale

Generally, dimensionless numbers in CFD, such as the Reynolds number, are based on the interaction length scale (ILS), which must be estimated. Usually, the Sauter mean diameter d_{32} can be used to specify the ILS (Zogg 1987). It is considered as an average of the particle size by volume and surface of the original bubble/droplet distribution (Siemens 2018, Zogg 1987). As it is difficult to determine d_{32} for an unknown flow, it is usually obtained from experiments. Additionally, the condensation process complicates the exact determination of d_{32} , since the amount of condensate increases with continuous flow and changes its Sauter diameter from inlet to outlet. One way of determining d_{32} is by assuming a logarithmic distribution of bubble/droplet diameters (Gnotke 2005). Hence, a simple logarithmic mean is calculated (see eq. 8). In conjunction with the maximum and minimum geometric boundaries, the ILS is given by:

$$l_{\alpha\beta} = d_{32} = \frac{(d_{max} - d_{min})}{\ln(d_{max}/d_{min})} \quad (8)$$

Here, $l_{\alpha\beta}$ is the ILS of the respective regime. For spherical bubbles, the maximum possible diameter $d_{max} = 4.9$ mm is limited by the width of the plate channel. The minimum bubble diameter $d_{min} = 0.049$ mm is assumed to 1 % of d_{max} . Then the interaction length scale for the bubbly regime is calculated according to equation 8. The procedure for calculating the droplet Sauter mean diameter is different from the bubble region. Because the Reynolds number of the droplet regime is much higher than the Reynolds number in the bubbly regime, the droplet Sauter diameter is assumed as one-hundredth of the bubble Sauter mean diameter.

RESULTS AND DISCUSSION

The PHE being examined in this paper consists of 80 stacked plates. All geometric parameters of each plate are listed in Table 3. With these parameters, the calculation according to Martin (Martin 1996) results in a hydraulic diameter of $d_h = 4.3$ mm for each plate channel.

For simplifications purposes, all roundings are removed from the 3D model. Furthermore, instead of a sinusoidal corrugated pattern, a trapezoidal one is modeled. All embossings in the distributor segments are removed and a uniform channel is assumed.

Table 3: plate parameters

Symbol	Meaning	Value
β	Chevron angle	60°
s	Plate thickness	0.4mm
Λ	Corrugated pitch	10mm
b	Amplitude	1.225mm
W_p	Plate width	192mm
L_p	Plate length	519mm
D_p	Port diameter	49mm

Figure 4 shows the cross-section of the modeled plate channel. The figure shows the locations in the cross-section where the liquid phase accumulates. Liquid collects due to gravity near the outlet at the lower end. It can also be seen in the figure that there are also small accumulations of the liquid phase in the main heat transfer area due to the enhanced condensation in that area.

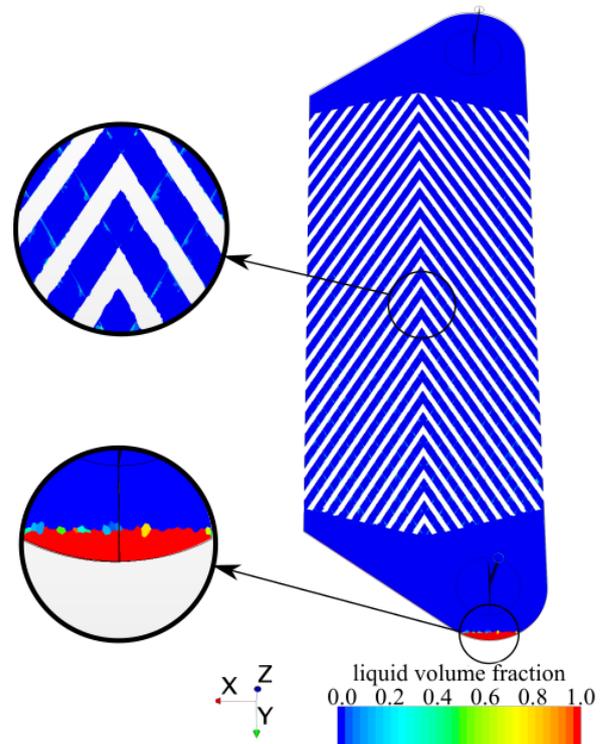


Figure 4: liquid volume fraction

The two characteristic parameters of every heat exchanger are the heat transfer and the pressure drop. Table 4 shows the comparison of the results of the simulation of the plate heat exchanger with the data obtained from Alfa Laval (Alfa Laval 2011). In the presented test case, the ammonia flows into the PHE at a temperature of $T_{f,in} = 75$ °C, condenses completely at $T_{f,sat} = 30$ °C and leaves the heat exchanger as a liquid at a temperature of $T_{f,out} = 28.1$ °C. At a system pressure of 1.16 MPa and a mass flow of 0.003 kg/s, this corresponds to a heat flow of $\dot{Q}_{real} = 3853$ W per plate channel (Alfa Laval 2011). In the simulation, the ammonia does not condense completely. This results in a

reduced heat flow of $\dot{Q}_{sim} = 1589 W$, see Table 4. This corresponds to a deviation of -58.7 % for the heat flow. The experimental data provide a pressure drop from inlet to outlet of $\Delta p = 779 Pa$ (Alfa Laval 2011). However, the result of the simulation is $\Delta p = 519 Pa$. Thus, the pressure drop is also below the experimental data with a deviation of -33.4 %.

Table 4: comparison of simulation and reality

Symbol	Simulation	Data from (Alfa Laval 2011)	Deviation
Δp	519Pa	779Pa	-33.4%
\dot{Q}	1589W	3853W	-58.7%

After the elimination of the systematic mistakes, the deviation of the mathematical model for the reduced heat flow is -53 % (see also mesh generation). An additional deviation of about 10 % can be expected due to the selected mesh. The deviation of -58.7 % in reduced heat flow between simulation and reality is therefore acceptable. The reason for this enormous deviation is the numerical model and the selected parameters. The correlations and benchmarks that are used for describing the Nusselt numbers of two-phase flow are fraught with uncertainty. As shown for example by Park and Hrnjak (Park and Hrnjak 2008), correlations are developed which can only predict numbers for the two-phase heat transfer within an accuracy of about ± 20 %. Such correlations are simply unsuitable for engineering applications. As already mentioned, the flow describing conservation equations are also not correct. The sum of the coarse mesh, inaccurate Nusselt numbers and imprecise conservation equations lead to a deviation of -58.7 % for the heat flow.

The pressure drop and the amount of condensate that forms are related. If the amount of condensate increases, the pressure drop increases as well. While in (Alfa Laval 2011) the entire gas-phase condenses, in the simulation only 35 % condenses. This reduced condensate mass flow influences the pressure drop. A linear drop in pressure can be observed in the main heat transfer area of the PHE. When comparing the deviation of the simulation of the PHE with the mesh independence study, which was carried out in advance, it is of note that the pressure drop of the PHE is 16.5 % (-33.4 % for PHE, -49.9 % for tube) above expectations, but still below the value given in the PHE specification (Alfa Laval 2011). Yet, there is one modeled parameter that has not been considered in the discussion. This parameter is the interaction length scale. The ILS has a significant impact on both pressure drop and heat transfer in the PHE. As mentioned above, ILS is related to the Reynolds number and thus to the pressure drop. The mass transfer is also influenced by the choice of the ILS.

In this study, only one constant is assumed to describe the ILS. However, this single constant is not enough to describe the entire flow in the PHE correctly. For that reason, a new model needs to be implemented to capture the local disparities.

Outlook

A numerical model for the simulation of condensation in the plate heat exchanger was created using the Eulerian Multiphase flow approach with the AIAD model for morphology detection. Necessary parameters, such as the Nusselt numbers and the ILS, were determined as a result of literature.

The results of the simulation show large deviations in the prediction of pressure drop (-33.4 %) and heat flow (-58.7 %). The ILS has a large impact on the pressure drop and the phase change. Therefore, the approach to determine the ILS should be redesigned and replaced by a model that detects local disparities.

The development of a new model for ILS is reasonable since condensation is an elementary component in highly efficient heat exchangers. For the design of highly efficient heat exchangers, a good prediction of the CFD is necessary to improve/optimize the design.

CFD can be a reliable alternative for the design of new heat exchangers once the mathematical simulation model has been validated. If this approach is applied in the design phase of new heat exchangers, another calculation model for the ILS should be used to advance its prediction accuracy as described in section "Related works". The CFD simulation is a more general approach in contrast to literature empirically models and can be assigned of different geometries. But the biggest advantage of CFDs is that local detailed information can be extracted. This allows conclusions to be drawn about fouling effects, dead volume, hot spots, phase transition position within the process apparatus. Therefore, CFD is the suitable method for a comprehensive analysis of the flow.

REFERENCES

- Alfa Laval. 2011. "AlfaFusion Plattenwärmeübertrager, Technische Spezifikationen" datasheet DEGLCPN-887(e)_2e Pos. 4.
- Bhramara, P., Rao, V. D., Sharma, K. V. and Reddy, T. K. K. 2009. "CFD analysis of two phase flow in a horizontal pipe—prediction of pressure drop." *Momentum*, 10, 476-482.
- Cartaxo, S. J. and Fernandes, F. A. 2011. "Counterflow logarithmic mean temperature difference is actually the upper bound: A demonstration." *Applied Thermal Engineering*, 31(6-7), 1172-1175.
- Castellano, S., Sheibat-Othman, N., Marchisio, D., Buffo, A. and Charton, S. 2018. "Description of droplet coalescence and breakup in emulsions through a homogeneous population balance model". *Chemical Engineering Journal*, 354, 1197-1207.
- Djordjevic, E. and Kabelac, S. 2008. "Flow boiling of R134a and ammonia in a plate heat exchanger." *International Journal of Heat and Mass Transfer*, 51(25-26), 6235-6242.
- Ferziger, J. H., and Peric, M. 2008. "Numerische Strömungsmechanik." Springer-Verlag.
- Focke, W. W., Zachariades, J. and Olivier, I. 1985. "The effect of the corrugation inclination angle on the thermohydraulic performance of plate heat exchangers" *International Journal of Heat and Mass Transfer*, 28, 1469-1479.
- García-Cascales, J. R., Vera-García, F., Corberán-Salvador, J. M. and González-Maciá, J. 2007. "Assessment of boiling and condensation heat transfer correlations in the modelling

- of plate heat exchangers." *International Journal of Refrigeration*, 30(6), 1029-1041.
- Gnotke, O. 2005. "Experimentelle und theoretische Untersuchungen zur Bestimmung von veränderlichen Blasengrößen und Blasengrößenverteilungen in turbulenten Gas-Flüssigkeits-Strömungen" (Doctoral dissertation, Technical University of Darmstadt).
- Hell, F. 1992. "Wärmeübertrager". Oldenbourg Industrieverlag, 2.
- Höhne, T. and Lucas, D. 2011. "Numerical simulations of counter-current two-phase flow experiments in a PWR hot leg model using an interfacial area density model." *International Journal of Heat and Fluid Flow*, 32(5), 1047-1056.
- Höhne, T. and Vallée, C. 2010. "Experiments and numerical simulations of horizontal two-phase flow regimes using an interfacial area density model." *The Journal of Computational Multiphase Flows*, 2(3), 131-143.
- Huang, J., Sheer, T. J. and Bailey-McEwan, M. 2012. "Heat transfer and pressure drop in plate heat exchanger refrigerant evaporators." *International Journal of Refrigeration*, 35(2), 325-335.
- Hyvärinen, A.-P., Raatikainen, T., Laaksonen, A., Viisanen, Y. and Lihavainen, H. 2005. "Surface tension and densities of $H_2SO_4+NH_3$ +water solutions" *Geophysical Research Letters*, 32, L16806.
- Ishii, M. and Hibiki, T. 2010. "Thermo-fluid dynamics of two-phase flow." Springer Science & Business Media.
- Ishii, M. and Kim, S. 2001. "Micro four-sensor probe measurement of interfacial area transport for bubbly flow in round pipes". *Nuclear Engineering and Design*, 205(1-2), 123-131.
- Khan, T. S., Khan, M. S., Chyu, M. C. and Ayub, Z. H. 2012. "Experimental investigation of evaporation heat transfer and pressure drop of ammonia in a 60 chevron plate heat exchanger." *International Journal of Refrigeration*, 35(2), 336-348.
- Kim, S. M. and Mudawar, I. 2013. "Universal approach to predicting heat transfer coefficient for condensing mini/micro-channel flow." *International Journal of Heat and Mass Transfer*, 56(1-2), 238-250.
- Kuo, W. S., Lie, Y. M., Hsieh, Y. Y. and Lin, T. F. 2005. "Condensation heat transfer and pressure drop of refrigerant R-410A flow in a vertical plate heat exchanger." *International Journal of Heat and Mass Transfer*, 48(25-26), 5205-5220.
- Lance, M. and Bataille, J. 1991. "Turbulence in the liquid phase of a uniform bubbly air-water flow." *Journal of Fluid Mechanics*, 222, 95-118.
- Legendre, D. and Magnaudet, J. 1998. "The lift force on a spherical bubble in a viscous linear shear flow." *Journal of Fluid Mechanics*, 368, 81-126.
- Lloyd, G. and Espanoles, A. 2002. "Best practice guidelines for marine applications of computational fluid dynamics." WS Atkins Consultants and Members of the NSC, MARNET-CFD Thematic Network: London, UK, 84.
- Lo, S., and Zhang, D. 2009. "Modelling of break-up and coalescence in bubbly two-phase flows". *The Journal of Computational Multiphase Flows*, 1(1), 23-38.
- Martin, H. 1996. "A theoretical approach to predict the performance of chevron-type plate heat exchangers." *Chemical Engineering and Processing: Process Intensification*, 35(4), 301-310.
- Méndez, C. G., Nigro, N. and Cardona, A. 2005. "Drag and non-drag force influences in numerical simulations of metallurgical ladles." *Journal of Materials Processing Technology*, 160(3), 296-305.
- Nowitzki, M. 2020. "Development and Validation of a Gas-Liquid Two-Phase Model for Industrial Computational Fluid Dynamics Application" Doctoral dissertation, Technical University Cottbus-Senftenberg.
- Numrich, R. and Müller, J. 2013. "Filmkondensation reiner Dämpfe" *VDI-Wärmeatlas*, 11, 1011-1027.
- Parekh, J. and Rzehak, R. 2018. "Euler-Euler multiphase CFD-simulation with full Reynolds stress model and anisotropic bubble-induced turbulence." *International Journal of Multiphase Flow*, 99, 231-245.
- Park, C. Y. and Hrnjak, P. 2008. "NH₃ in-tube condensation heat transfer and pressure drop in a smooth tube" *International Journal of Refrigeration*, 31(4), 643-651.
- Porombka, P. and Höhne, T. 2015. "Drag and turbulence modelling for free surface flows within the two-fluid Euler-Euler framework." *Chemical Engineering Science*, 134, 348-359.
- Ramkrishna, D. 2000. "Population balances: Theory and applications to particulate systems in engineering". Elsevier.
- Schiller, L. and Naumann, A. 1933. "Über die grundlegenden Berechnungen bei der Schwerkraftaufbereitung" *Zeitschrift Verein Deutscher Ingenieure*, 77, 318-32.
- Siemens. 2018. "STAR-CCM+ Dokumentation: Version 13.02".
- Soler, J. M. 1996. "Cluster diffusion by evaporation-condensation." *Physical Review B*, 53(16), R10540.
- Stephan, P., Kabelac, S., Kind, M., Mewes, D., Schaber, K., & Wetzel, T. 2019. "VDI-Wärmeatlas." Springer-Verlag.
- Štrubelj, L. and Tiselj, I. 2011. "Two-fluid model with interface sharpening." *International Journal for Numerical Methods in Engineering*, 85(5), 575-590.
- Thonon, B., & Bontemps, A. 2002. "Condensation of pure and mixture of hydrocarbons in a compact heat exchanger: experiments and modelling." *Heat Transfer Engineering*, 23(6), 3-17.
- Tomiya, A., Tamai, H., Zun, I. and Hosokawa, S. 2002. "Transverse migration of single bubbles in simple shear flows" *Chemical Engineering Science*, 57, 1849-1858.
- van de Sand R., Falk C., Corasanti S. and Reiff-Stephan, J. 2019. "A data-driven fault diagnosis approach towards oil retention in vapour compression refrigeration systems," *International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)*.
- Wang, L., Manglik, R. M. and Sundén, B. 2007. "Plate Heat Exchangers: Design, Applications and Performance." *International series on developments in heat transfer*.
- Wu, Q., Kim, S., Ishii, M. and Beus, S. G. 1998. "One-group interfacial area transport in vertical bubbly flow". *International Journal of Heat and Mass Transfer*, 41(8-9), 1103-1112.
- Yan, Y. Y., Lio, H. C. and Lin, T. F. 1999. "Condensation heat transfer and pressure drop of refrigerant R-134a in a plate heat exchanger." *International Journal of Heat and Mass Transfer*, 42(6), 993-1006.
- Yao, W. and Morel, C. 2004. "Volumetric interfacial area prediction in upward bubbly two-phase flow". *International Journal of Heat and Mass Transfer*, 47(2), 307-328.
- Zhang, J., Kærn, M. R., Ommen, T., Elmegaard, B. and Haglind, F. 2019. "Condensation heat transfer and pressure drop characteristics of R134a, R1234ze (E), R245fa and R1233zd (E) in a plate heat exchanger" *International Journal of Heat and Mass Transfer*, 128, 136-149.
- Zogg, M. 1987. "Einführung in die Mechanische Verfahrenstechnik". B. G. Teubner Stuttgart.

DISCRETE EVENT SIMULATION – MODEL OF A CALL CENTER IN SIMUL8 SOFTWARE

Martina Kuncová, Jan Fábry and Anna Marie Klímová
Department of Econometrics
University of Economics in Prague
W.Churchill Sq. 4, 13067 Prague 3, Czech Republic
E-mail: martina.kuncova@vse.cz; jan.fabry@vse.cz; ania.e3@seznam.cz

KEYWORDS

Discrete event simulation, call center, SIMUL8.

ABSTRACT

Simulation modelling is usually used when mathematical models and calculations are hard to apply on a system with stochastic behavior. This contribution deals with the application of simulation program SIMUL8 to the analysis of a call center. The main aim is to create a model based on the real data and afterwards to conduct two experiments to see the impact of changes on the functionality of the call center and on the number of customers lost. More suitable software can be used for the analysis of call centers. Because of our good experience we decided to use SIMUL8.

INTRODUCTION

In case of real systems that cannot be studied and analyzed using standard analytical tools, computer simulation is applied. Due to probabilistic and dynamic aspects of processes, a realization of experiments with the simulation model helps the decision-maker implement the improvement solution. In the paper, the call center system of the real company is analyzed. Mehrotra and Fama (2003) analyzed operations to increase demand for the call center service. They simulated the different scenarios with the right levels of cross-training to meet service level goals with the current staffing levels. Van Buuren et al. (2015) presented a detailed discrete event simulation model for call centers of emergency medical services. Their model includes two classes of centralists: call takers and dispatchers. The model discriminates between multiple types of applicants which differ in priorities. The model was made for general emergency medical services call centers, but it can also be used for other applications such as firefighter call centers. Mathew and Nambiar (2013) offer a straightforward tutorial on modelling call centers using discrete event simulation with MS Excel-based input and reporting. Ibrahim et al. (2016) provide a literature survey of modeling and forecasting call center arrivals. Call centers with uncertain non-stationary arrival rate and flexibility are analyzed in detail by Liao et al. (2012). Munoz and Brutus (2013)

deal with the question of trade-offs in a call center. Kuncová and Wasserbauer (2007) created a simulation model for the optimization of the number of helpdesk's operators, and for the optimization of the operator's working time.

In the following text we present the simulation model of a call center in the real company. As many authors show (e.g. Banks 1998; Montecvecchi et al. 2007), simulation experiments representing different scenarios have significant meaning for a company's decision making.

In the paper, we offer two experiments aimed at changes in the number of customers (acceding to a new advertisement) and in the number of operators (based on the closure of one call center building). The main aim is to analyze the impact of these changes on the percentage of lost customers and on the operators' utilization. The information obtained from call center staff says that the current percentage of lost customers is around 1-2% and the company accepts a maximum of 5% of unserved customers.

SIMUL8

SIMUL8 is a software package designed for Discrete Event Simulation or Process Simulation and developed by the American firm SIMUL8 Corporation (www.simul8.com). The software started to be used in 1994, and every year a new release has come into being with new functions and improved functionality. A visual 2D model of an analyzed system can be created by placing objects directly on the screen. SIMUL8 belongs to the simulation software systems that are widely used especially in industry (Greasley 2003). This software is suitable for discrete event simulation. Model of a call center as a set of activities of calling and answering can be taken as a typical discrete event simulation model. SIMUL8 uses 2D animation only to visualize the processes, but for the given problem, this view is sufficient. It is similar to SIMPROCESS, which is also aimed at the discrete event simulation (Dlouhý et al. 2011), but we decided to use SIMUL8 because of the easier way of queue modelling.

Pisaniello et al. (2018) used SIMUL8 to develop the simulation model of the call center in the children's hospital. On the case study, they demonstrate the meaning of the application of validation and verification

techniques as the most critical aspects of the simulation modelling process.

Similarly, Vermeulen (2017) shows the exiting application of SIMUL8 to call center staffing and performance in the video presentation.

Fousek et al. (2017) tried to find out the total time needed for the increased production given by the new contract and also to show the bottleneck of the production system.

SIMUL8 main components

SIMUL8 operates with 6 main parts out of which the model can be developed: Work Item, Work Entry Point, Storage Bin, Work Center, Work Exit Point, Resource (Concannon et al. 2007).

Work Item: dynamic object(s) (customers, products, documents or other entities) that move through the processes and use various resources. Their main properties that can be defined are labels (attributes), an image of the item (showed during the animation of the simulation on the screen) and advanced properties (multiple Work Item Types).

Work Entry Point: an object that generates Work Items into the simulation model according to the settings (distribution of the inter-arrival times). Other properties that can be used in this object are batching of the Work Items, changing of the Work Items! Label or setting of the following discipline (Routing Out).

Storage Bin: queues or buffers where the Work Items wait before the next processes. It is possible to define the capacity of the queue or the shelf life as time units for the expiration.

Work Center: main object serving for the activity description with the definition of the time length (various probabilistic distributions), resources used during the activity, changing the attributes of entities (Label actions) or setting the rules for the previous or following movement of entities (Routing In / Out).

Work Exit Point: an object that describes the end of the modeled system in which all the Work Items finish its movement through the model.

Resource: objects that serve for modelling of limited capacities of the workers, material or means of production that are used during the activities.

All objects (except resources) are linked together by connectors that define the sequence of the activities and also the direction of movement of Work Items.

After the system is modelled, the simulation run follows. The animation shows the flow of items through the system and for that reason the suitability of the model can be easily assessed. When the structure of the model is verified, several trials can be run and then the

performance of the system can be analyzed statistically. Values of interest may be the average waiting times or utilization of Work Centers and Resources (Shalliker and Ricketts 2002). SIMUL8 can be used for various kinds of simulation models (Concannon et al. 2007). The case studies can also be seen on the website www.simul8.com.

Our experience shows that SIMUL8 is easy to learn when only the main components are used (without the necessity to use Visual Logic with different programming functions). It can serve not only for the modelling of different services (Dlouhý et al. 2011) but also for the simulation of various production processes (Ficová and Kuncová 2013; Fousek et al. 2017).

PROBLEM DESCRIPTION

The main aim of this article is to create a simulation model of a call center of one unnamed telecommunications company and to test the influence of the number of customers' changes (increase) on the call center functionality. The type of the system can be described as an open queueing system with multiple parallel service lines and with an unlimited number of requests. In this case, however, classical mathematical models of queueing theory cannot be used to determine the average queue length nor the optimal number of call center operators since the system contains more variables than it is usual for the mathematical model. Therefore, it is preferable to use a simulation model, and discrete event simulation is a suitable solution to the problem.

In general, two kinds of requirements can enter a given call center. The first one is a customer (households, companies with a contract, companies without contract) requesting information; the second one is the called customer who has been chosen by the company itself for the questionnaire survey. In the case of customers calling to the call center, the inter-arrival times have a stochastic distribution (statistical analysis of the data revealed that the distribution should be exponential) with different parameters for each day period, whereas for customers who are called by the company the distribution can be taken as normal for the whole week. These customers are pre-selected for the given week by the company, and afterwards, the free operators could call them to ask for an opinion on the product or fill out a questionnaire.

Data for the simulation model was obtained on the basis of monthly traffic monitoring and discussions with call center staff. A more detailed description of data collection can be found in Klímová (2019).

CALL CENTER DATA

The objective of the simulation is to model the real call center with a defined number of operators for each day

period and afterwards to describe the impact of selected changes on the operation of this call center (also called “Infoline”). A call center is usually a group of employees that obtain the requirements of the customers and try to solve them. The requirements are reported by telephone/mobile phone. In the selected call center 3 groups of customers usually call to ask for any advice or information. The first type of customers can be described as individuals or households and it is unimportant whether they have a contract with the company or not. The other two types of customers are companies: both companies belonging to customers, i.e. with a valid contract, and other companies that do not yet have a contract. The call center has 3 buildings (called Infoline 1, Infoline 2, Infoline 3) where the workers/operators answer the customers’ questions or call to customers.

Table 1 summarizes necessary information about the intervals between customer calls in each time window. Distributions’ fittings were tested in Crystal Ball to find the estimations of parameters (detailed data analysis is presented in Klímová, 2019). For the inter-arrival times’ generation, the exponential distribution is used. Three different types of customers contacting the call center can be identified: individual customers or households (we call them B2C), companies that already have a contract (B2B) and companies without any contract (other). Data analysis of the call lengths showed that even for this case, the exponential distribution should be used to generate the time length of each call. The number of the customers called by the call center can be estimated as a random variable from the normal distribution with the mean value equal to 200 and standard deviation equal to 20. The average length of the call can be estimated by exponential distribution with the mean value equal to 416 seconds (nearly 7 minutes).

Table 1: Mean value of exponential distribution of inter-arrival times in seconds

Time of the day	B2C	B2B	Other	Call length
8:00-12:00	5.9	6.3	5.9	338
12:00-16:00	6.0	6.4	6.1	340
16:00-21:00	14.6	16.4	14.8	286
21:00-7:00	155	193	188	181

Table 2: Number of operators on infolines during a working day

Time of the day	Infoline 1	Infoline 2	Infoline 3
7:30-9:00	90	74	28
9:00-12:00	120	88	38
12:00-16:30	140	88	45
16:30-20:00	70	32	22

The analyzed call center has more buildings where all the operators work. For the purposes of the simulation model, four daily time windows (see Table 2) with different numbers of employees on three workplaces (infolines) were identified, followed by the emergency team, which is used to strengthen the standard daytime and nighttime capacity of the call center (see Table 3). The most numerous is the Infoline 1 team, where the temporary workers work, which is also reflected in staff turnover and work efficiency.

Table 3: Number of operators on emergency line during a working day

Time of the day	Emergency line
1:30-7:30	5
16:30-19:30	6
19:30-1:30	10

MODEL IN SIMUL8

The simulation model was developed in SIMUL8 software. The main entity (Work Item Type) that moves within the system is the customer, with the Type label indicating the customer (Type = 1 for the B2B customer, Type = 2 for the B2C customer, Type = 3 for other customers). Based on the data in Table 1, the new named time dependent distribution (see Figure 1) consisting of inter-arrival times distributions for all time windows had to be created for each type of the customer. Figure 2 shows the settings for the “other” customers and Figure 3 the settings of one new named distribution as a part of the time dependent distribution. Similar types of named distributions were prepared for inter-arrival times of customers B2B and B2C.

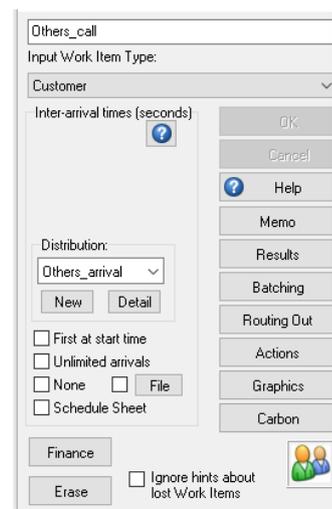


Figure 1: New distribution usage for the inter-arrival times of “other customers”

Aside from the customer calls, there should also be calls to customers done by an operator. According to the

information of the call center’s workers, the number of selected customers should be around 200 per week – it can be estimated by a normal distribution with the mean value equal to 200 and standard deviation equal to 20. For the model we defined the generation of these called customers once a week in the beginning – that is why the interarrival time was set as a fixed value equal to the length of the simulation run (1 week = 5 working days = 120 hours = 432000 seconds) – see Figure 4. The call is usually made when the operator is not busy. He/she then chooses a pre-selected customer and conducts an interview with him/her. Based on real data, it was estimated that the length of this interview could be approximated by an exponential distribution with a mean value of 416 seconds.

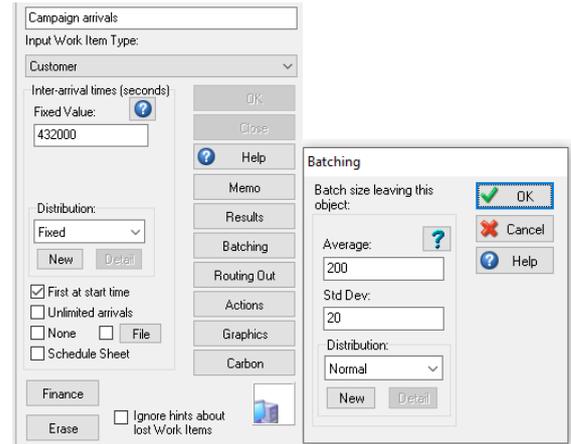


Figure 4: Settings of the generation of pre-selected customers for a call

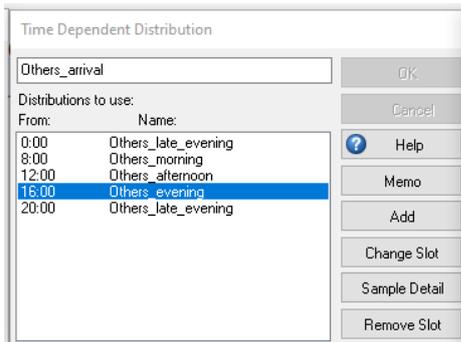


Figure 2: Creation of new time dependent distribution

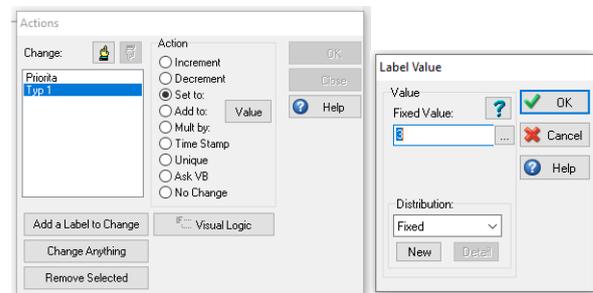


Figure 5: Label settings for each type of customer

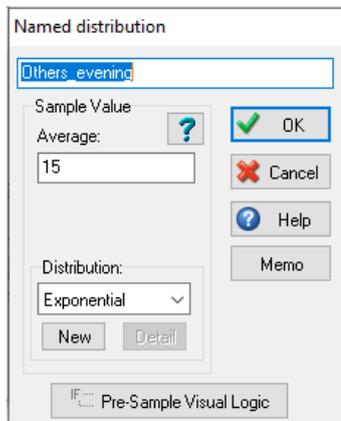


Figure 3: Creation of new named distribution

As the generation of customers is separated into 3 different inputs and so we do not have to distinguish them with different entity names, but it would be suitable for the results’ description. In this situation, the program prefers to create labels in which this distinction can be made. Label Type1 was used to monitor the number of customers served by type (see Figure 5). Each customer has been assigned a unique number identifying the customer group (1 for B2B, 2 for B2C and 3 for other).

After receiving calls, the customer is redirected to a free operator, first to Infoline1, then to Infoline2 or Infoline3, at night to Emergency calls. A minimum of 8 seconds elapses from the call to the connection. If no operator is free, the customer waits on the line. The customer’s patience while waiting may vary, but it has been inferred from the experience of the company’s employees that after about 200 seconds of waiting, the customer hangs up. Therefore, the queue parameters were set as follows: a minimum wait time of 8 seconds and a maximum patience of 200 seconds (see Figure 6). After 200-second waiting time expiration, the customers renege (see Figure 7).

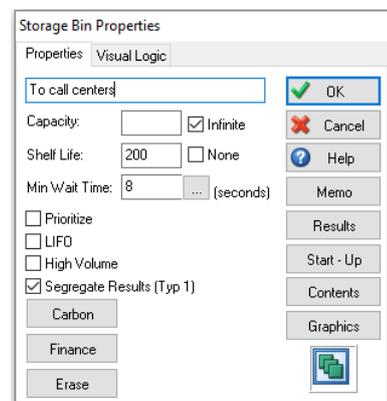


Figure 6: Queue to operators’ settings

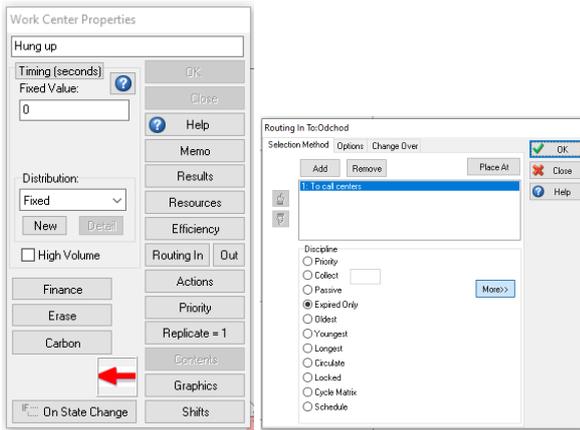


Figure 7: Early exits settings

Based on the data from Table 2 and Table 3 the shifts with a different number of operators were created for each Infoline (see Figure 8).

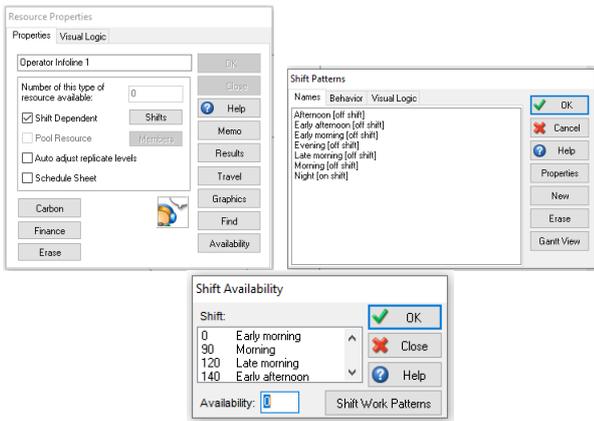


Figure 8: Shifts settings (Infoline1)

The setting of the number of operators is different for the campaign, i.e. for calls to customers by the call center. Here all operators from Infoline1, Infoline2 and Infoline3 could be used if they are currently free. In SIMUL8 a pool resource is selected (Figure 9). Work centers that require a pool resource can be given any resource from the list of members.

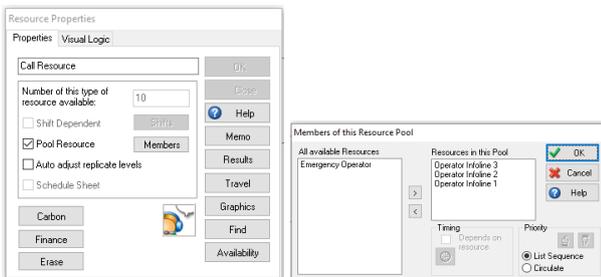


Figure 9: Pool resource settings

The whole model with 3 customer entrances, 1 calls to customer, 3 infolines, 1 emergency calls center, 1

campaign center, 1 exit for served customers and 1 exit for those who hung up early (see Figure 10).

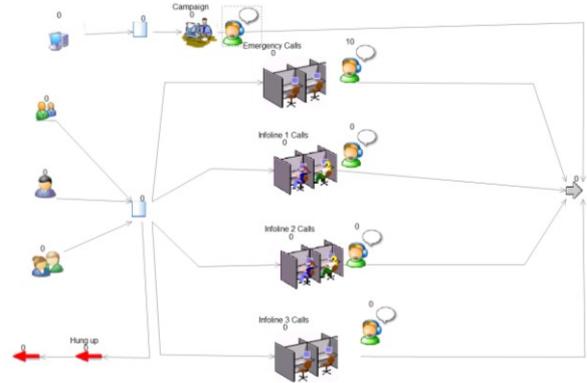


Figure 10: Final model in SIMUL8

Finally, the run settings are necessary to make. The model was tested on the simulation time of 5 days run, 24 hours per day. First, the warm-up period was set to 1 day but the results showed that it is not necessary to set a warm-up as it has no influence on them.

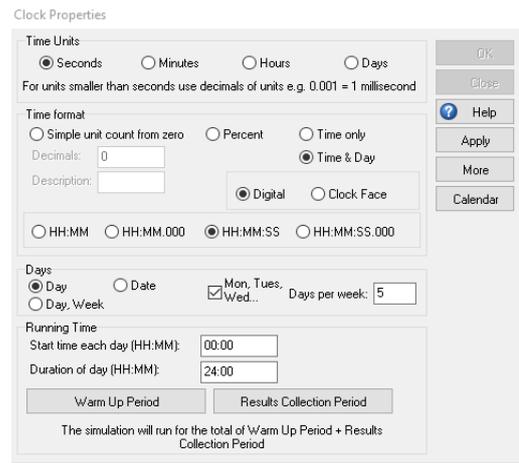


Figure 11: Simulation model run settings

RESULTS

After 1 run (5 days experiment) more than 87000 of customers were served, and only 1187 customers (1.34%) were lost (see Figure 12). The results (see Table 4, Table 5) corresponds with information and data obtained from call center workers (expected hang ups were about 1-2%).

The operators' usage for all infolines is high enough, but it corresponds with the information we have from the call center workers. Every available operator is used every day (Table 5) and its number corresponds to fluctuations in incoming calls. Figures 13, 14 and 15 illustrate the number of busy operators during 5 days on Infoline1, Infoline2 and on Emergency calls.

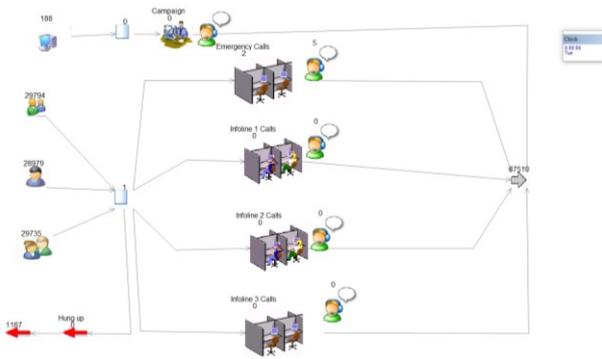


Figure 12: Five days simulation run results

Table 4: Customers generated and lost

Customer	No. of generated	lost (%)
B2B	28979	397 (1.37)
B2C	29735	427 (1.44)
other	29794	363 (1.22)
Pre-selected calls	188	0

Table 5: Operators' usage

Operator	% usage	Avg.no. of used operators	Max. no. of used
Infoline1	82	49.1	140
Infoline2	85	32.2	88
Infoline3	87	16.6	45
Emergency	61	2.8	10

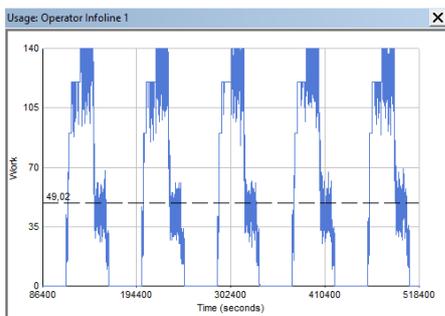


Figure 13: Infoline1 operators' usage

The maximum number of waiting customers was 127 (see Figure 16), but on average only 7.5 customers were waiting with an average waiting time of 36.7 seconds. About 64% of waiting calls were connected till 10 seconds (see Figure 17).

The maximum was higher on the first observed day (see Figure 16) because some operators were used for campaign and calls to selected customers. A campaign with 188 called customers with a maximum of 20 operators was handled during the first day of the simulation. In this situation, the campaign calls could be spread out over multiple days, but the impact on

operators' usage is small, so there is no problem managing the campaign on any day.

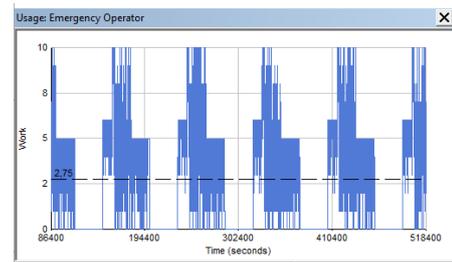


Figure 14: Emergency operators' usage

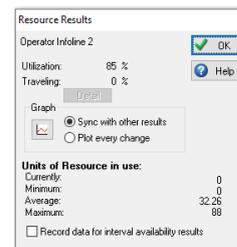


Figure 15: Infoline2 operators' usage

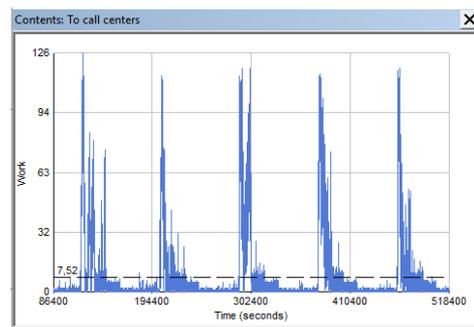


Figure 16: Number of customers in a queue to operators

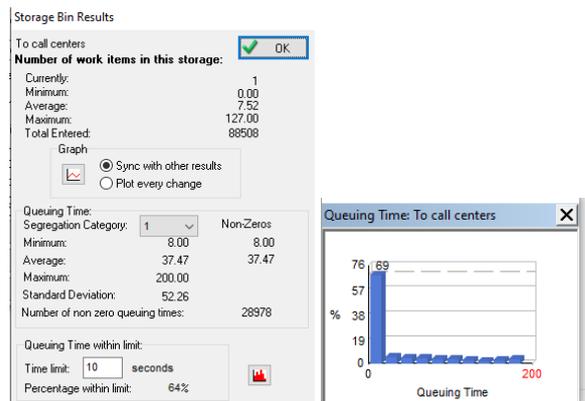


Figure 17: Queue to operators results

EXPERIMENTS WITH MODEL

Since the simulation model was validated and corresponded to reality, it could be used for experiments. There were two types of changes. The first experiment was to test how the percentage of lost

customers will be affected by the introduction of new advertising spots on a new product. The second experiment concerns the analysis of the call center operation during the reconstruction of one of the buildings, i.e. of the number of the operators' decrease.

The first experiment assumes adding new calling customers based on the advertisement action. For this type of arrivals, a new time dependent distribution was created with seven time slots and two probability distributions (exponential) related to the scheduled launch of the ads (see Table 6).

Table 6: New customers – inter-arrival times distributions

Day time	Distribution (avg. time in seconds)
8:00-10:30	Exp(45)
10:30-11:00	Exp(8)
11:00-14:00	Exp(45)
14:00-14:30	Exp(8)
14:30-20:00	Exp(45)
20:00-20:30	Exp(8)
20.30-21:00	Exp(45)

In the second experiment, we tested the closure of one call center building. It was a reconstruction of Infoline3. Operators cannot move to other buildings because these info lines are all in relatively remote locations. As a result, there is no activity at Infoline 3 throughout the week.

In both experiments, 10 trials (10 weeks) were tested. Average results of all trials and both experiments are summarized in Table 7. The usage of operators in both experiments increased as expected (except of Emergency where no additional calls were set). The number of waiting customers to be served has also increased and so has the average waiting time. The percentage of lost customers in the first experiment was about 3.5% which can be accepted by the call center, but in the second experiment, it is 11.2% which is very high.

Table 7: Results of experiments

Simul. object	Experiment 1	Experiment 2
Infoline1 usage	85%	89%
Infoline2 usage	88%	90%
Infoline3 usage	88%	0%
Emergency usage	64%	64%
Max. queue length	148	134
Avg. waiting time (seconds)	73.7	136.1
Lost customers	3516	9894
% of lost customers	3.5%	11.2%

The call center could be successful in the case of a promotional event with minimal impact on operators' usage and with acceptable impact on the percentage of the customers lost. In the case of short-term closure of Infoline3, the impact on lost customers would be more significant and unacceptable for the company. It is, therefore, necessary to consider not to close one building as a whole and attempt reconstruction while maintaining partial operation.

CONCLUSION

The aim of the contribution was to demonstrate the applicability of SIMUL8 on the call center modelling. The model was based on available information given by employees of the call center. The simulation model should have shown the impact of an increasing number of customers and a decreasing number of operators. First, the impact of a promotional event was tested to monitor the utilization of operators. Second, a short-term closure of one infoline was examined to analyze the number of lost customers. While the first strategy seems to be suitable for the human resource management, the second one is not quite acceptable because of too high percentage of the customers lost. Although the situation was slightly simplified, the model corresponds with the real situation and proves a possibility to satisfy increased demand (number of calls) and problems when one part of the call center is closed. We verified that for this type of the real system analysis, a simulation model is the suitable tool to indicate the impact of planned changes in the system processes.

ACKNOWLEDGEMENTS

This work was supported by the grant No. F4/66/2019 of the Faculty of Informatics and Statistics, University of Economics, Prague.

REFERENCES

- Banks, J., 1998. *Handbook of Simulation*. USA, John Wiley & Sons
- Concannon, K. et al. 2007. *Simulation Modeling with SIMUL8*. Visual Thinking International, Canada.
- Dlouhý, M., Fábry, J., Kuncová, M. and T. Hladík. 2011. *Business Process Simulation* (in Czech). Computer Press.
- Ficová, P. and M. Kuncová. 2013. „Looking for the equilibrium of the shields production system via simulation model“. In *Modeling and Applied Simulation 2013*. (Athens, Sept. 25-27). Genova : DIME Università di Genova, 50–56.
- Fousek, J., Kuncová, M. and Fábry, J.: Discrete Event Simulation – Production Model in SIMUL8. In: Proceedings of the 31st European Conference on Modelling and Simulation ECMS 2017, Budapest,

- May 2017. Dudweiler: Digitaldruck Pirrot, pp 229–234. ISBN 978-0-9932440-4-9
- Greasley, A. 2003. *Simulation modelling for business*. Innovative Business Textbooks, Ashgate, London.
- Ibrahim, R., Ye, H., L'Ecuyer, P., and H. Shen. 2016. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*. Vol. 32, pp. 865 – 874.
- Kuncová, M. and P. Wasserbauer. 2007. „Discrete Event Simulation – Helpdesk Model in SIMPROCESS“. In *ECMS 2007* (Prague, June 4-6). Dudweiler : Digitaldruck Pirrot, 105–109.
- Liao, S., Delft, C.V., Koole, G., and Jouini, O. 2012. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum*, Vol. 34, pp. 691–721.
- Mathew, B., and M.K. Nambiar. 2013. A tutorial on modelling call centers using discrete event simulation. In *Proceedings of the 27th European Conference on Modelling and Simulation*, Vol. 4, pp. 315 – 321.
- Mehrotra, V., and J. Fama. 2003. Call center simulation modeling: methods, challenges, and opportunities. In *Proceedings of the 2003 Winter Simulation Conference*, pp. 135-143, USA.
- Montevecchi, J. A. B., et al. 2007. Application of design of experiments on the simulation of a process in an automotive industry. In *WSC'07 Proceedings of the 2007 Winter Simulation Conference*, IEEE Press Piscataway, NJ, USA, 1601-1609.
- Munoz, D., and M.C. Brutus. 2013. Understanding the trade-offs in a call center. In *Proceedings of the 2013 Winter Simulation Conference*, pp. 3992-3993, USA.
- Shalliker, J. and C. Ricketts. 2002. *An Introduction to SIMUL8, Release nine*. School of Mathematics and Statistics, University of Plymouth.
- Simul8.com – SIMUL8 software. [online], [cit. 2020-02-20]. Available: <https://www.simul8.com/>
- van Buuren, M.R., Kommer, G.J., R. van der Mei, and S. Bhulai. 2015. A simulation model for emergency medical services call centers. In *Proceedings of the 2015 Winter Simulation Conference*, pp. 844-855, USA.
- Vermeulen, S. 2017. *Using Simulation Software to Optimize Call Center Staffing and Performance*. https://www.youtube.com/watch?v=U_aQPcClISY, accessed March 31, 2018.

AUTHOR BIOGRAPHIES

MARTINA KUNCOVÁ was born in Prague, Czech Republic. She has got her degree at the University of Economics Prague, at the branch of study Econometrics and Operational Research (1999). In 2009 she has finished her doctoral study at the University of West Bohemia in Pilsen (Economics and Management). Since the year 2000 she has been working at the Department of Econometrics, University of Economics Prague, since 2007 also at the Department of Economic Studies of the

College of Polytechnics Jihlava (since 2012 as a head of the department). She is a member of the Czech Society of Operational Research, she participates in the solving of the grants of the Grant Agency of the Czech Republic, she is the co-author of four books and the author of many scientific papers and contributions at conferences. She is interested in the usage of the operational research, simulation methods and methods of multi-criteria decision-making in reality. Her email address is: martina.kuncova@vse.cz

JAN FÁBRY was born in Kladno, Czech Republic. In 1993 he was graduated in Operational Research at the University of Economics Prague (UEP) and in 2006 he received his Ph.D. in Operational Research and Econometrics at the UEP. In 2015 he successfully completed the habilitation procedure in Econometrics and Operational Research at the UEP. Since 2002 he has been working at the department of econometrics at the UEP, initially as an assistant professor, later (since 2015) as an associate professor. In 2016 he joined SKODA AUTO University (SAU) in Mlada Boleslav as a member of the Department. of Logistics, Quality and Automotive Technology. He participates in projects founded by Grant Agency of the Czech Republic; he is the author or co-author of three books and many papers and contributions at conferences. At SAU, he is a supervisor of the courses of Operations Research and of Computer Simulation in Logistic Processes. At the UEP, he is a supervisor of the Czech courses of Discrete Models Case Studies in Operations Research, and of the English courses of Combinatorial Optimization and Operations Research. He is interested in Vehicle Routing Problems and the application of mathematical methods and simulation in production and logistics. Since 2002 he has been the secretary of the Czech Society for Operations Research. His email address is: jan.fabry@vse.cz

ANNA MARIE KLÍMOVÁ was born in Kralupy nad Vltavou, Czech Republic. She studied at the University of Economics Prague, study programme Quantitative Methods in Economics, study field Econometrics and Operational Research. She has Master's degree Econometrics and Operational Research. Her email address is ania.e3@seznam.cz

BALANCING ASSEMBLY LINE IN THE FOOTWEAR INDUSTRY USING SIMULATION: A CASE STUDY

Virginia Fani, Bianca Bindi, Romeo Bandinelli
Department of Industrial Engineering
University of Florence
Florence, Viale Morgagni 40/44, 50134, ITALY
E-mail: virginia.fani@unifi.it

KEYWORDS

Simulation, Optimization, Balancing, Footwear.

ABSTRACT

Fashion is one of the world's most important industries, driving a significant part of the global economy representing, if it were a country, the seventh-largest GDP in the world in terms of market size. Focusing on the footwear industry, assembly line balancing and sequencing represents one of the more significant challenges fashion companies have to face. This paper presents the results of a simulation-optimization framework implementation in such industry, highlighting the benefits of the use of simulation together with a finite capacity scheduling optimization model. The developed simulation-optimization framework includes the conduction of a scenario analysis that compares production KPIs (in terms of average advance, delay and resource saturation) related to different scenarios that include or not one or more type of stochastic events (i.e. rush orders and/or delays in the expected critical components delivery date).

INTRODUCTION

Assembly line balancing and sequencing represent one of the most important challenges widely discussed in the literature. Even if several classifications and optimization models can be found, as a matter of fact, in non-traditional industries, such as the fashion one, where quality and craftsmanship are the main Critical Success Factors (CSFs), empirical rules and non optimal solution are still applied (d'Avolio et al., 2015b).

According to this, the work aims to present the result of a case study, where a structured framework able to optimize the production planning and scheduling of the production has been applied, with the use of a solver and a simulator.

The paper is organized as follows. In section 2, we present a brief literature review on balancing and sequencing models, with a focus on the fashion industry. The proposed model has been detailed in section 3, and its application in a case study has been shown in section 4. Finally, in the last section we discuss the main conclusions of this work.

PRODUCTION OPTIMIZATION IN THE FOOTWEAR INDUSTRY

Balancing assembly line review

The problem of the line balancing has been discussed several times in the literature. The first published paper of the Assembly Line Balancing (ALB) problem has been the one of Salveson (1955), who suggested a linear programming solution. After that, two articles by Scholl and Becker (2006) and Becker and Scholl (2006) provide the state-of-the-art about exact and heuristic solution procedures for Single Assembly Line Balancing (SALB) problems and a survey on problems and methods in Generalized Assembly Line Balancing (GALB) respectively. SALB problems refer to the assembly lines configured as single-model, while the GALB refers to the ones configured as multi- or mixed-models.

As reported by Pachghare et al. (2014), SALB problems can be divided into the following categories: SALBP-1 Assigning tasks to stations minimizing the number of stations themselves for a given production rate (i.e. fixed cycle time), SALBP-2 Minimizing the cycle time (i.e. maximizing the production rate) for a given number of stations, SALBP-E: Maximizing the line efficiency minimizing, at the same time, the cycle time and the number of stations, considering their interdependency, SALBP-F: Establishing whether or not a feasible line balancing exists for a given combination of number of stations and cycle time, SALBP-3: Maximising the workload smoothness for a given number of stations, SALBP-4: Maximising workload relatedness and SALBP-5: Taking into account multiple objectives.

Among the GALB problems, the leather footwear assembly line can be described as a Mixed Assembly Line Balancing (MALB) problem. MALB problems can be classified in the same way as the previous one, having: MALBP-1: Assigning tasks to stations minimizing the number of stations themselves for a given production rate (i.e. fixed cycle time), MALBP-2: Minimizing the cycle time (i.e. maximizing the production rate) for a given number of stations, MALBP-E: Maximizing the line efficiency minimizing, at the same time, the cycle time and the number of stations, considering their interdependency, MALBP-F: Establishing whether or not a feasible line balancing exists for a given combination of number of stations and cycle time.

According to the literature, any of the GALB problems can be classified according to two dimensions: the

Objective Function (OF) that has to be optimized and the methodology used in order to solve it.

Looking at the first dimension, it is possible also to optimize more than a single OF simultaneously, moving from a single- to a multi-OFs. The OFs that can be taken into account are: Minimization of the number of stations, once fixed the desired output, specifying the cycle time, Minimization of the cycle time, once determined the number of stations, Maximization of the line efficiency, Minimization of the costs, Maximization of the profit, calculated as the difference between the revenues and the costs, Minimization of the deviation between the production time of every different type of item for every single station (i.e. horizontal balancing), Minimization of the deviation of the production time in every single station (i.e. vertical balancing) and Minimization or maximization of different scores related to line bottle necks, efficiency and quality of components.

The methodologies that can be used in order to solve ALB problems are: Linear optimization, Non-linear optimization, Limit value, Heuristic procedure, Analytic value, Simulation, Iterative procedure and Metaheuristic procedure (Battaia, 2013; Becker, 2006; Faccio, 2008; Pachghare, 2014).

Most of the publications in line balancing deal with SALB problems, in which only one type of product is processed in the assembly line (Sewell and Jacobson, 2012). On the other hand, as reported by Sivasankaran and Shahabudeen (2014), most of the papers dealing with MALB problems are academic, and few deals with a real-world environment. Moreover, in order to solve MALB problems on real assembly lines they are usually translated into SALB problems, assuming a single "equivalent item" to be produced having as processing time the average value of the different processing times of the original items.

Regarding the fashion industry, the footwear market segment is the analysed one where the balancing problems are applied and, according to this, where most of the academic contribution for the fashion industry have been found. For example, in their work Guimarães et al. (2014) talk about workers' macro-ergonomic evaluation, while Zangiacomini et al. (2004) dealing with production planning and scheduling for mass customisation. Concerning the design of assembly lines, Chen et al. (2014) use simulation to configure the layouts of stitching lines, Ulutas and Islier (2015) work on the layout problem and Dang and Pham (2016) design an assembly line using simulation. Other works are the ones of Chen et al. (2012), that propose a heuristic approach for scheduling problems in parallel sewing lines, and Quyen et al. (2017), that study the resource constrained assembly line balancing problem in a single model line.

In conclusion, there is an extensive literature about ALB problems, but few articles include applications in the fashion industry (Sadeghi et al., 2018).

Together with the long-term balancing problem, there is also the Mixed-Model Sequencing Problem (MSP) which goal is to define the better sequence of the items (Baybars, 1986; Boysen, 2006; Scholl and Becker, 2006)

in order to maximize the productivity of the assembly line.

MSP regards the optimization of the sequencing of mixed-models according to a specific OF, assuming as already defined the balancing problem and the layout of the conveyors. As assumptions, jobs are considered to be equally divided among the different employees in the stations, the line is considered to move at a fixed speed and the operator is free to start a new job when it has finished the previous one if there are, otherwise he waits for the next job.

Independently from the techniques adopted, objective function of sequencing problems can be classified as: Minimization of processing time, Minimization of processing cost and Minimization of the stocks (e.g. using JIT techniques).

Within the first category (Schneeweiß and Söhner, 1991), some examples include the minimization of the number of additional resources or the minimization of the workers' free time (i.e. the time occurring when an operator is waiting for the next item after he finished to process the previous one).

In the second category, a first objective that can be defined is the total labour cost, defining a regular cost for the operators working inside their station and an extra cost for the operators that work outside their station. Costs can differ depending on the type of jobs (Ziegler, 1990), the station (Thomopoulos, 1967) or the time needed to move outside the stations (Vrat and Virani, 1976).

In the third category, the availability of the material at the station is taken into consideration, in order to quantify and reduce the relative stock per station.

Research on this topic has been increased with the development of new technologies, like the AI techniques, that enabled the possibility to solve complex problems. Nevertheless, few papers deal with the fashion industry (Sivasankaran and Shahabudeen, 2014), whilst most of them are referred to traditional industries like automotive, especially when techniques, like JIT, are applied (Inman, 1991).

The footwear Industry

This footwear industry is one of the most critical of the fashion ones, due to complexity of the product and of the Supply Chain (SC). Most of the production phases are commonly outsourced, especially cutting and stitching but, sometimes, also the final assembly. In fact, subcontracting in footwear is a common practice, due to the high specialization required for the production of each component of shoes. This is one of the reasons why the footwear SC is really fragmented, with a lot of SMEs working along it, each one of them highly specialized on one of the steps described above.

These evidences can be translated in a high complexity to be managed in terms of information and production flows exchanged between different companies.

In this way, as highlighted in the work of Bord and Dulio (2007), investments on ICT solutions in terms of software integration between different SC partners but

also higher performance of the ones used at the single-company level represent a key to gain competitive advantages within the industry, with the main purpose of being able to monitor real-time each production process step in order to guarantee the flexibility needed to quickly respond to the unpredictable changes in demand.

Due to the fact that most of the companies along the footwear SC, and in the fashion SC in general, are SMEs, using an open-source software, as the optimization one integrated into the proposed framework, positively impacts their effectiveness and efficiency in working on the market, as demonstrated by Chituc et al. (2008) in their work.

MODEL DESCRIPTION

Problem description

Suppliers working in the footwear market segment have to develop their production plan according to their strategic objectives, guaranteeing the compliance to the requested delivery date, that is the main KPIs that brand owners use for evaluating their supply base performances.

The main objectives these companies take into account are related to maximize their performances, like more or less every supplier working in the fashion SC, but also the production mix balancing and sequencing, that represent a peculiarity of this market segment that has to be managed.

Footwear manufacturing encompasses major processes such as cutting, stitching and assembly.

Looking at the production process, the labour-intensive production steps followed to realise shoes can be summed up as suggested by Carpanzano and Ballarino (2008).

The pilot regards the assembly line process. Because of the fixed cycle time, the availability of raw materials, first of all leather, is an important variable in managing production plans. It represents one of the main constraints that has to take into account in the production of leather goods.

According to this, as in the leather goods pilot, it is needed to take into account another stochastic events during the simulation runs, that is the analysis of the impact that delays in the expected critical components delivery date have on KPIs value and the combined impact considering rush orders too.

Moreover, if compare with other pilots, modeling companies working in the footwear SC requires to include balancing and sequencing problems in the optimization and simulation models respectively.

This way, the MSP approach, taking into account that some items need major labour time in comparison with other ones, determines the right alternation of different type of products on the line, in order to guarantee the minimization of free time in every station of the assembly line. Then, the distributed simulation is used as empirical technique to validate the result.

Model overview

The simulation-optimization framework utilized within this work has been previously published by the authors in (Fani et al., 2017; Fani et al., 2018).

The model is composed by an optimization tool, developed using an open source solver named OpenSolver (www.opensolver.org) and a commercial simulator named AnyLogic® (www.anylogic.com).

The optimization model has been developed in order to fit the different companies' peculiarities including an OF defined as a combination of weighted parameters chosen by the single company and reflecting its CSFs. In fact, the weighted sum OF reflects the commercial agreement between these companies and the brands: different weights for different sub-objectives. Moreover, a solution implementable with an open source solver and a commercial spreadsheet has been chosen according to their low IT investment capability. Anylogic has been chosen for the possibility to implements different type of simulation approaches and for the easy interface with commercial databases (i.e. Microsoft SQL Server).

CASE STUDY

Optimization model in the footwear industry

Starting from the literature review previously described, the proposed framework, as reported in Section 3.2, has been used in order to resolve MALB problem of type F (i.e. MALBP-F), using the parameters rpbw (the resources balancing-related weight considering the whole resources pool considering the whole production plan) and rbw (the resources balancing-related weight considering the single resource $r \in RR$ considering the whole production plan) in the linear model optimization and including the objective function to minimize the horizontal balancing.

The elementary objectives included in the OF (i.e. the ones having positive weight) have been chosen because they better fit the CSFs of companies working in the footwear industry, and the results of the optimization model implementation have been validated comparing themselves to both the historical data and the production manager's experience.

The pilot has been carried out in a footwear company producing leather shoes for a big Italian Luxury brand, and the working phase analysed has been the conveyor.

Using the MALB problem approach, shoes have been classified into three types: "easy", "medium" and "difficult". In this company the number of products assembled is 8, with a total number of tasks equals to 42, comprising 91 elementary jobs. Every station can do one or more tasks. Taking the data from the balancing schema, the association between tasks and station has been done. The names of the tasks have not been reported because the company has not permitted to publish them, together with the names of both the stations and the items codes. Starting from the production cycle of the 8 different products, every code of the single item has been associated to one of the three categories ("easy", "medium" and "difficult").

Once defined this association, the binary diagram of the tasks done for every type of product in every station has been defined.

Whilst in the leather pilot the processing time of the product mix has been assumed by the experience of the company's production manager, in this case a production time data collection has been done together with the company, in order to find the processing time of every task and, consequently, the cycle time of each product.

The technique utilized to collect the data has been the one named Bedaux (Weatherburn, 2014). Every processing time has been recorded 10 times and then the standard time has been evaluated.

In the end, the standard time has been defined as the registered time plus an extra-time considering: Increases for physiologic factors, Increases for fatigue and Increases for unexpected events.

Once the cycle time of every category of the products has been defined, the optimize plan has been evaluated according to the following input data: Item code, Stock Keeping Unit (SKU) type, Requested quantity, Requested date.

Consider a production launch of 4,890 shoes of the different 8 skus, the optimized assembly line has been evaluated starting from a balancing plan declared by the company's management and, according to this, not included in the optimization model. In fact, the requested quantities for the items xxxxx1-8 included in the production plan received from the brand owner have been previously balanced according to the number of the stations and the binary diagram of the tasks.

Moreover, the constraint of the raw material availability has been previously taken into account. In fact, all the raw materials were available before the first day of production. This way, the constraint has not been included into the OF.

As a result, the balanced production plan has been optimized through the proposed model including only the daily mix of products in terms of "easy", "medium" and "difficult" items and taking into account the delivery date of each order.

On the other hand, the resolution of the sequencing problem has been demanded to the simulation model implementation, in order to evaluate the feasibility of the production plan changing the sequencing rules.

Simulation model in the footwear industry

In order to run the proposed simulation model, it has been set in a really different way if compared to the pilots on metal accessories and leather goods companies. In fact, the model moves from a job shop to an assembly line configuration, requiring a different set of input data such as the length of the assembly line and the constant speed it moves at. The company's assembly line moves 87 boxes, each of them with a maximum capacity of 4 pairs of shoes to be assembled, and 18 stations and relative machineries are located in the perimeter.

Moving solidly to the assembly line, the items have to pass in front of all the 18 stations but, according to the

items' classification between "easy", "medium" and "difficult" shoes, each of them can be or not processed on a single station and the workers will do only the tasks of the station that are included in the item's production cycle. If no tasks have to be done for processing an item on a specific station, the related worker has to skip the item and look for the next one in the assembly line that has to be processed in that station. According to this, in the modeled system workers can move from the station they have been associated to the assembly line, in order to take the first item that needs to be processed on the station and put again the item itself on the box where it was once it has been processed.

RESULTS

The first runs of the simulation model have been done in order to validate the processing time measured and assigned to each SKU type (i.e. "easy", "medium" and "difficult") considering a single worker per station. In particular, runs of simulation have been done using as input only the "easy" shoes, only the "medium" shoes and only the "difficult" ones respectively. According to the expected results, 700 pairs of "easy" shoes, 360 pairs of "medium" shoes and 280 pairs of "difficult" shoes can be processed per day.

Due to the fact that the scheduled production usually refers to few SKUs per day, the feasibility has been checked through second runs of the simulation model considering different sequencing empirical rules, represented by the different combination of "easy", "medium" and "difficult" shoes according to the products mix defined by the daily scheduled production plan.

Because of the fact that the simulation model starts with an empty conveyor, a warm-up period of 2 hours has been taken into account in order to achieve the steady-state situation.

In order to check the feasibility of the simulation model, the KPI that has been evaluated is the average daily assembly line productivity, especially the average percentage of the assembled products and the daily scheduled production detailed in Table 1. Moreover, the saturation of all the active stations (i.e. "Station 6" and "Station 16" are the excluded ones) for the SKUs to be produced has been taken into account, in order to compare the feasible solutions.

Table 1 - KPIs dashboard per sequencing empirical rules: overall values

KPI code	KPI	S_1	S_2	S_3
Prd_W_Avg	Average daily productivity	100%	100%	100%
Sat_W_Avg	Average daily saturation	29,48 %	30,08 %	29,62 %

Mks_W_Sum	Makespan [hh:mm:ss]	11:43:54	11:29:44	11:40:32
-----------	---------------------	----------	----------	----------

The column “KPI code” in Table 1 links the analysed KPIs. In particular, the KPI analysed in the footwear pilot all refer to the efficiency dimension and have been calculated at the end of the process (i.e. “Sink” block). First of all, the average value per day has been calculated for both the productivity (i.e. “Prd_W_Avg”) and the saturation (i.e. “Sat_W_Avg”) to obtain an overview of the flexibility and reactivity that the system can guarantee to perform extra-orders requested by the customers. In addition, the time between first item entering and last item exiting from the model (i.e. “Mks_W_Sum”) has been calculated in order to identify the sequence that enables to process the whole production plan in the shortest time. More in detail, looking at Table 2, all the sequencing rules confirm the feasibility of the daily scheduling plan (i.e. “Average daily productivity” equals to 100%), enabling the company to process all the scheduled SKUs. Considering the other KPIs, the average daily saturation has been calculated including only the active stations and refers to the makespan (i.e. the difference between the last exit date from a processing block and the first enter date on a processing block). For these two KPIs, the values differ considering the implementation of one or another sequencing rule, highlighting how the “Sequence_2” results in a higher average daily saturation and a shorter makespan.

Table 2 - KPIs dashboard per sequencing empirical rules: overall values (including reworking)

KPI code	KPI	S_1	S_2	S_3
Prd_W_Avg	Average daily productivity	99,34%	98,40%	99,24%
Sat_W_Avg	Average daily saturation	29,15%	29,55%	29,21%
Mks_W_Sum	Makespan [hh:mm:ss]	11:43:54	11:29:44	11:40:32

Moving from Table 1 to Table 2, the implementation of the sequencing rules allows the company to process all the daily scheduled SKUs, and this is related to the fact that a percentage of reworking (i.e. 2%) has been introduced according to the management requirements. On the other hand, the implementation of the simulation model including this type of stochasticity shows how the “Sequence_3” is the worst sequencing rule in terms of

KPIs. In fact, its implementation results neither in the higher values for average daily productivity and the average daily saturation or the shorter makespan. On the other hand, the best sequencing rule between the “Sequence_1” and “Sequence_2” depends of the company’s CSF: implementing “Sequence_2” results in the higher average saturation and shorter makespan, while “Sequence_1” guarantee the higher average daily productivity. Table 3 shows the detailed saturation per station, highlighting what is the bottleneck station for the analysed assembly line and production plan. The related KPI code is “Sat_S_Avg”, that measures the average saturation per resource.

Table 3 - KPIs dashboard per best sequencing empirical rules: average saturation per station

	Sequence 1	Sequence 2	Sequence 3
St 01	57,88%	58,9%	57,97%
St 02	46,18%	46,9%	46,25%
St 03	42,02%	42,7%	42,09%
St 04	25,40%	25,7%	25,46%
St 05	10,95%	10,9%	11,00%
St 06	0,00%	0,0%	0,00%
St 07	5,20%	5,1%	5,23%
St 08	5,19%	5,1%	5,22%
St 09	84,24%	85,7%	84,37%
St 10	1,73%	1,7%	1,73%
St 11	57,88%	58,9%	57,97%
St 12	46,18%	46,9%	46,25%
St 13	42,02%	42,7%	42,09%
St 14	25,40%	25,7%	25,46%
St 15	10,95%	10,9%	11,00%
St 16	0,00%	0,0%	0,00%
St 17	5,20%	5,1%	5,23%
St 18	5,19%	5,1%	5,22%

Once the feasibility has been checked and the KPIs for the balanced assembly line have been evaluated, the optimization of the number of workers per station has been the object of another scenario analysis conducted through simulation, assessing how the KPIs changes varying the number of workers associated to one or more stations. According to this, starting from the results in Table 5, one more worker has been associated to the station with the higher saturation independently from the implemented sequencing rule (i.e. “Station 9”). Moreover, the sequencing rule chosen to conduct this scenario analysis has been the one that results in better performances in (i.e. “Sequence_2”). The compared scenarios have been listed in Table 4.

Table 4 - Scenarios for simulation model in the footwear case study

	Description
S_1	No reworking; 1 resource for each station (see “Sequence_2” in Table 1)

S_2	Reworking; 1 resource for each station (see “Sequence 2” in Table 2)
S_3	No reworking; 2 resources per “Station 9”
S_4	Reworking; 2 resources per “Station 9”

For each one of the scenarios described in Table 4, the KPIs values used to the comparison have been listed in Table 5.

Table 5 - KPIs dashboard per sequencing empirical rules: overall values

KPI code	KPI	S_1	S_2	S_3	S_4
Prd_W_Avg	Average daily productivity	100%	98,40 %	100%	99,67 %
Sat_W_Avg	Average daily	30,08 %	29,55 %	30,20 %	29,91 %
Mks_W_Sum	Makespan [hh:mm:ss]	11:29:44	11:29:44	10:24:56	10:24:56

Looking at the results in Table 5, comparing the scenarios with no stochasticity (i.e. “Scenario_1” and “Scenario_3”), their implementation results in a shorter makespan (-9.4%) and a slightly higher average saturation (+0.4%) considering 2 workers on the “Station 9”. Comparing the other two scenarios that include reworking (i.e. “Scenario_2” and “Scenario_4”), moving from 1 to 2 workers on the “Station_9” the makespan has been reduced in the same way of the previous comparison (-9.4%) while the average saturation increases (+1.2%) in the “Scenario_4”. In addition, also the average daily productivity increases (+1.3%).

Table 6 shows the detailed saturation per station (i.e. KPI equals to “Sat_S_Avg”, as listed in Table 5) for each one of the three scenarios described in Table 5.

Table 6 - Scenario analysis for the best sequencing rule: average saturation per station

	Seq_2 Sce 1	Seq_2 Sce 2	Seq_2 Sce 3	Seq_2 Scen 4
St 01	59,81%	58,9%	66,01%	65,29%
St 02	47,72%	46,9%	52,67%	52,09%
St 03	43,40%	42,7%	47,90%	47,40%
St 04	26,19%	25,7%	28,91%	28,65%
St 05	11,17%	10,9%	12,33%	12,33%

St 06	0,00%	0,0%	0,00%	0,00%
St 07	5,31%	5,1%	5,86%	5,86%
St 08	5,30%	5,1%	5,85%	5,85%
St 09	87,07%	85,7%	48,05%	47,52%
St 10	1,78%	1,7%	1,97%	1,95%
St 11	59,81%	58,9%	66,01%	65,29%
St 12	47,72%	46,9%	52,67%	52,09%
St 13	43,40%	42,7%	47,90%	47,40%
St 14	26,19%	25,7%	28,91%	28,65%
St 15	11,17%	10,9%	12,33%	12,33%
St 16	0,00%	0,0%	0,00%	0,00%
St 17	5,31%	5,1%	5,86%	5,86%
St 18	5,30%	5,1%	5,85%	5,85%

CONCLUSION

The present work describes the results of the application of a framework that combines simulation and optimization into a model for supporting production planning and scheduling in a fashion footwear company. In detail, once the optimized plan has been chosen, several sequencing rules have been simulated firstly in a deterministic and considering four different stochastic environments. Analyzing the deterministic scenario, one sequencing rule has been chosen and then it has compared with the four different stochastics scenarios. The results show how the presented simulation-optimization framework can be applied in not-traditional sectors (i.e. the fashion one), where quality and craftsmanship are the main Critical Success Factors (CSFs), and empirical rules and not optimal solution are still applied.

REFERENCES

- Amen, M. (2000). Heuristic methods for cost-oriented assembly line balancing: A survey. *International Journal of Production Economics* 68, 1–14.
- Amen, M. (2001). Heuristic methods for cost-oriented assembly line balancing: A comparison on solution quality and computing time. *International Journal of Production Economics* 69, 255–264.
- Amen, M. (2006). Cost-oriented assembly line balancing: Model formulations, solution difficulty, upper and lower bounds. *European Journal of Operational Research* 168, 747-770.
- Bard, J. F., Dar-Elj, E. Z. E. Y., and Shtub, A. (1992). An analytic framework for sequencing mixed model assembly lines. *The International Journal of Production Research*, 30(1), 35-48.
- Bautista, J., Pereira, J., (2002). Ant algorithms for assembly line balancing. *Lecture Notes in Computer Science* 2463, 65–75.
- Bautista, J., Suarez, R., Mateo, M., and Companys, R., (2000). Local search heuristics for the assembly line balancing problem with incompatibilities between tasks. In: *Proceedings of the 2000 IEEE International Conference on Robotics and Automation*, San Francisco, CA, 2404–2409.
- Baykasoglu, A., and Özbakir, L., (2006). Stochastic U-line balancing using genetic algorithms. *International Journal of Advanced Manufacturing Technology*.
- Becker, C., and Scholl, A. (2006). A survey on problems and methods in generalized assembly line balancing. *European journal of operational research*, 168(3), 694–715.

- Carpanzano E, and Ballarino A. (2008). Collaborative networked enterprises: a pilot case in the footwear value chain. *Innovation in Manufacturing Networks*, 57-66.
- Chituc, C., Toscano, C., and Azevedo, A. (2008). Interoperability in Collaborative Networks: Independent and industry-specific initiatives-The case of the footwear industry. *Computers in Industry*, 59(7), 741-757.
- Cortez, P. M. C., and Costa, A. M. (2015) Sequencing mixed-model assembly lines operating with a heterogeneous workforce. *International Journal of Production Research*, 53(11), 3419-3432.
- d'Avolio, E., Bandinelli, R., and Rinaldi, R. (2015a). Improving new product development in the fashion industry through product lifecycle management: A descriptive analysis. *International Journal of Fashion Design, Technology and Education*, 8(2), 108-121.
- d'Avolio, E., Bandinelli, R., Pero, M., and Rinaldi, R. (2015b). Exploring replenishment in the luxury fashion Italian firms: evidence from case studies. *International Journal of Retail and Distribution Management*, 43(10-11), 967-987.
- Dörmer, J., Günther, H. O., & Gujjula, R. (2015). Master production scheduling and sequencing at mixed-model assembly lines in the automotive industry. *Flexible Services and Manufacturing Journal*, 27(1), 1-29.
- Drex A., Kimms A., (2001). Sequencing JIT mixed model assembly lines under station load and part usage constraints. *Management Science*, Vol. 47, No. 3, 480-491.
- Fani, V., Bandinelli, R., and Rinaldi, R. (2017). A simulation optimization tool for the metal accessory suppliers in the fashion industry: A case study. *Proceedings - 31st European Conference on Modeling and Simulation, ECMS 2017*, 23-26 May 2017; pp. 240-246.
- Fani, V., Bandinelli, R., and Rinaldi, R. (2018). Optimizing production allocation with simulation in the fashion industry: a multi-company case study. *Proceedings - Winter Simulation Conference, Part F134102*, pp. 3917-3927.
- Germanes, J. S., Puga, M. F., Sabio, R. B., Sanchez, E. M., & Hugo, J. C. (2017). Improving Efficiency of Shoe Manufacturer through the Use of Time and Motion Study and Line Balancing. *Journal of Industrial and Intelligent Information Vol*, 5(1).
- Jahangirian, M., Eldabi, T., Naseer, A., Stergioulas, L., K., Young, T. (2010). Simulation in manufacturing and business: a review. *European Journal of Operation Research*. 203(2010), 1-13.
- Jayaprakash, J., Reddy, K. M., K., and Ambedkar, P. (2015). Simulation of mixed model assembly line sequencing using PRO-Model software. *International Journal of Applied Engineering Researc*. 10(68), 854-856.
- Jeon, S. M., and Kim, G. (2016). A survey of simulation modeling techniques in production planning and control (PPC). *Production Planning & Control*, 27(5), 360-377.
- Kim, Y.K., Kim, J.Y., Kim, Y., (2000b). A coevolutionary algorithm for balancing and sequencing in mixed model assembly lines. *Applied Intelligence* 13, 247-258.
- Kim, Y.K., Kim, J.Y., Kim, Y., (2006). An endosymbiotic evolutionary algorithm for the integration of balancing and sequencing in mixed-model U-lines. *European Journal of Operational Research* 168, 838-852.
- Kim, Y.K., Kim, S.J., Kim, J.Y., (2000c). Balancing and sequencing mixed-model U-lines with a co-evolutionary algorithm. *Production Planning & Control* 11, 754-764.
- Kim, Y.K., Kim, Y., Kim, Y.J., (2000a). Two-sided assembly line balancing: a genetic algorithm approach. *Production Planning and Control* 11, 44-53.
- Kucukkoc, I., & Zhang, D. Z. (2014). Mathematical model and agent based solution approach for the simultaneous balancing and sequencing of mixed-model parallel two-sided assembly lines. *International Journal of Production Economics*, 158, 314-333.
- Levitin, G., Rubinovitz, J., Shnits, B., (2006). A genetic algorithm for robotic assembly line balancing. *European Journal of Operational Research*. 168, 811-825.
- Nazar, K. A., and Pillai, V. M. (2018). Mixed-model sequencing problem under capacity and machine idle time constraints in JIT production systems. *Computers & Industrial Engineering*, 118, 226-236.
- Sadeghi, P., Rebelo, R.D., Ferreira, J.S. (2018), Balancing mixed-model assembly systems in the footwear industry with a variable neighbourhood descent method, *Computers and Industrial Engineering*, 121, pp. 161-176.
- Sadeghi, P., Rebelo, R.D., Soeiro Ferreira, J. (2017), Balancing a Mixed-Model Assembly System in the Footwear Industry, *IFIP Advances in Information and Communication Technology*, 513, pp. 527-535.
- Scholl, A., Becker, C. (2006). State-of-the-art exact and heuristic solution procedures for simple assembly line balancing. *European Journal of Operations Research* 168, 666-693.
- Scholl, A., Becker, C., (2005). A note on an exact method for cost-oriented assembly line balancing. *International Journal of Production Economics* 97, 343-352.
- Scholl, A., Becker, C., (2006). State-of-the-art exact and heuristic solution procedures for simple assembly line balancing. *European Journal of Operations Research* 168, 666-693.
- Vrittika Pachghare, R. S. Dalu (2014), Assembly Line Balancing – A Review, *International Journal of Science and Research (IJSR)*, 3(3) 2014
- Zamami Amlashi, Z., & Zandieh, M. (2011). Sequencing Mixed Model Assembly Line Problem to Minimize Line Stoppages Cost by a Modified Simulated Annealing Algorithm Based on Cloud Theory. *Journal of optimization in Industrial Engineering*, (8), 9-18.

AUTHOR BIOGRAPHIES

VIRGINIA FANI is a PhD student at the Industrial Engineering Department of the University of Florence. The issues she is dealing with are related to production processes optimization along the supply chain, with particular focus on the peculiarities that the fashion companies have to face. Her e-mail address is virginia.fani@unifi.it

BIANCA BINDI is a PhD student at the Industrial Engineering Department of the University of Florence. During her PhD, she carried out both research and consulting activities in the field of Supply Chain optimization and radio-frequency technologies (e.g. RFID, NFC). Her e-mail address is bianca.bindi@unifi.it

ROMEO BANDINELLI is a researcher at the University of Florence, Department of Industrial Engineering. He is member of the Observatory GE.CO. of Politecnico di Milano, program chair of the IT4Fashion conference and member of the IFIP 5.1 "Global Product development for the whole lifecycle". His e-mail address is romeo.bandinelli@unifi.it

Finance and Economics and Social Science

THE EUROPEAN STABILITY MECHANISM AND SOVEREIGN BOND YIELDS: AN ANALYSIS IN LIGHT OF NEW DEBATES

Eszter Boros and Gábor Sztanó
Department of Finance
Corvinus University of Budapest
Fővám tér 8, HU-1093, Budapest, Hungary
Email: eszter.boros@uni-corvinus.hu
gabor.sztano@uni-corvinus.hu

KEYWORDS

eurozone, bailout, European Stability Mechanism, reform, Italy, bond yields

ABSTRACT

The sovereign debt crisis revealed that there was a need for a bailout mechanism in the then prevailing framework of the euro area (EMU). In 2010, bond spreads of troubled periphery sovereigns started to soar relative to the core countries, threatening monetary policy transmission. Ever since, the ways of crisis management and EMU institutional reforms have been sparking a conflict between a German view of country-level responsibility and French-Italian calls for more risk sharing. The most recent chapter of this debate is about the ongoing reform of the European Stability Mechanism (ESM). This paper focuses on the evolution of the EMU financial assistance framework, up until the latest concerns of Italy. Our key question is whether policy steps resulting in a permanent bailout mechanism have played a role in driving sovereign yields. By using an event study approach and panel regressions, we find that related announcements have significantly contributed to a decrease in periphery sovereign bond yields. This result suggests that markets reacted positively, and their expectations moved toward a more integrated and resilient European financial market. For debates on the ESM overhaul, this contribution to financial stability should serve as a common ground. A “package approach” bundling multiple key reforms together, as stressed by Italy, may well also need to be taken into account.

INTRODUCTION

Tensions in Italy in December 2019 have revealed concerns about one of the ongoing institutional reforms of the euro area. Doubts over the intended changes in the European Stability Mechanism have been voiced by Italy in the final phase of approval. This sheds light on a deeper disagreement among EMU members on the directions that key reforms should take. The debate over the ESM overhaul is all the more crucial because this change could be the first major one completed since 2014. However, Italy’s reservations may cause a delay

and prompt EMU leaders to speed up progress on other pending reforms.

As developments on the ESM can be a turning point, we deem it essential to review the underlying arguments and more broadly, the ESM’s “track record” in enhancing financial stability. This issue is closely related to sovereign bond markets, not just because the ESM is mandated to assist states, but also because Italian critics of the reform envisage a rise in bond yields (due to changes facilitating the restructuring of privately held debt). Thus, our paper seeks to examine the relationship between yields and bailout-/ ESM-related events in the crisis- and post-crisis period. This research can help reveal whether the creation of a bailout facility had an impact on the evolution of EMU sovereign yields, and also deliver insights for the debate on the current reform. Applying an event study approach, we first create an event set which includes announcements on the EMU bailout framework, along with actual cases of financial assistance. This event set is then used in a panel regression to check its possible impact on yields of periphery bonds with different maturities. (As conventional in the literature, the term “periphery” is used to refer to the Mediterranean member states and Ireland.)

To provide a preliminary insight into our results, we find that ESM-related events have a significant negative effect on yields, indicating an overall stabilizing role of the EMU-level bailout arrangements. This result may serve as a common ground for debates on ESM settings, as well as on a set of broader financial reforms. Our contribution can be regarded as new especially in terms of its coverage of ESM-related events. It does not only examine bailouts themselves, but a much wider range of announcements on the respective institutional progress.

This paper is structured as follows. The next section provides background, with a focus on the evolution of motivations leading up to the creation and reform of the ESM. Our interpretation also aims to spot the key events as a basis for the detailed event set created subsequently. The third section then describes this set, along with the methodology, data and variables used to get model outputs. Results and limitations are thereafter discussed, with the last section summarizing our conclusions.

THE EUROPEAN STABILITY MECHANISM: BACKGROUND, CREATION AND REFORM

Like all major changes in the EMU's institutional architecture during the last couple of years, the creation of a common bailout fund was also triggered by the euro crisis. At the time when Greece's severe debt problem quickly raised market tensions in 2010, the euro area had neither any crisis management functions/framework nor dedicated financial resources to assist its troubled members (Pisani-Ferry 2010, 2012; Christova 2011; Gocaj and Meunier 2013; Baldwin et al. 2015). In fact, up until the crisis, several years had passed in the belief that sovereigns in a monetary union could not go bankrupt (Surányi 2012). If for no other reason than because their peers would help them "get through". This view was indeed not in line with the "no bail-out clause" (enshrined in EU treaties) and it was also inconsistent with the lack of a proper "lender of last resort" function on the part of the European Central Bank (ECB) (see De Grauwe 2012, 2013). Nonetheless, the introduction of the euro prompted a substantial reduction in perceived risks, compressing bond spreads for all member countries. This kind of euro-related confidence, together with macro and policy factors, contributed to a prolonged period in which EMU sovereigns could borrow at historically low interest rates. Some of them (Greece and Italy) went on piling up huge debts during this period. (In other cases, like Spain, private debts increased and ultimately became a threat for sovereigns, in a deadly embrace between banks and states [Acharya et al. 2014].)

Market sentiments then drastically changed in late 2009 (Giordano et al. 2013, Schwendner et al. 2015). Yields for Southern states and Ireland soared, not least because prospects of any bailout or fresh liquidity seemed totally uncertain. Anxiety was fuelled by an extremely fragile EMU framework which left sovereigns without any backstop for their debts akin to FX denominated liabilities. As De Grauwe (2012, 2013) points out membership in a monetary union involves losing full control of the legal tender. In other words, there is not even an implicit guarantee that money will always be available to pay off creditors. This is because influence on money supply is in the hands of a single central bank of the area, allowing no substantive room for unilateral action. In contrast, "standalone" states can practically repay any amounts of debts denominated in their own currency at any time. This certainly does not mean that a sovereign default is impossible in this case (as a state in serious trouble would ultimately default on its foreign liabilities and lose access to foreign markets). Nevertheless, at certain critical moments like those in 2010 and thereafter, the ability for limitless debt service may be decisive for markets' risk perception (see Mehrling 2000 for a discussion on this).

Regarding the sudden stop of market finance in 2010, the problem was not merely that countries in distress had no room for creating euros on their own. Rather, it was about their inability to apply for a last resort. The ECB (or any other institution) was namely not mandated to fulfil this

role (De Grauwe 2012, 2013, Surányi 2012). As regards banks, access to emergency liquidity was possible, but also uncertain because it was conditional on ECB action. Certainly, it is a general rule that applicants need to be solvent to be eligible for last resort. It is also clear that at least one EMU member (Greece) was bankrupt at the time when an emergency loan could have been contemplated. Thus, we do not argue that the spike in bond yields was only attributed to a missing "lender of last resort" function or a fragile institutional framework more broadly. It is instead suggested here that the pace and amplitude of market reactions were largely aggravated by this weakness. As already mentioned, deficiencies also included the lack of an EMU-wide crisis management framework and any financial resources explicitly available for bailouts. There was in fact no agreement on what a "bailout" could mean in the EMU framework at all. Furthermore, no procedure had been laid down for exiting the monetary union as membership had been meant to be irreversible (De Grauwe 2018). When exit then emerged as a (partly rhetoric) option during the crisis, it threatened with unbearable economic and political costs (Eichengreen 2010).

Against this background, the ECB eventually took action in 2012 to bring relief to EMU sovereigns. After some previous purchases limited in size and time, the central bank decided to commit to infinite intervention in bond markets if needed. This was hallmarked by the speech of ECB President Mario Draghi who promised to do "whatever it takes to preserve the euro" (Draghi 2012). Through the subsequent Outright Monetary Transactions (OMT) program, the ECB practically offered to fulfil a "lender of last resort" role for sovereigns. This, however, did not come without a price tag in a political sense. Related debates about risk sharing and intra-area "transfers" were well on the rise. Due to this controversy, the OMT program soon found itself before the German Federal Constitutional Court, being challenged as violating the ban on monetary financing (Várnay 2017). (German reservations were later rejected by the European Court of Justice.)

This ECB case is important here as it reflects differences between a German and a French-Italian stance regarding risk sharing. Bénassy-Quéré et al. (2018) show that Germany, along with some other countries of the EMU core, stresses responsibility at the level of the individual member states and a fear of moral hazard. In contrast, some countries led by France and Italy argue for more risk sharing and stronger cooperation at the union level.

Such differences have also been leaving a mark on the evolution of bailout policies till today. Market panic in 2010 first caused EMU leaders to envisage ad hoc bailout commitments (February 2010). This was well before the realization that the ECB's unlimited intervention power is also unavoidable (De Grauwe 2012). A one-time lifebelt for Greece in the form of loans with strict conditionality could actually be regarded as being in line with the "no bailout" (no fiscal transfers) rule (Micossi et

al. 2011). For Germany, it could seem well-constrained while also well-suited to stop contagion threatening with bringing down German banks with large Southern exposures. But bond yields signalled that markets would not be calmed by such a standalone solution. The approach had to be scaled up as soon as the Greek rescue package was signed in May 2010. Thus, the European Financial Stability Facility (EFSF) was created as a temporary vehicle granting emergency credit to troubled sovereigns (up until 2013). The EFSF framework included €60 billion in loans and credit lines to be provided by the EU budget (an amount also known as the European Financial Stabilisation Mechanism, EFSM) and further bilateral credit guarantees by members up to €440 billion euros. Moreover, the IMF granted a contribution of €250 billion (Christova 2010; Gocaj and Meunier 2013). Guarantees were provided by member states on a pro-rata basis, according to their ECB capital keys, meaning that Germany became the biggest potential contributor, able to set the parameters of the institution. The EFSF issued bonds in financial markets to finance bailout commitments. These securities gained high credit rating thanks to the underlying state guarantees with Germany at the first place. Subsequently, EFSF extended loans to Ireland and Portugal in 2011. Assistance came not just with widely criticized austerity-based “reform programs”, but also with rather punitive interest rates. This was largely due to the fact that Germany wanted to make sure assistance was not a “subsidy”, and no “Eurobond” was created (Gocaj and Meunier 2013). EFSF principles and procedures draw upon those of the IMF which, along with IMF participation, seemed to be an appropriate “deterrent” to rule out moral hazard (Pisani-Ferry 2010).

Albeit certainly not unlimited (like an ECB intervention could be), this temporary rescue vehicle was supposed to put an end to market tensions. However, favourable reactions were short-lived, due to concerns about the actual size of contributions, especially the amount of money readily available (i.e. not only guaranteed). “The EU was once again grossly underprepared to deal with the burgeoning crisis” (Gocaj and Meunier 2013, p. 247). Therefore, in March 2011, the European Council decided to establish a permanent stability mechanism, the ESM, replacing the EFSF from 2013 onwards. The ESM’s lending capacity was set to €500 billion. Most of its subscribed capital came in the form of guarantees and “callable capital”, besides a paid-in part of €80 billion (Manasse 2011; Minenna and Aversa 2019). That is, a guarantee-based approach was maintained, along with the size of the overall rescue capacity. An innovation compared to the previous EFSF setting was the concept of private sector involvement (Christova 2011). It was envisaged that in case of irreversible debt dynamics, the recipient state would have to start a renegotiation of its debts with private creditors. Although such an outcome remained very unlikely (as it would have to be preceded by an official declaration of unsustainability in the frame of the ESM procedure), some further steps were taken to make way for private involvement. Starting from 2013,

so-called Collective Action Clauses (CACs) have been included in new government securities with a maturity over 1 year, issued by any member state (ESM 2020a). These CACs foresee that a change in bond terms (such as a restructuring) can happen if approved by a qualified majority of creditors at the levels of each bond series and all series combined.

In terms of tools, the ESM was at start able to extend loans with strict conditionality, and also to buy government bonds in primary markets (up to 50% of the final issued amount to reduce the risk of a failed auction, ESM 2020b). Interest rates of loans were decreased to some extent (by 100 basis points). Shortly after its inception, further enhancement of ESM powers took place, allowing the institution to recapitalize banks and make purchases in secondary markets (Christova 2011). The former was used to rescue Spanish banks in 2012, and Cyprus also entered a program thereafter in 2013. (Apart from loans and indirect bank recapitalization, no other instruments have been used yet [ESM 2020b].)

As it can be seen from the above, markets indeed pushed EMU leaders to adopt more far-reaching bailout solutions amid a sustained euro area debt crisis. Schwendner et al. (2015) argue that a consolidation in sovereign bond markets (a dissolution of negative correlations between daily changes of yields of core versus periphery countries) can be attributed to the new rescue and stability mechanisms. Similar conclusions are drawn by Kiss et al. (2019) who find that EFSF/ESM loans contributed to the observed decline in long-term yield premia in the aftermath of the euro crisis. In contrast, Gödl and Kleinert (2016) establish that announcements of financial assistance and fiscal measures (as conditions of rescue packages) prompted no significant change in long-term periphery bond yields (except for Ireland). These authors conclude that other events played a more decisive role (e.g. ECB action and separate country-specific episodes like the Greek haircut in 2012). The evolution of yields is the key question of our paper. Our findings are thus presented in the next section, based on an event set extended to cover the current ESM reform, as well.

A very recent debate about the ESM is related to its ongoing reform. In late 2019, Italy announced that it would seek changes in the approach regarding the amendment of the ESM Treaty, which had come to the last phase of approval (Fonte and Jones 2019, ANSA 2019). It is essentially this turmoil which has directed our attention to the ESM and its overhaul. The situation, resembling a deadlock, namely points to the fact that meaningful EMU reforms have stalled since 2014, due to different visions of the member states (Bénassy-Quéré et al. 2018; Minenna and Aversa 2019). Italy’s key fear is that new ESM rules may raise its debt servicing costs. Such an outcome would certainly hurt Italy whose public debt-to-GDP ratio is the second largest in the euro area. (The figure was 137.3% in the third quarter of 2019 according to Eurostat data, surpassed only by a Greek

ratio of 178.2%.) We close this section by reviewing the reform and the potential rationale for Italian doubts.

Changes in the ESM framework were initiated a couple of years ago when proposals like a non-paper by then German Finance Minister Wolfgang Schäuble (2017) raised the question of debt restructuring. Schäuble suggested that ESM assistance should come with an automatic extension of the maturities of sovereign bonds of the recipient state. Moreover, there should be a mandatory debt restructuring mechanism to be used if deemed as necessary to restore debt sustainability. The non-paper also points to the need to modify current CACs to facilitate private sector involvement in risk sharing. Despite such proposals, current revisions do not include an automatic obligatory restructuring as a condition for ESM help (ESM 2020a). Amendments in CACs, however, have become a part of the reform, and these are the main source of Italian concerns about rising yields. Current CACs are so-called “double-limb CACs”. As already mentioned, they require two separate majorities to approve a change in bond terms: one at the level of each series and one at the level of all series combined (ESM 2020a). This feature benefits “holdout” investors who can delay a debt restructuring by acquiring majority in a single series. Similar CACs have caused misery during the 2012 Greek restructuring when debt burdens from 18 series (out of 35) could not be eased due to “holdout” investors (Bénassy-Quéré et al. 2018).

To prevent such an outcome, “single-limb” CACs are now to be introduced. These “allow the majority vote to take place at the level of all (...) series combined, without the need for a majority at the level of the holders of each individual series” (ESM 2020a). The amendment makes debt restructuring (if needed) clearly easier. This can in turn decrease the attractiveness of government bonds, especially in case of highly indebted sovereigns like Italy. Italian critics argue that higher yields will follow, making it “more likely that Italy will have to restructure or even default on its debt” (Fonte and Jones 2019). Market reactions are also examined in the next section.

Other key points of the ESM reform include a backstop role for the Single Resolution Fund (SRF) and enhancing the effectiveness of precautionary credit lines (PCLs). A part of the banking union, the SRF is an EMU-level fund for the resolution of failing banks. Granting a backstop is aimed at strengthening its stabilization function. Notwithstanding, such a commitment will necessarily put a strain on the resources of the ESM. (Even though potential ESM loans are capped [ESM 2020a].) As regards precautionary credit lines, the innovation lies in more standardization. PCL requests will be processed on the basis of standardized eligibility criteria, enhancing speed and transparency. To the best of our knowledge, neither the backstop function nor the change regarding PCLs have been challenged by recent Italian critics.

Italy’s reservations have had a remarkable impact on the approval process so far. Although in December 2019, the

Eurogroup agreed in principle on a revised ESM Treaty text, its final adoption by national parliaments has been postponed (possibly) until spring this year. It is not by accident that Eurogroup President Mário Centeno hinted at the importance of a “package approach” (Centeno 2020). Italy has namely been stressing this as a solution that could make the ongoing ESM reform more feasible. A “package” here refers to the need for advancing other reforms parallelly. An EMU-wide bank deposit guarantee, a common unemployment insurance mechanism and progress on a eurozone budget are among those preferred by Italy (Fonte and Jones 2019). However, these plans are much less advanced as they involve even more disagreement among the members. In what follows, we examine the potential impact of the evolution of the EMU bailout framework on the yields of the most vulnerable sovereigns of the currency area.

EMPIRICAL ANALYSIS

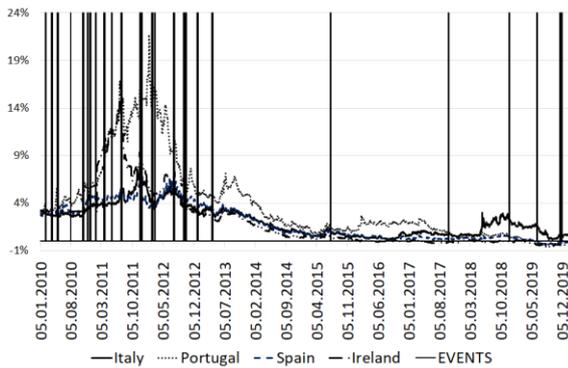
In the empirical part of our paper, we analyse the impact of news related to the EFSF/ESM on government bond yields in the eurozone periphery, namely in Italy, Spain, Portugal and Ireland. We disregard Greece because of a notable shrinkage in its marketable debt, implying a less reliable reflection of the prevailing market sentiments.

We assess market reactions by employing an event study approach, explaining the change in bond yields with event dummies and control variables. It is commonly called in the literature, as ‘intervention analysis’: we measure the impact of bailout policy events by regarding the change in sovereign bond yields. We also apply a set of control variables to support our estimation.

Data and Variables

We use benchmark government bond yields on different maturities (1, 5 and 10 years). Figure 1 presents the evolution of 5-year benchmark yields since the early period of the euro crisis. Benchmark yields enable us to compare roughly similar securities in terms of maturity, although the shift from one bond to another as benchmark might lead to an unexplained rise in yields. As it often happens with time series financial data, the chosen yields are not stationary, which was confirmed by the augmented Dickey-Fuller test. Furthermore, using first differences later allows us to disregard fundamental determinants of yields which do not change in the very short run (Gödl and Kleinert 2016). Thus, we decided to calculate the first differences of bond yields. This eliminates the problem of changing benchmarks, as well, given that unexplained rise due such shifts occurs only rarely and may not mislead us. We use daily observations, so daily changes of yields (*dyield*) appear as the left-hand side variable (in the model discussed below). The time frame covers every working day between January 2010 and January 2020. All bond yields are retrieved from Bloomberg.

Figure 1: Government Bond Yields (5-year maturity) and Dates of Events



Source: Bloomberg, own calculations.

Prior to model calculations, we created a set of announcements related to EFSF/ESM funding to distressed countries and relevant institutional changes, ranging from early indications of a bailout scheme in 2010 up until to the recent reform proposal in 2019. These Events are collected from literature quoted above and double-checked by using official EU website. In total, 29 events are considered as marked in *Figure 1*. Note that we omitted country-specific fiscal announcements and ECB decisions from the set as they are not directly connected to our scope. (The possible effects of these events are captured by control variables discussed below.) We created a dummy variable for the events (*Events*), with 1 marking the day of announcements, as well as the days preceding and following them. (All other days have 0.) Surrounding days are taken into account as anticipated and lagged effects are likely to be captured this way. As for control variables, we collected several indicators, some of which have different values for each periphery country, while others are common indicators. Country-specific indicators include the uncertainty index (*Unc_index*) and Bloomberg political risk index (*BBG_pol_risk*) as provided by Bloomberg. In order to proxy underlying euro area processes and global risk sentiments, we selected VIX index (*VIX_index*), Citi macro risk indicator (*Citi_macro*), Citi short-term indicator (*Citi_short*) and Citi CPI surprise index (*Citi_CPI*) for the eurozone, along with EU uncertainty index (*EU_unc_index*). In addition, German yields (*Ger_XY*) and so-called ‘inflation breakevens’ (*CPI_breakeven*) were selected. The latter embodies information about the market pricing of inflation-linked bonds, i.e. this variable could be regarded as a market perception of future inflation rates. We also use periphery CDS data as a control variable (*PER_CDS*) as downloaded from Bloomberg. Note that we used 5Y contracts here as other maturities were not liquid at all times for all countries. Daily data are used as first differences (in accordance with the time series of bond yields), while monthly data such as Citi-indices, political risk and uncertainty indices remained unchanged.

As a robustness check, we resort to the ‘popularity’ of Google keywords related to the EFSF and ESM. We rely on Google’s own scale for measuring ‘popularity’. (The scale ranges from 1 to 100, with higher values indicating more searches in case of a keyword. The underlying methodology is accepted for the purposes of this paper.) As the potential ‘popularity’ of these words overlap, we decided to use the form ‘ESM OR EFSF’ in the first place, but we looked at separate popularity indices, as well. ‘Popularity’ indices may give a more nuanced picture about the importance and durability of the relevant effects. However, values are only available on a monthly basis, so differences in times of consecutive events are not reflected.

Model

In the literature, modelling the impact of policy announcements is common as it may be interpreted easily, even though the method is quite limited in capturing longer-term effects. Our model roughly follows the one used by Falagiarda and Reitz (2015). In their work, they employed a similar set of panel data to examine the impact of the ECB’s quantitative easing programs.

We estimate the following model (*Equation (1)*):

$$d(\text{yield}_{i,t}) = \alpha + \beta_1 \text{Events} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \varepsilon \quad (1)$$

where $d\text{yield}$ is the first difference of the respective government bond yield in country i at time t . Regarding the three different maturities involved, we decided to repeat the same regression for each, using the respective maturities of the risk free (German) rate. Remember that *Events* is a dummy for policy announcements described earlier. This is a common variable with the same content for all countries in our panel. \mathbf{X} and \mathbf{Z} are sets of control variables, with \mathbf{X} being country-specific, while \mathbf{Z} not. (\mathbf{X} includes *Unc_index*, and *BBG_pol_risk*, while \mathbf{Z} comprises the rest of our control variables.) Intercept α is common and time invariant. (Chow test showed that country-specific fixed effects are not needed in the model.) Finally, ε denotes the error term.

Note that our approach is similar to that of Gödl and Kleinert (2016) and Kiss et al. (2019) as both studies apply variations of the event study method and regressions. However, we resort to a single, more detailed and comprehensive event set regarding the evolution of bailout policies. This goes beyond focusing exclusively on the fact whether financial assistance was received by a country or not (Kiss et al. 2019). We also strove to overcome limitations arising from a small number of events of separate types, and we pooled announcements on EMU bailout funds exclusively (two features different from Gödl and Kleinert (2016)).

Results and Discussion

Estimating the model, we find that the announcements had a significant negative impact on the bond yields of EMU periphery countries between January 2010 and 2020 (Tables 1-3). In other words, news about the establishment and use of the European bailout funds have tended to contribute to the moderation of contemporaneous sovereign yields since the outbreak of the euro crisis. After several iterations, we decided to omit most of the indices (*Unc_index*; *BBG_pol_risk*; *Citi_macro*; *Citi_short*; *Citi_CPI*; *EU_unc_index*) as they did not improve the explanatory power of the model. Thus, the final version of the model includes the change in the CDS price of a given periphery country, the change in VIX index and the benchmark German government bond yield on the respective maturity. The latter proxy variables allow us to capture global, regional trends affecting the bond yields, while CDS is supposed to reflect the default risk of the respective periphery country (also as perceived by markets).

Table 1: Panel Outputs: Impacts of EFSF/ESM-Related Events on Periphery Bond Yields on 1Y Maturity

	Estimates			
	coeff.	std.err	t-ratio	p-value
const	0.0002	0.0017	0.10	0.9133
Events	-0.0599	0.0101	-5.45	4.92e-08***
d_PER_CDS	0.0008	6.63e-05	11.88	2.6e-032***
d_VIX_index	0.0022	0.0011	2.07	0.0380**
d_GER_1Y	-0.2619	0.0915	-2.86	0.0042***
F-statistic: 50.89 p-value: 1.78e0-42*** R-square: 0.02				

Table 2: Panel Outputs: Impacts of EFSF/ESM-Related Events on Periphery Bond Yields on 5Y Maturity

	Estimates			
	coeff.	std.err	t-ratio	p-value
const	-0.0006	0.0013	-0.52	0.6027
Events	-0.0214	0.0080	-2.67	0.0076***
d_PER_CDS	0.0012	5.03e-05	23.53	2.3e-119***
d_VIX_index	0.0045	0.0008	5.37	8.05e-08***
d_GER_1Y	-0.0740	0.0352	-2.10	0.03560**
F-statistic: 165.47 p-value: 1.304e0-137*** R-square: 0.06				

Table 3: Panel Outputs: Impacts of EFSF/ESM-Related Events on Periphery Bond Yields on 10Y Maturity

	Estimates			
	coeff.	std.err	t-ratio	p-value
const	-0.0009	0.0010	-1.00	0.3142
Events	-0.0111	0.0061	-1.81	0.0698*
d_PER_CDS	0.0009	3.8e-05	24.55	1.7e-129***
d_VIX_index	0.0058	0.0006	9.18	4.9e-020***
d_GER_1Y	0.1245	0.0245	5.07	3.92e-07***
F-statistic: 179.48 p-value: 5.42e0-149*** R-square: 0.06				

Source: Own estimations.

In case of 1-year government bond yields, we find that our event variable is statistically significant at every conventional significance level ($p < 1\%$). Therefore, we may establish that EFSF/ESM announcements significantly decreased the 1-year government bond yields of the periphery, with approximately 6 basis points on average each time.

Regarding 5-year yields, the event set also proved to be significant at a 1% level. Nonetheless, the estimated coefficient is smaller which means that events decreased bond yields by about 2 basis points on average each time. In case of 10-year bonds, events are still significant, although only at a higher (but still conventional) threshold (10%). The coefficient shows that the impact was also the smallest in this case: we may identify a 1-basis point average change connected to the events presented above. In sum, we can conclude that EFSF/ESM-related announcements contributed to a decrease in periphery bond yields in the aftermath of the euro crisis.

Although the overall explanatory power of these models could be upgraded, this is not key here insofar as our event variable is statistically significant even together with strong indicators of default risk and market uncertainty (as control variables). So, we deem these results to be useful insights, especially what regards different maturities. According to the literature, short-term yields are more likely to be affected by market sentiments and expectations about short-term monetary policy decisions, while long-term yields are reflecting long-run structural factors. Furthermore, long-term yields are more likely to be affected by quantitative easing programs, by nature. Our results suggest that market expectations about further turbulence eased mostly in case of shorter maturities. This is in line with the background described in the previous section. It can be assumed that policy steps before the eventual creation of a permanent fund (the ESM) were mostly judged by markets as a temporary fix. That is, investor confidence returned in case of short-term periphery bonds, but they were not strongly convinced that changes at hand will help to cut the divergence among eurozone countries in the long run. This may be due to the incremental and largely uncertain nature of the institutional progress and also to views about the overall size and structure of the dedicated funds. On the other hand, an alternative explanation is that the ECB's QE programs may have muted the impact of the institutional changes in case of 10-year maturities.

Regarding the robustness of our results, we carried out two types of alternative estimations. First, choosing and rotating different sets of the above control variables did not change the significance levels (p-values) of the event dummy in a sizable manner, on either of the maturities involved. Coefficients remained negative and they did not change substantially, ranging between 0.015 and 0.023. Second, we substituted the event set with Google popularity statistics, and we found that 'popularities' of

EFSF and ESM keywords proved to be significant. It is a thought-provoking result, raising questions that go beyond the scope of our paper (e.g. how market sentiments are indeed reflected in Google searches). Nevertheless, we argue that this alternative outcome confirms our results that markets reacted positively to the news on EFSF/ESM. In other words, the observed decrease in sovereign bond yields of the periphery is less likely to be attributable to an unknown, omitted variable. This is because searches for EFSF/ESM are rightly supposed to have been prompted by actual events and delivered information about (potential) additional funds for troubled EMU member states.

Regarding the market reactions after the debates in November 2019, Italian concerns about the ESM overhaul seem not to be underpinned. That is, we could not observe an obvious unilateral rise in sovereign bond yields for Italy (neither for Greece with similarly high debts). This can be due to the fact that the ESM reform does not include an automatic obligatory debt restructuring as a condition for financial assistance. Although the modification of CACs can be regarded as disadvantageous for private investors in future cases, Italian fears of capital flight have not come true so far.

Our results are in line with those of Schwendner et al. (2015) and Kiss et al. (2019) as we also found that the EFSF/ESM played a significant role in decreasing periphery bond yields. (At least what regards prompt market reactions in case of related news.) Our finding adds to the literature by demonstrating that the significant relationship can also be established when considering a broad set of EFSF/ESM-related events (not only disbursements of bailout loans themselves). This result may be a common ground for current debates. That is, in case of every proposal, it should be assessed whether they are in line with preserving the contribution of the ESM to financial stability.

Limitations

Our methodology has the usual limitations of event study methods which do not allow to assess longer-term impacts of events and are not suited to uncover connections between explanatory variables. Therefore, we could offer just limited insights into the root causes and underlying dynamics of the perceived decrease in periphery bond yields. Note also that event study outcomes may be sensitive to changes in the event set. In this paper, we strove to reach a reasonable coverage of EFSF/ESM-related events.

We shall add two further remarks to the regressions above. First, as we have just noted, the persistence of the reducing effect could not be handled in this type of model. Other factors also influence bond yields, such as particular data releases, deterioration in investment environment or business confidence, changes in domestic policies, shifts in the global economy etc. That is, the direct effects of announcements examined in this paper are likely to disappear after a couple of days. According

to our calculations for the event set used here, the decreasing effect lasted on average for 2-4 days after the announcement, with some heterogeneity in the different countries (2 days in Italy, 3 in Portugal, Spain and Ireland). (In this respect, the date of disappearance is the first day when the yield rose for the first time after the announcement.)

Second, although the institutional reform of the ESM received serious criticism from some Italian politicians at the end of 2019, there are at this time too few data points to assess these announcements alone. Thus, we could only provide a qualitative assessment as part of our paper's broader context. With more hindsight, however, the impacts of the reform (including the introduction of "single-limb" CACs) need to be separately evaluated in the future. A question whether new CACs may provoke higher yields in times of market turbulence, can only be answered later, too.

CONCLUSIONS

This paper examined the relationship between sovereign bond yields and events related to the EMU bailout framework. The relevance of the topic has lately been underlined by concerns about the ongoing ESM overhaul. As Italy, one of the most indebted countries of the euro area, has been afraid of a resulting increase in its bond yields, the larger issue of pending EMU reforms has been in the spotlight again.

Therefore, we deemed it essential to investigate the background of the concerns and more broadly, the role of Europe's bailout arrangements in driving sovereign yields. An event study approach and panel regressions were applied for this purpose. Our aim was to create a comprehensive set of major EFSF/ESM-related events after the outbreak of the euro crisis.

Our empirical findings show that the announcements related to EFSF/ESM institutional progress significantly decreased government bond yields in the eurozone periphery. That is, the creation of an EMU bailout framework is likely to have contributed to restoring financial stability. However, we could establish that long-term government bond yields reacted less strongly, which might be attributed to quantitative easing programs and/or the relative scepticism of investors regarding the long-run effects of the bailout mechanisms.

Regarding the most recent evolution of yields since late 2019, we found no obvious unilateral rise for Italy. Time will certainly tell whether a possibly less favourable position of private investors (due to new CACs) prompts capital flight and a rise in periphery yields. Nonetheless, a "package approach" stressed by Italy may well need to be taken into account. That is, EMU financial reforms should at best come together in order to reduce the chance of unintended outcomes for the most fragile members.

REFERENCES

- Acharya, V.; Drechsler, I.; and Schnabl, P. 2014. "A Pyrrhic Victory? Bank Bailouts and Sovereign Credit Risk." *The Journal of Finance*, Vol. 69, No. 6 (Dec 2014), 2689-2739.
- ANSA 2019. "Conte Defends ESM, Tells Opponents to Say if They Want Euro Exit." ANSA News Agency, 12.12.2019.
- Baldwin, R.; Beck, T.; Bénassy-Quéré, A.; Blanchard, O.; Corsetti, G.; De Grauwe, P.; Haan, d. W.; Giavazzi, F.; Gros, D.; Kalemli-Ozcan, S.; Micossi, S.; Papaioannou, E.; Pesenti, P.; Pissarides, C.; Tabellini, G.; and Weder di Mauro, B. 2015. "Rebooting the Eurozone. Step I. Agreeing on a Crisis Narrative." *CEPR Policy Insight*, No. 85 (Nov).
- Bénassy-Quéré, A.; Brunnermeier, M.; Enderlein, H.; Farhi, E.; Fratzscher, M.; Fuest, C.; Gourinchas, P.-O.; Martin, P.; Pisani-Ferry, J.; Rey, H.; Schnabel, I.; Véron, N.; Weder di Mauro, B.; and Zettelmeyer, J. 2018. "Reconciling Risk Sharing with Market Discipline: A Constructive Approach to Euro Area Reform." *CEPR Policy Insight*, No. 91 (Jan).
- Centeno, M. 2020. "Remarks by Mário Centeno following the Eurogroup Meeting of 20 January 2020", available at the official website of the European Council, as of 03.12.2020.
- Christova, A. 2011. "The European Stability Mechanism: Progress or Missed Opportunity?" *Baltic Journal of European Studies*, Vol. 1, No. 2, 49-58.
- De Grauwe, P. 2012. "The Governance of a Fragile Eurozone." *The Australian Economic Review*, Vol. 45, No. 3, 255-268.
- De Grauwe, P. 2013. "Design Failures in the Eurozone – Can They Be Fixed?" *European Economy, Economic Papers* 491 (Apr 2013). European Commission, Brussels.
- De Grauwe, P. 2018. "Political Economy of Deconstructing the Eurozone." In "Economics of Monetary Union", P. de Grauwe, Chapter 8. Oxford University Press, 12th Edition, 149-164.
- Draghi, M. 2012. "Verbatim of the Remarks Made by Mario Draghi." Speech by Mario Draghi, President of the European Central Bank at the Global Investment Conference in London, 26 July 2012. European Central Bank Release.
- Eichengreen, B. 2010. "The Euro: Love It or Leave It?" *VOX CEPR Policy Portal*, 4 May 2010.
- ESM (2020a): "ESM Treaty Reform Explainer", available at the official website of the ESM, as of 2 February 2020.
- ESM (2020b): "Lending Toolkit", available at the official website of the ESM, as of 07.12.2020.
- Falagiarda, M. and Reitz, S. (2015) "Announcements of ECB Unconventional Programs: Implications for the Sovereign Spreads of Stressed Euro Area Countries." *Journal of International Money and Finance*, 53, 276–295.
- Fonte, G. and Jones, G. 2019. "Italy PM Defends Reform of Eurozone Bailout Fund but Seeks Concessions." *CNBC Online News*, 2 December 2019.
- Giordano, R.; Pericoli, M.; and Tommasino, P. 2013. "Pure or Wake-Up-Call Contagion? Another Look at the EMU Sovereign Debt Crisis." *Temi di discussione* (Economic Working Papers) 904, Bank of Italy, Economic Research and International Relations Area
- Gocaj, L. and Meunier, S. 2013. "Time Will Tell: The EFSF, the ESM, and the Euro Crisis." *Journal of European Integration*, Vol. 35, No. 3, 239-253.
- Gödl, M. and Kleinert, J. 2016. "Interest Rate Spreads in the Eurozone: Fundamentals or Sentiments?" *Review of World Economics / Weltwirtschaftliches Archive*, Vol. 152, No. 3 (Aug), 449-475.
- Kiss, G. D.; Csiki, M.; and Varga, J. Z. 2019. "Comparing the IMF and the ESM Through Bond Market Premia in the Eurozone." *Public Finance Quarterly*, 2019/2, 277-293.
- Manasse, P. 2011. "The Trouble with the European Stability Mechanism." *VOX CEPR Policy Portal*, 5 April 2011.
- Mehrling, P. 2000. "The State as Financial Intermediary." *Journal of Economic Issues*, Vol. 34, No. 2 (June), 365-368.
- Micossi, S.; Carmassi, J.; and Peirce, F. 2011. "On the Tasks of the European Stability Mechanism." *CEPS Policy Brief*, No. 235, 08.03.2011.
- Minenna, M. and Aversa, D. 2019. "A Revised European Stability Mechanism to Realize Risk Sharing on Public Debts at Market Conditions and Realign Economic Cycles in the Euro Area." *Economic Notes by John Wiley & Sons, Ltd*, Vol. 48, No. 1-2019.
- Pisani-Ferry, J. "Euro-Area Governance: What Went Wrong? How to Repair It?" *Bruegel Policy Contribution*, No. 2010/05
- Pisani-Ferry, J. 2012. "The Known Unknowns and the Unknown Unknowns of the EMU." *Bruegel Policy Contribution*, No. 2012/18 (Oct).
- Schäuble, W. 2017. "Non-Paper for Paving the Way Towards a Stability Union." Proposal for the Eurogroup (Oct).
- Schwendner, P.; Ott, T; Schüle, M.; and Hillebrand, M. 2015. "European Government Bond Dynamics and Stability Policies: Taming Contagion Risks." *ESM Working Paper Series*, 8/2015.
- Surányi, Gy. 2012. "The Global Crisis: Have We Learned the Right Lessons?" *CASE Network E-Briefs*, No. 07/2012.
- Várnay, E. 2017. "The European Central Bank Amid the Crisis – The OMT Case" (in Hungarian language). In "State – Crisis – Finance" (in Hungarian language), Kálmán, J. (ed.). Gondolat, Budapest, 368-394.

DATA SOURCES

- Eurostat. General government gross debt, quarterly data (Table code: teina230). Downloaded on 08.02.2020
- Bloomberg. Used tickers are available upon request. Downloaded from the terminal on 28.01.2020
- Google Trends. Popularity indices. Downloaded on 02.02.2020

AUTHOR BIOGRAPHIES

Eszter BOROS, MSc is a PhD candidate at Corvinus University of Budapest (CUB), Doctoral School of General and Quantitative Economics. She earned her master's degree in Economics from CUB, specializing in Bank and Public Finance. She is a lecturer at the Department of Finance. Her main field of research is monetary unification, with special regard to the euro area. Her email address is eszter.boros@uni-corvinus.hu

Gábor SZTANÓ, MSc is a PhD candidate at Corvinus University of Budapest (CUB), International Relations Multidisciplinary Doctoral School. He earned his master's degree in Economics from CUB, specializing in Bank and Public Finance. He is a lecturer at the Department of Finance, teaching Finance. His main field of research is monetary policy in emerging countries. His email address is gabor.sztano@uni-corvinus.hu

MODELLING THE RELATIONSHIP BETWEEN DEMOGRAPHIC STRUCTURES OF THE RUSSIAN POPULATION

Anna Bagirova

Oksana Shubat

Ural Federal University

620002, Ekaterinburg, Russia

Email: a.p.bagirova@urfu.ru

Email: o.m.shubat@urfu.ru

KEYWORDS

Econometric modelling, demographic structures, daily childcare, regression analysis

ABSTRACT

Russia is now facing serious demographic problems. Several reasons cause the decrease in the number of children and their share in the population. We analyzed the possible relationship between two different demographic structures in Russia, which characterize the share of children in the population and the share of population involved in a daily childcare. We applied correlation analysis, methods of econometric modelling and analytics infographics. The results of our study as follows: 1) the young-age dependency ratio in the Russian regions is very slightly correlated with the share of the population involved in daily childcare; 2) the number of children of different age groups in the population is either unrelated or very slightly correlated with the share of the population involved in daily childcare; 3) the relationship between the fertility rate in the region and the degree of population involvement in childcare cannot be confidently confirmed either; 4) in Russia there is a group of regions with atypically high levels of the studied variables. This is a special phenomenon in the Russian demographic space.. Our results make it possible to strengthen the information base necessary for developing more effective demographic measures.

INTRODUCTION

Many European countries are now facing problems of low fertility and population ageing. Several reasons cause the decrease in the number of children and their share in the population. Sociologists primarily speak of a decreased level of the need for children, an increase in the "family-work" conflict (Feeney and Stritch 2019; Aisenbrey and Fasang 2017; Moran and Koslowski 2019), which entails a reduction in the desired number of children. Economists talk about the influence of the material level and insufficient standards of living (Rangelova and Bilyanski 2019); medical professionals - about the deterioration of the population's reproductive health (Woods and Hensel 2020; Janevska 2017); psychologists - about the spreading phenomenon of intense motherhood and the growing sense of guilt among mothers who do not have time to devote the

"ideal" amount of attention to their children (Forbes, Donovan and Lamar 2019; Milkie, Nomaguchi and Schieman 2019), etc. In our opinion, a number of the listed reasons are related to the concept of parental labour - the work that parents do by giving birth, raising, caring for, and developing their children. Indeed, in modern society, parenthood is a labour one needs to be prepared for (from a medical and psychological point of view) and to know how to combine with professional work. It is also a labour which should be rationally organized using various resources and implemented efficiently.

In countries experiencing low fertility problems, there are usually some forms of parental leaves - periods after the birth of the child, which parents (mother and / or father) can devote entirely to childcare. For example, in Russia this period lasts until the child is 3 years old, in Germany - up to 6 years, at the public sector enterprises in Greece - up to 10 years, in Sweden - up to 1.5 years, and part of this period must be used by the child's father (Koslowski et al. 2019). After this period, parents usually return to the professional labour market, and daily childcare is delegated to other persons or institutions (relatives, hired professionals, kindergarten teachers, etc.). The time parents spend on parental labour usually decreases with the age of the child. At the same time, the number of people engaged in the child's care, education and development increases. For example, in Russia, grandparents traditionally help parents to look after their children. While parents are busy in the labour market during the day, grandparents spend time with their grandchildren, accompany them to various classes, cook food for them, take them for a walk, etc. According to Russian official statistics, about 20% of women aged 55 and older provide daily childcare (to their children or someone else's) (Comprehensive monitoring of living conditions 2018).

The composition of family members involved in childcare may vary by region, type of culture and lifestyle. Older brothers and sisters and more distant relatives may look after younger children. Such care can be carried out on an ongoing (daily) basis or periodically, from time to time. Thus, the implementation of the parental labour can vary across a range of parameters. In the context of our study, these basic parameters are as follows:

1) the set of functions that a labour subject performs in relation to a child (a complete or incomplete list of parental functions);

- 2) the number of children that an adult takes care of simultaneously (one, two or more);
- 3) the frequency with which the subject performs these functions (daily or less frequently);
- 4) daily time expenditure on the parental labour (from 0 to 24 hours a day);
- 5) the presence and proximity of family ties with the child / children in relation to which the functions of parental labour are realized (parents, grandparents, siblings, distant relatives).

DATA AND METHODS

We analyzed the possible relationship between two different demographic structures, which characterize the share of children in the population and the share of population involved in a daily childcare. Based on our concept of parenthood as a labour activity, we can say that these people perform at least one function of parental labour on an ongoing daily basis. They may or may not be relatives of the child.

The hypothesis of the study was as follows: the degree of involvement of the Russian population in parental labour is determined by the structural features of the population, which reflect the share of children in the population and the fertility rate. It should be noted that the consideration of the problem in this aspect, and, even more so, the use of the statistical modelling to solve it is a new approach for Russian science and practice. The demand for this approach is primarily associated with the negative dynamics of the birth rates, which has been observed in recent years in Russia despite the active state policy of stimulating the fertility.

It is important that current Russian official statistics required to test the hypothesis are extremely limited. Countrywide studies aimed at researching parental labour are not conducted. Some relatively valid indicators can be found in sample population surveys conducted by the Russian Federal State Statistics Service. However, these data do not cover a long period of time; it is impossible to form time series of the length (duration) necessary for the statistical analysis.

At the same time, an important feature of the demographic development of Russia, which allowed us to conduct our study, is regional diversity: the variability of fertility rates and its potential determinants, the existence of different “demographic modernities”, models of demographic space, in the country (Shubat, Bagirova and Akishev 2019). For example, in 2018, the share of the population involved in daily childcare varied from 13.2% to 46.9% in the Russian regions (Comprehensive monitoring of living conditions 2018). The total fertility rate ranged from 1.12 to 2.97 (Total Fertility Rate 2019). A sufficient number and a wide variety of regional situations made it possible to create the necessary sample size to apply statistical modelling methods.

Thus, we analyzed the following indicators characterizing the demographic situation in 85 regions of Russia in 2018:

Var 1: The share of the population involved in daily childcare without payment. In the first approximation, this variable characterizes the involvement of the population in parental labour (The structure of the resident population 2019). This indicator reflects the population structure by engagement in the parental labour. The higher it is, the higher the share of the population performing certain functions of parental labour.

Var 2: Young-age dependency ratio. This indicator reflects the demographic structure of the population by age, since it shows the number of children aged 0-14 per 1000 people of working age (Russian Regions 2018).

Var 3: The share of children aged 0-4 (Var 3₀₋₄), 5-9 (Var 3₅₋₉), 10-14 (Var 3₁₀₋₁₄) in the total population (The structure of the resident population 2019). These indicators also characterize the demographic structure of the population.

Var 4: Total fertility rate (Total Fertility Rate 2019). This is the most accurate statistical indicator of the fertility. It is defined as the total number of children that would be born to each woman if she were to live to the end of her child-bearing years and give birth to children in alignment with the prevailing age-specific fertility rates.

To conduct the study, we used econometric modelling methods. We calculated the Pearson correlation coefficients and Spearman rank correlation and estimated the parameters of the regression equations (ordinary least squares method) for the studied variables. We also used analytics infographics: we studied the distribution of data using boxplot and identified outliers in order to adjust the results of econometric modelling.

RESULTS

1. Our study showed that the young-age dependency ratio in the Russian regions is very slightly correlated with the share of the population involved in daily childcare. The parameters of the regression model showed that the relationship between these indicators is positive (the share of the population involved in parental labour increases with the increasing load of children), but rather weak (Table 1-3, model 1). The coefficient of determination is slightly more than 11%, which indicates the unsatisfactory quality of the estimated model. At the same time, we excluded errors in the model specification. Visualization of the source data allowed us to suggest a linear model of the relationship as the most probable (Figure 1).

Table 1: Model Summary*

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.335	0.112	0.101	3.8272
2	0.310	0.096	0.084	3.8620
3	0.282	0.079	0.067	4.0012

Table 2: ANOVA*

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	141.1	1	141.1	9.632	0.003
	Residual	1113.2	76	14.7		
	Total	1254.3	77			
2	Regression	120.8	1	120.8	8.098	0.006
	Residual	1133.5	76	14.9		
	Total	1254.3	77			
3	Regression	103.5	1	103.5	6.465	0.013
	Residual	1200.7	75	16.0		
	Total	1304.3	76			

Table 3: Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error			
1	Constant	19.466	3.815		5.103	0.000
	Var 2	0.034	0.011	0.335	3.103	0.003
2	Constant	21.927	3.298		6.649	0.000
	Var 3 ₀₋₄	1.477	0.519	0.310	2.846	0.006
3	Constant	19.628	4.543		4.320	0.000
	Var 4	7.308	2.874	0.282	2.543	0.013

* Dependent variable: Var 1 – Involvement in unpaid daily childcare;
 Var 2 – Young-age dependency ratio;
 Var 3₀₋₄ – The share of children aged 0-4;
 Var 4 – Total fertility rate.

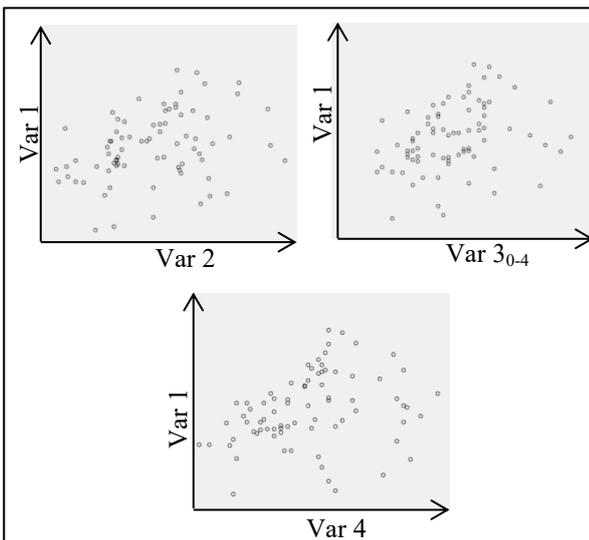


Figure 1: The correlation of demographic indicators and the share of the population involved in daily childcare (Var 1 – Involvement in unpaid daily childcare; Var 2 – Young-age dependency ratio; Var 3₀₋₄ – The share of children aged 0-4; Var 4 – Total fertility rate)

2. We also found out that the number of children of different age groups in the population is either unrelated or very slightly correlated with the share of the population involved in daily childcare. This relationship was not confirmed for age groups of 5–9 years and 10–14 years. For the age group of 0–4 years, the parameters of the regression model showed that the relationship is positive (the share of the population involved in parental labour increases with the increasing share of children aged 0–4 in the population of the region). However, this correlation is rather weak (Table 1-3, model 2). In this case, we also excluded errors in the model specification (Figure 1).

3. The relationship between the fertility rate in the region and the degree of population involvement in childcare cannot be confidently confirmed either. The parameters of the regression model were significant and indicated a positive correlation: the share of the population involved in daily childcare increased with the increasing total fertility rate. However, this model proved to be of unsatisfactory quality. (Table 1-3, model 3 and Figure 1).

4. Our study showed that there is a group of regions in the Russian demographic space which differs significantly from the average Russian level and the levels of other regions. This group’s visualization of variables, for which statistically significant models were obtained, is presented in Figure 2. Generalized statistics with all the studied variables are presented in Table 4.

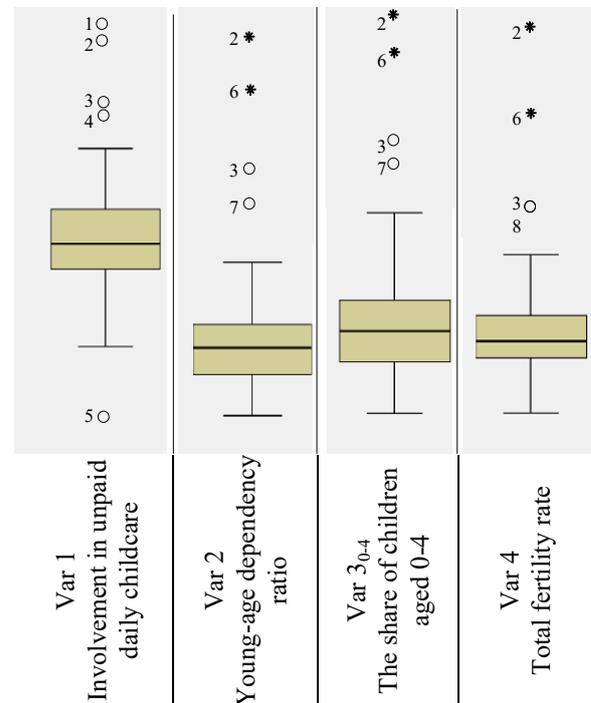


Figure 2: Boxplots of the studied variables (1 – Chukotka autonomous okrug; 2 – Tyva Republic; 3 – Altai Region; 4 – Republic of Buryatia; 5 – Magadan Oblast; 6 – Chechen Republic; 7 – Republic of Ingushetia; 8 – Nenets autonomous okrug)

Table 4: The frequency of a region identification as outlier or extreme outliers in the distribution of the studied variables

Russian Region	Frequency
Altai Region	6
Tyva Republic	6
Chechen Republic	5
Republic of Ingushetia	4
Chukotka autonomous okrug	2
Nenets autonomous okrug	1
Republic of Buryatia	1
Republic of Dagestan	1

In terms of statistical modelling, these regions are considered either outliers (more than 1.5 interquartile range distance from the third quartile) or extreme outliers (more than 3 interquartile range distance from the third quartile). Therefore, they are subject to exclusion from the studied set of regions in the modelling process. However, we think that it is advisable to conduct a separate special study for these regions.

DISCUSSIONS

The obtained results indicate a weak relationship between the two demographic structures of the population: in terms of age (reflecting, on the one hand, the share of children in the population, and, on the other hand, the fertility rate) and in terms of the share of population involved in parental labour. In our opinion, the following explanations are possible.

Primarily, the number of births per 1 adult is higher in Russian regions with a high number of children and a high young-age dependency ratio (Table 5).

Table 5: Correlation between the total fertility rate and the young-age dependency ratio

Indicator		Value
Pearson Correlation	Coefficient	0.897
	Sig. (2-tailed)	0.000
Spearman's rho	Coefficient	0.845
	Sig. (2-tailed)	0.000

Therefore, the birth of a second, third and subsequent child does not have a direct proportional effect on the growth in the number and the share of population implementing parental labour. The labour functions associated with these children are carried out by the same parents who are already engaged in caring for the first child. Hence, in this case we can speak of an increase in the intensity of parental labour, but not of an increase in the number of subjects involved in it.

Regions with atypically high levels of the studied variables found during the analysis are a special phenomenon in the Russian demographic space. These regions (Chechnya, Tuva, Ingushetia, Altai, Dagestan)

have a high fertility rate, but they are characterized by a rather low standard of living. High birth rates in these regions can be explained by sociocultural (prevailing reproductive norms, marital behavior of young people, traditional family structure) and religious factors. The existence of such demographic phenomenon requires a special study. This may become the basis for the development of new measures of state support for fertility. These measures, unlike those currently implemented in the country, will not prioritize purely economic incentives for fertility.

The limitation of our study is the inability to analyze the population involved in childcare differentially for a number of parameters of parental labour: the completeness of the performed parental functions; the number of children that an adult takes care of simultaneously; daily time expenditure on parenting; the proximity of family ties with the child / children in relation to which the parental labour functions are realized. This kind of information can only be collected through specially organized sociological survey. It would allow us to acquire a more complete picture of the Russian population involved in parental labour.

CONCLUSIONS

Our research is one of the first attempts to study the determinants of the population involvement in parental labour. The obtained results show that, firstly, there is a high differentiation by demographic structures of the population reflecting the share of children and the share of people involved in daily childcare in the Russian regions. Secondly, there is no strong relationship between the two studied demographic structures. Thirdly, there are unique regions in terms of these indicators in Russia. A more thorough study of the situation in these regions will provide an idea of the extreme forms of this phenomenon.

Another important conclusion is the need to strengthen the information base of this type of research. The recent negative demographic trends in Russia obviously require the development of more effective measures of state support and stimulation of the fertility rate. This is impossible without strengthening the information base of the decision-making process. Our study, the obtained results and the described limitations present one of the possibilities of this development.

ACKNOWLEDGMENTS

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

REFERENCES

- Aisenbrey, S. and A. Fasang. 2017. "The interplay of work and family trajectories over the life course: Germany and the United States in comparison". *American Journal of Sociology*, Vol 122(5), 1448-1484. DOI: 10.1086/691128
- Comprehensive monitoring of living conditions. 2018. Rosstat, Moscow. URL:

- https://gks.ru/free_doc/new_site/KOUZ18/index.html
(access date 16.02.2020).
- 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, DOI: 10.1177/0734371X17733789
- Forbes, L.K., Donovan, C., and M.R. Lamar. 2019. "Differences in Intensive Parenting Attitudes and Gender Norms Among U.S. Mothers". *Family Journal*, Vol 28, issue 1, 63-71.
- Janevska, R. 2017. "Improving reproductive health of Roma women". *European Journal of Public Health*, Vol 27, Issue 3, suppl 3, cdx189.077, <https://doi.org/10.1093/eurpub/ckx189.077>
- Koslowski, A. et al. 2019. International Review of Leave Policies and Research 2019. URL: <https://www.leavenetwork.org/annual-review-reports/> (access date 16.02.2020).
- Milkie, M.A., Nomaguchi, K. and S. Schieman. 2019. "Time Deficits with Children: The Link to Parents' Mental and Physical Health". *Society and Mental Health*, Vol 9(3), 277-295. DOI: 10.1177/2156869318767488.
- Moran, J. and A. Koslowski. 2019. "Making use of work-family balance entitlements: how to support fathers with combining employment and caregiving". *Community, Work and Family*, Vol 22, Issue 1, 111-128. DOI: 10.1080/13668803.2018.1470966
- Rangelova, R. and V. Bilyanski. 2019. "Economic aspects of demographic changes in the European Union and in Bulgaria". *Ikonomicheski Izsledvania*, Issue 5, 25-54.
- Russian Regions. Socio-Economic Indicators. 2018. Statistical Book. Rosstat, Moscow. URL: https://gks.ru/bgd/regl/b18_14p/Main.htm (access date 16.02.2020).
- Shubat, O. M., Bagirova, A. P. and A.A. Akishev. 2019. "Methodology for Analyzing the Demographic Potential of Russian Regions Using Fuzzy Clustering". *Ekonomika regiona [Economy of Region]*, Vol 15(1), 178-190. DOI: 10.17059/2019-1-14
- The structure of the resident population. Single inter-departmental information and statistical system (SIDIS). Rosstat, Moscow. 2019. URL: <https://fedstat.ru/indicator/43219> (access date 16.02.2020).
- Total Fertility Rate data. Single inter-departmental information and statistical system (SIDIS). 2019. Rosstat, Moscow. URL: <https://fedstat.ru/indicator/31517> (access date 16.02.2020).
- Woods, J. and D. Hensel. 2020. "What female adolescents value for improving male sexual and reproductive health". *Journal of Adolescent Health*, Vol 66, Issue 2, S116-S117.

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/>

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

RUSSIAN GRANDPARENTING: DEMOGRAPHIC AND STATISTICAL MODELLING EXPERIENCE

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

KEYWORDS

Statistical modelling, Russian grandparenting, parental labour, independent samples tests, nonparametric tests

ABSTRACT

The demographic situation in Russia has been quite complicated for many years. Its important negative manifestations are low birth rates and low life expectancy at birth. In this situation, studies on the role of older people (primarily grandparents) in the birth, upbringing, and development of grandchildren become especially important. The purpose of our exploratory research is to model the objective and subjective characteristics of Russian grandparents who realize the functions of parental labour in relation to their grandchildren. Statistical modelling of differences between the target group and the alternative group was carried out primarily on the basis of the parametric and nonparametric independent samples tests: t-test, Mann-Whitney U test, median test. Specific features related to age, level of education, social activity and subjective assessments of health of the grandmothers involved in daily childcare were revealed. The model of a typical grandmother actively involved in parental labour was presented. This will be the basis for developing a full-scale survey and determining the best research design.

INTRODUCTION

The demographic situation in Russia has been quite complicated for many years. Low fertility – Russia is 186th in the world in terms of the total fertility rate (Country Comparison: Total Fertility Rate data 2020), low life expectancy at birth – 158th in the world (Country Comparison: Life Expectancy at Birth data 2020) force the country's leadership to make serious efforts to change the prevailing trends. In 2019, the Russian government adopted a new state program - the so-called national project “Demography” (National project “Demography” 2018). This project pays special attention to measures aimed at increasing the birth rates and the life expectancy (including active life) of the population.

These two problems are studied separately in Russia. At the same time, modern demographers, sociologists, psychologists and economists from around the world are studying the role of older people (primarily grandparents) in the birth, upbringing, and

development of grandchildren (Sichimba et al. 2017; Nedelcu 2017; Coall et al. 2018). In particular, the family-work conflict stands out as one of the most important problems of the working population in developed economies in recent years (Aisenbrey and Fasang 2017). Attracting grandparents to the upbringing of grandchildren is one of the potential mitigation tools. Scientists note the various positive effects of this intergenerational interaction for the elderly people: increased life expectancy (Chapman et al. 2018); strengthening of positive motivation and improved mental and psychological well-being (Coall and Hertwig 2010); improved physical health (Kim et al. 2017), reduced risk of death (Hilbrand et al. 2017); increased level of happiness (Danielsbasca and Tanskanen 2016), etc. There are also studies that identify positive aspects of intergenerational interaction for children: the impact on their learning outcomes (Del Boca et al. 2018), mitigation of situations of family crises, in particular, parental divorces (Attar-Schwartz and Buchanan 2018), etc.

The inclusion of grandparents in the process of raising grandchildren is influenced by various objective and subjective factors. Objective factors are, for example, the socio-demographic characteristics of the elderly: age, gender, health status, marital status, education, work experience, geographical distance from the place of residence of grandchildren, etc. Apparently, the role of subjective factors is also significant: personal activity, hierarchy of life values, satisfaction with relationships with children, self-assessment of health status, etc.

We consider the participation of grandparents in raising their grandchildren as a type of labour activity. They carry out part of the parental functions by taking care of their grandchildren. The approach to parenthood as a labour activity is quite common in the scientific literature (Erickson 2005; Oakley 1974; Daniels 1987; Pedersen et al. 2011; Bagirova et al. 2014; Veress and Bagirova 2018).

The purpose of our study is to model the objective and subjective characteristics of grandparents who realize the functions of parental labour in relation to their grandchildren. In our opinion, it is advisable to conduct an analysis separately for grandmothers and grandfathers, since the intensity and reasons for involvement in the process of parental labour may differ

in these gender groups. In this paper, we focus on the study of grandmothers.

DATA AND METHODS

1. A full-scale sample survey is necessary to identify the specifics and factors of parental labour of grandparents. To develop it, we conducted a pilot / exploratory study based on the data available in the Russian official statistics. We used data from the Federal statistical survey “Comprehensive monitoring of living condition” (Comprehensive monitoring of living conditions 2018). This survey is carried out by the Federal State Statistics Service in all regions of Russia once every two years. It covers 60 thousand households. Its results are considered representative of the whole country, urban and rural settlements, and of individual socio-demographic groups of the population. For our research we used the data of the 2018 survey. Some of its questions allowed us to model the parental labour of grandparents in the first approximation. In Russia, there are no surveys that use the category of “parental labor” in the Russian official statistics.

2. We used variable: “Does your daily routine include taking care of children, your own or other people’s (without payment)?” as the most valid indicator of the population’s involvement in the process of parental labour. Answer options to this question included: yes, no, difficult to answer, no answer.

3. Since there are no data in Russian official statistics that allow us to unambiguously identify the socio-demographic groups of elderly people as people who have grandchildren, we modeled this group based on the most valid indicator - the respondent's age. We focused on the official statistical indicator of the average age of a first-time mother: 28.7 years old in 2018 (Mean age of mothers at childbearing (years) 2019) and 25.5 years old in 1989 – the previous generation of mothers (Mean age of mothers at childbearing (years) 2002). Thus, we selected women aged 55 and older for our research. On average, a Russian woman becomes a grandmother at this age.

4. In the process of forming the target group of respondents, we considered the fact that the intensity, nature, factors and reasons for involvement in the process of parental labour may differ in two groups of grandmothers – those who live with their grandchildren (in the same household) and those who live separately. Therefore, we considered it appropriate to study these groups differentially. In this study, we focused on the grandmothers who live separately from their grandchildren.

The general scheme for the formation of the target group of respondents is presented in Figure 1. We selected women aged 55 and older, living separately from their children and taking care of children daily (in this case, they are most likely to be grandchildren). Grandmothers who were not involved in the process of daily childcare were an alternative group which the first group was compared to.

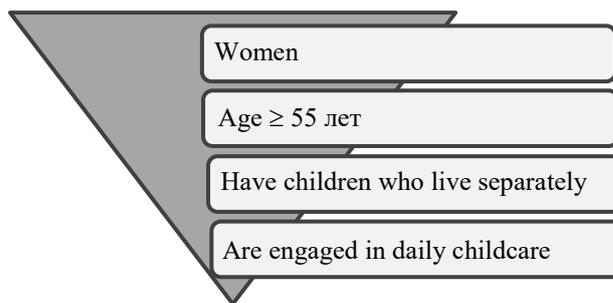


Figure 1: Formation of a Target Group of Respondents for the Study

5. The following variables were selected to model the differences between the target and the alternative groups and to create a socio-demographic portrait of grandmothers actively involved in the process of parental labour:

Var 1: age (years);

Var 2: educational level (number of years spent on education);

Var 3: marital status;

Var 4: place of residence (urban or rural area);

Var 5-8: health variables - objective indicators and subjective assessments:

- Var 5: frequency of applying for outpatient care in a medical organization;

- Var 6: frequency of ambulance calls;

- Var 7: self-assessment of health (from 1 – «very bad» to 5 – «very good»);

- Var 8: self-assessment of the opportunity to lead an active life;

Var 9: social activity variable (visits to the theater, cinema, sports and religious events, cafes and restaurants, trips around the country and abroad over the last year). During the modelling process, this group of indicators served as the basis for creating an aggregate variable which reflects the total number of the respondents' activities over the last year.

The choice of these variables was justified by research hypotheses, which consist of the influence of a combination of objective and subjective characteristics on the respondents’ involvement in childcare. Subjective characteristics included personal activity in various fields (presumably, the higher the activity, the higher the desire and ability to realize the functions of parental labour in relation to grandchildren) and self-assessment of health status (a similar relationship was assumed: the higher the self-assessment of health, the higher the involvement in processes of caring for grandchildren).

6. Statistical modelling of differences between the target group and the alternative group was carried out primarily on the basis of the following parametric and nonparametric independent samples tests:

- t-test;

- Mann-Whitney U test;

- median test.

We found it necessary to use these tests simultaneously for two reasons. Firstly, they allow us to evaluate differences between various types of data with

different distribution types. Secondly, the scientific literature does not present an unequivocal opinion on the possibilities and usefulness of a particular test. The authors prove the greater effectiveness of using various tests based on their own experiments and simulation studies.

T-test is often used to determine if the means of two sets of data are significantly different from each other provided that the test samples are characterized by normal distribution. In our study, we used this test in conjunction with Levene's test for equality of variances for two possible cases – when equal variances are assumed and not assumed. In studies, the use of these tests is most often considered exclusively for continuous data. However, Zar, based on the experiments of several authors, concludes that interval-scale or ratio-scale measurement are not intrinsically required for the application of the parametric testing procedures. These methods may be considered for ordinal-scale data if the assumptions of such methods are met (Zar 2014).

Additionally, we used the Mann-Whitney U test to model differences. This is a nonparametric analogue of the t-test. We used it because assumption of normality was often not met. U test is the most powerful nonparametric alternative to the t-test for independent samples. Several studies have shown that in certain cases Mann-Whitney U test may offer even greater power to reject the null hypothesis than the t-test (see, for example, Hollander, Wolfe and Chicken 2013). Alternatively, Zimmerman shows on the basis of a series of experiments that replacement of the t-test by a nonparametric alternative under violation of homogeneity of variance does not necessarily maximize correct decisions (Zimmerman 1987). Gibbons and Chakraborti, based on the results of a simulation study, conclude that very little power will be lost if the Mann-Whitney U test is used instead of tests that require the assumption of normal distributions (Gibbons and Chakraborti 1991).

Mann-Whitney U test was originally designed to work with continuous data (Mann and Whitney 1947). Later it was corrected to enable using it on data containing ties (see, for example Hollander, Wolfe and Chicken 2013). This is important because some variables in our study are numerical, however, they contain ties. Other variables are ordinal-scaled data (which means that they have ties).

For greater reliability of the results of our modelling, we used the median test, which compares the medians of two samples. Mood shows that the median test is about 64% as powerful as the two-sample t-test (Mood 1954). Zar notes that this test is about 67% as powerful as the Mann-Whitney test (Zar 2014). Fligner and Policello proved that if the two sampled populations have equal variances and shapes, then the Mann-Whitney U test is a test for difference between medians (Fligner and Policello 1981).

Thus, we compiled a socio-demographic portrait (model) of the studied group based on the positive results of at least two independent samples tests.

We also used other methods to analyze the relationship of individual variables – crosstabs analysis with Phi coefficient calculation (if both variables are dichotomous) and Cramer's V (for two categorical variables). However, according to the analysis results, the variables that were analyzed using these methods did not add any significant characteristics to the final model.

RESULTS

3476 people were selected as the target group. These were grandmothers involved in the daily process of parental labour. 19771 people were selected for the alternative group, which the target group was compared to. These were grandmothers who were not involved in the daily process of parental labour.

The tests for the significance of differences (tables 1-5) shows that the target group differs from the alternative group in the following parameters:

- Var 1: age (target group respondents are younger);
- Var 2: educational level (respondents of the target group have a higher level of education);
- Var 7: self-assessment of health status (target group respondents assess their health higher);
- Var 9: social activity in various areas of life (respondents of the target group go to the cinema, theater, cafes / restaurants, to sports events more often).

Table 1: Group Statistics (t-test)

Variable	Is childcare a part of daily activities?	Mean	Std. Deviation
Var 1	Yes	63.78	6.195
	No	68.06	8.943
Var 2	Yes	12.32	2.399
	No	11.64	2.813
Var 7	Yes	2.92	0.507
	No	2.77	0.602
Var 9	Yes	1.61	1.585
	No	1.14	1.401

Table 2: Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	Sig. (2-tailed)
Var 1	1*	808.1	0.000	-27.147	0.000
	2*			-34.910	0.000
Var 2	1*	109.6	0.000	13.390	0.000
	2*			14.962	0.000
Var 7	1*	555.4	0.000	13.961	0.000
	2*			15.745	0.000
Var 9	1*	202.4	0.000	17.851	0.000
	2*			16.374	0.000

* 1 – Equal variances assumed

2 - Equal variances not assumed

Table 3: Ranks (Mann-Whitney Test)

Variable	Does your daily routine include taking care of children, your own or other people's	Mean Rank	Sum of Ranks
Var 1	Yes	8954.28	31125090.50
	No	12093.37	239098037.50
Var 2	Yes	12823.15	44457846.50
	No	11361.13	223723373.50
Var 7	Yes	12822.19	44557127.50
	No	11405.63	225363867.50
Var 9	Yes	13417.73	46640028.00
	No	11308.64	223583100.00

Table 4: Mann-Whitney Test Statistics

	Var 1	Var 2	Var 7	Var 9
Mann-Whitney U	2.51E+07	2.98E+07	3.01E+07	2.81E+07
Wilcoxon W	3.11E+07	2.24E+08	2.25E+08	2.24E+08
Z	-25.451	-11.993	-13.959	-17.966
Asymp. Sig. (2-tailed)	0.000	0.000	0.000	0.000

Table 5: Median Test Statistics

	Var 1	Var 2	Var 7	Var 9	
Median	66.00	12.00	3.00	1.00	
Chi-Square	545.0	54.0	14.7	266.9	
df	1	1	1	1	
Asymp. Sig.	0.000	0.000	0.000	0.000	
Yates' Continuity Correction	Chi-Square	544.2	53.7	14.4	266.2
	df	1	1	1	1
	Asymp. Sig.	0.000	0.000	0.000	0.000

The differences between the two compared groups in the following parameters were insignificant:

- Var 3: marital status (the share of respondents engaged in daily childcare did not differ among women who were married – officially or not, never married, widowed or divorced)

- Var 4: place of residence (involvement in daily childcare did not differ between grandmothers living in urban and rural areas);

- Var 5 and Var 6: objective parameters of health status (the frequency of ambulance calls and the frequency of seeking outpatient care in a medical organization did not differ between the two compared groups of grandmothers; interestingly, the aforementioned subjective assessments of health turned out to be a significant factor differentiating the two groups).

Thus, the test results allowed us to create the following model of a typical grandmother actively involved in the process of parental labour (Figure 2).

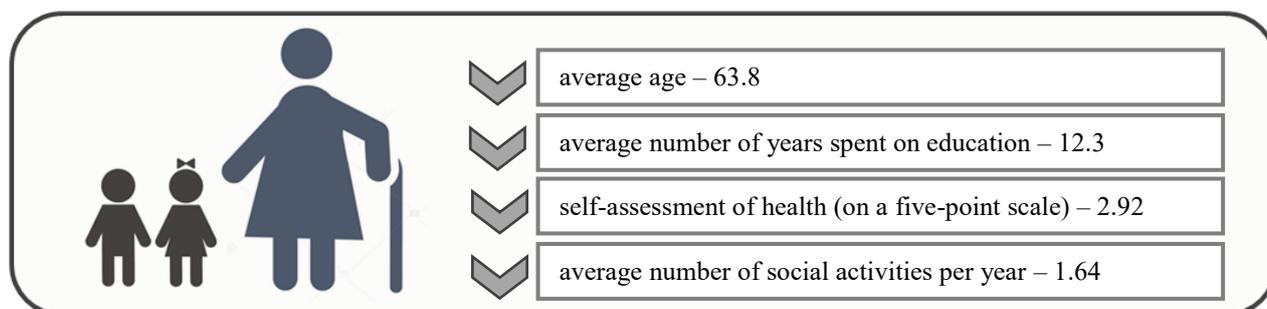


Figure 2: Model of a Typical Grandmother Actively Involved in the Process of Parental Labour

DISCUSSIONS

Understanding the characteristics of grandmothers who are involved in the upbringing of their grandchildren raises questions about the reasons for the non-participation of others. In our opinion, the motivation for parental labour of grandparents in relation to their grandchildren can be enhanced by special state demographic measures aimed at stimulating this type of labour. We largely mean material incentives, i.e. payments for the time that grandparents spend with their children, thereby replacing parents. In our opinion, the

establishment of such payments could be based on documents from institutions of preschool, school and additional education, which would indicate who brings the child to this institution and who picks him/her up after classes, what is the duration and frequency of these classes. In addition, some of the documents can be provided by the healthcare system, which has the ability to record information about the caregiver of a sick child. An increase in the time spent by grandparents with grandchildren should lead to an increase in the amount of corresponding payments for their labor.

An increase in the number of grandparents engaged in parental labour would help solve a number of important Russian socio-economic problems:

1) insufficient provision of kindergartens. For example, the average availability of preschool education in Russia for children aged 1.5 to 3 is 85.8% (Availability of preschool education for children aged 1.5 to 3 data. 2019), however, this indicator does not exceed 70% in 9 Russian regions. The increase in the number of grandparents involved in parental labour is primarily relevant for regions with low levels of kindergarten availability. Perhaps these regions could become a pilot platform for the development of incentive systems for the inclusion of grandparents in the implementation of the parental labor functions;

2) low standard of living of seniors. In 2019, the average amount of an appointed old-age pension was 15096 rubles (219.6 euros) (Average amount of an appointed old-age pension data 2019). State-paid parental labour could raise the living standards of pensioners. In addition, the very fact of payments (even small ones) for caring for grandchildren would indicate the recognition of the importance and significance of this activity by the state, which could become an independent incentive for participation in this type of labor;

3) low fertility. In 2018, the total fertility rate in Russia was 1.58, which is 25% lower than the level of simple reproduction (Total Fertility Rate data 2019). The help of grandparents in raising grandchildren could encourage the family to have the next child and to reduce the breaks between the births of children. Both would contribute to the increase in the fertility rate in Russia.

CONCLUSIONS

The study led to the following conclusions. Firstly, the results, which were obtained on the basis of non-specialized (and therefore, possibly not fully valid) data, nevertheless, revealed specific features of the group of grandparents involved in daily childcare and formed the model of a typical grandmother actively involved in parental labour. This will become the basis for the development of a full-scale survey, which will allow us to investigate the features more fully and to model the behavior of this population group. Conducting this survey is extremely relevant, given the importance of increasing the size of this population group to solve the problems of reproduction of the Russian population.

A significant part of this survey will be devoted to the study of socio-psychological variables related to the motivation and other subjective factors of participation of grandparents in the parental labour. The use of such variables will allow us to create a more complete model of a typical grandmother actively involved in the process of parental labour.

In addition, we consider it appropriate to conduct further analysis on the basis of another methodological approach, the results of which will complement and potentially expand the previously obtained results. The

approach involves assessing the level of “active grandparenting” in different groups of grandmothers identified by various socio-psychological, demographic and economic variables. This will enable us to create a detailed model of a typical grandmother actively involved in the process of parental labour. Such approach will allow us to identify specific “growth potentials”- those groups of grandmothers for whom special incentive measures can be developed to increase their involvement in parental labour.

Secondly, the resulting model of a typical grandmother engaged in parental labour allows us to create a certain “portrait” of the opposite category - those who are not involved in this labour. This is a kind of target group for which special state measures can be developed aimed at enhancing the parental labour of grandparents.

Thirdly, we have shown that there are several problems in Russia that can be solved by expanding the circle of grandparents involved in parental labour. This seems possible through the introduction of state incentives for the labour related to the grandchildren care.

ACKNOWLEDGMENTS

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

REFERENCES

- Aisenbrey, S. and A. Fasang. 2017. “The interplay of work and family trajectories over the life course: Germany and the United States in comparison”. *American Journal of Sociology*, Vol 122(5), 1448-1484. DOI: 10.1086/691128
- Attar-Schwartz, S. and A. Buchanan. 2018. “Grandparenting and adolescent well-being: evidence from the UK and Israel”. *Contemporary Social Science*, Vol 13(2), 219-231.
- Availability of preschool education for children aged 1.5 to 3 data. Single inter-departmental information and statistical system (SIDIS). 2019. Rosstat, Moscow. URL: <https://fedstat.ru/indicator/59578> (access date 16.02.2020).
- Average amount of an appointed old-age pension data. Single inter-departmental information and statistical system (SIDIS). 2019. Rosstat, Moscow. URL: <https://fedstat.ru/indicator/31455> (access date 16.02.2020).
- Bagirova, A., Shubat, O. and V. Dorman. 2014. “Employees’ parental labor stimulation in Russian companies: socioeconomic view”. In *Proceedings of the 8th International Days of Statistics and Economics* (Prague, Czech Republic, September 11th-13th, 2014), 53-62. URL: <http://msed.vse.cz/msed/2014/article/238-Bagirova-Anna-paper.pdf> (access date 16.02.2020).
- Chapman, S.N., Pettay, J.E., Lahdenperä, M. and V.C. Lummaa V. C. 2018. “Grandmotherhood across the demographic transition”. *PLoS ONE*, 13(7), e0200963.
- Coall, D.A. and R. Hertwig. 2010. “Grandparental investment: Past, present and future”. *Behavioral and Brain Sciences*, Vol 33(1), 1-19.
- Coall, D.A., Hilbrand, S., Sear, R. and R. Hertwig. 2018. “Interdisciplinary perspectives on grandparental investment: a journey towards causality”. *Contemporary Social Science*, No. 13(2), 159-174.

- Comprehensive monitoring of living conditions. 2018. Moscow: Rosstat. URL: https://gks.ru/free_doc/new_site/KOUZ18/index.html (access date 16.02.2020).
- Country Comparison: Life Expectancy at Birth data. 2020. Central Intelligence Agency. URL: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html> (access date 16.02.2020).
- Country Comparison: Total Fertility Rate data. 2020. Central Intelligence Agency. URL: <https://www.cia.gov/library/publications/the-world-factbook/fields/356rank.html> (access date 16.02.2020).
- Daniels, A.K. 1987. "Invisible Work". *Social Problems*, No 34, 304-415.
- Danielsbasca, M. and A.O. Tanskanen. 2016. "The association between grandparental investment and grandparents' happiness in Finland". *Personal Relationships*, 23, 787-800.
- Del Boca, D., Piazzalunga, D. and C. Pronzato. 2018. "The role of grandparenting in early childcare and child outcomes". *Review of Economics of the Household*, Vol 16(2), 477-512.
- Erickson, R.J. 2005. "Why emotions work matters: Sex, gender, and the division of household labor". *Journal of Marriage and Family*, No 67, 337-351.
- Fligner, M. A. and G. E. Policello. 1981. "Robust rank procedures for the Behrens-Fisher problem". *Journal of the American Statistical Association*, 76(373), 162-168.
- Gibbons, J.D. and S. Chakraborti. 1991. "Comparisons of the Mann-Whitney, Student's t, and Alternate t Tests for Means of Normal Distributions". *The Journal of Experimental Education*, 59:3, 258-267, DOI: 10.1080/00220973.1991.10806565
- Hilbrand, S., Coall, D.A., Gerstorf, D. and R. Hertwig. 2017. "Caregiving within and beyond the family is associated with lower mortality for the caregiver: A prospective study". *Evolution and Human Behavior*, No. 38(3), 397-403.
- Hollander, M., Wolfe, D. A. and E. Chicken. 2013. "Nonparametric Statistical Methods". New York: John Wiley & Sons.
- Kim, H.-J., Kang, H. and M. Johnson-Motoyama. 2017. "The psychological well-being of grandparents who provide supplementary grandchild care: a systematic review". *Journal of Family Studies*, No. 23(1), 118-141.
- Mann H.B. and D.R. Whitney. 1947. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics*. 18(1), 50-60, doi:10.1214/aoms/1177730491.
- Mean age of mothers at childbearing (years). In *The Demographic Yearbook of Russia 2019*. Moscow: Rosstat. URL: https://gks.ru/bgd/regl/B19_16/Main.htm. (access date 16.02.2020).
- Mean age of mothers at childbearing (years). In *The Demographic Yearbook of Russia 2002*. Moscow: Rosstat. URL: https://gks.ru/bgd/regl/B02_16/IssWWW.exe/Stg/d010/i010330r.htm (access date 16.02.2020).
- Mood, A. M. 1954. "On the asymptotic efficiency of certain non-parametric 2-sample tests". *Annals of Mathematical Statistics*. 25(3), 514-522.
- National project "Demography". Official web-site of the Ministry of Labour of the Russian Federation. URL: <https://rosmintrud.ru/uploads/editor/dd/11/Ministry-0-1095-src-1545749719.7683.doc> (access date 16.02.2020).
- Nedelcu, M. 2017. "Transnational grandparenting in the digital age: mediated co-presence and childcare in the case of Romanian migrants in Switzerland and Canada". *European Journal of Ageing*, No. 14(4), 375-383.
- Oakley, A. 1974. "The sociology of housework". New York: Pantheon.
- Pedersen, D.E., Minnotte, K.L., Susan, E. and G. Kiger. 2011. "Exploring the relationship between types of family work and marital well-being". *Sociological Spectrum*, 31, 288-315.
- Sichimba, F., Mooya, H. and J. Mesman. 2017. "Predicting Zambian Grandmothers' Sensitivity Toward Their Grandchildren". *International Journal of Aging and Human Development*, No. 85(2), 185-203.
- Total Fertility Rate data. Single inter-departmental information and statistical system (SIDIS). 2019. Rosstat, Moscow. URL: <https://fedstat.ru/indicator/31517> (access date 16.02.2020).
- Veress, J. and A. Bagirova. 2018. "Parental Labour: Labour Market Acceptance Driven by Civil Activism". In *Proceedings of the 14th European Conference on Management, Leadership and Governance ECMLG 2018* (Utrecht, Netherlands, October 18th-19th, 2018), 307-314.
- Zar, J. H. 2014. "Biostatistical Analysis". Essex: Pearson Education Limited.
- Zimmerman, D.W. 1987. "Comparative Power of Student T Test and Mann-Whitney U Test for Unequal Sample Sizes and Variances". *The Journal of Experimental Education*, 55:3, 171-174, DOI: 10.1080/00220973.1987.10806451

AUTHOR BIOGRAPHIES

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/>

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/>

WHAT IS THE BEST WAY TO HELP? CENTRAL BANK STRATEGIES AND THE INTERBANK MARKET

Gábor Kürthy
Ágnes Vidovics-Dancs
János Száz
Péter Juhász
Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest Hungary
email: gabor.kurthy@uni-corvinus.hu

KEYWORDS

interbank market, central bank, liquidity, Markov process, asymmetric random walk

ABSTRACT

The paper proposes an analytical framework for addressing liquidity problems in cases where the expected value of the liquidity flow is zero, but negative deviations from the normal liquidity stance can trigger crises. Crises can be avoided by effectively signalling the temporary nature of the problem with the help of a central counterparty. The theoretical models and the numerical evaluations apply to the interbank market; however, the framework can also support the analysis of similar problems such as international payments or corporate liquidity.

INTRODUCTION

Every family, company, municipality, bank, and country has a liquidity situation, even if they do not monitor it. The specific reasons that change a liquidity position may be so varied and complex that it is advisable to treat liquidity developments as a random process for certain types of analysis. However, this is only for the convenience of modeling, and it is not a matter of general rule that blind luck controls the ability to meet instant payment obligations.

Three things can happen to the liquidity situation of a person or institution: it improves, worsens, or does not change. We will not complicate this in the first step. It does not matter if it gets a little better or significantly better if it gets better. The analysis is radically simplified by the assumption that successive changes are always caused by new impulses, that is, these changes are independent in time.

Therefore, the tool of our study will be a trinomial tree, a well-known species of Markov chains. Let 0 denote the state in which we have the required liquidity and categorize the liquidity surplus and deficit states into discrete classes. In the simplest case, the liquidity position remains unchanged with probability $p_m = 1/2$, improves with $p_u = 1/4$ and worsens with $p_d = 1/4$. We cannot go beyond the edges, hence, there is a $1/2$ chance of turning back or staying there.

It is more exciting to assume that the evolution of liquidity is a mean-reverting process. That is, increasing liquidity surpluses (deficits) increase the probability of a

downward (upward) movement. Furthermore, let us suppose that if we do not have the liquidity needed, we can get in trouble; the higher the liquidity shortage, the higher the probability of this trouble.

With all these assumptions, the random walk between liquidity states becomes not only mean-reverting but asymmetric as well. And now we can start asking the real questions.

Should A give a liquidity loan to B at the expense of its own liquidity? Does the network of mutual assistance improve the average liquidity situation in the system?

Another interesting question in this framework is, if there is a central counterparty (CCP) with a limited rescue power, are the big ones or the small ones worth saving? Are those in big trouble (who need more money) worth leaving to their fate? On which parameters depends the effectiveness of the rescue strategies?

Let us examine a place where liquidity surpluses and deficits meet literally every day. This is the interbank market. Let us model this much-studied market through the glasses of the analytical framework outlined above.

HISTORICAL BACKGROUND

It has been long known that central banks can contribute to financial stability by lending to banks that lack liquidity. When the Overend-Gurney Bank got into a liquidity crisis in 1866, the Bank of England (BoE), despite being asked to do so, was in no hurry to help. The consequence was a general, systemic liquidity crisis on the London money market. The lessons of the case were summarized by Walter Bagehot, offering a recipe to avoid similar future situations: the central bank should immediately provide a liquidity loan to a bank in need, with adequate collateral and high interest rate (Bagehot 1873). BoE, that had been existing for almost 200 years by that time, became a real central bank by embracing the Bagehot proposals.

In contrast, the foundation of the Fed was based precisely on the idea that situations of liquidity shortage and crisis, which were frequent due to the geographic structure and cyclicity of the economy and the banking system, could be prevented if there was a bank that could provide credit in time (Mehrling 2002).

A bank's liquidity crisis is particularly dangerous because it can cause an infection. Therefore, it is worth examining, on the one hand, what leads to a bank run, that is to say, depositors withdraw too much liquidity

from a particular bank, and, on the other hand, how it becomes a general panic when several players in the banking system are attacked simultaneously.

The decision of depositors depends on several factors according to the classical DD model (Diamond-Dybvig, 1983) and to its later versions (Bhattacharya-Gale 1987, Diamond-Rajan 2001, Freixas-Holthausen 2005). One such factor is what the depositor thinks about the quality of the bank's illiquid assets. If he suspects (or may know) that the market value of the assets is declining, he will withdraw the deposit. However, it is not only individual belief that counts. The depositor may know for sure that the bank's assets are excellent, if others do not think so, it is worth joining the crowd and deciding to liquidate. These considerations can lead to an extremely unfavorable Nash-equilibrium where the optimum strategy for all depositors is liquidation.

Regardless of the quality of the bank's assets, a crisis may occur. This happens when the number of impatient depositors (who want to consume earlier) increases. Here, too, the above equilibrium situation may occur: the choice of the time path of consumption is influenced by what the depositor thinks of the consumption decision of the others. If he believes that too many market players have become impatient, then he should not hesitate, because in this case, postponing the consumption may lead to the inability to consume at all because of the bank's crisis.

One of the suggestions to avoid similar situations is the introduction of deposit insurance. However, this raises several problems: on the one hand, with deposit insurance, depositors do not control the bank, and on the other, the bank tends to finance riskier projects (Anginer-Demirguc 2018). In addition, deposit insurance is in practice confined to retail depositors, while the onset of a bank crisis may be the result of decisions made by large investors such as money market funds. That is, the behavior of institutional investors can still lead to the unfavourable equilibrium of the DD model.

DD-like models have some critical assumptions that in practice may not be valid. Depositors, especially the small ones, are unable to value the bank's assets. The processes leading up to the 2008 crisis prove that even the official institutions (supervisors, credit rating agencies), the interbank market, and sometimes even the bank managers themselves are unable to do so. Given the decision already made by other players, the depositor can easily decide on liquidation, but an ex-ante strategic game between retail depositors is hardly imaginable.

If something is easily observable for depositors, it is the size of the bank's liquidity. The magnitude or proportion of this relative to the bank's liabilities may indicate to the depositor the likelihood of a crisis. However, since the depositor is unfamiliar with the operation of the bank, misinterpretation of the temporary liquidity shortage and the consequent withdrawal of the deposit may create a self-fulfilling liquidity problem. To prevent such situations, the bank must credibly signal to depositors that the deficit is only temporary.

The liquidity situation of depositors or of the bank can be modeled in several ways. The solution we have

chosen, according to which we describe the development of liquidity using a Markov process, is rare in the literature. In Deaton's (1991) model, consumers' labor income, and thus their current liquidity, is a Markov process. This can be used to explain the aggregate savings rate, but the model does not examine the impact on the liquidity situation of the banking system. Temzelides (1997) examines banking systems with different structures, where the strategy choice of clients - as a repeated game - results in a Markov process of the liquidity of banks. The model has the ability to reflect the liquidity crises experienced in U.S. banking history. In Csercsik and Kiss's (2018) study, the liquidity situation of depositors follows a Markov process, for which banks need to come up with an optimal strategy that keeps the probability of a bank run below a certain level.

The literature on the subject does not take into account that the central bank as an institution is not only a monetary authority but also a financial supervisor. This is not the case everywhere, but, especially since the 2008 crisis, on the one hand, the two institutions have been merged in several countries, and on the other hand, even if they have not been merged, there is much closer cooperation than before. This, in addition to mitigating systemic risks (Borio - White 2004), can help the central bank, as a lender of last resort (LLR), provide liquidity loans to solvent banks only. This has been for long a prerequisite for the LLR function (Bagehot 1873, Mehrling 2011), but it lacked the institutional background.

Even if the bank is solvent, it does not need to be rescued immediately and at all costs from a liquidity shortage (when there is no liquidity crisis yet, but there is a good chance for it) or from the crisis. That is, the central bank may adjust the decision to provide banks with liquidity to some strategy and not necessarily to the price offered (interest rate). The question is what should be this strategy if the central bank wants to discipline the market (constrained liquidity supply) and, at the same time, also wants to minimise the portion of banks in liquidity shortage and/or in a liquidity crisis.

MODELING ASSUMPTIONS

Other than those listed in the previous section, we are not aware of approaches that describe bank liquidity as a random walk. Thus, the assumptions below are arbitrary and will obviously need to be refined later.

Let the market size of the bank-accounts be 1, the share of the observed bank (henceforth B) is λ , the share of other banks is $(1 - \lambda)$. Suppose that there is a large number of clients, and they are evenly distributed among the banks. (That is, there are no special banks that keep accounts with e.g., public servants, car dealers, etc. So if the size of the public servants' accounts is x , then $\lambda * x$ of these accounts are at B).

The balance sheet of a bank looks like this:

R (liquity reserves)	D (deposits)
L (loans)	E (equity)

Capital (equity / net worth) is an important part of a bank's balance sheet, but it does not play a role in our model because we deal with liquidity.

There are two components of the liquidity reserve:

$$R = R_C + R_D$$

where

R_C - cash reserves

R_D - account at the central bank

We assume that the bank can exchange cash and central bank account money reserves without transaction costs. Customers can pay to each other with cash and with deposit accounts. With deposit payment, D and R_D decrease, with cash payment, D and R_C decrease. In the first case, the liquidity outflow of the buyer's bank equals the liquidity inflow of the seller's bank. This is not always the case when making a cash payment: the customer may just withdraw the cash and keep it; or make a purchase with it, but the seller does not deposit the full amount in his/her bank.

We assume that customers have some cash holdings, so there are two types of systemic shocks: net withdrawal and net cash depositing.

First, let us assume that p percentage of customers purchase in a given period; and suppose that if a cash purchase is made, the seller will immediately deposit the proceeds in her/his bank. As a result of the latter assumption, we do not need to distinguish between the two types of liquidity reserves.

In this case, the liquidity flow of bank B is as follows:

liquidity outflow: $p*\lambda*(1 - \lambda)$

($p*\lambda$ is spent by depositors of B, but due to the assumed even distribution of costumers ($p*\lambda$)* λ liquidity is returned).

liquidity inflow: $(p*(1 - \lambda))*\lambda$

(($p*(1 - \lambda)$) is spent by the depositors of other banks, λ times this amount flows into B)

Thus, the expected change in the size of the bank's liquidity is just zero.

Of course, this does not mean that the change is zero at all times. It is up to the bank's Treasury to analyze the further properties of this flow, that is, to evaluate its higher order statistical moments over time. For example, there may be a net liquidity outflow on the first 5 days of the week because customers withdraw cash that is spent at various stores. Sooner or later, these stores deposit the money in the bank. More cautious stores deposit at the end of the day. Other stores will wait until Friday. And there may be forgetful or careless stores that do not deposit even at the end of the week. In any case, it can be assumed that the more cash a store has, the more likely it deposits it. For the bank, this means that the greater the liquidity shortage, the more likely the deficit decreases.

Customers usually spend their money evenly in time (so one-fifth of customers does the shopping on Monday, one-fifth on Tuesday, and so on). But shopping can be hampered by anything, so there may be days when liquidity at the bank is increasing: cautious stores deposit their cash revenue (from impatient costumers of

other banks) at end of the day, patient or lazy costumers of the bank postpone their shopping. But the longer clients postpone, the more likely they withdraw money, and the bank's liquidity surplus is reduced.

So an average 5-day week, when everyone is behaving normally, looks like this:

	M	T	W	Th	F	sum
liquidity out	1	1	1	1	1	5
liquidity in	0	0	0	0	5	5
<i>liquidity flow</i>	-1	-1	-1	-1	4	0

Table 1: A possible liquidity flow of the bank

But if the store is careless, or if customers are impatient (spending more than the average), it can turn negative; on the other hand, if stores are cautious, customers are patient (or costumers of other banks are impatient), it can turn in a positive direction.

Independently of the process described, the bank may suffer a negative liquidity shock. We assume that customers are well informed and aware of the bank's current liquidity situation. And the worse the liquidity situation is, the more likely they are to decide to withdraw their deposits. It is important to understand that there is no reason to panic because liquidity sooner than later returns to the normal level, i.e., negative deviations are only temporary. However, customers are unfamiliar with the liquidity process, and they can only monitor the size of liquidity.

THE MODEL

The liquidity position of banks is modeled on a discrete probability field.

Let $\underline{y} = (v_{-(n+1)}, v_{-n}, \dots, v_0, \dots, v_n)$ denote the probability distribution of the bank's liquidity situation at a certain period, where the indices of the elements of the vector represent the liquidity status of the bank:

- in states $-n \leq j \leq n$ the size of the bank's liquidity is $n + j$:

the normal state is $j = 0$;

in states $j < 0$ the bank is in liquidity deficit

in states $j > 0$ the bank is in liquidity surplus.

- *the state $j = -(n + 1)$ is the liquidity crisis when the bank is unable to meet its short-term payment obligations.*

The bank randomly walks between these states for several periods. For the sake of simplicity, a period should be a day. During the first half of the day, the bank's costumers change the bank's liquidity status.

Ignoring the possibility of shocks, the bank jumps from state j to states $(j - 1)$, j , $(j + 1)$ with the probabilities pd_j , pm_j , pu_j respectively.

In line with the model assumptions, the bank knows the following about transition probabilities:

$$pu_i \geq pu_j \quad \text{if} \quad i < j$$

$$pd_i \geq pd_j \quad \text{if} \quad i > j$$

$$pu_n = 0$$

$$pd_{-n} = 0$$

The bank is aware of this part of the model and, therefore, cannot get into a crisis. However, it is unaware that customers may withdraw liquidity from the bank as a precaution, not for financing their purchases, but for prudential reasons. More precisely, the bank is aware of the possibility of shocks, but it has no means to calculate the probabilities. If it had, it could adjust its liquidity strategy accordingly.

The probability of liquidity withdrawal, that is, entering the $-(n+1)$ state in the j th state is ps_j , where

$$\begin{aligned} ps_j &= 0 & \text{if } j &\geq 0 \\ ps_j &> 0 & \text{if } j < 0 \\ ps_i &> ps_j & \text{if } i < j < 0 \end{aligned}$$

The interbank market opens in the second half of the day, where banks can lend or borrow money from each other or the central bank. The consequence of interbank activity is that the next day the shock probability changes, but the transition probabilities remain unchanged. The shock probability changes because customers decide to withdraw their deposits (the next day) depending on the observed size of the bank's liquidity at the beginning of the day. However, they do not know where this liquidity comes from. Clients decide on "normal" withdrawals and depositings based on their own situation, which is not influenced by the interbank market.

If a bank falls in a liquidity crisis during the day, it will not be able to apply for a loan on the interbank market, so it will stay in state $-(n+1)$. The next day, the central bank can rescue banks in crisis by restoring their liquidity status to normal (state 0). The probability of rescue is pr .

We first examine the random walk process, assuming no interbank lending. In this case, the transition probabilities are described by the following matrix.

$M \in R^{(2n+2) \times (2n+2)}$, where the rows and columns of M are numbered (similar to \underline{v}) from $-(n+1)$ to n . The element $m_{i,j}$ denotes the transition probability from state i to state j . Thus:

if $-n \leq j \leq n$:

$$\begin{aligned} m_{j,-(n+1)} &= ps_j \\ m_{j,j-1} &= (1 - ps_j) * p d_j \\ m_{j,j} &= (1 - ps_j) * p m_j \\ m_{j,j+1} &= (1 - ps_j) * p u_j \end{aligned}$$

furthermore:

$$\begin{aligned} m_{-(n+1),-(n+1)} &= 1 - pr \\ m_{-(n+1),0} &= pr \end{aligned}$$

Other elements of the matrix are zeros.

The dynamics of the model are described by the following equation:

$$\underline{v}^{t+1} = \underline{v}^t M \quad (1)$$

The steady state equilibrium of the model is the state vector \underline{v}^* for which:

$$\underline{v}^* = \underline{v}^* M \quad (2)$$

Since M is not positive, the equilibrium solution cannot be determined by eigenvalue search, but it can be found numerically.

We model the interbank market with the following constraint: banks that take liquidity loans are given exactly enough liquidity to have n units (to get into the state $j=0$), so the next day the bank is faced with $ps_0 = 0$ shock-probability.

The function $f: R^{2n+2} \rightarrow R^{2 \times (2n+2)}$ represents the interbank market, i.e.:

$$f(\underline{v}) = (\underline{v}_1, \underline{v}_2)^T,$$

where:

$$\underline{v}_1 + \underline{v}_2 = \underline{v}$$

and \underline{v}_1 represents banks who took out a liquidity loan, while \underline{v}_2 represents those who did not.

Since banks in the first group have zero probability of shock, so they face the \underline{M} transition matrix in the first half of the following period:

$$\underline{M} \in R^{(2n+2) \times (2n+2)}$$

if $-n \leq j \leq n$:

$$\begin{aligned} \underline{m}_{j,j-1} &= p d_j \\ \underline{m}_{j,j} &= p m_j \\ \underline{m}_{j,j+1} &= p u_j \end{aligned}$$

Other elements of the matrix are zeros.

The dynamics of the model is as follows. At the end of period t , the banks are in state $\underline{v}^t = \underline{v}_1^t + \underline{v}_2^t$. After the first half of the next day, the state vector is:

$$\underline{v}^{t+1} = \underline{v}_1^t \underline{M} + \underline{v}_2^t M \quad (3)$$

The equilibrium of the model is the probability vector \underline{v}^* for which:

$$\underline{v}^* = \underline{v}_1^* \underline{M} + \underline{v}_2^* M \quad (4a)$$

and

$$\underline{v}^* = \underline{v}_1^* + \underline{v}_2^* \quad (4b)$$

CENTRAL BANK STRATEGIES

Prior to the opening of the interbank market, the aggregate liquidity deficit and surplus are:

$$LD = \sum_{-n \leq j < 0} (-j) * v_j \quad (5)$$

$$LS = \sum_{j > 0} j * v_j \quad (6)$$

In the interbank market, it is assumed that the central bank acts as a central counterparty, i.e., the banks do not contract with each other directly. The central bank distributes liquidity according to a predefined procedure, which has two components:

- actual liquidity supply:

$$\overline{LS} = \min[LS, (1 - c) * LD] \quad (7)$$

where $0 \leq c < 1$ denotes the cut constraining the available liquidity. The size of the cut expresses central bank discipline and flexibility.

- the principle of liquidity allocation implemented by the central bank according to the f_S strategy functions.

As in the basic model:

$$f_S(\underline{v}) = (\underline{v}_1, \underline{v}_2)^T, \text{ and}$$

$$\underline{v}_1 + \underline{v}_2 = \underline{v}$$

At a given level of central bank discipline / flexibility (c), we examine four strategies (S).

$S = 0$ - in this case, the central bank does not lend at all, i.e.: $\underline{v}_1 = \underline{0}$. This gives us the basic model without the interbank market.

$S = 1$ - in this case, the central bank randomly selects a \underline{v}_1 vector that satisfies:

$$\sum_{-n \leq j < 0} (-j * v_{1,j}) = \overline{L\overline{S}}$$

$S = 2$ (LTH - Low To High) in this case, banks with low liquidity shortages will receive loans first, i.e.:

$$v_{1,-1} = \min[\overline{L\overline{S}}, v_{-1}], \text{ and for } -n \leq j < -1$$

$$v_{1,j} = \min\left[\frac{-1}{j}\{\overline{L\overline{S}} - \sum_{j < i \leq -1} (-i * v_{1,i})\}, v_j\right]$$

$S = 3$ (HTL - High To Low) in this case banks with high liquidity shortages will receive loans first, ie:

$$v_{1,-n} = \min\left[\frac{1}{n}\overline{L\overline{S}}, v_{-n}\right], \text{ and for } -n < j \leq -1$$

$$v_{1,j} = \min\left[\frac{-1}{j}\{\overline{L\overline{S}} - \sum_{i < j} (-i * v_{1,i})\}, v_j\right]$$

NUMERICAL SOLUTIONS

To evaluate the basic model and the central bank strategies, let $n = 5$, and the probabilities be as follows.

	-5	-4	-3	-2	-1	0	1	2	3	4	5
p_s	0.9	0.7	0.5	0.3	0.1	0	0	0	0	0	0
p_d	0	0.25	0.25	0.25	0.25	0.25	0.3	0.35	0.4	0.45	0.5
p_m	0.5	0.3	0.35	0.4	0.45	0.5	0.45	0.4	0.35	0.3	0.5
p_u	0.5	0.45	0.4	0.35	0.3	0.25	0.25	0.25	0.25	0.25	0

Table 2: Transition probabilities of the model

The probability of rescuing banks in crisis is $p_r = 0.1$

The results shown by Figure 1 can be interpreted in two ways:

- the individual bank will find itself in the various liquidity situations over the long-term with the probabilities shown in the figure. The expected liquidity of a bank based on probabilities belonging to fields $-5, -4, \dots, 5$ is 4.52, which is lower than the ideal 5 - this is caused by negative shocks.
- the participants of the banking system are distributed among the liquidity conditions in the long run according to the probabilities shown in the figure.

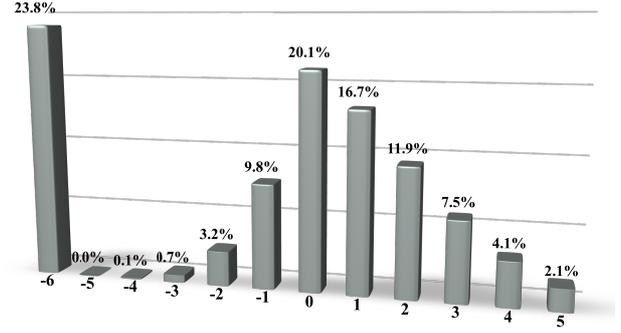


Figure 1: Steady State Equilibrium of the Basic Model

Banks in a liquidity crisis are rescued (set to 0) by a third party in the base model with a 10% probability. In practice, this is usually the central bank and/or the government. The capacity of the central bank in the event of a liquidity crisis is unlimited, i.e., it can decide to rescue all banks in crisis immediately ($p_r = 1$), but it may also have the attitude of never saving anyone ($p_r = 0$) like the pre-Bagehot Bank of England. Depending on the probability of rescue, the proportion of banks in crisis (or more precisely, the equilibrium probability of the crisis) is shown in Figure 2. From this, it seems that the central bank does not have to be "overly" committed. On the one hand, even if it tries to save everyone right away, there will always be banks in crisis because they have just got there. (In the model, rescues occur in t , for banks that got in trouble in $(t-1)$ or earlier. But in t , newer banks get into this state.) On the other hand, the flattening of the curve shows that the central bank is devoting ever greater resources to monitoring the market as a whole in vain, because after a while it can barely reduce the proportion of banks in crisis. Third, the more the central bank is willing to bail out the banks, the less they will pay attention to their liquidity situation, thus moral hazard is increasing. (The latter two factors, i.e., central bank cost and moral hazard are not included in the model).

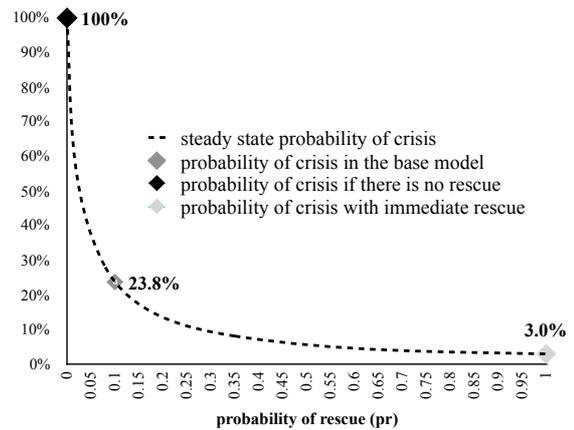


Figure 2: Steady State Probability of Crisis as a Function of the Probability of Rescue

Before comparing the different central bank strategies, let us look at the results of the following simulation. In the first case (Figure 3), three banks are randomly walking over 100 periods, and there is no interbank

market. In the second case (Figure 4), there is an interbank market, where, as described in the model, banks with liquidity shortage receive overnight credit. It is clear that in the latter case, the volatility of the size of (individual) liquidity is much lower.

A complete evaluation of central bank strategies would involve an examination of the statistical moments of liquidity flows. However, this goes beyond the scope of the study. Besides, our goal is not to choose the best strategy, but to demonstrate that the model presented may be suitable for this. Therefore, in the following, we focus only on the edge of the equilibrium distribution.

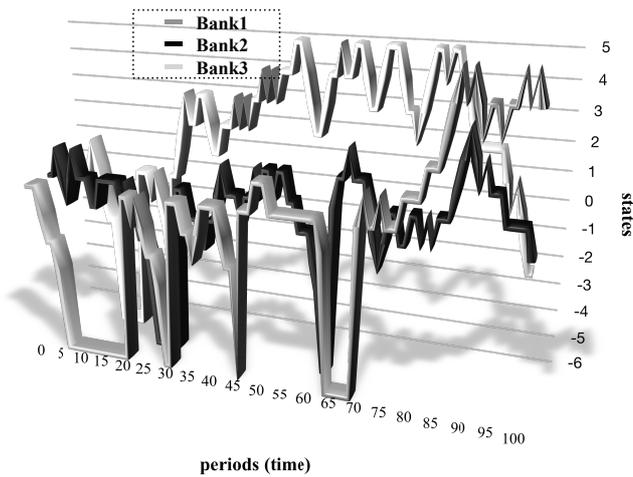


Figure 3: Random Walk without Interbank Market

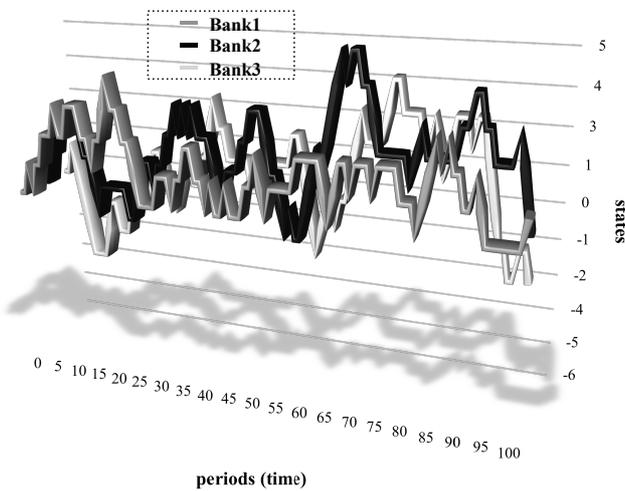


Figure 4: Random Walk with Interbank Market.

Figure 5 shows the equilibrium crisis probability at a given level of the central bank rigor. The figure has several lessons:

If the central bank does not apply a cut ($c = 0$), then regardless of its strategy, there will be no bank in crisis in the long run. The explanation for this is as follows. Whichever field the bank is on (even $j=-6$), the probability of getting to $j=0$ (in a reasonably long time) is 1. For a group of banks that reach state zero, the

expected liquidity demand is the same as the expected liquidity supply since transition probabilities are symmetric around this state. Thus (since $c = 0$), the shock probability is zero for the members of this group. And since all banks are expected to touch the zero field in the long run, the equilibrium crisis-probability is zero.

By operating a liquidity market, a lower equilibrium crisis-probability can be achieved than by increasing the chances of a rescue by the central bank (pr). We have seen earlier that even in the case of immediate rescues, the share of banks in crisis cannot be reduced below 3%. That is, banks are better off with nets (access to liquidity, central bank guarantees, IMF credit line, etc.) than with fish (rescues in times of crisis).

Out of the three possible liquidity-providing strategies, with the central bank's increasing tightening (cut), the HTL strategy seems to be the best. However, with a low cut, the LTH strategy performs better. It also appears that the first strategy, when the central bank randomly distributes liquidity, sometimes dominates the other two. The result of the first strategy is actually partly the result of a simulation, since the central bank (the program that runs the simulation) randomly chooses the vectors that follow the rules of the strategy. Therefore, the realisation of this may be different from what is shown in the figure: if the central bank distributes liquidity randomly, depending on what is pulled out of the hat, the equilibrium distribution (including the likelihood of an equilibrium crisis) will be different.

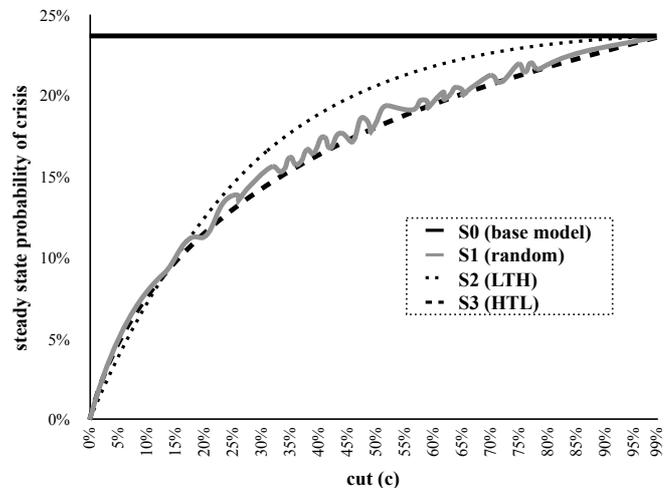


Figure 5: Equilibrium Crisis-Probability in Case of Different Central Bank Strategies

Given the shock probabilities, it would be possible to calculate, in a given market situation (distribution), which liquidity allocation would lead to the best result, i.e., to the least equilibrium crisis-probability. However, the practical implementation of such a sophisticated strategy would be too demanding for the central bank since it would need to know exactly the probabilities of shock. If the central bank does not have this information and only knows (which is plausible) that banks with lower liquidity have a higher probability of shock, it should choose Strategy 3 (HTL).

CONCLUDING REMARKS

The previous section demonstrated, through a rather limited example, that the model may be suitable for analysing liquidity problems. The study can be extended to other areas. Small open economies that use foreign currency for their international transactions are facing the similar problem as banks do. The external balance of the economy - at best - fluctuates around zero, but in times when the external deficit is reducing the country's international reserves, panic among investors and sudden capital outflows can cause payment difficulties. It may be helpful if someone (such as the IMF) is ready to provide liquidity loans. The same can be said for companies where some subsidiaries are experiencing temporary payment difficulties from time to time. In this case, it may be handy if other affiliates or the parent company help, even by creating a cash pool.

The model can be refined to tell more about reality. There are three possible directions. With the introduction of players of different sizes, much more can be explored about the CCP's strategy choices. In this case, the question is not only what is the appropriate strategy based on the size of the individual liquidity shortage, but also whether e.g., can someone be big enough to be saved anyway. As we have pointed out, choosing the right strategy can be based on accurate calculations, provided that the probabilities of shock are known. Similarly, proper monitoring of the market can reduce the number of actors in the crisis to a sufficiently low level. But in practice, all this comes at a cost. It may be worth comparing these costs with the benefits of a well-chosen liquidity allocation. That is, rather than through some property of the equilibrium distribution optimisation could be done by introducing some sort of social utility function into the model. Last but not least, the model is currently not capable of demonstrating infections. This would require that the elements of the transition matrix are influenced by the current liquidity position of the participants.

REFERENCES

- Anginer, D. - Demircuc-Kunt, A. 2018. "Bank Runs and Moral Hazard : A Review of Deposit Insurance". Policy Research working paper; No. WPS 8589. Washington, D.C. : World Bank Group.
<http://documents.worldbank.org/curated/en/548031537377082747/Bank-Runs-and-Moral-Hazard-A-Review-of-Deposit-Insurance>
- Bagehot, W. 1873. "Lombard Street: A Description of the Money Market". Henry S. King & Co., London.
- Bhattacharya, S. - Gale, D. 1987. "Preference Shocks, Liquidity and Central Bank Policy". *New Approaches to Monetary Economics*, 69–88. doi:10.1017/cbo9780511759628.005
- Borio, C. - White, W. R. 2004. "Whither Monetary and Financial Stability? The Implications of Evolving Policy Regimes". BIS Working Paper No. 147.
<https://ssrn.com/abstract=901387> or <http://dx.doi.org/10.2139/ssrn.901387>
- Csercsik, D. - Kiss, H. J. 2018. "Optimal Payments to Connected Depositors in Turbulent Times: A Markov Chain Approach" *Complexity*. doi: <https://doi.org/10.1155/2018/9434608>
- Deaton, A. 1991. "Saving and Liquidity Constraint." *Econometrica*, Vol. 59, No. 5: 1221-1248
- Diamond, D. W. - Dybvig, P. H. 1983. "Bank Runs, Deposit Insurance, and Liquidity". *Journal of Political Economy*, Vol. 93, No. 3: 401-419
- Diamond, D. W. - Rajan, G. R. 2001. "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking." *Journal of Political Economy*, Vol. 109, No. 2: 287-327
- Freixas, X. - Holthausen, C. 2007. "Interbank Market Integration Under Asymmetric Information". *Review of Financial Studies*, Vol. 18, No. 2: 459-490
- Mehrling, P. 2002. "Retrospectives: Economists and the Fed: Beginnings". *Journal of Economic Perspectives*, Vol. 16, No. 4: 207-218
- Mehrling, P. 2011. "The New Lombard Street: How the Fed Became the Dealer of Last Resort". Princeton: Princeton University Press, 2011
- Temzelides, T. 1997. "Evolution, Coordination and Banking Panics" *Journal of Monetary Economics*, Vol 40, No. 1: 163-183

AUTHOR BIOGRAPHIES

PÉTER JUHÁSZ works as an Associate Professor for the Department of Finance at Corvinus University of Budapest (CUB). He is a CFA charterholder and holds a PhD from CUB. His research topics include business valuation, financial modelling, and performance analysis.

His e-mail address is: peter.juhasz@uni-corvinus.hu

GÁBOR KÜRTHY, PhD, is an associate professor and the head of the Department of Money, Banking and Public Finance at Corvinus University of Budapest. His main research areas are theory of money, banking and external imbalances.

His e-mail address is: gabor.kurthy@uni-corvinus.hu

JÁNOS SZÁZ, CSc is a full professor at the Department of Department of Money, Banking and Public Finance at Corvinus University of Budapest. He was the first academic director and then president of the International Training Center for Bankers in Budapest. Formerly he was the dean of the Faculty of Economics at Corvinus University of Budapest and President of the Budapest Stock Exchange. Currently, his main field of research is financing corporate growth when interest rates are stochastic.

His e-mail address is: janos.szaz@uni-corvinus.hu

ÁGNES VIDOVICS-DANCS, PhD, CIIA is an associate professor at the Department of Department of Money, Banking and Public Finance at Corvinus University of Budapest. Her main research areas are government debt management in general and especially sovereign crises and defaults. She worked as a junior risk manager in the Hungarian Government Debt Management Agency in 2005-2006. Since 2015, she is the chief risk officer of a Hungarian asset management company.

Her e-mail address is: agnes.dancs@uni-corvinus.hu

CLUSTERING EU COUNTRIES BASED ON DEATH PROBABILITIES

Kolos Csaba Ágoston

Institute of Mathematics and Statistical Modelling
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
Centre for Economic and Regional Studies
Tóth Kálmán u. 4. Budapest 1097, Hungary
E-mail: kolos.agoston@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

Mortality, Clustering, Death Probabilities

ABSTRACT

Background Our research is conducted to identify certain grouping of 24 European countries based on their death probabilities. Gathering 2014 data from Human Mortality Database our research *objective* was twofold. First, we wanted to find homogeneous groups of countries where mortality is similar and for a financial institution they could be grouped as risk communities. Second, we wanted to identify the optimal number of groups as a basis for strategy making. Two different clustering *methods* were used in our research, k-means and k-median clustering. We applied asymmetric measure (QDEV) in k-median method to handle the differences in country sizes and age groups. Our *results* are stable but different in k=3 clusters, k-means clustering resulted in a big Western-European cluster and two small-medium Eastern groups; however, k-median clustering gave a homogeneous Eastern group and besides a bigger Western cluster Spain, Italy, and France formed a separated group of countries.

INTRODUCTION

In our globalizing world it is highly important for all industries to shift from individual customer service to grouped solutions. Defining homogeneous groups of customers decreases production and service costs of companies and it could create value for the customers at the same time. Global insurance companies and financial institutions active in several countries might be aware of life expectancy and death probability differences across countries in order to decrease their longevity risk. Nevertheless, insurance companies place great emphasis on the establishment of homogeneous risk communities. The heterogeneity of the insureds is well known and is further amplified by the phenomenon of anti-selection. Insurance companies aim to establish homogeneous risk communities, however in terms of the price calculation it is not favorable if some risk communities get too fragmented. The allocation of the heterogeneous insureds into (somewhat) homogeneous groups is called risk classification in the literature (see Crocker and Snow (1986)). Risk classification might be the most often used actuarial method that can be

supported by the adequately chosen cluster analytical methods.

In the case of life insurances, some factors of heterogeneity have been known for centuries, for example the difference between the male and female death probabilities. In the last decades deeper analyses had also been revealed – thanks to the evolution of computing technologies. The difference between the mortality pattern of the white and blue collars became general knowledge, and the difference in death probabilities based on educational attainment is increasingly recognized. The territorial diversities are also more and more obvious. Kovács and Vaskövi (2019) also used cluster analysis to group European countries base on their life expectancy and retirement age patterns. In this paper we investigate the national differences of unisex death probabilities in 24 European countries based on their 2014 data and give possible classification using different clustering methods. We include death probabilities separately for former East and West-Germany to identify possible remaining differences.

DATA: DEATH PROBABILITIES OF EU COUNTRIES

Individual longevity is uncertain, thus the length of human lifetime can be described by a random variable. Although there were specific attempts to describe the distribution of this variable with a functional form (see Marshall and Olkin (2007)), sufficient result is still missing. Instead, age specific death probabilities are calculated: q_x gives the probabilities that a living x -year-old person will die within a year. These rates can be calculated based on institutional (insurance companies or pension funds) data or based on nationwide statistics. The estimation process differs in some extent for the two cases: for institutional data mostly the Kaplan-Meier method is used; however, for nationwide statistics the so called Lexis-diagram is used. Both methods belong to nonparametric statistical methods, i.e. neither Kaplan-Meier method nor Lexis diagram assumes data fits normal or any well-understood distributions.

We used unisex crude death probabilities (a certain averaging is used to calculate unisex rates from male and female death probabilities) for EU countries available in Human Mortality Database (HMD). In Table 1 unisex death probabilities in year 2014 of 3 chosen countries

are shown in every 5 years (former East and West-Germany are described separately):

Table 1: Death probabilities (q_x) of 3 European countries from age 1 to 110 in 2014

years	AUT	DE-E	DE-W	HUN
1	0.00019	0.00028	0.00024	0.00036
5	0.00009	0.00007	0.00006	0.00011
10	0.00009	0.00007	0.00005	0.00011
20	0.00038	0.0004	0.0003	0.00038
30	0.00042	0.00049	0.00041	0.00051
40	0.00096	0.00105	0.00085	0.00161
50	0.00257	0.00345	0.00266	0.00588
60	0.00699	0.00828	0.0074	0.01485
70	0.01626	0.01667	0.01682	0.02904
80	0.04379	0.04816	0.04536	0.06906
90	0.14823	0.1515	0.15042	0.17988
100	0.3746	0.3684	0.37224	0.37025
110	1.0000	1.0000	1.0000	1.0000

For example, bold figures in Table 1 mean the probability that an East-German individual at the age of 50 dies within one year is 0.35% and at the age of 90 is 15.15%. Crude death probabilities are significantly different in ages, moreover for smaller countries we face particular ages without deaths. For this reason, the crude death probabilities are smoothed (Ágoston, 2003). Smoothing method differ from country to country, for child ages and for young adults a polynomial function (with high degree) is fitted or some kind of moving average method is applied. For adults and old ages crude probabilities often smoothed based on Gompertz-Makeham (Gompertz, 1825 and Makeham, 1867) mortality law.

Standardized data is used in order to reduce the dispersion of elderly death probabilities since the latter can be a magnitude higher than the ones in younger ages. If we did not standardize data, the clusters would only be formed based on the elderly mortality.

Infant death probabilities are omitted from the database considering three main reasons, (i) infant mortality (probability of death in the age group 0 to 1) is significantly higher than it is in other child age groups, (ii) the intrauterine death is not clearly classified, (iii) the death probability of children between 0 to 1 years has changed significantly in the last decades and this change was not in line with the change of any other age groups.

In the next section we describe the clustering methodologies used to classify EU countries based on their death probabilities. Vékás (2019) found empirical evidence of changing mortality curves in European countries, and in our paper we also attempt to identify mortality patterns among the examined countries.

METHODOLOGY

Cluster analytical methods on mortality data appear in Ágoston, Majstorović and Vaskövi (2019). They used parametric methods where first the Makeham mortality law (Makeham, 1867) is fitted to the data then the clustering method is applied on the fitted survival curves. The Makeham mortality law fits well on data of age groups 30 to 100 years; however, in this paper we wanted to analyze also death probabilities of younger ages, thus crude death probabilities were clustered here. K-means and k-median clustering methods were applied on data and in k-median method we used an asymmetric similarity measure ($QDEV$) described by Arató et al. (2009).

K-means Method

One of the first, and most popular clustering method until today is the k-means method (McQueen, 1967). The method can classify the observations in such a way, that the squared sum of distances in the groups are minimal. The method is very fast but highly dependent on the selection of initial cluster center (it is a computed center not a real data), thus it can be viewed as a heuristic approach.

If we use clustering method based on death probabilities the most straightforward way is that we take the death probabilities for ages $1..N$, we consider it as an N dimensional vector in the N dimensional space and use a k-means algorithm for these vectors.

Using k-means method when (re)calculated the cluster centers, the *mean* of death probabilities is calculated. It can happen that we simply take the average of Germany and Luxemburg, although German data is based on 80.77 million people and Luxembourgian data is based on 0.55 million people. In these specific cases where the size of countries is very much different, the mean would be problematic. But we can move further: when we calculate the distance between cluster center and instances we use Euclidean distance, i.e. all ages play the same important role although the sample size can differ greatly from ages to ages; old age probabilities are calculated based on significantly less data than middle age probabilities.

There would be a straightforward option to give weights for ages. The problem is if the sample size (even its relative value) differs for countries, it would not be fortunate to use the same weights for a given age for all countries. Another approach is to specify a distance or similarity measure that can consider the number of entities in an age group. Arató et al. (2009) define three different similarity measures for life tables, for our purpose the $QDEV$ measure is relevant. We consider two countries (a and b) and the similarity between them is defined by:

$$QDEV = \sum_{i=0}^{100} \frac{e_i^a (q_i^a - q_i^b)^2}{q_i^b} \quad (1)$$

where e is the so called exposure (the time while an x -year-old person was alive, often quite close to the number of individuals alive). In expression (1) the term $(q_i^a - q_i^b)^2$ is squared therefore the major difference between death probabilities has significantly higher importance than many small differences together. The previous term is normalized by probability of country b , meaning greater difference can be tolerated if the probability itself is a higher value, and multiplied by the exposure, meaning that for small communities even high differences can be tolerated.

We can see that expression (1) is not symmetric in a and b . In cluster analysis similarity measures are usually symmetric but we can find examples for asymmetric measures, as well (see Okada, 2000). We have two options: keep the measure asymmetric and chose a method which can handle asymmetric measure or symmetrize it somehow (for instance the average of the two direction). We tried both ways but in calculations of this paper kept the measure asymmetric.

For asymmetric measure it is not straightforward how cluster centers should be calculated in k-means method (even if we symmetrize the QDEV measure, it is still problematic). Although Olszewski (Olszewski 2011) gives a modified k-means algorithm for asymmetric measures, the algorithm was not tested on real data; therefore, we suggest to use k-median clustering method that is suitable for the asymmetric QDEV similarity measure.

K-median Method

The so-called k-median (or also p-median) problem was already researched at the dawn of the appearance of cluster analytical methods. The idea was, that if we minimize the absolute deviation instead of the squared sum, then in case of a one-dimensional problem the cluster centers will be data points (or at least can be chosen to be data point). Unlike the average, the median is not defined in multi-dimensional space but the name stayed with the method. In case of a multi-dimensional k-median problem a distance or similarity matrix calculated first, and the task is to decide which data points should be cluster centers and which cluster center and data point should be matched to in order to minimize the total distance from the centers.

We calculated the QDEV equation for every possible country pairs and a similarity matrix is produced. This matrix is the input for the k-median method.

RESULTS

Results of K-means Method

If the number of clusters is 2, we get reasonable but trivial groups: Central-European and Baltic countries form the first cluster, Western countries forms the second. k=2 clustering is stable, while all 10,000 runs with random initial cluster center gave the same result of 1326.15 between groups distance.

Table 2: k-means clustering with 2 clusters

Cluster Id	Countries in the cluster
1	Czech Republic, Slovakia, Hungary, Poland, Estonia, Lithuania, Latvia, Bulgaria, Croatia,
2	Austria, East-Germany, West-Germany, Belgium, Netherlands, Luxemburg, France, Great-Britain, Ireland Denmark, Finland, Sweden, Portugal, Spain, Italy, Slovenia

Slovenia belongs to the second cluster together with all developed countries; however, the three Baltic countries are grouped to the eastern block. This clustering suggests that rapid economic development of Baltic countries was not accompanied by a significant improvement of demographic processes. Based on clustering of 2014 death probabilities we did not find difference between East and West-Germany.

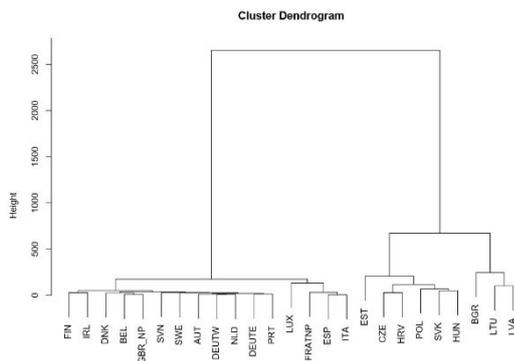
When we raised the number of clusters up to 3 then the first cluster is divided into two separated groups where Bulgaria, Lithuania and Latvia form one cluster and the remaining 6 countries from the first cluster form the second. Clusters are still very stable (6,017 out of 10,000 runs come to the same between groups distance of 1,661.922).

We raised the number of clusters up to k=6 to gain more detailed grouping. In Table 3 clusters of countries are shown. Raising the number of clusters up to 4 Bulgaria would form an individual group; then Estonia leaves its cluster. The biggest cluster of Western countries comes apart only at k=6 where Spain, Italy, France, and Luxemburg form the sixth cluster. East and West-Germany remain in the same cluster meaning there is no significant difference between death probabilities of the two parts of the country.

Table 3: k-means clustering with 4, 5 and 6 clusters

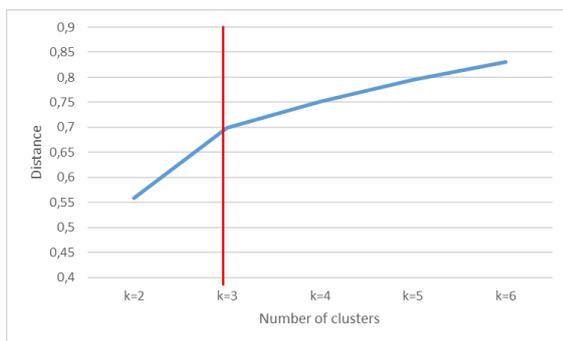
Cluster Id	k=4	k=5	k=6
1	BGR	BGR	BGR
2	LTU, LVA	LTU, LVA	LTU, LVA
3	CZE, EST, SVK, HRV, HUN, POL	CZE, SVK, HRV, HUN, POL	CZE, SVK, HRV, HUN, POL
4	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, ESP, ITA, LUX, FRA	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, ESP, ITA, LUX, FRA	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT
5		EST	EST
6			ESP, ITA, LUX, FRA

Since similarity matrix is applied in k-means clustering we could also assign hierarchical clustering method on our data. Figures 1 shows the clustering steps and distances of the 24 countries visualizing the results of the k=6 k-means clustering by a dendrogram (Ward method). The 12 West-European countries forming cluster No 4 are the most similar with the minimum dissimilarity measure. Cluster No 6 is formed by 4 countries (Luxemburg, Italy, Spain, and France) joining cluster No 4 in the second step. The East-European countries form two main groups (cluster No 2 and 3); however, Estonia and Bulgaria have significantly different death probabilities both forming one distinct group (cluster No 1 and 5).



Figures 1: Dendrogram of hierarchical clustering, Ward method

We also applied cluster elbow method which is a heuristic way to define the optimal number of clusters. The variance explained is calculated as a ratio of between-group-VAR and total-VAR and this percentage is plotted. Increasing the number of clusters would raise the variance explained; however, the marginal gain is flattening. Where the slope of the curve decreases there is the “cluster elbow”, i.e. the optimal number of clusters. Figures 2 shows the cluster elbow of k-means clustering equals to 3.



Figures 2: Cluster elbow diagram of k-means clustering

Over k=4 the k-means clustering could not be considered stable since the random cluster centers result in wide variety of between groups distances.

Results of K-median Method with QDEV Similarity Measure

In k-median method the cluster centers are real data points, in our case they are countries that the most typical countries are in the cluster.

When k=2 the same clusters were produced as with k-means method, i.e. blocks of 15 Western and 9 Eastern European countries. Belgium is the cluster center (the most typical country) in West-Europe and Slovakia in East-Europe.

In case of k=3 the Western-block is divided into two smaller groups where Spain, Italy and France formed a new group of countries. The same clustering result is to be identified at higher number of clusters in k-means method (k=6). Belgium as cluster center in West-Europe is replaced by (former) West-Germany, and in the new cluster Spain is pointed as cluster center. Slovakia remained the center of East-European countries. From this cluster number, we could observe that big countries become cluster centers.

In Table 4 results of $4 \leq k \leq 6$ clustering are summarized. Countries indicated bold and underlined are the cluster centers of each group. At k=4 the eastern-block is divided; Hungary, Bulgaria, Lithuania, and Latvia form cluster No 1. In k-median clustering, Hungary does not connected to other Visegrad countries but to Bulgaria and the Baltic countries. When we further increased the number of clusters, Eastern blocks were not changed, but France and Sweden were moved from their cluster. The former East and West-Germany remain in the same cluster meaning there is no difference of death probabilities in the two halves of the German country.

Table 4: k-median clustering for $3 < k \leq 6$ clusters

Cluster Id	k=4	k=5	k=6
1	HUN, <u>BGR</u> LTU, LVA	HUN, <u>BGR</u> LTU, LVA	HUN, <u>BGR</u> LTU, LVA
2	CZE, <u>POL</u> , EST, SVK, HRV	CZE, <u>POL</u> , EST, SVK, HRV	CZE, <u>POL</u> , EST, SVK, HRV
3	AUT, DEUE, <u>DEUW</u> , BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, LUX	AUT, DEUE, <u>DEUW</u> , BEL, NLD, DNK, FIN, GBR, IRL, SVN, PRT, LUX	AUT, DEUE, <u>DEUW</u> , BEL, NLD, DNK, PRT, SVN,
4	<u>ESP</u> , ITA, FRA	ESP, <u>ITA</u> , SWE	ESP, <u>ITA</u> , SWE
5		<u>FRA</u>	<u>FRA</u>
6			FIN, LUX, <u>GBR</u> , IRL

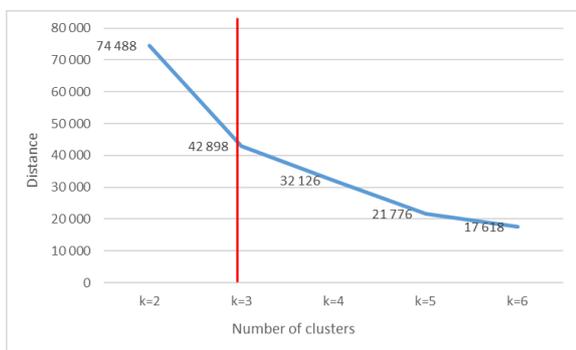
In the case of a k-median problem (or also by k-means problem), the number of clusters is an input parameter; therefore, selection of the optimal cluster number must

be part of the analysis. In k-median method the objective function is calculated which is the sum of within-group distance measures, shown in Table 5.

Table 5: values of objective functions at k-median

	Objective functions
k = 2	74,487.61
k = 3	42,897.50
k = 4	32,125.75
k = 5	21,775.68
k = 6	17,617.9

Figures 3 shows the optimal cluster number of k-median clustering using the values of objective functions.

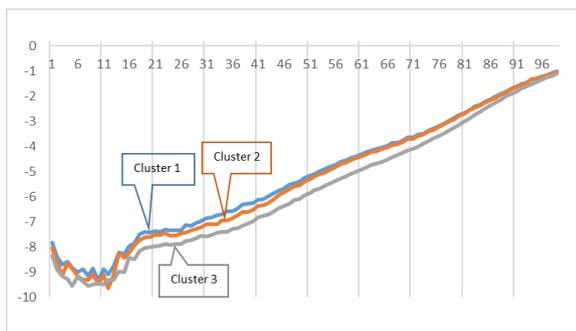


Figures 3: Cluster elbow diagram with objective functions of k-median clustering

Comparison of K-means and K-median Results

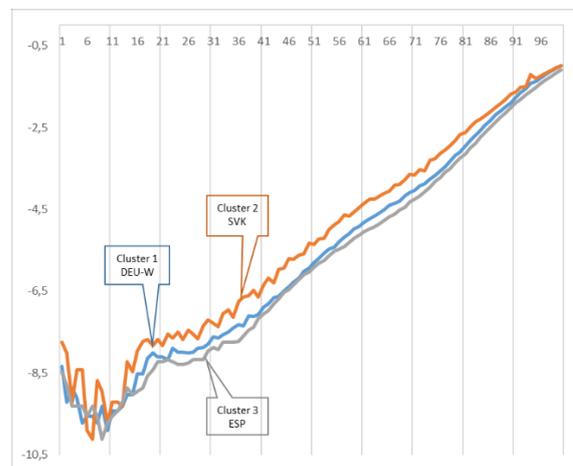
In k-median method the cluster centers are real data points, in our case these are the most typical countries of each cluster. On the contrary, cluster centers in k-means method are computed centers and very rare coincide with certain data point.

Figures 3 shows the logarithmic death probabilities of k-means clusters (k=3) on a log scale. Cluster 1 (Bulgaria, Lithuania and Latvia) has the highest death probabilities in most of the age groups; however, from age group 70 the death probability curves of Cluster 1 and 2 overlap. Cluster 3 (15 Western-European countries) has the lowest age-specific probability of death for the total 1-100 years. There is one age group (12-13 years) where the probability of death is lower in cluster 2 (Eastern countries) than is cluster 3.



Figures 4: logarithmic death probabilities for age groups, k-means clustering (k=3)

We explained in the previous section that using k-median clustering method the optimal number of clusters is again three. Thus, on Figures 5 we represent logarithmic death probabilities of k-median method of 3 clusters where the most typical county of each cluster (the cluster center) is shown.



Figures 5: death probabilities for age groups, k-median clustering (k=3)

From age 15 cluster No 3 (Spain, Italy, and France) has the lowest death probabilities; however, in younger ages its top position is not prevalent. The variance in Cluster 2 (East European countries) in younger ages is outstanding.

CONCLUSIONS

In our research we investigated the 2014 death probabilities of 24 European countries and attempted to group them by specific clustering methods. K-means clustering was applied where k=2 gave trivial result, i.e. East and West Europe were differentiated. We increased the cluster number and found that Bulgaria and Estonia are significantly different from other countries of the Eastern block. The Western countries in our research are mainly similar; nevertheless, Spain, Italy, France, and Luxemburg set up a new cluster at k=6 clustering.

We found that k-means method could be problematic while in (re)calculation of cluster centers the mean of death probabilities is applied. Since we have significantly different country sizes and also age groups in each country we suggest to use QDEV as an asymmetric similarity measure. K-means method is not suitable to be applied on asymmetric measure; therefore, k-median clustering method was also applied on our dataset. This method gave us exact solution even for higher cluster numbers. K-median clustering drove to the same result as k-means at k = 2 clusters but concerning 3 ≤ k ≤ 6 results are fairly different. K-median clustering rather divided the group of Western European countries, while k-means separated the Eastern block. We examined former East and West-

Germany separately to identify potential differences left, but we did not find any (even with higher cluster numbers East and West Germany stayed in the same cluster).

For further possible research we might analyze longitudinal changes in death probabilities and find different mortality patterns in the investigated 24 European countries.

ACKNOWLEDGEMENTS

This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00017, titled "Sustainable, intelligent, and inclusive regional and city models".

REFERENCES

- Ágoston, K. Cs. (2003). Death rates and their estimation (in Hungarian). In Banyár, J.: Life insurance pp. 377–390. Aula, Budapest.
- Ágoston, K. Cs. and Majstorović, S. and Vaskövi, Á. 2019. "Spectral Clustering of Survival Curves". In Proceedings of the 15th International Symposium on Operations Research in Slovenia (ISBN 978-961-6165-55-6), pp. 81–86.
- Arató, M., and Bozsó, D. and Elek, P. and Zempléni, A. 2009. "Forecasting and simulating mortality tables." Mathematical and Computer Modelling. Vol. 49, pp. 805–813.
- Crocker, K. J., - Snow, A. 2000. The Theory of Risk Classification. In: Dionne, G. Handbook of Insurance. Kluwer Academic Publishers. Boston / Dordrecht / London.
- Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London (Series A), 115:513–585.
- Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [22/01/2020]).
- Kovács, E. 2012. "Living Better, Living Longer? Is Ageing in Line with Economic Performance?". Hungarian Statistical Review, Vol. 90, pp. 79-95.
- Kovács, E. and Vaskövi, Á. 2019. "Living Longer. Working Longer? Life Expectancy and Retirement Age Trends in OECD Countries". In Proceedings of the 33rd International ECMS Conference on Modelling and Simulation in Italy, pp. 103-108.
- Makeham, W. (1867). On the law of mortality. Journal of the Institute of Actuaries, 13(6):325–358.
- Marshall, A.W., Olkin, I. 2007. "Life Distributions. Structure of Nonparametric, Semiparametric, and Parametric Families", Springer, New York.
- Organization for Economic Cooperation and Development (OECD). 2018. *OECD Pensions Outlook 2018*. OECD Publishing, Paris
- Olszewski D. 2011. "Asymmetric k-Means Algorithm". In: Dobnikar A. - Lotrič U. - Šter B. (eds) Adaptive and Natural Computing Algorithms. ICANNGA 2011. Lecture Notes in Computer Science, vol 6594. Springer, Berlin, Heidelberg.
- Okada, A. 2000. "An Asymmetric Cluster Analysis Study of Car Switching Data". In: Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg.
- Vékás, P. 2019. "Rotation of the age pattern of mortality improvements in the European Union". Central European Journal of Operations Research <https://doi.org/10.1007/s10100-019-00617-0>

AUTHORS' BIOGRAPHIES

KOLOS CS. AGOSTON, graduated as an actuary and wrote his PhD thesis in insurance markets. He is now an associate professor at Corvinus University of Budapest where he teaches various subjects in operational research and actuarial sciences. He is also the head of Institute of Mathematics and Statistical Modelling. His research topics belong to optimization problems such as cash management, cutting problems and recently college admission problem. His email address is kolos.agoston@uni-corvinus.hu

ÁGNES VASKÖVI, MSc is a PhD candidate at Corvinus University of Budapest, and an 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

Circular Economy: a Coloured Petri Net based discrete event simulation model

Marco Gribaudo
Dip. di Elettronica, Informazione e
Bioingegneria
Politecnico di Milano
via Ponzio 34/5, 20133, Milano
marco.gribaudo@polimi.it

Daniele Manini
Dip. di Informatica
Università di Torino
Corso Svizzera 185, 10149, Torino
daniele.manini@unito.it

Marco Pironti
Paola Pisano
ICxT Innovation Center
Università di Torino
Lungo Dora Siena, 100A, 10153, Torino
{marco.pironti|paola.pisano}@unito.it

Veronica Scuotto
Department of Management
Università di Torino
Corso Unione Sovietica, 218bis 10134 Torino
veronica.scuotto@unito.it

KEYWORDS

Performance evaluation; Circular economy; Petri Nets

ABSTRACT

Transition from linear economy to circular economy has been so fast due to worsening circumstances stem from climate change and pollution environmental effect. The circular economy has provoked by the emergent need to cope with resources scarcity (e.g. water and food). This economy generates a new way of re-thinking and re-design process by re-using the same material, minimise the impact of waste and pollution, and revamp the economy. In this context, cities offer a desirable place for the evolution of the circular economy thanks to their closeness to citizens, companies, and service suppliers. Such cycles, are very interesting systems from a performance evaluation point of view: in this paper we offer an holistic case study by employing Coloured Petri Nets to describe a real case scenario of circular economy. We focus on a very current example which considers the circular production of chitin by bioconversion of municipal waste and we show how we can describe and analyse it by using standard approaches to performance evaluation. The results allow to focus on interesting metrics and performance indicators, which may not be easily obtainable with conventional techniques used in the economic domain.

I. INTRODUCTION

The global environmental warn is provoking a sense of responsiveness among businesses, governments, no profit organisations and so on. These actors are coping with waste, pollution and resource scarcity, by re-thinking, re-designing, and re-using components of existing products. This new process is known as circular economy which is an alternative process to reduce and reuse waste to be converted into a new product. In addition, “recycling what cannot be reused, repair-

ing what is broken, remanufacturing what cannot be repaired” is enclosed in the circular economy process [29]. The life span of each component of a product increases supplying renewable energy. New jobs are offered and new skills are requested, involving the entire ecosystem. [10] point out the fact that the current economy is formed of five helix, that is “university-industry-government, media-based and culture-based public, and environment” which are intertwined in a circular loop of resources and energy for a high level of efficiency [25], [26]. Moving beyond the linear economy which aims to converge natural resources into cheap products driven by the logic of “take-make-dispose” economic model, the circular economy develops a re-thinking, re-designing, and re-using approach. This approach seeks to adopt a bottom up approach involving citizens for a more sustainable development. Accordingly, urban contexts are becoming more responsiveness than the past to the effect of climate change, over-population, and pollution and so are embracing the circular economy approach. They have a strong experience in offering a “sustainable waste management”? [24]. This also allows a frequent dialogue with territorial actors such as businesses, suppliers, government, citizens, etc. in producing more “durable, repairable and recyclable products”. In addition, the universities are playing a relevant role in training and nurturing new professional skills. These are good initiatives but there still is a concern how those are generated avoiding the deterioration of reused material [7]. Cities so are called to stimulate the pursuance of circularity and achieve sustainability goals thanks to their proximity with all territorial actors. The link between circular economy and cities is becoming popular among scholars [26], [33], [14], [3], [8] but a lack of empirical studies persists [28].

In this paper we present our experience in modelling a case study of Circular Economy using Coloured Petri Nets (CPNs) [17]. The preliminary goal is to provide

an approach that allows us to analyse the qualitative behaviour of the system under study, i.e., the circular manufacturing of chitin via bioconversion of urban waste. This analysis takes advantage of the benefits of modelling by CPNs [13], [6]. In particular, in this case we were able to quickly define the network representing the real scenario (half an hour). Moreover, the time needed to obtain the results that we show in this work is negligible, that is to say a few seconds. Using other simulation approaches, e.g. ARENA [1] to name one, model development may take days and results can require minutes or even hours to be computed. Other modelling approaches, such as fluid approximations [21], [12] are not suitable since they cannot explicitly consider the combination of entities of different types.

The rest of the paper is organised as follows. In Sec. II, we introduce Circular Economy. Sec. III first provides an example by presenting the role of chitin cycle in urban waste recycling, in line with the case study analysed. Sec. IV first describes the model developed to show the real system analysed and its parameters. Additionally, results are provided leading on metrics and performance indicators. Finally, Sec. V draws some conclusions.

II. CIRCULAR ECONOMY

Although it is growing popularity, the phenomenon of circular economy started in 1862 by Simmonds who declares the need of generate innovation from the waste which was not efficiently reused. Such need became more pervasive with the rise of environmental issues [27] which induces the shift from a linear economy to a circular economy [2]. This shift is happening gradually. Indeed, nowadays the number of companies which are fully embracing the circular approach is increasing. Living and producing in a circular flow of tangible and intangible resources, the development of the worldwide economy seats on four blocks: 1. re-thinking design model; 2. Circular business model; 3. “recycling and reverse value chain”, and 4. Stimulating Eco-innovation. The re- thinking design model (1) employs already used material in a circular loop, enlarging the life span of each single components. It gets closer local, national and global actors aimed at achieving sustainability goals [19]. This stimulates new circular business models (2), that is new architectures developed horizontally enclosed in high territorial proximity. Individuals work in the same environment scaling the reuse of material from a local to a global space. Value chains are based on recycling and reverse (3), where the production starts from the bottom to the up. Citizens are empowering for the reuse of the waste which is delivered the chain of production in order to made a new product. The value creation is not generated linearly but entails a closed- loop system for industries in which reverse value chain activities (rescue, repair, refurbishing, recycling, remanufacturing, or redesign of returned products from the end user) [31]. Finally, innovations are sustainable and eco-friendly. Refer-

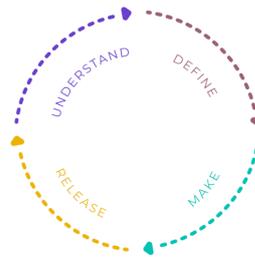


Fig. 1. Circular Design Process.

ring to the concept of incremental innovation [15], eco-innovations concern the re-use of existing components of a product “rather than the full product” to generate new goods. Metaphorically, the economic system is “cradle to cradle” restoring cycle of nature by the reuse of waste [20] which assumes value for a company [11]. In addition, a bottom up approach is employed which empowers citizens and encourages cities in being more sustainable and facilitators of such change. For instance, in China the government tends to facilitates sustainable businesses mainly at their early stage [32]. In a nutshell, undesirable products become desirable and fruitful for the productivity of natural resources. They are re-used and converted in new selling-goods. Stahel claims reusing waste to offer new services. This takes place on the development of six activities “Regenerate, Share, Optimise, Loop, Virtualise and Exchange” [28] which are employed at a local, national and global level. With a focus on a local level, we draw the attention on cities which are playing a relevant role in the circular economy. They are the new hubs for eco-innovations, striving to improve citizens’ life quality [24]. Cities takes up the responsibility of developing smart projects. For example, the smart mobility which aims to minimise the level of pollution. To achieve smart projects, cities can adopt Circular Design Process (CDP), see figure 1. The CDP is formed of four steps referring to the design thinking and the human centred mode: 1. Comprehension of material to explore new re-use; 2. Associating a meaning to material to better define the scope of a designer; 3. Use material to generate new prototypes; 4. Build a narrative storytelling to launch a new material so as to engage customers.

III. CASE STUDY: CIRCULAR MANUFACTURING OF CHITIN

In this paper, we selected a circular economy case based on the work presented in [23] where the circular manufacturing of chitin via bioconversion of urban waste is analysed. The problem of global waste is rapidly increasing due to the people trend to move in urban area. Urban citizens generate an amount of waste four times greater than country side residents. For this reason, the management of gathering, storing, and destroying waste is becoming one of the most expensive items for municipalities, both for developing and developed countries. The 2018 global municipal solid waste production was at least 2 billion tons per year and the is estimated to

reach 3.4 billion tons by 2050. With this projection the world needs to switch to new models of production and consumption based on sustainable and circular paradigms [30]. [23] specifies that, generally speaking, waste is approximately composed of 60% organic matter from food, vegetables, and garden material; 30% paper, cardboard, textiles, and other cellulosic materials; 12% plastics; 3% metals; and 3% glass, and that for organic streams the main aim is to reduce and valorise. In particular, valorisation of food waste is primarily performed by bioconversion using microorganisms, enzymes, and animals. The extraction of proteins from these entities allows to produce matter for animal or human consumption as well as the production of energy. The example studied in [23] takes into account the black soldier fly (BSF, *Hermetia illucens*), a popular insect globally for its efficient conversion of a wide variety of organic materials, such as urban and agricultural waste, into biomass. In order to quantify this process it is highlighted that in the average of two weeks required for the instar, the time between egg hatching and the prepupal stage, BSFs process 20 times their own weight in waste. Furthermore, Chitin, constituting 6-9% of BSFs' dry weight, is the second-most abundant organic polymer on earth with annual worldwide bio-production estimated at 1011 tons across every ecosystems.

Therefore, we opted to model this process since it provides for the first time a general route to embed manufacturing within its surrounding ecosystem. An example of the circular manufacturing of chitinous biocomposites via bioconversion of urban refuse is showed in figure 2.

IV. MODELLING, SETTINGS, AND RESULTS

In line with the example of circular economy showed in the figure 2, we draw a new model combining the CPN approach with a real scenario. Petri Nets (PNs) [22] are a modelling tool for the representation of concurrent asynchronous systems. PNs are bipartite graphs where two types of nodes, called places and transitions respectively, are connected by directed arcs. Each place can contain a finite integer number of tokens. The state of the system is given by the distribution of tokens over the places, which is called *marking*. The dynamics of the model is defined by state changes due to firing of transitions, which move tokens over the places. The main advantage is that they provide a graphical representation of complex behavioural such as concurrency, conflict, synchronization, etc. Coloured Petri Nets are an extension that allows the representation of large systems where tokens are augmented with attributes. Attributes are called colours, and are divided in classes called types. Each token has associated a set of types that defines its attributes. A transition is enabled to fire if all the places connected with incoming arrows have enough tokens of the colour associated with the respective arc. When a transition fires, it removes the tokens from its input places and inserts tokens into

its output places according to the colours associated with the outgoing arrows. For a tutorial on CPN, the reader can refer to [16].

A. THE COLOURED PETRI NET MODEL

The resulting net is reported in figure 3. We set five different token colours corresponding to the system entities, namely *Products*, *Cellulosic*, *Chitin*, *Insect* and *Protein*. The arcs are labeled only if the transition among the arcs fires tokens of a different color with respect to the one of the incoming arc. For instance, transition *Production* fires a token of color *Product* when 1 token of color *Product*, 1 token of color *Chitin*, and 1 token of color *Cellulosic* are available in place *Plant*, and transition *EndOfLife* fires 1 token of color *Product* and one token of color *Cellulosic* for each incoming token of color *Product*. Otherwise, transitions firing tokens of the same color of the incoming arc are labelled with the name of the color, e.g., see transition *Biodegradation*.

Our CPN model is implemented with JMT [18]. One interesting feature of CPN is that they are supported by multiple tools [4], [9], [5], we chose JMT for the modern graphical interface. The model shown in figure 4 was developed through some hypothesis and simplifications compared to the original Petri Net of figure 3, which are reported below. For sake of simplicity and in order to focus the analysis on chitin impact, we neglected the protein cycle depicted with greater dashed lines and circles. Furthermore, the chitin path drawn with smaller dashed lines and circles has an immediate transition only, therefore in the implemented CPN it is reduced with only one arc that goes from *Prepupa* place to *Plant* place. The arc in the Production Loop (see figure 2) that goes from Research&Development to the Plant is not reported in the CPN as it represents a conceptual step, but does not involve transfer of entities. However, the product cycle is closed by drawing the arc from the place *End Of Life Products* to the transition *Recycle* that accounts for product matter directly recycled by the the plant. Finally, we added a source *Food* modelling the food provisioning to the city and a sink¹ place *FoodOut* accounting for the food that comes out of the system.

B. MODEL PARAMETERS

In this scenario, we chose to model each entity required by each stage of the circular production cycle with a single token of the corresponding color: i.e. one token represents the daily amount of a given entity (insects, products, food waste, etc.). All timed transitions k have an exponential firing time distribution and their respective firing rate is indicated with λ_k . Rate λ_{Food} , for source *Food*, is set to 1 food token per day. Given that adult BSF lives for 5-8 days and in this time they must find a mate and lay eggs, we set λ_{Eggs} equal to 1/4 insect per day accounting for 4

¹In JMT a Source is used to model entities entering the system and a Sink collects entities leaving the system to allow computation of specific performance metrics.

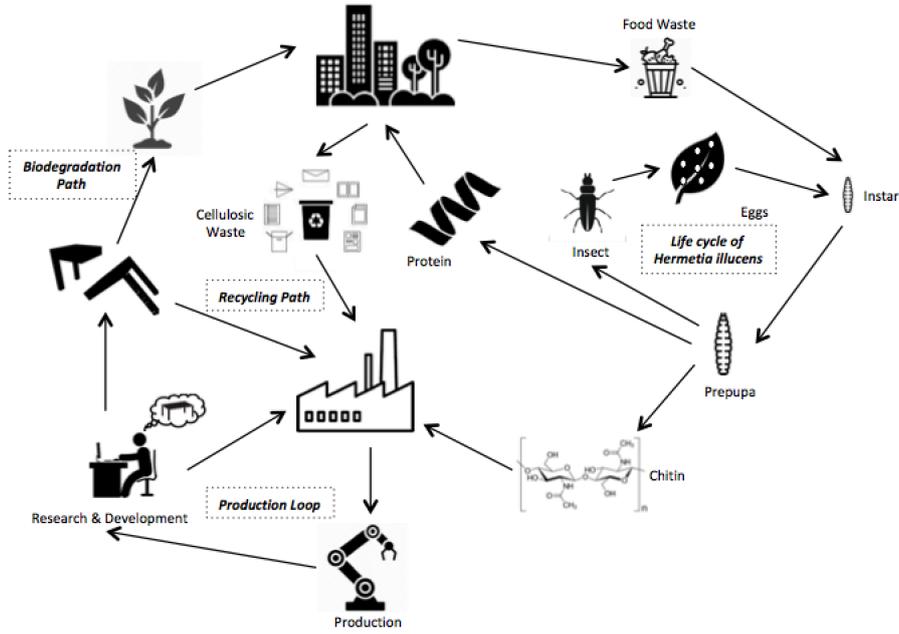


Fig. 2. Chitin cycle in the recycling of municipal waste

days required to hatch the eggs. Then the larvae will take roughly 2 weeks before they are ready to pupate, hence $\lambda_{Prepupa}$ is set to 1/15 insect per day. Both these transitions are set as infinite server since each insect evolves independently. The initial marking of PLACE *Insect* has been set to 20 to allow the insect evolution process to start. Transitions *Recycle*, *End Of Life*, and *Biodegradation* have their transition rates set to 0.5 per day with infinite server policy. *Food Waste*, *Production*, and *Cellulosic* have their transition rates set to 1.5 per day, and are all single server, i.e., each transition serves one token at a time. These rates have been arbitrarily set since more accurate values require focused studies that are planned for future works. Place *Product* has the initial marking set to 10 to start the production loop.

Another important extension will concern the comparison of results presented in this work with the ones existing in literature derived by different techniques, to assess the strength and limitations of the proposed CPN approach.

C. RESULTS

We have analysed the model with JSimGraph, the discrete event simulation component of JMT, and computed the 99% confidence interval for each metric and index. Since results were characterised by very tight intervals, that would have not been clearly visible in the pictures, we decided not to show them for clarity purposes. We first studied the system behaviour as function of the initial marking of place *Insect* that we denote with N . Note that in order to obtain the insect evolution process, given the parameter setting

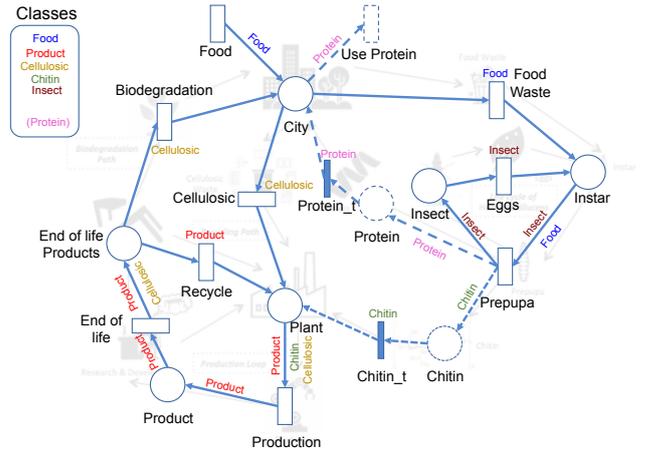


Fig. 3. The CPN based model.

presented above, N has to be greater at least 20 times the daily requirement, since BSF life cycle is set as an average of 19 days where 4 days are for hatching eggs and 15 days are for the instar. We run 7 simulations with N ranging from 10 to 40 tokens with step 5. Indeed, as can be seen in figures 5 and 6 the system is not stable for $N < 20$ resulting in unbounded accumulation of food waste and a simultaneous reduction in production. Figure 6 shows an interesting evolution of food waste in place *Instar* for $20 \leq N \leq 28$ while the system stabilises after $N \geq 30$. We then performed a detailed study of the food waste in place *Instar*, and of the chitin in the *Plant* place, for $20 \leq N \leq 28$, with results shown in figure 7. As expected the accumulated food waste decreases with the number of insects, while

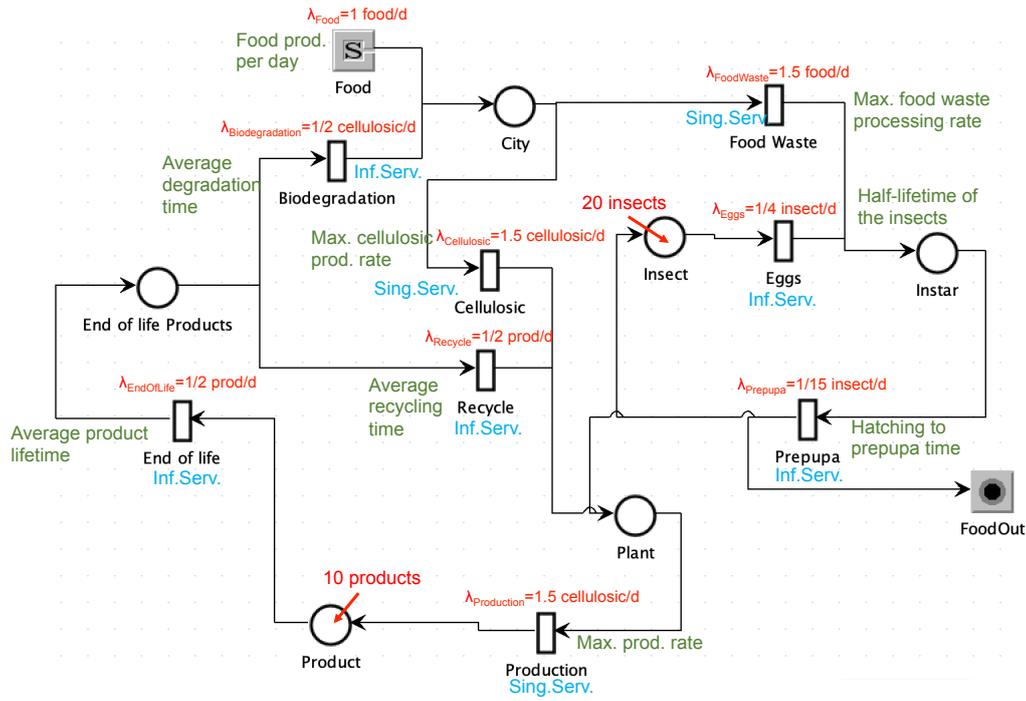


Fig. 4. The resulting CPN implemented on JMT.

the chitin production increases. The figure shows that $N = 25$ gives a good trade-off between chitin production and food waste accumulation.

We further studied the system by setting N to 25 and computed the probability density function (pdf) of the number of tokens in place *Product* (see figure 8), of the number of tokens of class *Chitin* in place *Plant* (see figure 9), of the number of tokens in place *Insect* (see figure 10), and of the number of tokens of class *Instar* in place *Food* (see figure 11). From figures 10 and 11 it is possible to note that the number of *Insects* and *Instar* are distributed with large variability centred around the average time required for hatching and instar. Figure 9 illustrates, as expected, that the Chitin is the most sensitive element of the whole production cycle with a 27% probability of having an empty stock. Figure 8 provides insights about the size of the warehouse to hold products that enter and exit the cycle.

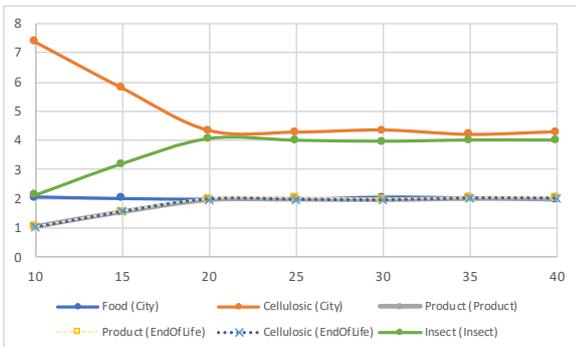


Fig. 5. Average number of *City*, *Product*, *Insect*, *EndOfLife* tokens vs initial marking of *Insect* place.

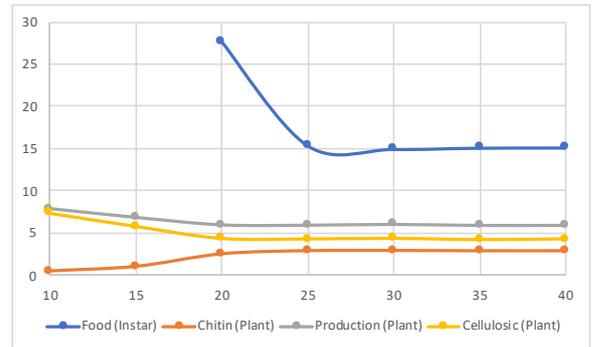


Fig. 6. Average number of *Plant* and *Instar* tokens vs initial marking of *Insect* place..

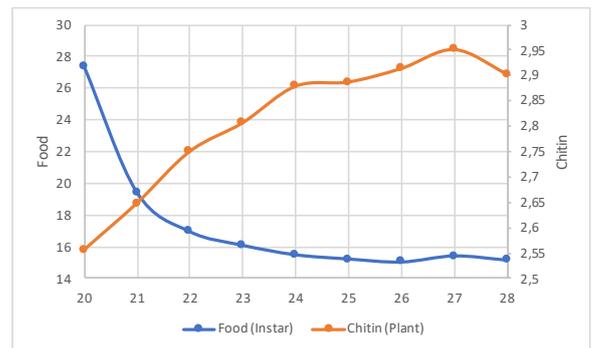


Fig. 7. Average number of tokens of class *Food* for place *Instar* and Chitin for place *Plant*.

V. CONCLUSIONS

This interdisciplinary work offers a real case scenario of recycling of municipal waste by employing a CPN model. Such model allowed us to get relevant result in a short time. As aforementioned, we were able to get a network of real scenario in half an hour. However, this study has also got some limitations. For instance,

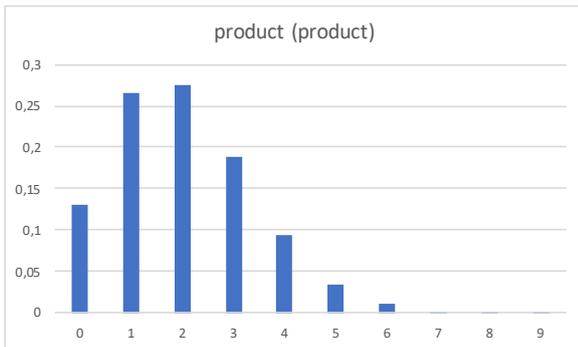


Fig. 8. Distribution of tokens in place *Product*.

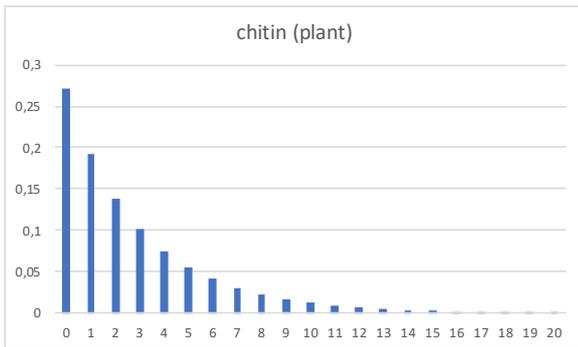


Fig. 9. Distribution of tokens of class *Chitin* for place *Plant*.

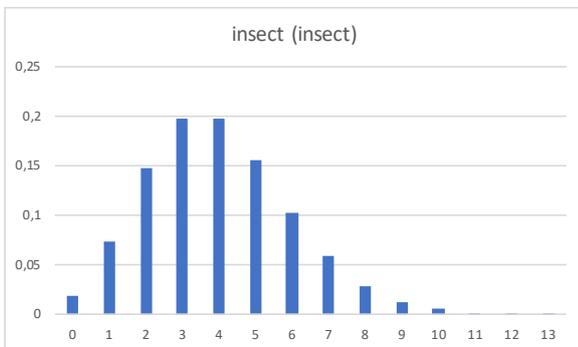


Fig. 10. Distribution of tokens in place *Insect*.

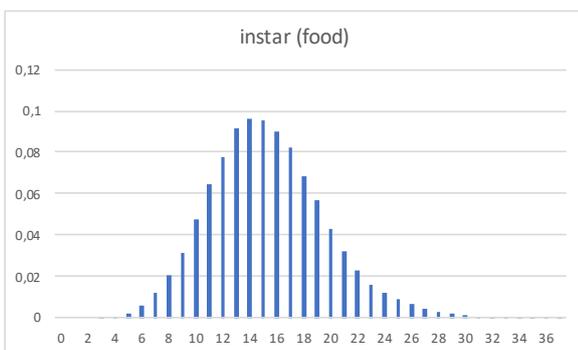


Fig. 11. Distribution of tokens of class *Food* for place *Instar*.

some parameters have been arbitrarily established and part of future developments will aim to focus on a deep investigation to derive a model as close as possible to the real scenario. In addition, this approach will be applied to other case studies to highlight the effectiveness of this technique and to provide relevant ideas for the application of this economic paradigm.

REFERENCES

REFERENCES

- [1] Arena simulation software. <https://www.arenasimulation.com>.
- [2] Frosch R. A. and Gallopoulos N. E. Strategies for manufacturing. *Scientific American*, 261(3), 1989.
- [3] Su B., Heshmati A., Y. Geng, and Yu X. A review of the circular economy in china: moving from rhetoric to implementation. *Journal of Cleaner Production*, 42.
- [4] E. Barbierato, M. Gribaudo, and M. Iacono. Defining formalisms for performance evaluation with simthesys. *Electronic Notes in Theoretical Computer Science*, 275(1):37–51, 2011.
- [5] Giovanni Chiola, Giuliana Franceschinis, Rossano Gaeta, and Marina Ribaudo. Greatspn 1.7: Graphical editor and analyzer for timed and stochastic petri nets. *Perform. Evaluation*, 24(1-2):47–68, 1995.
- [6] F. Cordero, A. Horváth, D. Manini, L. Napione, M. De Pierro, S. Pavan, A. Picco, A. Veglio, M. Sereno, F. Bussolino, and G. Balbo. Simplification of a complex signal transduction model using invariants and flow equivalent servers. *Theoretical Computer Science*, 412(43):6036–6057, 2011.
- [7] Cullen. Circular economy: theoretical benchmark or perpetual motion machine? *Journal of Industrial Ecology*, 21(3), 2017.
- [8] Zhijun F. and Nailng Y. Putting a circular economy into practice in china. *Sustainability Science*, 2(1), 2007.
- [9] G. Franceschinis, M. Gribaudo, M. Iacono, N. Mazzocca, and V. Vittorini. Drawnet++: Model objects to support performance analysis and simulation of systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2324 LNCS:233–238, 2002.
- [10] Carayannis E. G. and Campbell D. F. Triple helix, quadruple helix and quintuple helix and how do knowledge, innovation and the environment relate to each other?: a proposed framework for a trans-disciplinary analysis of sustainable development and social ecology. *International Journal of Social Ecology and Sustainable Development (IJSESD)*, 1(1):41–69, 2010.
- [11] Pauli G. The blue economy. *Paradigm Publications*.
- [12] R. Gaeta, M. Gribaudo, D. Manini, and M. Sereno. Fluid stochastic petri nets for computing transfer time distributions in peer-to-peer file sharing applications. *Electronic Notes in Theoretical Computer Science*, 128(4):79–99, 2005.
- [13] Rossano Gaeta, Marco Gribaudo, Daniele Manini, and Matteo Sereno. On the use of petri nets for the computation of completion time distribution for short TCP transfers. In Wil M. P. van der Aalst and Eike Best, editors, *Applications and Theory of Petri Nets 2003, 24th International Conference, ICATPN 2003, Eindhoven, The Netherlands, June 23-27, 2003, Proceedings*, volume 2679 of *Lecture Notes in Computer Science*, pages 181–200. Springer, 2003.
- [14] B. Guo, Geng Y., Ren J., Zhu L., Liu Y., and Sterr T. Comparative assessment of circular economy development in china's four megacities: The case of beijing, chongqing, shanghai and urumqi. *Journal of Cleaner Production*, 162.
- [15] Schumpeter J. The theory of economic development. *Cambridge: MA: Harvard University Press*.
- [16] Kurt Jensen. *Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use. Vol. 2, Analysis Methods*. Monographs in theoretical computer science: an EATCS series. Springer, 1995.
- [17] Kurt Jensen and Lars Michael Kristensen. *Coloured Petri Nets - Modelling and Validation of Concurrent Systems*. Springer, 2009.
- [18] Bertoli M., Casale G., and Serazzi G. Jmt: performance

engineering tools for system modeling. *SIGMETRICS Perform. Eval. Rev.*, 36(4):10–15, 2009.

- [19] Bocken N. M., De Pauw I., Bakker C., and van der Grinten B. Product design and business model strategies for a circular economy. *Journal of Industrial and Production Engineering*, 33(5):308–320, 2016.
- [20] Braungart M. and McDonough W. Remaking the way we make things. *North Point Press*.
- [21] D. Manini and M. Gribaudo. Modelling search, availability, and parallel download in p2p file sharing applications with fluid model. pages 449–454, 2006.
- [22] T. Murata. Petri Nets: Properties, Analysis and Applications. *Proceedings of the IEEE*, 77(4):541–580, April 1989.
- [23] Sanandiya N.D., Ottenheim C., Phua J.W., Caligiani A., S. Dritsas, and Fernandez J.G. Circular manufacturing of chitinous bio-composites via bioconversion of urban refuse. *Scientific Reports*, 10(4632), year = 2020, issn = 2045-2322,
- [24] Ghisellini P., Cialani C., and Ulgiati S. A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems. *Journal of Cleaner productions*, 114:11–32, 2016.
- [25] D. W. Pearce and R. K. Turner. Economics of natural resources and the environment. *JHU Press*.
- [26] A. Petit-Boix and S. Leipold. Circular economy in cities: Reviewing how environmental research aligns with local practices. *Journal of Cleaner Production*, 195.
- [27] Stahel W. R. and Reday G. The potential for substituting manpower for energy. *Report to the Commission of the European Communities*.
- [28] Prendeville S., Cherim E., and Bocken N. Circular cities: mapping six cities in transition. *Environmental Innovation and Societal Transitions*, 26.
- [29] W. R. Stahel. The circular economy. *Nature*, 531(7595), 2016.
- [30] Keijer T., Bakker V., and Slootweg J. C. Circular chemistry to enable a circular economy. *Nature Chemistry*, 11.
- [31] Jayaraman V. and Luo Y. Creating competitive advantages through new value creation: a reverse logistics perspective. *Academy of management perspectives*, 21(2):56–73, 2007.
- [32] Scuotto V., Tarba S., Messeni Petruzzelli A., and Chang V. International social smes in emerging countries: Do governments support their international growth? *Journal of World Business*.
- [33] Geng Y., Zhu Q., Doberstein B., and Fujita T. Implementing china's circular economy concept at the regional level: A review of progress in dalian, china. *Waste Management*, 29(2), 2009.



Marco Gribaudo is an Associate Professor at the Politecnico di Milano, Italy. He works in the performance evaluation group. His current research interests are multi-formalism modelling, queueing networks fluid models, mean field analysis and spatial models. The main applications area are applied comes from Big Data applications, Cloud Computing, Multi-Core Architectures and Wireless Sensor Networks. Email marco.gribaudo@polimi.it.



Email daniele.manini@unito.it.

Daniele Manini He is Researcher and Assistant Professor at Università degli Studi di Torino, Italy. He was a Visiting Researcher at BME Budapest, Hungary. His research interests include Performance Evaluation of Complex Systems in Communication Networks, Biology, and Economic. He has been involved in national and International research projects, including the COST Action Random Network Coding and Designs over GF(q).



Marco Pironti He was a Visiting scholar at the Center for Computational Research and Management Science, MIT, Boston (MA), at the Institute of Management, Innovation and Organization, Haas School of Business, Berkeley and at the CEbiz of Columbia University and Visiting Professor at Westminster Business School (UK). He is a Full Professor of Innovation Management and Entrepreneurship at the University of Torino Computer Science Department, Director of ICxT Interdepartmental Innovation Center and member of Scientific Committee of PhD program in Innovation for Circular Economy . He is an author of more than 90 articles and other publications. His main research interests are relating to strategy, innovation management and business modeling and planning. Email marco.pironti@unito.it.



Paola Pisano She is visiting scholar at Westminster University and assistant professor of Innovation and Entrepreneurship at the University of Torino Computer Science Department. She is involved in several projects on innovation and start up with national and international companies and research groups as ICxT Innovation Center. She is author of books and several articles. Email paola.pisano@unito.it.



Veronica Scuotto Prof., Dr Veronica Scuotto (PhD, FHEA, MBA, BA-Honour) after working at the University of the West of Scotland (UK) and then at the Pôle Universitaire Léonard de Vinci in Paris (France) as an Associate Professor in Entrepreneurship and Innovation, she joined the University of Turin (Italy). Her research interests are focused on SMEs, entrepreneurship and digital technologies. Her work has been featured in several peer to peer international journals and books. Email veronica.scuotto@unito.it.

FORECASTING RESIDENTIAL ELECTRICITY CONSUMPTION BASED ON URBANIZATION AND INCOME PROJECTIONS

Emília Németh-Durkó
Péter Juhász
Fanni Dudás
Department of Finance
Corvinus University of Budapest
H-1093, Budapest, Hungary, Fővám tér 8.
E-mail: durko.emilia@uni-corvinus.hu

KEYWORDS

residential electricity consumption, income deciles, rural-urban life, urbanization, scenarios

ABSTRACT

This study presents an alternative method for predicting residential electricity consumption. Based on electricity expenditures data from Hungary, we calculated consumption patterns for people living in different types of settlements by income deciles. We developed 12 scenarios of expected urbanization and income effects for the period 2020 to 2050. Our forecast outputs are consistent with scenarios in the literature proving that in the absence of long-term panel data, future electricity consumption can still be adequately estimated based on the evolution of electricity expenses.

INTRODUCTION

The economic growth of the developed economies, the population growth and the urbanization are likely to cause a robust increase in electricity consumption of the households (Ramírez-Mendiola et al. 2017; Taale and Kyeremeh 2016). Thus, the residential sector has immense potential regarding the improvement of energy-saving and efficiency (Sun et al., 2018). As the sustainable energy management is a top priority in the EU as well (Bianco et al. 2019), the efforts to identify the crucial factors of households electricity consumption have become prominent (Bianco et al. 2019; Wiedenhofer et al. 2013; Curtis and Pentecost 2015; Rosas et al. 2010). The energy consumption simulations and the forecasts are vital in shaping the proper energy policies and in making the right decisions to maintain sustainable development. Earlier papers prefer time-series models to forecast annual electricity consumption (Daut et al. 2017; Wei et al. 2019). Although the traditional models do not require a lot of historical data (Wei et al. 2019; Deb et al. 2017), In the case of Hungary even the needed minimum is not available for modelling (Jebli and Youssef 2015).

This paper develops a new model for estimating the residential electricity consumption based on

historical income and urbanization data while the detailed regional distribution of income remains unknown. This forecast seeks to answer the question of whether urbanization helps the achievement of the sustainability goals and whether the increasing rate of the rural and urban population will contribute to the reduction of the long-term residential electricity consumption. We used Hungary as an example in our model because electricity consumption decreased in the past decades, while the rate of urbanization grew (KSH 2018).

SOCIO-ECONOMIC AND SPATIAL FACTORS

We may divide the factors affecting electricity into two groups depending on how those influence electricity consumption directly or indirectly. The direct factors are the characteristics of the households, like the age of the members, their qualifications, the efficiency of the appliances they use or the dwelling characteristics. The indirect factors are income and urban form (Wiedenhofer et al. 2013; Yang et al. 2019; Taale and Kyeremeh 2019). This paper builds on the effects of the latter group.

Income proved to be a significant factor in several studies to affect electricity consumption. (Wiedenhofer et al. 2013; Santamouris et al. 2007; Hussain and Asad 2012). However, the direction and the extent of its impact on electricity consumption varies widely (Table 1). A 1% growth in the income can increase electricity consumption even by 11% (Brounen et al. 2012), though, the same change may also cause a significant decline (Borozan et al. 2017). Not only the consumption of the poorest and most rich differ, but there are also significant differences across all the income deciles (Rosas et al. 2010; Szép 2013). Researchers agree that the change of the electricity price does not affect the consumption significantly (Atalla and Hunt 2016; Yang et al. 2019; Wang et al. 2019) but the income remaining after electricity expenditures (Gomez et al. 2013) and the ratio of the expenses per income affect electricity consumption (Rosas et al. 2010).

Table 1: The effects of income on electricity consumption

Effect	Extent	Source
Increase	+0.13%	Ye et al. (2017)
	+0.40%	Wiedenhofner et al. (2013)
	+0.57%	Yang et al. (2019)
	+11.00%	Brounen et al. (2012)
Decrease	-0.50%	Borozan et al. (2017)
	-0.44%	Contreras et al. (2009)
	-0.47%	Gomez et al. (2013)

Santamouris et al. (2007) examined the relationship between the annual electricity expenditures and the income and found that wealthy households used nearly 38% more electricity than poor ones. Gomez et al. (2013) identified spatial differences (Table 2).

Table 2: The effects on electricity expenditures

Factor	Effect	Extent	Source
Income	increase	0.27%	Gomez et al. (2013)
		0.13%	Rosas et al. (2010)
		0.38%	Santamouris et al. (2007)
		0.15%	Hussain and Asad (2012)
	decrease	-0.20%	Curtis Pentecost (2015)
Price	-	-	Atalla and Hunt (2016)
			Yang et al. (2019)

Population determines the electricity use since each 1% growth of it boosts the latter by nearly 1.5% (Yang et al. 2019; Sun et al. 2014). However, population growth differs across urban forms (Liddle and Lung 2014; Larson and Yezer 2015).

The spread of cities is often regarded as the leading cause of increased energy use (Poumayong and Kaneko 2010; Yang et al. 2019; Larivière and Lafrance 1999) and the emission of harmful substances (Yang et al. 2019; Taale and Kyeremeh 2016; Jonas et al. 2015;). However, it has also been proven that urbanization has its advantages coming from the economies of scale (Shammin et al. 2010; Poumayong and Kaneko 2010; Yang et al. 2019).

Table 4: The effects of urbanization on electricity consumption

Effect	Direction	Extent	source
rate of urban residents	decrease	-0.10%	Shammin et al. (2010)
	increase	+0.19%	Yang et al. (2014)
		+0.23%	Yang et al. (2019)
urban areas	increase	+1.06%	Yang et al. (2019)
country-side	increase	+2.55%	Yang et al. (2019)
	decrease	-0.13%	Poumayong (2010)
	increase	+0.14%	Yang et al. (2019)
depending on the income	decreasing	-0.03%	Yang et al. (2019)
	increasing	+0.50%	Yang et al. (2019)
	increasing	+0.90%	Poumayong (2010)
		+0.07%	

The rural electricity consumption is up to 10% higher than the use of the urban population (Yang et al. 2019; Poumayong and Kaneko 2010, Lenzen et al. 2004; Shammin et al. 2010). Thus, the urbanization rate might have a favourable effect on electricity consumption (Chen et al. 2008) (Table 4). According to Poumayong and Kaneko (2010),

urbanization has an increasing impact only in electricity consumption of the poorest people. In the case of the medium and high-income households, urbanization reduces electricity consumption.

INCOME INEQUALITIES

Electricity consumption is a widely used index of economic development. The amount of consumed electricity reflects the population's income conditions quite well (Liddle and Lung 2014). According to Szép's (2013) discriminant analysis, electricity consumption of the lowest and the highest income deciles show a significant gap in Hungary. Figure 1 shows that the highest growth in electricity consumption can be seen in the case of the wealthiest households. The lowest five deciles spent 5 to 10 percent of their income on electricity, while the higher deciles spent only 2% in 2010. By 2018, these percentages had decreased in the case of both groups (KSH 2018). (Figure 1). For the top deciles, most of the savings on energy expenditures were lost due to a large number of electronic equipment in rich households.

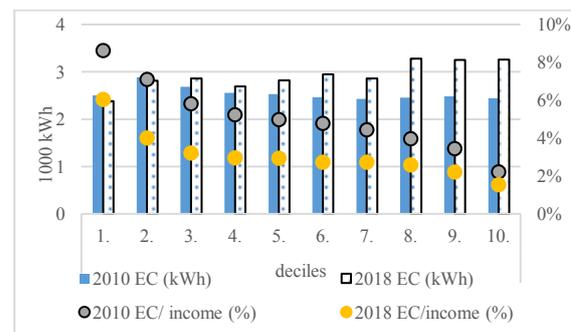


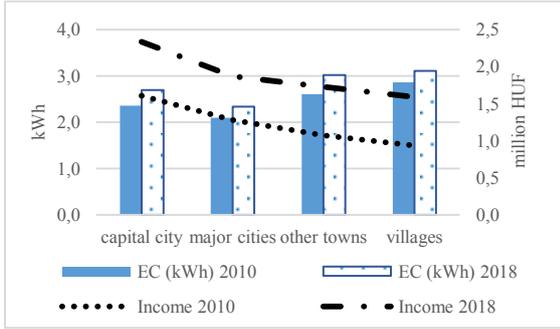
Figure 1: The annual electricity consumption (EC) of the households (1,000 kWh) and the electricity/income ratio (EC; %) by deciles in 2010 and 2018

The poorer households are not so abundant in modern appliances, their financial conditions are well below the average, so they cannot afford energy-saving renovations or investments (Bíró-Szigeti 2011). Thus, there is no significant energy saving in their case, either. Finding a way to reduce the electricity consumption of poor households is an often-raised issue in the literature (Rosas et al. 2010).

The per capita electricity consumption is also affected by the number of people living in a household. In 2018, the average number of people in the least wealthy household was 2.8, while this figure was 1.8 in the wealthiest households (KSH 2018). The per household electricity consumption peaked in the richer households in 2010, while it was there that the per capita consumption was the lowest. By 2018, per household consumption topped in the upper three deciles (KSH 2018).

SPATIAL INEQUALITY

In 2000, there were more than 200 towns registered in Hungary, while in 2014 almost 350 settlements enjoyed town status, with 70% of the country's population. (Gerse and Szilágyi 2015). If we move downward in the hierarchy of the settlements, the income per household continuously decreases. The people living in rural areas earn only 89.5% of the national average (KSH 2018). However, the per household electricity consumption is the highest there (3,110 kWh) (Figure 2).



*1 euro=335 HUF, March 1, 2020

Figure 2: The annual consumption of households (1,000 kWh) and the annual gross income (HUF) according to settlement type in 2010 and 2018

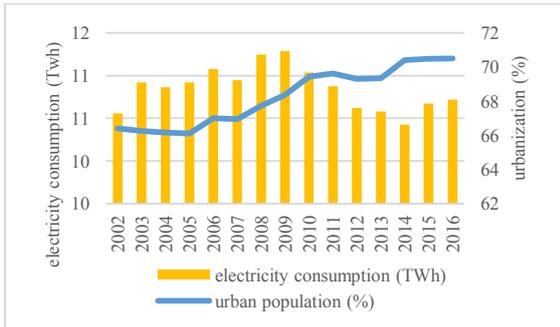


Figure 3: Residential electricity consumption (TWh) and the urban population (%) in Hungary between 2000 and 2018

During the examined period, electricity consumption increased in every type of settlement. In 2018, the lowest consumption was in the major cities and the highest consumption was in the village households. The growth rate was the highest in the medium towns. After the global economic crisis, electricity consumption began to plummet. (Figure 3) During recent years, the consumption increased again, while the urban population grows at a lower rate.

MODEL DESCRIPTION

We have prepared a forecast of electricity consumption of the Hungarian population for the period of 2019 to 2051. Our first hypothesis is that electricity consumption differs in the various stages of the urbanization (urban area, countryside). Income changes not only between the two extremes (the rich and the poor), but it also differs by the

income deciles. Thus, our second hypothesis is that future electricity consumption can be predicted quite reliably from the urbanization (regional) distribution of the electricity expenditures and the electricity expenditures of the income deciles. In the case of Hungary, detailed information on household electricity consumption has only been available since 2010. Thus, long-term historical regressions offer limited aid in predicting future amounts of consumption. We used data on Hungarian household electricity consumption for the period between 2010 to 2018 in two breakdowns: income deciles and types of settlement. As in the case of deciles, only data on electricity expenditures were available; we used the annual historically established selling prices for standard electricity consumption (EC) to estimate electricity consumption (kWh per capita) per decile (d). As no data on the settlement breakdown of the population from deciles were available, we calculated the consumption of an average household at the national level by equally weighting electricity consumption of the ten deciles.

$$EC_d = \frac{\text{Electricity expenditures by deciles } \left(\frac{\text{HUF}}{\text{capita}}\right)}{\text{Average consumer price of electricity } \left(\frac{\text{HUF}}{\text{kWh}}\right)} \quad (1)$$

Besides the consumption per decile, we also mapped the household distribution across settlement types (EC_s).

$$EC_s = \frac{\text{Electricity expenditures by settlement type } \left(\frac{\text{kWh}}{\text{capita}}\right)}{\text{average consumer price of electricity } \left(\frac{\text{HUF}}{\text{kWh}}\right)} \quad (2)$$

Then contrasting the average household consumption for each settlement type to the estimated national average, we received a correction factor (CF):

$$CF = \frac{EC_s \left(\frac{\text{kWh}}{\text{capita}}\right)}{EC_d \left(\frac{\text{kWh}}{\text{capita}}\right) * \text{income deciles } (\%)} \quad (3)$$

The correction factor (%) would cover two effects: (1) based on the literature, the different settlement types imply different consumption figures even in the same income deciles and (2) deciles are not evenly represented in all settlement types. Thus, the average correction factor value of 103.34 percent for the capital city means that the average household in Budapest consumed 3.34 percent more electricity (kWh) than the national average. That may be due to (1) the average household having higher than the national average income and (2) citizens in major cities using more electricity per capita because of the smaller than average household size. Estimating the future change of these three inputs made it possible for us to give an

estimate of the total electricity consumption of Hungarian households (EC_T):

$$EC_T = \frac{\text{Electricity expenditures by settlement type} \left(\frac{kWh}{\text{capita}} \right)}{\text{average consumer price of electricity} \left(\frac{HUF}{kWh} \right)} \quad (4)$$

Luckily, in the literature, we found robust professional estimates for these three factors separately that allowed us to build scenarios by using the combination of the alternative paths of the three factors. By processing literature and statistics together, we estimated both low (L) and high (H) outcomes for the future electricity expenditures and the urbanization. In the case of the population, we allowed for a medium (M) level, too. Table 5 shows the 12 scenarios made for forecast future residential electricity consumption in Hungary.

Table 5: Scenarios examined

#	1	2	3	4	5	6
I	L	L	H	H	L	L
U	L	H	L	H	L	H
P	L	L	L	L	M	M
#	7	8	9	10	11	12
I	H	H	L	L	H	H
U	L	H	L	H	L	H
P	M	M	H	H	H	H

I=income, U=urbanization, P=population

In this paper, the urbanization level is measured by the proportion of urban residents in the total population. These categories meet the requirements of the international standards and the interpretation of the Hungarian Central Statistical Office (KSH, 2018). The population is the entire population of Hungary (number of people). The data originate from the website of the KSH.

BASE SCENARIO

In the base scenario (HLL), the income is high (H), urbanization is low (L), and the population is also low (Table 6). The gap between the income extremes is growing because in the case of the deciles of 6 to 10, the growth can even be over 80% (GKI, 2019) and this is also reflected by electricity consumption (Szép 2013). In our forecast, similarly to the UN (2020), we project increasing social inequalities, so we calculated a higher rate of income growth (4%) In the case of the top deciles and a lower rate of income growth (1-3%) In the case of the lower deciles. The extremely high income of the rich is expected to increase electricity consumption up until about the middle of the period. However, the electricity consumption of the wealthiest deciles will begin to decline by 0.5%-1% from 2040 on. The rate of poor people will grow, but this will not significantly reduce or affect the level of the electricity use (Borozan et al. 2017;

Rosas et al. 2010), or the price of electricity either (Atalla and Hunt 2016; Yang et al. 2019)

Table 6: Base (HLL) Scenario (yearly changes, rounded)

	2018 (fact)	2030	2051
Electricity use decile 1	10%	2%	2%
Electricity use decile 2	17%	1%	1%
Electricity use decile 3	8%	2%	2%
Electricity use decile 4	0%	2%	2%
Electricity use decile 5	4%	3%	3%
Electricity use decile 6	5%	3%	3%
Electricity use decile 7	-1%	3%	3%
Electricity use decile 8	3%	5%	4%
Electricity use decile 9	-3%	6%	4%
Electricity use decile 10	8%	6%	5%
urbanization rate	67%	75%	84%
capital	17%	19%	14%
major cities	19%	21%	17%
small and medium towns	31%	35%	54%
villages	33%	25%	16%
Population	-0.2%	-0.5%	-0.6%
Total consumption (TWh)	12,40	17,50	29,01

The low (L) urbanization estimate envisions suburbanization toward the medium towns. In Hungary, even 160,000 people could move to new places of living by 2051 (Lennert 2019). The population of the capital will decrease by 10%, and the people will migrate to urban areas where the climate is more pleasant, and the workplaces are easily accessible. The villages will suffer the consequences of the population growth in the towns, and by 2051, 84% of the population will live in towns and cities (Lennert 2019). The correction factor shows the consumption deviating from the national average by the types of settlements and as the average of the past years with the income and settlement type effects. Compared to small and medium-sized cities, large cities consume less electricity. (Table 7). In the capital, we assume a minimum-level growth of the electricity expenses because wealthier people will use energy-efficient appliances (Rosas et al. 2010; Santamouris et al. 2007). The major cities will consume less electricity compared to the small and medium towns (Table 7).

Table 7: An estimate of the correction factor

	Capital	Major cities	Small and medium towns	Villages
Urbanization rate (%)	17.5	19.18	30.74	32.59
Consumption (kWh/capita)	1313	1078	1316	1315
Relative consumption*	105	90	100	110
% of the national average	1.03	0.88	0.98	1.08

*Small and medium towns=100

We only allowed for three trends for the population. According to the pessimistic (L), the medium (M) and the optimistic (H), estimates, the population of the country in 2051 maybe 8.35 million, 8.75 million or 9.14 million people respectively. Most scenarios (Eurostat 2020; Lennert 2019; Obádovics 2018; Tagai 2015; Hablicsek 1998) agree that the population of Hungary will undergo an unprecedented decline. We think the pessimistic case is most likely to happen (8.35 million people).

MEGATRENDS

The second urbanization scenario was prepared following the megatrends (HHH). HHH is less by the Hungarian visions, but we created this scenario under the urbanization rate of the developed countries (Eurostat, 2020; UN, 2020). In this case, the entire population (and the population of every urban settlement type) will grow, while the population of the villages will drastically decline. Depending on the nature of the city, we predict a yearly growth of 0.5% to 2%. The population growth in the capital will continue up until 2032, and then it will drop to 0.5%. The growth of the major cities will be 2% all along, and most of the new growth of the capital is included here. The small and medium towns will have a minimum-level, 0.5% growth. The decline of the population in the villages will be at an average level of 5% each year. In the case of the urbanization, the high (H) scenario forecasts significant growth in the ratio of major cities. The income gap will keep growing, but according to our estimates, the wealthiest people will surely pass the income level by 2035 at which their electricity expenditures will drop. The increase in electricity consumption in the upper deciles is 1% higher than the average of previous years. From 2040, we planned a 0.5% reduction because of the disappearance of the rebound effect. The difference in energy consumption between the income deciles will be more significant from time to time (Rosas et al., 2010).

SMALL AND MEDIUM TOWNS

According to our forecast, the Hungarian settlements will have undergone a significant transformation by 2051 (Figure 4). In the coming decades, the dominant settlement type will be the small and medium town due to the Hungarian suburbanization peculiarities. The population decline of the villages will not slow down, and the gap will be wider and wider between the population of the towns and the villages.

By 2051, 84% of the Hungarian population will live in towns and cities, following the UN (2020) forecast. The population of the villages will drop to half of the current 33% rate. The population of the capital and the major cities will not significantly

change due to the Hungarian urbanization peculiarities.

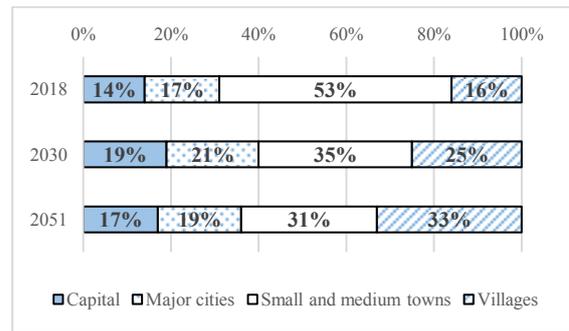


Figure 4: Settlement structure of Hungary between 2018 and 2051

BIDIRECTIONAL SCENARIOS

Our forecast shows a growing level of electricity consumption. By considering all the possible cases, electricity consumption might be around 18 to 25 TWh in Hungary by 2040 and 22 to 33 TWh by 2051. Compared to the base scenario (HLL), the other scenarios deviate downward. The first five scenarios predict a little more than half of the electricity consumption of the base scenario (Figure 5).

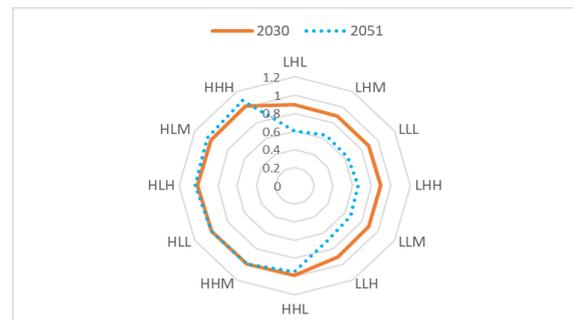


Figure 5: Deviations from the base scenario

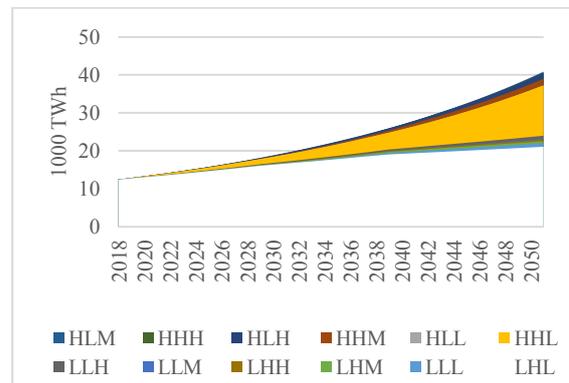


Figure 6: Scenarios for electricity consumption (1000 TWh) of Hungary between 2018 and 2051

The scenarios forecasting low energy consumption all calculate low energy expenditures. The deviation will grow with time, so the widening gap between the income levels, the more and more different consumption habits and expenditures and the

changes of the settlement structure all have an effect on electricity consumption (Figures 5 and 6). The scenarios grow together with small differences until 2025 (Figure 7), which is the starting year of the trend-deviations in the suburbanization migration process. The population of the villages will drop to nearly to the half of the current level. The lowest electricity consumption is forecast by the LHL scenario, the LHM is only slightly in advance, so income has a greater effect on electricity consumption than the number of the population. The conditions of the HHH scenario show the highest level of consumption. The extremes are not identical with the extreme scenarios, since the LLL does not produce not the lowest consumption level. When the suburbanization starts, the scenarios with lower income levels show declining curves, while the scenarios with higher income predict higher consumption.

CONCLUSIONS

Results show that urbanization reduces electricity consumption of the more deprived deciles (LHL and LLL) in the long run, and increases it in the richest deciles (HLH and HHH). Our results are consistent with Yang et al. (2019) and Wiedenhofer et al. (2013), but only partially confirm the Hungarian results of Poumayong and Kaneko (2010). Both hypotheses were validated. (1) In our estimation the residential electricity consumption differs at the levels of urbanization, and (2) the outcomes based on the income and urbanization data are in line with the forecasts of other scenarios (NES, EUCO). The 20-22000 Twh (NES) forecast corresponds to the BASE scenario and the 18000 Twh (EUCO) forecast to LHL. However, the results of the high income and urbanization paths (HHX and HLH) overestimate the consumption of known scenarios by 5-10%, which may be due to an overestimation of the rebound effect. Our model might be distorted due to various factors. Estimations used the average consumer electricity price as the ratio of the daytime and night electricity consumption is unknown. The time series are only available since 2010, and we assumed a balanced ratio of the deciles for all regions because of the lack of detailed data.

From a general perspective, the policy conclusions of our study highlight the risk of environmental pollution because of rising electricity consumption. However, residential electricity consumption has dropped significantly due to the pandemic. In Europe demand fell by an average of 4.4% and in Hungary the decline in electricity consumption was 1% in March compared with the same month last year (MEKH 2020). The impact of this shock, such as increased home office activity, can fundamentally change long-term estimates of residential electricity demand.

REFERENCES

- Atalla, T.N. and L.C. Hunt. 2016. "Modelling residential electricity demand in the GCC countries". *Energy Economics*, 59, pp. 149-158.
- Bianco, V.; F. Cascetta; A. Marino; and S. Nardini. 2019. "Understanding energy consumption and carbon emissions in Europe: A focus on inequality issues". *Energy*, 170, pp. 120-130.
- Bíró-Szigeti, Sz. 2011. "Mikro- és kisvállalkozások marketingfeltételeinek vizsgálata az energiamegtakarítás lakossági piacán". PhD Thesis, Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem
- Borožan, D. and L. Borožan. 2018. "Analyzing total-factor energy efficiency in Croatian counties: evidence from a non-parametric approach". *Central European Journal of Operations Research*, 26(3), pp. 673-694.
- Brounen, D.; N. Kok; and J. M. Quigley. 2012. "Residential energy use and conservation: Economics and demographics". *European Economic Review*, 56(5), pp. 931-945.
- Contreras, S.; W. Smith; T.P. Roth; and T.M. Fullerton Jr. 2009. "Regional evidence regarding US residential electricity consumption". *Empirical Economics Letters*, Vol. 8, No. 9 : pp. 827-832.
- Curtis, J. and A. Pentecost. 2015. "Household fuel expenditure and residential building energy efficiency ratings in Ireland". *Energy Policy*, 76, pp. 57-65.
- Daut, M. A. M.; M.Y. Hassan; H. Abdullah; H. A. Rahman; M.P. Abdullah; and F. Hussin. 2017. "Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review". *Renewable and Sustainable Energy Reviews*, 70, pp. 1108-1118.
- Deb, C.; F. Zhang; J. Yang; S.E. Lee; and K.W. Shah. 2017. "A review on time series forecasting techniques for building energy consumption". *Renewable and Sustainable Energy Reviews*, 74, pp. 902-924.
- Eurostat 2020. *Population on 1st January by age, sex and type of projection*. Retrieved from <http://data.europa.eu/88u/dataset/g33Nsv6Vud3AmmX9JEOW>, Accessed on April 15, 2020
- Gerse, J. and D. Szilágyi. 2015. *Magyarország településhálózata 2.*, Központi Statisztikai Hivatal, Budapest.
- GKI Gazdaságkutató Zrt. 2019. *Az „elmúlt 8 év” és a „majdnem elmúlt 8 év” A háztartások fogyasztása*. Retrieved from <https://www.gki.hu/wp-content/uploads/2018/01/GKI-Fogyaszt%C3%A1s-2002-2018.pdf>, Accessed on February 11, 2020
- Gomez, L. M. B.; M. Filippini; and F. Heimsch. 2013. "Regional impact of changes in disposable income on Spanish electricity demand: A spatial econometric analysis". *Energy economics*, 40, S58-S66.
- Hablicsek L. 1998. "Öregedés és népességsökkenés. Demográfiai forgatókönyvek, 1997– 2050." *Demográfia*, 41(4), pp. 472–495.
- Hussain, I. and M. Asad. 2012. "Determinants of residential electricity expenditure in Pakistan: Urban-rural comparison". *Forman J. Econ. Stud*, 8, pp. 127-141.

- Jebli, M. B. and S.B.Youssef. 2015. "The environmental Kuznets curve, economic growth, renewable and non-renewable energy, and trade in Tunisia". *Renewable and Sustainable Energy Reviews*, 47, pp. 173-185.
- KSH. 2018. *A háztartások életszínvonalá*. Retrieved from <https://www.ksh.hu/docs/hun/xftp/idoszaki/hazteletszin/2018/index.html#section-3>, Accessed on February 15, 2020
- KSH. 2019. *Népesedési világnap*. Retrieved from <https://www.ksh.hu/docs/hun/xftp/stattukor/nepesedesi19.pdf>, Accessed on February 15, 2020
- Lariviere, I. and G. Lafrance. 1999. "Modelling electricity consumption of cities: effect of urban density". *Energy economics*, 21(1), pp. 53-66.
- Larson, W. and A. Yezer. 2015. "The energy implications of city size and density". *Journal of Urban Economics*, 90, pp. 35-49.
- Lennert, J. 2019. "A magyar vidék demográfiai jövőképe 2051-ig, különös tekintettel a klímaváltozás szerepére a belső vándormozgalom alakításában". *Területi Statisztika*, 59(5), pp. 498-525.
- Lenzen, M.; C. Dey; and B. Foran. 2004. "Energy requirements of Sydney households". *Ecological Economics*, 49(3), pp. 375-399.
- Liddle, B. and S. Lung. 2014. "Might electricity consumption cause urbanization instead? Evidence from heterogeneous panel long-run causality tests". *Global Environmental Change*, 24, pp. 42-51.
- MEKH. 2020. A COVID-19 járvány hatása a magyar villamosenergia piacra. Retrieved from <http://www.mekh.hu/> Accessed on April 20, 2020
- Obádovics, Cs. 2018. "A népesség szerkezete és jövője". In: Monostori, J.–Óri, P.–Spéder, Zs. (szerk.): *Demográfiai portré 2018*, pp. 271–294., KSH Népeségtudományi Kutatóintézet, Budapest.
- Portfolio.hu. 2018. *Csökken a szegénység Magyarországon, de egyre jobban szétszakad a társadalom*. Retrieved from <https://www.portfolio.hu/gazdasag/20181130/csokken-a-szegenyseg-magyarorszagon-de-egyre-jobban-szetszakad-a-tarsadalom-306215>, Accessed on March 5, 2020
- Poumanyong, P. and S. Kaneko. 2010. "Does urbanization lead to less energy use and lower CO2 emissions? A cross-country analysis". *Ecological Economics*, 70(2), pp. 434-444.
- Rosas, J.; C. Sheinbaum; and D. Morillon. 2010. "The structure of household energy consumption and related CO2 emissions by income group in Mexico". *Energy for sustainable development*, 14(2), pp. 127-133.
- Santamouris, M.; K. Kapsis; D. Korres; I. Livada; C. Pavlou; and M.N. Assimakopoulos. 2007. "On the relation between the energy and social characteristics of the residential sector". *Energy and Buildings*, 39(8), pp. 893-905.
- Shammin, M. R.; R.A. Herendeen; M.J. Hanson; and E.J. Wilson. 2010. "A multivariate analysis of the energy intensity of sprawl versus compact living in the US for 2003". *Ecological Economics*, 69(12), pp. 2363-2373.
- Sun, C.; X. Ouyang; H. Cai; Z. Luo; and A. Li. 2014. "Household pathway selection of energy consumption during urbanization process in China". *Energy Conversion and Management*, 84, pp. 295-304.
- Szép, T. 2013. "Energiafelhasználás és energiahatékonyság". *Energiagazdálkodás* 54(4), pp. 18–21
- Taale, F. and C. Kyeremeh. 2019. "Drivers of households' electricity expenditure in Ghana". *Energy and Buildings*, 205, 109546.
- Tagai, G. 2015. "Járési népesség-előrejelzés 2051-ig". In: Czirfusz, M.–Hoyk, E.–Suvák, A. (szerk.): *Klimaváltozás – társadalom – gazdaság. Hosszú távú területi folyamatok és trendek Magyarországon*, Publikon Kiadó, Pécs, pp. 141-166.
- United Nations. 2018. *2018 Revision of World Urbanization Prospects*. Retrieved from <https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html>, Accessed on March 1, 2020
- Yang, Y.; J. Liu; Y. Lin and Q. Li. 2019. "The impact of urbanization on China's residential energy consumption". *Structural Change and Economic Dynamics*, 49, pp. 170-182.
- Ye, H.; Q. Ren; X.Hu; T. Lin; L. Xu; X. Li; and B. Pan. 2017. "Low-carbon behavior approaches for reducing direct carbon emissions: Household energy use in a coastal city". *Journal of Cleaner Production*, 141, pp. 128-136.
- Wang, Q.; M.P. Kwan; K. Zhou; J. Fan; Y. Wang and D. Zhan. 2019. "Impacts of residential energy consumption on the health burden of household air pollution: Evidence from 135 countries". *Energy policy*, 128, 284-295.
- Wiedenhofer, D.; M. Lenzen; and J.K. Steinberger. 2013. "Energy requirements of consumption: Urban form, climatic and socio-economic factors, rebounds and their policy implications". *Energy policy*, 63, pp. 696-707.
- Wei, N.; C. Li; X. Peng; F. Zeng; and X. Lu. 2019. "Conventional models and artificial intelligence-based models for energy consumption forecasting: A review". *Journal of Petroleum Science and Engineering*, 181, 106187.

AUTHOR BIOGRAPHIES

EMILIA NÉMETH-DURKÓ holds an assistant lecturer position of the Department of Finance at the Corvinus University of Budapest, where she teaches Corporate Finance and Business Valuation. Her field of research covers energy economics and environmental finance. Her e-mail address is durko.emilia@uni-corvinus.hu

PÉTER JUHÁSZ serves an associate professor at the Department of Finance at Corvinus University of Budapest (CUB). He holds a PhD from CUB and his research topics include business valuation, financial modelling, and performance analysis. His e-mail address is: peter.juhasz@uni-corvinus.hu

FANNI DUDÁS is a PhD student at Corvinus University of Budapest since 2019. Her research interest is focused on energy economics. Her e-mail address is: fanni.dudas@uni-corvinus.hu

SALES FORECASTING AND NEWSBOY MODEL TECHNIQUES INTEGRATED FOR MERCHANDISE PLANNING AND BUSINESS RISK OPTIMIZATION

Tomasz Brzeczek
Poznan University of Technology
Engineering Management Department
2 Jacek Rychlewski Str., Poznan 60-965, Poland
E-mail: tomasz.brzeczek@put.poznan.pl

KEYWORDS

Merchandise planning, newsboy model, risk, sales forecasting, profit maximization.

ABSTRACT

We consider a discrete newsboy problem of a supply quantity optimization to maximize profit and minimize risk objectives. Results show an advantage of introduction of time series data modelling and forecasting into simulation of demand distribution.

INTRODUCTION

Merchandising means the way in which the flow of merchandise is planned and managed from the supplier to the distribution centre (Jackson and Show 2001). Further we focus at merchandise planning defined as planning and control of merchandise inventory of the retail firm, in a manner, that balances between the expectation of target customer and the strategy of saler. Merchandise planning has a lot in common with assortment planning. Main objective of business' strategy is usually profit or revenue maximization.

Paper starts with the review of research conducted and techniques on modeling of demand and solving out supply quantity. Theory and models review implies that there is needed research on optimization techniques capable to fully integrate historical sales data analysis into profit maximization and decision risk analysis. Such approaches were focused at developing of multiple products expected profit maximisation or price management applications. We focus at research on simulating risk of profit using econometric time series of sales. Than we solve supply quantity using risk assessment.

Sales summary data and its time series modeling has been very popular in businesses for many years (Dalrymple 1987). Its handicap is wide choice of qualitative methods differing with complexity level. Econometric, statistical, but also heuristic methods like neural networks are developed (Sastri 1992). Moreover, econometric modelling of time series is applicable for wide product and businesses range and captures sales dynamics in time. An alternative is to use logit choice model based on sales transaction data. Its developed to capture substitution of alternatives of a product.

There is proposed a technique of analysis that integrates econometric modelling and forecasting of sales into a discrete newsboy problem of a supply quantity optimization. The technique uses forecasted expected value of sales and residuals estimates to simulate an empirical distribution of demand needed for a newsboy problem. Owing to this technique we process aggregate multi-product data. According to theory and practice merchandise planning needs data driven techniques with formality and aggregation level adjustable by a business and feasible for decision supporting system implementation (Hubner 2011).

In further section we discuss conditions for application of the technique. We also present results of our empirical research at the enterprise's retail merchandising. Pros and cons of the technique are discussed. Paper ends with conclusions and recommendations. We notice directions of further development of the technique in order to deepen profitability and risk analysis of product assortment and product variety decisions. The technique is extendable by a portfolio analysis of products sales and their correlations.

LITERATURE REVIEW

Merchandise planning is widely applied for retail sales and operations management. Operational research of merchandise planning consists of following aspects:

1. Customers' demand analysis and sales forecasting. Forecast of sales for entire organization, department and product wise is to be made. Product variety (width), breadth and depth as well as pricing and margin policy of a firm should be determined.
2. Determining supply quantity in order to maximise expected profit and subject to economic, financial risk and budget limit.
3. Merchandise stock control and stock keeping costs optimization.
4. Merchandise is assorted and presented usually product category wise and product competitive relations wise (Kok et al. 2009).

Regarding point 1 two main alternatives are time series modelling and microeconomic binomial or multinomial customer choice models (Kok et al. 2009). Choice is simple if business do not collect personal data of customers. Collecting such data is costly and

unwanted by some customers. Moreover, it is very difficult to communicate with consumers for salers of fast moving consumption goods and for producers that are not being distributors. Even if data was collected there would be many problems with individual declarations: errors, incompleteness of anonymous answers or face to face communication manipulations. Earnings or spending can be claimed concerning various time scale, currency, credit or number of households members. Moreover, even such complex transactions data base does not include potential customers that resign from buying. If data is replaced or supplemented with public statistics, the specifics of a company's market position, its customers demand and its market risk are missed.

Further in the paper we consider sales time series analysis. This choice impedes individual customers segmentation. However, product portfolio dimensions and product competitive relations are not missed if sales product categories' time series or total sales that can be disaggregated. Disaggregation can be done using results of Principal Component Analysis, ABC/XYZ analysis of stocks indicators or product share-growth matrix techniques. These are only examples because the literature that studies the economics of product variety is vast (Kok et al. 2009).

Hence, aggregate sales time series analysis is chosen we resign from revenue management technique (Bitran and Caldentey 2003). Its theoretical and formal complexity is high. It is used jointly with logit choice models and sales transaction data. Revenue management techniques at least theoretically have the advantage of prices optimization. In practice many stationary retailers can implement price changes only from period to period and rather in order to make adjustments to market demand and supply determinants or to sale out seasonal products or products with close expiration date. Under such circumstances newsboy model is appropriate. Prices should be calculated from market prices. In case of different life-cycle point for market and given product comparative pricing or analogies forecasting can be used (Vinod 2005).

The standard newsboy problem is a wide known basic exemplification of the profit optimization depending on supply quantity. It is also called a newsvendor model. Demand has random distribution. There are two variants: with discrete random variable of demand and with continuous one. Supply quantity, called also order quantity, is discrete or continuous appropriately as demand is. The model was extended or tested in order to:

- optimize alternative objectives of expected utility and budgeted profit (Lau 1980),
- prove optimality of the ordering rule based on the mean and the variance of demand with unknown distribution (Gallego and Moon 1993)
- capture risk influence on regular price and order quantity Agrawal and Seshadari (2000),
- minimize the cost of product variety choice (Rajaram 2001),
- introduce multi-stage supply chain dynamic programming (Kogan and Lou 2003),

- analyse benefits of risk pooling of individual demands having different level of variability (Gerchak and He 2003),
- model risk-sharing between the newsboy and the supplier (Cachon and Lariviere 2005),
- perform multi-product profit risk optimization by Vaagen and Wallace (2008),
- optimize a wholesale regular price treated as insured value in case of salvage (Watt and Vazquez 2015),
- capture advertising and marketing influence on demand and order quantity (Hrabec et al. 2017).

Gallego and Moon (1993) prove that expected value and standard deviation is all the information needed about demand distribution to figure out ordering quantity formula. Vaagen and Wallace (2008) provide an analysis of a few product variants order quantities with optimization of total expected profit, its variance and semivariance. They analyse theoretical two-state uniform distributions of variants demand with positive and negative correlations. We follow this works. However, we simulate demand distribution using empirical data about aggregate sales and results from its econometric modelling and forecasting procedure. Expected error of forecast is used instead historical mean.

The maximization of expected profit objective is the most effective decision under assumptions of stability of demand distribution and about many repetitions of the same decision which average result nearing the expected value. Lets assume that economy fluctuates dynamically or changes trend? Data modelling should result with a few following forecasts that are underestimated or overestimated. Therefore we analyse also usage of the measure of forecasting ex post error to correct the expectations and decision. Such a simulation of risk is in accordance with theory of scenario planning theory (Bishop et al. 2007).

THE TECHNIQUE OF SUPPLY OPTIMIZATION

The technique's concept is presented at Figure 1. Technique solves one of assortment planning problems and is determined by other assortment issues, business strategy and market conditions. Therefore assortment planning issues are an oval in the centre of the figure. Diamonds contain data or theory knowledge that should be help to aquire parameters of the problems. Arrows show the sequence of analysis and are signed with analysis names. One of them is theoretical and empirical analysis of preferences of business's target customer group. Aquired parameters are product width, product lines length and product assortment depth that constitute core of the business. Further analysis should concern at least these products if we want to achieve results that are viable for business and concern business risk. For currently operating undiversified company actual dimensions of merchandise can be taken into account.

Parameters of product prices and margins or overall trading margin are carried out from historical data about profitability. Possible adjustment for future profitability

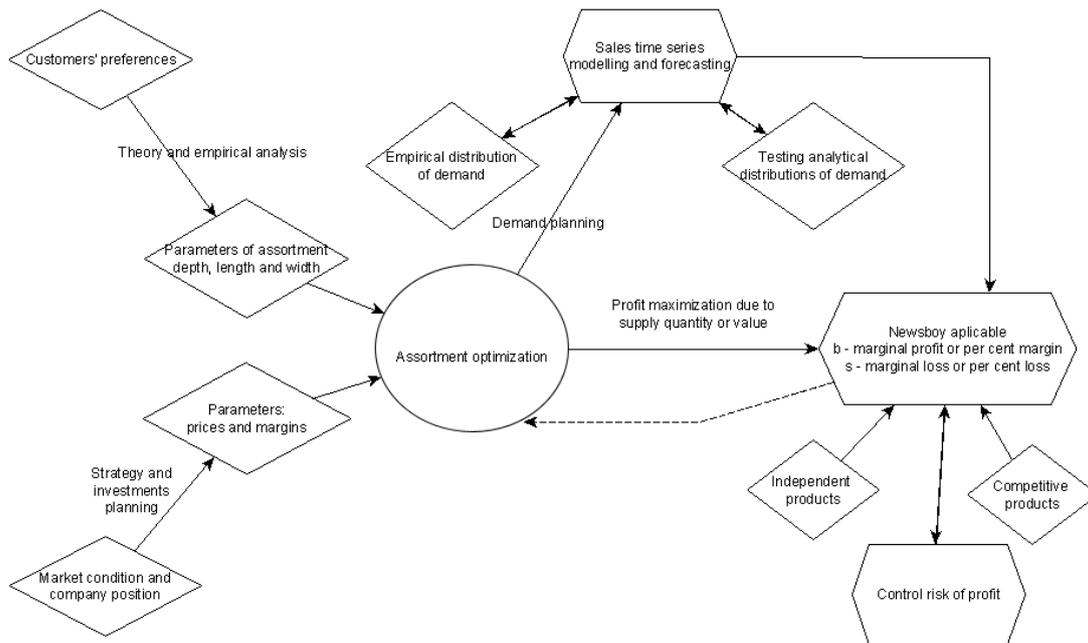


Figure 1: The Scheme of Proposed Assortment Planning Technique

can be made due to known or forecasted dynamics of market, demand, company position and strategy, and the finished investments in production or trading capacity. The technique concept is to integrate econometric modelling and forecasting of sales with optimization model. Hexagonal frames mark techniques.

Prior to maximize profit owing to solving supply quantity we need analytical or historical distribution of demand. Often analytical one is assumed after its validation with usage of historical data. We propose to simulate demand distribution using results of econometric modelling. In particular to use the forecast of expected value of sales. We use the standard error of the model and also expected forecast error as variability parameters of demand distribution. Validation sample of observations would allow also to calculate risk using forecasts errors and their measures.

Finally we solve newsboy problem for simulated demand distribution or distributions. Knowing margin and salvage costs parameters we apply newsboy model to one product demand and sales distribution or to multi-product aggregate sales. Appropriately we figure out optimal supply quantity of a given product or total supply value for multi-products. Total supply value and its proportional disaggregation into products' supply quantities is feasible for complements that do not compete. Aggregated analysis holds also for substitutes if their relative prices differ slightly and price substitution is outperformed by out of stock substitution. Regarding income effects aggregate sales analysis is valid if concerns fmcg goods but not luxury goods and high value goods. In case of independent product categories their demand distributions should be analysed separately.

The dotted line shows assortment dimensions adjustment implied by the result of the technique application. Especially if additional objective is the minimization of

risk of the expected profit due to competitive relations and estimated correlations. Such extension of the technique needs an appropriate risk function that takes into account product competitive relations. Modern portfolio theory is an appropriate technique.

RESEARCH

Data

We chose gastronomy and retail trade provision company located at sea ferry by a ferryman. Stocks cannot be supplied on every day basis as the analysed ferry sails regularly for a few days distance and goes back. Small orders can be done once a week. Company prefers big delivery once a month during a few days stay at main harbor. Crew members have free time.

As sales per cruise are not available in accounting reports we chose quarterly sales time series. Monthly sales analysis of supply would be the most accurate as such is delivery lead time. Additionally quarterly sales smooth weather anomalies in sales data and risk analysis is incomplete. We collected only quarter sales data and we apply the technique to this data. At least long durability merchandise can be ordered once quarterly. Other goods supply can be rescaled monthly or weekly.

Business's range consists of 3 product groups: restaurants and bars food provision, restaurants and bars alcoholic and non-alcoholic beverages, and FMCG goods provision for retailing shops. As these are generally compliments we will focus at total sales data.

Operating return on total sales fluctuated between 1,8 and 10%. So we assume that the highest value of 10% is the estimate of business's mark-up in relation to cost of good, service and trading costs. Offcourse if whole supply is sold. We assume also 10 per cent of lost value of merchandise that is supplied and doesn't find demand.

It is an alternative to salvage price and discount rate. It would mean that profitability of 1,8% was achieved when 55% of merchandise supply value was sold with 10% mark-up and 45% was unsold and generated loss that accounts for 10% of oversupply:

$$\frac{0,1 \cdot 0,55 \cdot \text{Supply} - 0,1 \cdot 0,45 \cdot \text{Supply}}{0,55 \cdot \text{Supply}} \cdot 100 = 1,8 .$$

Total sales from quarter 4 of 2014 to quarter 2 of 2019 is analysed time series. Last quarter is excluded to validation sample. Equation (1) presents first order differences model that was estimated:

$$\Delta y_t = -72\,500 - 6\,312\,510q_1 + 13\,728\,500q_2 + 3\,061\,170q_3 - 10\,487\,000q_4, \quad (1)$$

where q_1 – q_4 are dummy variables of quarters of the year. All parameters are significant at error probability level of at most 0,01. R square equals 0,97. Forecast error is by 5% higher than standard deviation of residuals. Forecast of expected value of total sales is calculated using the model of increment and previous quarter observation.

Further we analyse three types of empirical distributions of demand around forecast of expected sales that is 34 390 050:

- (1) modelled with sales model errors, possible demanded amounts differ from historical sales time series,
- (2) uniform distribution between pessimistic, neutral and optimistic scenarios of demand, where a mean of negative errors, the mean of all errors and the mean of positive errors are calculated and added to a forecast,
- (3) uniform distribution between pessimistic, neutral and optimistic scenarios of demand, where from the forecast of expected value we deduct root of average square of negative errors of the model and we add root of average square of its positive errors.

Results

Calculations of total profit were made in accordance with formula (2):

$$P(z, d) = 0,1\min(z; d) - 0,1\max(0; z - d) \quad (2)$$

where z means supply amount and d means demand amount. For different distributions of demand amounts D was calculated optimum supply z^* that gives maximum of expected profit value:

$$\pi = \max_z E[P(z, D)]. \quad (3)$$

To analyse risk we calculate standard deviation $s(z)$. We use it to calculate maximum of low profit:

$$\pi^- = \max_z \{E[P(z, D)] - s(z)\}. \quad (4)$$

Results are presented in Table 1. Realised profit is calculated for one observation in the validation sample using supply quantity that maximizes expected profit given in the row entitled z .

Table 1: Results for Simulated Distributions

	Distribution		
	1	2	3
z	34677366	34390050	34677366
Π	3312025	3347864	3348439
Realised profit	2951924	2980656	2951924
s	202171	128893	212697
Π^-	3109854	3218971	3135742
$\text{Arg}[\Pi^-(z)=\max]$	33047325	33022934	33534602

Comparing distributions 1, 2 and 3 the last two give more actual data about risk owing to grouping negative errors into pessimistic scenario with pessimistic realisation and grouping positive errors into optimistic scenario with optimistic realisation. If errors are not grouped their values compensate themselves to 0 so if negative errors are less numerous they are on average bigger in absolute terms. Should their risk be flattened by their smaller frequency in all errors number? This is the case that occurs in the analysed data. In the analysed time series there were less negative errors but with higher average in absolute terms. Therefore standard deviation of errors is replaced with average errors in positive errors group and in negative errors group separately for distribution 2. Root of mean squares in groups are used is distribution 3. Although demand distributions 1 and 3 differ they result with the same supply quantity decision. If errors in one or in both groups would be more variable usage of third distribution gives an advantage. Distribution 2 accounts for variability of errors levels and is also the most robust to difference in negative and positive errors frequency. Hence, we recommend usage of distribution 2 in case of empirical data without known analytical distribution. It is valid especially for small sample that can face high error of estimation. Finally the technique allowed to simulate different distributions of demand using the same time series of sales. Using distribution 2 and solving newsboy problem resulted with lower optimal supply and with higher realized profit. Using less risky objective distribution 2 resulted with the smallest supply order among distributions. The smallest supply would result with the highest profit in next quarter as sales were lower than expected.

REFERENCES

- Agrawal, V. and S. Seshadari. 2000. "Impact of Uncertainty and Risk Aversion on Price and Order Quantity in the Newsvendor Problem". *Manufacturing & Service Operations Management* 2, 410-423.
- Bishop, P.; Hines, A. and T. Collins. 2007. "The Current State of Scenario Development: an Overview of Techniques". *Foresight* 9, No. 1, 5-25.
- Bitran, G. and R. Caldentey. 2003. "An Overview of Pricing Models for Revenue Management". *Manufacturing & Service Operations Management* 5, No. 3, 203-229.
- Cachon, G.P. and M.A. Lariviere. 2005. "Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations". *Management Science* 51, No. 1, 30-44.

- Dalrymple, D.J. 1987. "Sales Forecasting Practices: Results from a United States Survey". *International Journal of Forecasting*, No. 3-4, 379-391.
- Gallego, G. and I. Moon. 1993. "The Distribution Free Newsboy Problem: Review and Extensions". *Journal of the Operational Research Society* 44, No. 8, 825-834.
- Gerchak, Y. and Q-M. He. 2003. "On the Relation Between the Benefits of Risk Pooling and the Variability of Demand". *IIE Transactions* 2003, No. 35, 1027-1031.
- Hrabec, D.; Haugen, K.K. and P. Popela. 2017. "The Newsvendor Problem with Advertising: an Overview with Extensions". *Review of Management Science* 11, 767-787.
- Hubner, A. 2011 *Retail Category Management. Decision Support Systems for Assortment, Shelf Space, Inventory and Price Planning*, Springer, Heidelberg.
- Jackson T. and D. Shaw. 2001. *Mastering Fashion Buying and Merchandising Management 2001*, T, Jackson and D. Shaw. Palgrave MacMillan, London.
- Kogan, K. and S. Lou. 2003. "Multi-stage Newsboy Problem: A Dynamic Model". *European Journal of Operational Research* 2003, No. 149, 448-458.
- Kok, G.A.; Fisher M.L. and R. Vaidyanathan. 2009. "Assortment Planning: Review of Literature and Industry Practice". In *Retail Supply Chain Management 2009*, N. Agrawal and S.A. Smith (Eds.), Springer Science+Business Media, 99-153.
- Lau, H-S. 1980. "The Newsboy Problem under Alternative Optimization Objectives". *Journal of Operational Research Society* 31, 525-535.
- Rajaram, K. 2001. "Assortment Planning in Fashion Retailing: Methodology, Application and Analysis". *European Journal of Operations Research* 129, No. 1, 186-208.
- Sastri, T. 1992. "Multiple-Step-Ahead Prediction by Hierarchical Neural Networks". In *Proceedings of a 1992 Joint German/US Conference Operations Research in Production Planning and Control* (Hagen, Germany, June 25-26), Springer-Verlag, 529-549.
- Vaagen, H. and S.W. Wallace. 2008. "Product Variety Arising from Hedging in the Fashion Supply Chains". *International Journal of Production Economics* 114, No. 2, 431-455.
- Vinod, B. 2005. "Practice Papers: Retail Revenue Management and the New Paradigm of Merchandise Optimisation". *Journal of Revenue and Pricing Management*, No. 3, 358-368.
- Watt, R. and F.J. Vazquez. 2015. "An Analysis of Insurance in the Newsboy Problem". www.semanticscholar/paper.

Tomasz BRZECZEK is an Assoc. Prof. in the Dept. of Engineering Management at Poznan University of Technology, Poland. He received his PhD in Economics from the Poznan University of Economics, Poland, where he was a PhD student in Operations Research Chair. He was a member of organizing committees at international conferences. He is the author and co-author of over 50 scientific publications, including monographs and papers with IF. He specializes in time series econometrics and operations research

INCOME INEQUALITY IN HUNGARY

Ildikó Gelányi (ildiko.gelanyi@uni-corvinus.hu)
Department of Banking and Monetary Finance

András Olivér Németh, PhD (nemeth.andras@uni-corvinus.hu)
Department of Economic Policy and Labour Economics

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

Income inequality, tax and benefit system, income taxation, social transfers, deciles, regional differences.

ABSTRACT

In this article, we describe the income inequality situation in Hungary from two different perspectives: inequalities among income deciles, and regional differences. Both types of inequalities have increased in the last few years due to changes in the tax and benefit system. An important contributing factor of increasing income inequalities was the introduction of a linear personal income tax, together with the increased role of tax allowances in the family support system. Regional differences have been traditionally significant in Hungary (despite the small size of the country), and the positions of the least developed regions of the country have continued to worsen in the last few years. After assessing the role of the tax and benefit system, we also briefly try to give some insight to possible interventions in order to decrease inequalities among income deciles.

INTRODUCTION

Income inequality is an important determinant of the well-being of a society. Besides promoting economic growth (i.e. the achievement of a higher level of overall welfare), economic policy should also contribute to a fair distribution of the fruits of this growth. The tax system has a vital role in this process; one of its main functions is to redistribute incomes from the richer to the poorer, therefore to decrease income inequality. However, according to the European Commission (2020) income inequality has not only increased in Hungary in the past few years, but “changes in the tax and benefit system ... contributed to the increased level of income inequality.” (pp. 29-30)

In this paper, we analyse income inequality in Hungary in two perspectives: inequality among high-income and low-income citizens, and regional inequality. We also examine the role of the tax and benefit system in decreasing inequalities. Finally, we calculate the rate of increase of

social incomes necessary in order not to have an upward trend in income inequality in the next few years.

We have used data from the Hungarian Central Statistical Office (CSO 2020a) and the National Tax and Customs Administration (NTCA 2019). The two sets of data are not entirely compatible with each other, which restricts our ability to combine them in our simulation. The reliability of these datasets is also somewhat questionable for different reasons. The data of CSO are based on self-reported answers of households, therefore may be subject to either intentional or unintentional distortions. The NTCA dataset reports information from tax files, and naturally it does not include illegal incomes. Another limitation of our calculations is that they are based on averages, therefore they cannot give proper information about the heterogeneity of the society.

INEQUALITY AMONG INCOME DECILES IN HUNGARY

Measuring inequality

We have calculated the Y90/10 and Y80/20 ratios given by Equations (1) and (2) for earned incomes (including pension benefits) and Equations (3) and (4) for disposable incomes.

$$Y90/10 = \frac{Y_{10}}{Y_1} \quad (1)$$

$$Y80/20 = \frac{Y_{10} + Y_9}{Y_1 + Y_2} \quad (2)$$

$$Y90/10_D = \frac{Y_{10} - t(Y_{10})}{Y_1 - t(Y_1)} \quad (3)$$

$$Y80/20_D = \frac{Y_{10} - t(Y_{10}) + Y_9 - t(Y_9)}{Y_1 - t(Y_1) + Y_2 - t(Y_2)} \quad (4)$$

where Y_i denotes the earned income of i -th decile (including pension), and $t(Y_i)$ denotes the net payment for the state (taxes and other charges minus family-related and other social transfers excluding pensions).

We have used data from the Hungarian Central Statistical Office (CSO 2020a). Y90/10 shows the ratio of incomes of the richest and poorest 10 percents of the population. Table 1 summarizes these ratios for disposable incomes.

Table 1: Social inequality after tax and transfers (own calculation based on data from CSO 2020a)

Year	Y80/20 _D	Y90/10 _D
2010	4.65	7.26
2011	4.62	7.39
2012	5.08	8.26
2013	5.21	8.47
2014	5.10	8.34
2015	5.01	8.22
2016	5.08	8.55
2017	4.89	8.19

In 2017, the earned income of the wealthiest 20 percent of the population was 7.74 times higher than that of the bottom 20 percent, meanwhile the difference was 7.63-fold in 2010. That is, inequality increased between these two years, although its value fluctuated during this period. The same ratios in the case of disposable income are 4.89 in 2017 and 4.65 in 2010 (see Table 1). Therefore, a larger increment can be seen in the inequality of disposable income. The Country Report of the European Commission (2020) used different data, but their result is similar: changes in the tax and benefit system have not decreased inequality. What is more, according to the Report, Hungary experienced the largest increase in inequality in the EU (from 3.6-fold to 4.4-fold difference between the top and bottom 20 percents of the population) between 2008 and 2018.

Measuring the redistributive effectiveness of the tax and benefit system

The ratio of the previously defined measures of inequality, $Y90/10_D$ and $Y90/10$, shows how the tax and benefit system affects inequality: how the difference between the incomes of the top and bottom 10 percents of the population is decreased by redistribution. A similar index can be defined to measure the role of redistribution between the top and bottom 20 percents of the population. These indices are given by Equations (5) and (6):

$$I_{90/10} = \frac{Y90/10_D}{Y90/10} \quad (5)$$

$$I_{80/20} = \frac{Y80/20_D}{Y80/20} \quad (6)$$

These indices can measure the success of the government's redistributive function: the tax and benefit system works better from this point of view, if it significantly decreases the difference between the top and bottom strata.

In Figure 1, we can see that this redistributive role of the Hungarian tax and benefit system weakened somewhat between 2010 and 2017. The ratios of inequality measures at the level of disposable and earned incomes have not shown a clear trend neither in the case of income deciles, nor in the case of income quintiles, but the indices were higher in 2017 than in 2010 in both cases.

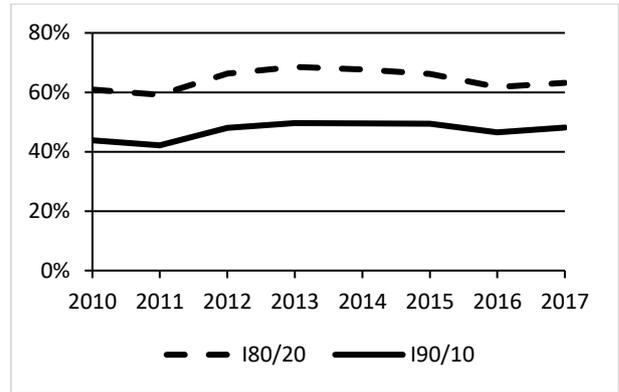


Figure 1: The change in the redistributive role of the tax and benefit system (own calculation and graph based on data from CSO 2020a)

Modelling the effects of a possible restructuring of the transfer system

In this subsection we analyse how the redistributive properties of the tax and benefit system would be affected by a change in the transfer system. We have calculated with a rearrangement of family-related and other social transfers: multiply the actual transfers of the different income deciles with the coefficients in Table 2. (Unfortunately, we cannot calculate with the family tax allowance since there is no data available.)

Table 2: Multiplier of transfer payment in the deciles

1	2	3	4	5	6	7	8	9	10
1.5	1.4	1.2	1	0.8	0.6	0.4	0	0	0

Figure 2 shows how such a restructured transfer system would affect income inequality. The calculated measures of inequality in the disposable incomes would be significantly lower in the modelled hypothetical case (marked with *) than their actual values. However, the level of inequality would not decrease through the examined period even in this hypothetical scenario.

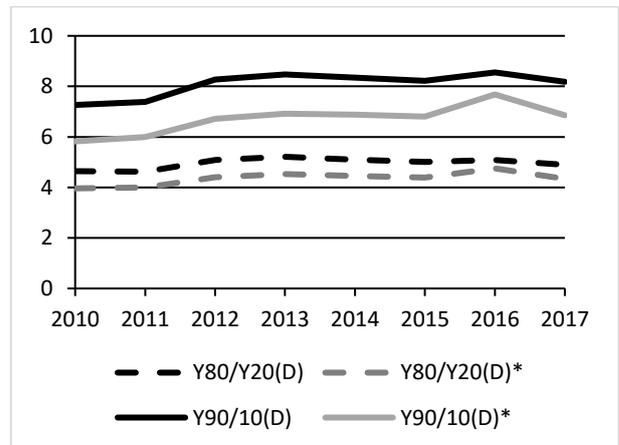


Figure 2: Actual and hypothetical inequality based on the modelled restructuring of the transfer system (own calculation and graph based on data from CSO 2020a)

According to this calculation, a significant improvement of the redistributive effectiveness of the tax and benefit system would make the rethinking of family tax allowances necessary. Table 3 contains the indices of the redistributive effectiveness in the case of the modelled restructuring of the transfer system.

Table 3: Redistributive effectiveness in the case of a restructured transfer system (own calculation based on data from CSO 2020a)

Year	$I_{80/20}^*$	$I_{90/10}^*$
2010	52%	35%
2011	51%	34%
2012	57%	39%
2013	60%	41%
2014	59%	41%
2015	58%	41%
2016	58%	42%
2017	56%	40%

Why does it seem essential to change the existing system of family tax allowances? In Hungary, it is increasing steeply with the number of children. In case of 1 or 2 children the first decile's average family cannot exploit the whole amount of the tax allowance. In the case of three children, the bottom 8 deciles' average families would not be able to exploit totally the tax base allowance. (The unusable tax allowance of average families for each deciles in case of 1, 2 or 3 children is summarized in Table 4.) However, averages do not represent the society properly. The average household size is 2.3 persons, while those families that are entitled to higher levels of allowances are naturally larger. Actually, a family of 2 earners and 3 children in and above the third income decile is able to exploit the whole tax allowance.

Table 4: Unused family tax base allowance for an average household at different number of children (estimated data for 2018)

Decile	Average earned income of the household	Unused tax base allowance 1 child	Unused tax base allowance 2 children	Unused tax base allowance 3 children
1	736 717	63 283	249 839	957 839
2	1 826 846	0	0	623 518
3	2 529 075	0	0	392 750
4	2 561 783	0	0	385 660
5	2 413 206	0	0	425 652
6	2 693 192	0	0	339 360
7	2 695 343	0	0	336 660
8	3 460 831	0	0	98 345
9	4 032 049	0	0	0
10	6 626 559	0	0	0

REGIONAL INEQUALITIES IN HUNGARY

Although Hungary is a relatively small country, significant regional differences can be seen within its borders. According to the latest Eurostat (2020) data, the per capita GDP in Budapest is approximately 4.5 times higher than in the poorest county (Nógrád). Taking into account purchasing power parity, Budapest is at 145 percent of the post-Brexit EU27 average, while Nógrád is at only 32 percent. Per capita GDP in Hungary as a country is at 71 percent of the EU27 average. If we concentrate on the level of regions (mostly meaning the NUTS-2 statistical regions), we can find Central Hungary (including Budapest) at 108 percent of the EU average, and the Northern Great Plain region at 46 percent.

Table 5: Regional differences in per capita GDP (source of data: Eurostat 2020)

Region	Per capita GDP in purchasing power standards (% of the EU27 average)
Central Hungary	108
Western Transdanubia	72
Central Transdanubia	66
Southern Great Plain	52
Northern Hungary	49
Southern Transdanubia	49
Northern Great Plain	46

As can be seen from Table 5, the regions of Hungary can be classified into three clusters. The first cluster contains only Central Hungary, which is much more developed than the other parts of the country. The second cluster includes two regions: Western and Central Transdanubia. In comparison, the third cluster consists of the remaining four regions, all of which are among the 25 poorest ones within the European Union.

Regional inequalities based on tax and income data

A similar picture of significant regional differences arises if we analyse the data from personal income tax statistics. The National Tax and Customs Administration of Hungary regularly publishes the most important tax statistics in a statistical yearbook. These yearbooks contain the main data about the personal income tax filers from a regional aspect as well. The latest available yearbook (NTCA 2019) covers the year 2017.

Table 6 shows the average yearly taxable income per filer in the different regions. The table also provides the data expressed as a percentage of the national average. Although the exact ranking of the regions is slightly different from what we have seen in Table 5, the general picture is similar. Average personal incomes in Central Hungary are significantly higher than in any other part of the country. Central and Western Transdanubia are around the national average, while the remaining four regions are clearly below it and relatively close to each other.

Table 6: Average taxable income per filer in the regions of Hungary, 2017 (source of data: NTCA 2019)

Region	HUF	% of national average
Central Hungary	3,384,168	125.0%
Central Transdanubia	2,704,610	99.9%
Western Transdanubia	2,556,155	94.4%
Southern Transdanubia	2,243,705	82.9%
Southern Great Plain	2,241,577	82.8%
Northern Hungary	2,205,389	81.5%
Northern Great Plain	2,109,121	77.9%

These regional differences have been fairly stable in the last few years. If we compare the data from the NTCA yearbooks, we can see, that in the period between 2010 and 2017, per capita taxable income in Central Hungary has always varied between 125 and 130 percent of the national average, while on the other hand, the Northern Great Plain region has always been between 77 and 80 percent.

According to the Hungarian tax legislation, taxable income consists of two main groups: consolidated incomes (consisting mainly of salaries and incomes from self-employment) and separately taxed incomes (including e.g. capital gains or income from private businesses). On average, more than 90 percent of the incomes are in the first group. Besides reporting average incomes, the NTCA yearbooks also provide data about the distribution of consolidated incomes. Figure 3 shows this distribution in the three aforementioned clusters. The first cluster (Central Hungary) is characterized by a significantly higher share of yearly incomes above 20 million HUF, and between 10 and 20 million HUF (9.6 and 14.4 percent, respectively). In the second cluster (Central and Western Transdanubia), the share of high incomes is lower, and the overall distribution is fairly similar to that of the national level. The third cluster (the remaining four regions) has still lower average incomes, and the share of yearly incomes below 2 million HUF is significantly higher than in the more developed parts of the country (25.2%).

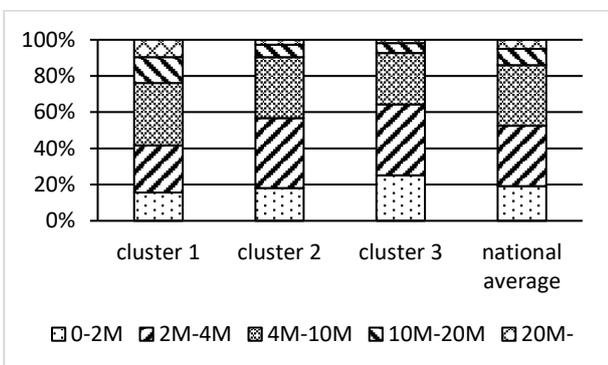


Figure 3: Distribution of consolidated incomes, 2017 (own calculation and graph based on data from NTCA 2019)

Income inequalities are frequently described by the Lorenz curve and the Gini coefficient. The available data

from the personal income tax statistics make it possible to graph a Lorenz curve using county-level aggregate data. Figure 4 shows this Lorenz curve for the 20 Hungarian counties (including Budapest). The Lorenz curve itself graphs the cumulative incomes of the counties as a function of the cumulative number of tax filers in the counties (to do that, counties have to be ranked from lowest to highest average income per filer). As a comparison, the 45-degree straight line also appears on the figure – this would be the Lorenz curve in the case of no difference among the counties in per capita incomes.

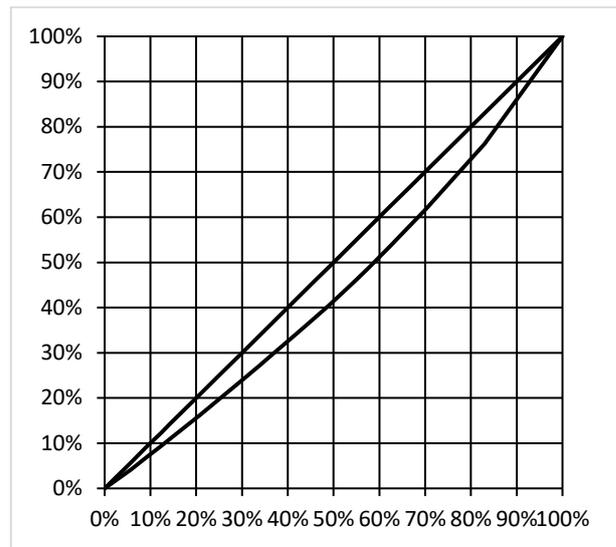


Figure 4: The Lorenz curve of Hungarian counties (own calculation and graph based on data from NTCA 2019)

The Gini coefficient measures the area between the 45-degree straight line and the Lorenz curve, relative to the area of the whole triangle below the 45-degree straight line. In the case of the Lorenz curve on Figure 4, the value of the Gini coefficient is 11.6 percent. According to World Bank data, Hungary’s overall Gini coefficient is 30.9 percent, i.e. we can say that the regional differences can explain a significant share of overall income inequalities in Hungary.

Taxes and social incomes

An important question about regional inequalities and the tax system is whether the tax and benefit system decreases the inequalities or not. One way to assess this issue is by comparing income tax payments and social incomes (including pension, child-related benefits etc.). The Central Statistical Office of Hungary publishes data about the latter.

Regions with higher taxable incomes pay more taxes, but also tend to receive more social incomes as well. However, if the ratio of income tax payment and social incomes (on regional level) is higher in regions that have higher taxable incomes, then we can say that the tax and benefit system decreases regional inequalities somewhat. As it can be seen in Figure 5, this is the case in Hungary.

In Central Hungary, where per capita taxable income is much higher than in other parts of the country, the ratio of tax payments and social incomes is 49.1%, significantly above the similar ratios in other regions. Generally, we can also see the positive relationship between the per capita average taxable income and the tax-to-social income ratio among the regions.

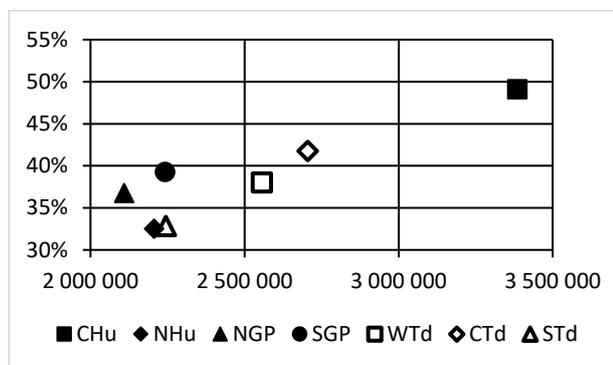


Figure 5: The ratio of income tax payments and social incomes in Hungarian regions, as a function of the average per capita taxable income (own calculation and graph based on data from NTCA 2019 and CSO 2020)

However, it is also true that the role of regional redistribution has significantly decreased since 2010. In that year, the tax-to-social income ratio was 57.9 percent in Central Hungary, 37-38 percent in Western and Central Transdanubia, and only 25-26 percent in the remaining four, more deprived regions of the country.

SIMULATION OF THE REDUCTION OF INEQUALITY

In the following simulation, we are looking for the necessary raise in social transfers to avoid a further increase in inequality. We calculate the future incomes based on those of 2017 as the latest reliable data. Taking into consideration that we are in 2020, but we do not have a long enough time series of comparable data, we determined the last year of our simulation to be 2024.

Before completing the simulations, we converted income data for 2010-2017 to 2018 prices by using the consumer price index reported by the CSO (2020b). Expressed in 2018 prices, we found that the real value of social transfers excluding pensions has decreased in the examined period regarding the lower deciles (see Table 7).

According to the statistical data, social transfers have decreased significantly in 2012 in the bottom 6 deciles, while they have increased in the 7th, 9th, and 10th deciles, especially in the wealthiest decile. The general increase of social incomes in the later years was not able to compensate for this change.

Table 7: Changes in the real value of social transfers (in 2018 prices) in the bottom 2 and top 2 deciles (own calculation based on data from CSO 2020a and 2020b)

	Deciles			
	1	2	9	10
2011	-8.2%	-0.3%	-0.5%	1.7%
2012	-14.1%	-23.7%	4.2%	13.5%
2013	0.8%	2.7%	2.2%	0.0%
2014	0.2%	-3.8%	1.9%	4.0%
2015	2.7%	2.2%	2.7%	2.4%
2016	0.8%	0.8%	1.5%	2.4%
2017	2.5%	-0.7%	1.0%	-0.6%
<i>mean</i>	-2.2%	-3.2%	1.9%	3.3%

Because of the abovementioned situation, we decided to simulate in two different ways: first, we use the average rate of change of social transfers for each decile experienced between 2010-2017 to determine the surplus rate by which the tendency has to be modified to reach our goal. Second, because the average rates of change are highly distorted by the data of 2012, i.e. they do not properly represent the whole time period, we also calculated a general annual growth rate for social transfers that would be necessary in the entire society in order to avoid the worsening of the inequality situation.

First of all, we calculated the rate of inequality in disposable income between the highest and the lowest quintiles. In the calculation of disposable income, we used the following equation:

$$Y_D = Y \cdot (1 - t_e) + TR \quad (7)$$

where Y is the earned gross income including pensions, t_e is an effective tax rate determined by different assumptions (see below), and TR denotes social transfers excluding pensions. Table 8 shows the $Y_{80}/20_D$ rates depending on whether we include the experienced trend of social incomes in the calculation, or not. In this case, we used the effective tax rate of 2017 in the calculation of Y_D of each decile.

Table 8: The projected changes in the rate of disposable income ($Y_{80}/20_D$) in two different approaches

	w/o tendency	with tendency
2017	4.89	4.89
2018	5.01	5.05
2019	5.12	5.20
2020	5.24	5.36
2021	5.36	5.53
2022	5.49	5.69
2023	5.61	5.86
2024	5.74	6.04

In the simulation, we calculated the following cases, while the growth rate of the gross income is fixed at the average level of the 2010-2017 period:

- the tax rate is at the level of 2017 in each decile;
- the tax rate is at the minimum level of the period in each decile;
- the tax rate is at the maximum level of the period in each decile;
- in the lowest five deciles, the tax rate is at the minimum level, while in the highest five deciles, it is at the maximum level (something like a system of progressive taxation).

We were searching for the necessary growth rate of social transfers to keep income inequality at its 2017 level. We calculated our results both if there is a trend in social transfers and if there is not. So, these results can be interpreted as a minimum goal of the government in influencing social transfers if its objective is not to let the Hungarian income inequality situation worsen.

Table 9: The result of the simulation in different cases (goal is $Y_{80/20_D}(2024) = Y_{80/20_D}(2017)$)

	Growth rate (no trend)	Growth rate (above trend)
t(2017)	7.64%	10.90 %
t _{min}	9.87%	13.21%
t _{max}	11.20%	14.56%
t(min ₍₁₋₅₎ ,max ₍₅₋₁₀₎)	7.64%	10.90%

CONCLUSION, FURTHER RESEARCH

We have examined the topic of income inequalities in Hungary from different perspectives: we analysed inequalities among income deciles and also regional differences. As we have seen, there are significant inequalities in Hungary, and the changes in the tax and benefit system since 2010 have contributed to increased inequalities. In our opinion, this tendency should be turned (or at least moderated), since larger inequalities lead to both social tensions and economic problems. In this article, we also briefly gave some insight to possible interventions in order to decrease inequalities.

We plan to further examine income inequalities, including a simulation about how regional differences can be expected to change and how they can be affected by government redistribution. We also would like to incorporate a more detailed analysis of economic policy measures, including the system of family tax allowances and family related social transfers.

REFERENCES

- CSO. 2020a. *Online database* (ksh.hu/engstadat), Data of total households by deciles, regions and type of settlements (2010-2018). Available at https://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_zhc014d.html?down=2314. Downloaded on February 7, 2020.
- CSO. 2020b. *Online database* (ksh.hu/engstadat), Time series of consumer price indices (1985-2019). Available at https://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_zhc014d.html?down=2314. Downloaded on February 10, 2020.
- European Commission. 2020. "Country Report Hungary 2020" Available at https://ec.europa.eu/info/sites/info/files/2020-european_semester_country-report-hungary_en.pdf. Downloaded on February 28, 2020.
- Eurostat. 2020. *Online database* (<https://ec.europa.eu/eurostat/data/database>), Gross domestic product (GDP) at current market prices by NUTS 2 regions (nama_10r_2gdp) data series. Downloaded on March 8, 2020.
- NTCA. 2019. *NAV évkönyv 2018 [NTCA Yearbook 2018]*. National Tax and Customs Administration, Budapest.

AUTHOR BIOGRAPHIES

Ildikó GELÁNYI received her master in economics at Corvinus University of Budapest and applied mathematics at Eötvös Loránd University. She is an assistant professor at the Department of Banking and Monetary Finance at Corvinus University of Budapest. Her e-mail address is ildiko.gelanyi@uni-corvinus.hu.

András Olivér NÉMETH, PhD is an assistant professor at the Department of Economic Policy and Labour Economics at Corvinus University of Budapest. His teaching portfolio includes several subjects from microeconomics to public economics and economic policy; his main research interests are economic growth and fiscal policy. His e-mail address is: nemeth.andras@uni-corvinus.hu.

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

THE NECESSARY SIZE OF THE SKIN-IN-THE-GAME TO STAY IN THE GAME

Kira Muratov-Szabó
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
KELER CCP Ltd
Rákóczi street 70-72. Budapest, 1074, Hungary
E-mail: muratov.kira@gmail.com

Melinda Szodorai
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
KELER Ltd
Rákóczi street 70-72. Budapest, 1074, Hungary
E-mail: szodorai.melinda@keler.hu

Andrea Prebuk
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
E-mail: prebuk.andrea@gmail.com

Kata Váradi
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
E-mail: kata.varadi@uni-corvinus.hu

KEYWORDS

Default waterfall, central counterparty, stress test, skin-in-the-game

ABSTRACT

The role of central counterparties (CCPs) is to manage counterparty risk during trading on exchanges. In order to assure resilience, CCPs are required to operate a multilevel guarantee system, the default waterfall (DW). Our paper focuses on the third layer of this DW, which is the own capital of the CCP, the skin-in-the-game (SITG). While EMIR (European Market Infrastructure Regulation), the European regulatory background for CCP, expects a 25% contribution to the guarantee system of its own capital, experts challenge the necessity of such a level of own capital inclusion. The size does not serve only as a line of defense against clearing member default, but it also creates incentives for both CCPs and its participants to adjust their risk-taking depending on the level of SITG contributed. In our model, we quantify the SITG and show the difference between covering the losses exclusively with the defaulting member's own resources and the SITG or also with non-defaulting members' contributions.

INTRODUCTION

Following the global financial crisis in 2008, CCPs gained a significant role since the crisis exposed the vulnerability of the whole financial system. Regulators and every actor of the market are interested in assuring the resilience of the clearinghouses. The reinforcement of financial stability has gone far since the crisis; however, further implementations and improvements are to be made.

The CCP becomes a "central" party between traders, becoming a buyer to the seller, and a seller to the buyer,

a process called novation. The two parties are, therefore, no longer exposed to each other, but only to the CCP, which provides insurance against bilateral default risk (Biais et al., 2016). Many researchers (Duffie, 2015, Lopez and Saeidinezhad, 2017, Markose et al., 2015) point out that the CCPs may bear systematic risks, but the benefit they provide is undeniable. For example besides taking over counterparty risk, the CCP takes market participants' trading exposures onto its balance sheet, relieving the counterparties of multilateral risk exposures.

In order to assure robustness to default, CCPs operate a default waterfall system. These are resources of the CCP, serving the assurance of absorption of losses generated by the default of a participant. EMIR requires three typical resources of the waterfall system: margins, skin-in-the-game, and default fund contributions.

Our research aims to analyze how much "skin" is enough for a CCP to withstand extreme shocks to protect the non-faulty members' default contribution to be exhausted. Researchers examined the loss mutualization role of the default waterfall, of which the CCP's own capital takes part as well. The dependency among members may have negative externalities when members take excessive risk in the hope of higher returns, even on the cost of the CCP. The trade-off regulators face while CCPs collect the default fund contributions is to prevent members from excessive risk-taking and, at the same time, keeping the funding cost low, pushing the incentives towards safer investments. The default fund designs and the capital a CCP is including in the fund affect the incentives of participants.

The paper proceeds as follows: at first, the default waterfall is briefly presented, followed by the detailing of the skin-in-the-game amount effect on incentives. In the third part, the model is presented. Lastly, results are detailed, and the paper closes with our conclusion.

LITERATURE

A CCP's leading role and purpose are to centralize counterparty risk management in the financial markets they operate (Pirrong, 2011). Hughes and Manning (2015) compare the risk profiles of CCPs and banks. In their perspective, the primary financial risk of a CCP stems from the likelihood that the CCP executes the replacement trades in its matched book at a disadvantageous price. Consequently, market liquidity risk arises in the case of a member's default. For banks, the most critical risk factor is the credit risk of their borrowers.

Researchers (including Murphy and Nahai-Williamson 2014; and Pirrong 2011, 2014, Hughes and Manning 2015) identify the vulnerable points by which a CCP could trigger or amplify systematic risk that includes liquidity risk, information, and incentive issues, too.

Regulatory background

Authorities have put great attention to strengthen the global safeguards for central clearing by adopting the CPMI-IOSCO Principles for Financial Market Infrastructures, dedicated Financial Stability Board guidance, and the EMIR regulation in 2012. EMIR requires CCPs to ensure the resilience and stability of the financial system.

CCPs must operate a default waterfall system (EMIR, Article 45, 2012) in order to have access to financial resources in the event of a clearing member's (CM) default. The main elements of the default waterfall, are the margin, the default fund, and the skin-in-the-game. EMIR Article 41 and chapter VI. of the regulatory technical standards (RTS, 2013) – supplementation of EMIR – require CCPs to have proper margining methodologies that enable CCPs to cover losses of the defaulting member. This is the first layer of defense; therefore, the regulatory background sets rigorous requirements for CCPs regarding the methodology. The aim is that the amount of the defaulting member's margin shall be enough to cover the losses it generates. It is also important to mention that the statistical model shall cover the losses *in normal market conditions*. The parameters for non-OTC financial assets have a confidence level of at least 99%; lookback period of 250 days that includes a stressed period, liquidation period is at least two days. EMIR also emphasizes the importance of the inclusion of procyclicality. Different procyclicality handling methods are analyzed by Berlinger et al. (2018), Szanyi et al., (2018), give an overview of the deficiencies and potential improvements of the procyclicality requirements.

While margins serve to cover losses from day-to-day business, the second layer, the default fund contribution sources (EMIR, 2012, Article 48), serve to protect the system in extreme market conditions. This is why it is required to be designed to withstand extreme but plausible market conditions. According to the practice (e.g., KELER CCP, 2020), EMIR (2012, Article 49.), the default fund contribution value is determined by

stress tests. RTS (2013) demands historical and hypothetical scenarios, and CCPs should cover the default of the CM with the highest, or the second and third highest ($\max(1;2+3)$) exposure. Loss-mutualization and cross guarantee between clearing members appear at this level. However, the rule is to exhaust the resources of the defaulting member's contribution first. The non-defaulting members' contributions are used only after the third layer of defense is proven to be inadequate.

The applicable regulations involve the CCP in the game as well. The third layer is the so-called skin-in-the-game, which is the own funds of the CCP. Depending on the policy and default waterfall model the CCP applies, there could be a second skin-in-the-game as a resource if the funds mentioned above are not enough to cover the losses.

For reaching the end of the default waterfall (R&R CCP, 2019), in case of the inadequacy of funds, the CCP must be prepared. A recovery plan is put into force if the CCP reaches its funds' limits (Cont, 2015).

Skin-in-the-game and the role it plays

The current regulatory framework requires a considerable fraction of the CCP's equity at 25 percents, according to Article 35 of EMIR, to provide as skin-in-the-game in the default waterfall system. It is related to incentives too that the CCP management and not just the shareholders should bear the consequences if it is inevitable to reach out for the CCP's capital buffer (Cont, 2015). Cont (2015), Murphy (2017) and McPartland and Lewis (2017) point out that in case the waterfall is exhausted, both contributions of faulty and non-faulty members' contributions are proven to be inadequate, and before entering the recovery phase of the CCP, there should be another tranche, which is known in the literature as another part of the skin-in-the-game, the senior tranche. The senior tranche is not mandatory, but several CCPs opt for it in order to avoid using more drastic tools of recovery.

However, Murphy (2017) also raises a prominent, but unanswered question in the debate: "*how much skin is enough to create good incentives for the CCP?*" He suggests that regulators are the ones who should answer. Cox (2015) suggests that supervising authorities "*should have the responsibility to ensure that a sufficiently objective and balanced decision is reached,*" and he does not give a precise answer to the question. The concern is critical because while in the European Union, the contribution is a pre-set percentage, according to CME, in 2018, in the United States, the exchange contributed about \$375m, or roughly 5.25% of its capital (Suprise, 2015).

Reasoning against high skin-in-the-game

Compared to the regulator's opinion, CCP experts have precisely the opposite judgment. CME highlights that "*Skin in the game doesn't protect end client.*" (Daly, 2015) In their view, concentration risk is the biggest fear

CCPs can have because it can encourage moral hazard of the clearing members as CCPs contribute more substantial financial resources to the default waterfall. To deal with concentration risk, they propose to handle the most significant exposures in a way, that the clearing member causing it, would pay for that risk by additional collaterals so that when they fail, those funds are available to resolve the default (Suprise, 2015). Otherwise, the end clients would also suffer from the exhaustion of the CCP's financial resources. On the long run, this would benefit neither the CCP nor the end clients, especially if the default events accumulate.

Incentives regarding for-profit CCPs

CCP capital contributions matter since the owners and users are distinct and so have responsibilities toward one another; moreover, they may also have very different interests as well. The for-profit CCP's primary purpose is to maximize its own expected utility. The most prominent aspect of a for-profit CCP is that its default waterfall can be funded either solely by clearing members or by the CCP. Both have their benefits and drawbacks, all affecting the incentives of every participant in the system.

A wholly clearing member-funded waterfall's risk is that it prefers a higher return on equity instead of safety, potentially leading to limited exposure to default risk, resulting in sloppy risk management practices. The improper management concusses and minimizes the credibility of the CCP, leading to disbelief in its role in fulfilling proper risk management responsibilities.

The amount of the skin-in-the-game can be harmful to be too high or too low. Researchers also point out that this layer does not only have a loss-absorption function, but it also indicates the risk profile and the incentives of the CCP that may alter the incentives and risk perception of clearing members. Murphy (2017) highlights, if a CCP's capital contribution to the waterfall is too small, clearing members would perceive high risk in the clearing activity, so instead of joining the system, they would seek to engage trades not subject to clearing, namely, OTC trade. Regarding the incentives of risk management in case of a default, Carter and Garner (2015) also argue the CCP's skin-in-the-game value determinant.

On the other hand, a for-profit must focus on the return on equity if the contribution of the default waterfall is substantial; resulting in increasing clearing activity fees, creating a disincentive to clear trades, increasing systemic risk. Furthermore, as Cox (2015) explains, the too-high contribution to skin-in-the-game included in the waterfall could endanger the CCP's long-term existence in case of an extreme default event. Clearing members would be encouraged to avoid helping the management of a defaulted member since the junior tranche absorbs the losses. The higher the junior tranche is, the probability of funds requested from the non-defaulting members' decreases, also resulting in a decrease in incentives to be part of the default management too.

Incentives regarding clearing member-owned (user-owned) CCPs

Huang (2019), Cox and Steigerwald (2017), and McPartland and Lewis (2017) study the different ownership structures of CCPs. Another type of ownership structure worth analyzing are the mutualized CCPs, which means that they are owned by the clearing members and exchanges that use their services. In line with their use of concepts, as being owned mainly by a small number of large clearing members, they are unanimously calling them user-owned CCPs. From this structure, there could be misaligned incentives between members, as smaller members would have less impact on the decision of the policy the CCP would apply.

Compared to the for-profit model, in this case, it maximizes the total welfare surplus. Another difference the two researchers point out is the fact that these CCPs hold additional capital, and the required collateral amount is low. Nevertheless, because the owners are the same as the users, the financial resources in the default waterfall system have the same source: the capital of the system participants.

Recovery and resolution

Besides the sound risk management practices, another significant feature the regulation has addressed are the recovery and resolution regimes for CCPs to assure continuity in the provision of clearing services for systemic stability and of an orderly resolution of the CCP (CPSS-IOSCO (2012), CPMI-IOSCO (2014), FSB (2014a)).

Peters and Wollny (2018) point out the importance of preparation. The recovery plan is a crucial tool for both CCPs and regulators to be prepared to identify the critical services. The stress scenarios in case of default and non-default events may stop the CCP from being able to provide its critical services. Both quantitative and qualitative criteria could trigger the application of all or part of the recovery plan; and the recovery tools in case of different events. The plan should outline the possibilities the CCP has not just in case of a participant's default, but for events related to operational deficiencies.

If the recovery tools are proven to be insufficient, the next step is the resolution plan activation. It can be put into force even if the CCP is materially breaching its core obligations.

Resolution authorities may, however, intervene before the defined trigger point. If so, the early intervention should be adequately justified and considered as a tool of last resort, and it should be used to protect the broader financial markets. A way too early intervention can be unnecessary and very likely would have disadvantages of shifting the responsibility of the losses to the public sector. Moral hazard can be created because of a premature resolution, that would weaken the reputation of the CCP, but its ability to conduct an orderly and useful loss allocation tool as well.

MODEL

We run simulations in order to show how the size of one level of the default waterfall – the so-called skin-in-the-game (SITG) – should be defined. Our model consists of one CCP with two different financial assets: stock and currency. The currency can be traded on the options and futures markets, while the stock can be traded on the spot, options, and futures markets. After 7500 days of simulation, we shock day no. 7501 to see which clearing member will have the largest exposure, namely, would have the largest loss if it would default. We assumed in our simulation that only that clearing member will default, which has this largest exposure. Our question was, that in this case, what the minimum size of the SITG should be in two different situation: 1) the CCP should not exhaust the default fund contribution of the non-defaulting members; 2) the default waterfall is just enough not to start the recovery and resolution process. For simulating the price evolution of the stock and currency, we use the simulation method of Illés et al. (2019). The basics of the simulation model are the following (more details can be found in Illés et al. (2019)):

1. 7500 days are simulated.
2. We run the simulation 100 times.
3. Price simulation:
 - The return of the stock and currency is simulated by arithmetic Brownian motion.
 - The correlation between the price of the two underlying is simulated with Cholesky decomposition.
 - The occurrence of stresses is modeled with a Poisson process, while the extent of the shock is modeled with lognormal distribution.
 - At the occurrence of stress, the correlation between the assets increases to 0.9 and decreased by 0.95 every day while the minimum value of their correlation is set to 0.5.
 - The applied shock parameters are summarized in Table 1.

Table 1: Parameters of the price simulation

Parameters for the price simulation with ABM		
	Stock (St)	Currency (Ccy)
α	10%	5%
σ	15%	10%
$S(0)$	1000 EUR	1000 EUR
dt	1 day	1 day
Shock parameters that affect the value of the shock		
μ	-20	-20.6
standard deviation	0.7	0.8
decrease of shock	0.97	0.99
Shock parameters that affect the date of the shock		
λ	0.005	0.0045

4. Four clearing members are present on the market, whose positions are summarized in Table 2.

Table 2: Number of positions per clearing members

Clearing members	CM1		CM2		CM3		CM4	
	St	Ccy	St	Ccy	St	Ccy	St	Ccy
LongPut	3	5	2					
ShortPut					5	5		
LongCall	3	5						
ShortCall					3	5		
LongForward				5				
ShortForward		5						
LongUnderlying	4		2		3			
ShortUnderlying							9	

5. Initial margin calculation:
 - The margin of the underlying assets is calculated by the method of Béli and Váradí (2016).
 - The portfolio level margin is defined by the SPAN (Standard Portfolio Analysis of Risk) margin calculation method.
6. Default fund calculation: We identify four different historical stress scenarios and calculate whether in case of each scenario's price change, the margin is sufficient to cover the potential losses of each clearing member in case of default, or not, on the last day of the simulation. The value of the DF is the scenario that has the highest loss of the max(1;2+3) exposures according to EMIR. The parameters of the four applied historical scenarios:
 - 1 & 2: Min/max stock: lowest/highest stock return during the 7500 days, and taking the currency return the same day.
 - 3 & 4: Min/max currency: lowest/highest currency return during the 7500 days, and taking the stock return the same day.
7. The CM level DF contribution is defined by the ratio of the CM level initial margin levels.

The simulation model of this paper furthermore assumes that the spot and derivative markets are merged, so it calculates the margins, the default fund (DF), and the skin-in-the-game accordingly.

To examine the minimum necessary SITG level, we need the next day's prices as well. So we continue to day 7501 stressing the stock and currency returns with a hypothetical scenario. For example, we increase the return of the stock by 5% (+5%) and decrease the return of the currency by 5% (-5%) for each simulation. Having the prices for day 7501, and having the IM and DF contributions based on the previous day, we can quantify the uncovered losses for each CMs. The DF contribution, the initial margin, and the loss per CMs give the use of the CCP's own resources – namely the SITG – and given all this information, we can quantify the rate of the SITG within the default waterfall system (DW).

RESULTS

In the following we show the result of the 100 simulation by assuming the mentioned (5%, -5%) hypothetical parameters, which are much higher than the historical parameters. In our 100 realizations the average historical stress parameters were the following according to Table 3:

Table 3: Average historical stress parameters

Min stock		Max stock	
Stock	Currency	Stock	Currency
-3.54%	-0.01%	3.63%	0.04%
Min currency		Max currency	
Stock	Currency	Stock	Currency
-0.05%	-2.37%	0.18%	2.41%

The skin-in-the-game is the third level of the default waterfall system to cover the members' losses, the defaulting member's initial margin and default fund contribution are the previous layers. As a first step, we examine the size of the SITG within the total waterfall by answering the following two questions:

1. What should be the size of the SITG to not have to use the non-defaulting member's DF contribution and to avoid reaching the fourth layer of the DW?
2. What should be the size of the SITG just to have a DW that covers precisely every loss of the defaulting member? That is, if we use all the four layers.

In every single realization, CM1 resulted as the defaulting member. Table 4 summarizes the size of the SITG within the DW.

Table 4: Summary of the size of SITG within the DW

	Size of the SITG within the total DW while excluding the application of the non-defaulting members' contribution	Size of the SITG within the total DW while including the application of the non-defaulting members' contribution
Min	12.57%	0.00%
Max	34.08%	19.99%
Average	22.18%	7.46%

If we use only the defaulting member's contribution to cover its losses, SITG should be at least the 12.57% of the total DW to recover, while in 27 times out of our 100 realisations, zero SITG is needed to cover all the losses if we include the non-defaulting members' contributions as well. In this case, the average SITG size is the third of the time when it does not involve the non-defaulting CMs.

Figures 1 and 2 present the differences between the built up of the DW system in the two cases. In Figure 1 and 2 we can see the value of the DW without the value of the SITG, and also the value of the SITG. Both of them are

in the percentage of the total DW. The difference is that in Figure 2 the SITG is smaller, since the non-defaulting members' DF contribution can be applied as well. We call the SITG in this case as adjusted SITG.

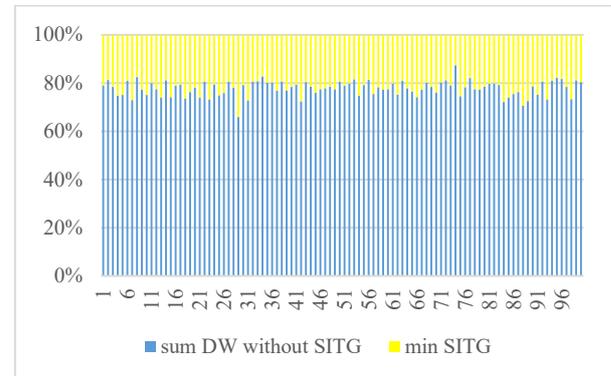


Figure 1: Default waterfall system setup covering losses without using non-defaulting members' contribution

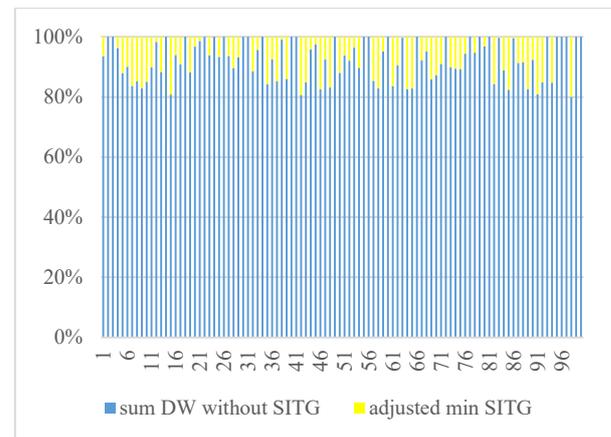


Figure 2: Composition of the default waterfall system including the non-defaulting members' contribution to cover losses

However, we ran the 7501st day's price change with the hypothetical parameters as +5% for the stock and -5% for the currency, we also study how the size of the SITG would change if we change these parameters from +/-15%.

Table 5: Sensitivity of the SITG and the adjusted SITG

Stock parameter	SITG		adjusted SITG	
	-15%	15%	-15%	15%
-15%	52.97%	52.97%	49.15%	49.15%
15%	60.22%	53.56%	56.63%	48.70%

Table 5 shows the two extreme values of our results, namely when the stress parameters are +/-15%. Looking at the results, it seems, SITG is a bit more sensitive to the change of the stress parameters if we try

to cover the losses including only the defaulting member's contribution; however, the difference is not significant. On the other hand, we can see that, based on our simulation, both SITG and adjusted SITG react symmetrically to the change in the parameter of the currency while the shock parameter of stock remains -15%. In contrast, if the stock stress parameter increases, the reaction on the negative side of the currency parameter is higher than on the positive side for both cases.

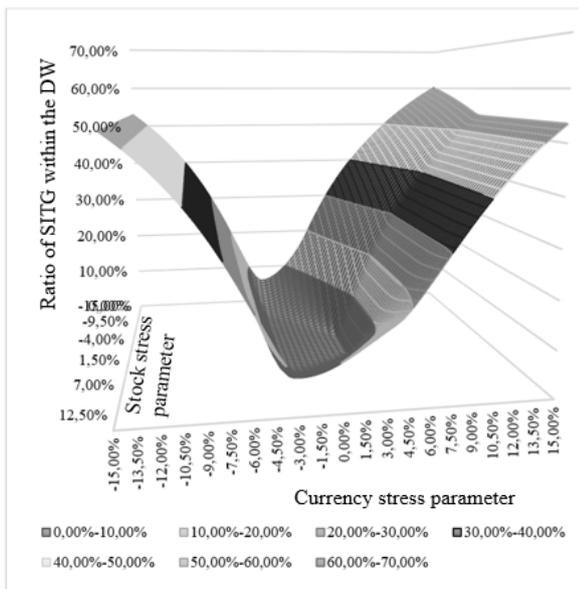


Figure 3: Sensitivity of the SITG

In Figures 3 and 4 we can see the SITG surface, which shows the minimum value of the SITG in the function of the stress parameter of the stock and the currency as well.

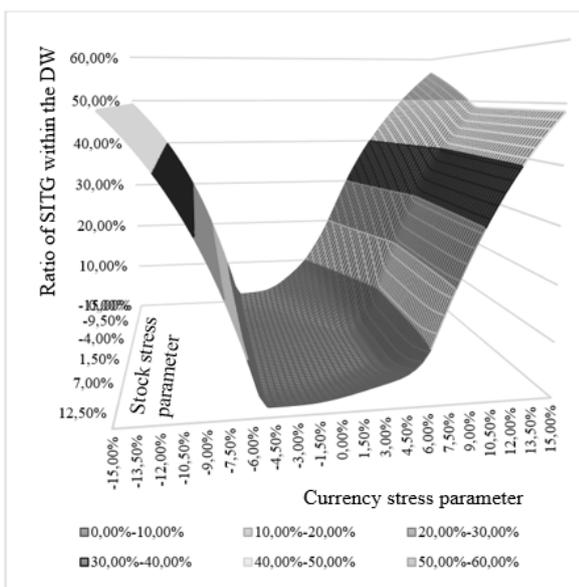


Figure 4: Sensitivity of the adjusted SITG

We can see, that till the shock parameters stay around the average shock parameters of the historical scenarios (based on the 100 simulations), the SITG stays very low, close to zero. As we start to increase the parameters, notably, the SITG will start to increase as well. The two most extreme values are presented in Table 5; we did not use parameters greater than +/-15%. Based on the historical stress parameters, we can suppose that these price changes are impossible to happen on a daily basis. Figure 3 and 4 highlight that the size of the SITG within the total DW system is less sensitive for the change of the shock parameters if we involve the non-defaulting members' contributions as well, since those can absorb the losses to a certain extent.

CONCLUSION

Assuming merged spot and derivative markets, we analyzed the size of the skin-in-the-game within the total default waterfall system from the point of view of how it can cover the defaulting member's losses. After running our model for 100 times, we quantified the SITG and showed the difference between the two cases: if we try to cover the losses excluding and including the non-defaulting members' default fund contributions as well. However, involving the non-defaulting members' money in the loss-absorption process may be unfair for them, but on the other hand, the CCP could maintain a much smaller SITG level. Our model demonstrates that the optimal level of SITG can highly depend on the risk-taking willingness of its clearing members, so the position it takes on the spot and derivatives markets. A CCP must operate a default waterfall system suitable for its risk profile and calibrate it, so it does not alter the CMs' incentives, mitigating the exacerbation of stress on the market(s) it clears.

REFERENCES

- Berlinger, E., Dömötör, B., Illés, F., Váradi, K. 2016. Stress indicator for clearing houses. *Central European Business Review*.
- Biais, B., Heider, F., Hoerova, M. 2012. Clearing, Counterparty Risk, and Aggregate Risk. *IMF Economic Review* 60, 193–222.
- Carter, L., Garner, M. 2016. Skin in the game: central counterparty risk controls and incentives. *JFMI* 4, 39–54. <https://doi.org/10.21314/JFMI.2016.056>
- Cont, R. 2015. The end of the waterfall: Default resources of central counterparties. *Journal of Risk Management in Financial Institutions* 8, 365–389.
- Cox, R.T., Steigerwald, R.S. 2017. A CCP is a CCP is a CCP. CPMI-IOSCO (Committee on Payments and Market Infrastructures-International Organization of Securities Commissions) 2014. *Recovery of Financial Market Infrastructures*.
- CPSS-IOSCO 2012. *Principles for Financial Markets Infrastructures*.
- Daly, R. 2015. Do CCPs Need More Skin in the Game? *Tradersmagazine.com* 1.
- Duffie, D. 2015. *Resolution of Failing Central Counterparties*, in Kenneth E. Scott, Thomas H. Jackson & John B. Taylor, *Making Failure Feasible, How Bankruptcy Reform Can End "Too Big to Fail"*. Stanford, CA: Hoover Institution Press) Working paper.

- EMIR – European Market Infrastructure Regulation: Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4th July 2012 on the OTC derivatives, central counterparties and trade repositories (EMIR - European Market Infrastructure Regulation) Available: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R0648&from=EN> downloaded: 8th February 2019.
- ESMA 2012. European Securities and Markets Authority, Consultation Paper on Anti-Procyclical Margin Measures. Available: <https://www.esma.europa.eu/pressnews/esma-news/esma-consults-ccp-anti-procyclical-margin-measures> downloaded: 8th January 2019
- ESMA 2019. EU-wide CCP stress test. ESMA, February 2019.
- FSB 2014. Key attributes of effective resolution regimes for financial institutions, October.
- R&R CCP. 2019. General Secretariat of the Council. 2019. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Framework for the Recovery and Resolution of Central Counterparties and Amending Regulations (EU) No 1095/2010, (EU) No 648/2012, and (EU) 2015/2365 and Directives 2002/47/EC, 2004/25/EC, 2005/56/EC, 2007/36/EC, 2011/35/EU and (EU) 2017/1132. <https://www.consilium.europa.eu/media/41614/st14540-ad01-en19.pdf>.
- Huang, W. 2019. Central Counterparty Capitalization and Misaligned Incentives, February. <https://www.bis.org/publ/work767.htm>.
- Hughes, D., Manning, M. 2015. CCPs and Banks: Different Risks, Different Regulations. Bulletin.
- Illés, F., Muratov-Szabó, K., Prebuk, A., Szodorai, M., Váradi, K. 2019. Together forever or separated for life: Stress test of central counterparties in case of merged and separated markets. 33rd ECMS conference paper DOI: <http://doi.org/10.7148/2019>
- Lopez, C., Saeidinezhad, E. 2017. Central Counterparties Help, But Do Not Assure Financial Stability. Munich Personal RePEc Archive.
- Markose, S., Giansante, S., Shaghghi, A. 2012. Too interconnected to fail”, Financial network of US CDS Market: topological fragility and systemic risk. *Journal of Economic Behavior and Organization* 83, 627–646.
- McPartland, J., Lewis, R. 2017. The Goldilocks problem: How to get incentives and default waterfalls “just right.” *Economic Perspectives*, Federal Reserve Bank of Chicago.
- Murphy, D. 2017. I’ve got you under my skin: large central counterparty financial resources and the incentives they create. *Journal of Financial Market Infrastructures* 5, 54–74. <https://doi.org/DOI: 10.21314/JFMI.2017.073>
- Murphy, D., Nahai-Williamson, P. 2014. Dear Prudence, won’t you come out to play? Approaches to the analysis of CCP default fund adequacy. *Bank of England Financial Stability Papers*.
- Pirrong, C. 2011. A Bill of Goods: CCPs and Systemic Risk. *Journal of Financial Market Infrastructures* 2, 55–85.
- RTS – Technical Standard: Commission delegated regulation (EU) 153/2013 of 19th December 2012 supplementing Regulation (EU) No 648/2012 of the European Parliament and of the Council with regard to regulatory technical standards on requirements for central counterparties. Available: <http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:052:0041:0074:EN:PDF> downloaded: 8th January 2019.
- Suprise, G. 2015. Skin in the game doesn’t protect end client, CME says. *GlobalCapital* 106–106.
- Szanyi, C., Szodorai, M., Váradi, K. 2018. A Supplement to the Regulation of Anti-Cyclical Margin Measures of Clearing Activities. Working paper SSRN 3242078.
- Váradi, K., Béli, M. 2017. Alapletét meghatározásának lehetséges módszertana. *Financial and Economic Review* 16.

AUTHOR BIOGRAPHIES

KIRA MURATOV-SZABÓ is a university student at the Corvinus University of Budapest, doing her Masters program in Finance. Besides she is a junior risk analyst at KELER CCP Ltd. Her main research field was during her Bachelor studies is market microstructure, with a special interest in order and quote driven markets.

ANDREA PREPUK is a university student at Corvinus University of Budapest doing her Masters program in Finance. Her main interest field is the application of game theory in real financial decisions.

MELINDA SZODORAI is a risk analyst at KELER Ltd. Her primary responsibilities are operational risk management and regulatory reporting. She majored in 2013 in finance and management at Babes-Bolyai University. Currently, she is also a Ph.D. student at the Corvinus University of Budapest. Her main research areas are stress tests and market infrastructures.

KATA VÁRADI is an Associate Professor at the Corvinus University of Budapest (CUB), at the Department of Finance. She also graduated at the CUB in 2009, and after it obtained a PhD in 2012. Her main research areas are market liquidity, central counterparties, capital structure, and risk management.

ACKNOWLEDGEMENT

EFOP-3.6.3.-VEKOP-16-2017-00007 számú “Tehetségből fiatal kutató” – A kutatói életpályát támogató tevékenységek. (“Talented young researcher” – Supporting activities of researcher career)

We thank for KELER CCP Ltd for excellent research assistance.

The views expressed here are those of the authors, and do not necessarily reflect those of the KELER CCP and its affiliates and employees.

COMPENSATION SCHEME WITH SHAPLEY VALUE FOR MULTI-COUNTRY KIDNEY EXCHANGE PROGRAMMES

Péter Biró,
Márton Gyetvai
Institute of Economics
CERS,
Inst. of Mathematics and Statistical Modelling
Corvinus University of Budapest
H-1097, Budapest, Hungary
E-mail: biro.peter@krtk.mta.hu,
gyetvai.marton@krtk.mta.hu

Xenia Klimentova
INESC TEC
4200-465, Porto, Portugal
E-mail: xenia.klimentova@inesctec.pt

João Pedro Pedroso
INESC TEC,
Faculdade de Ciências
Universidade do Porto
4169-007 Porto, Portugal
E-mail: jpp@fc.up.pt

William Pettersson
School of Computing Science
University of Glasgow
G12-8QQ, Glasgow, United Kingdom
E-mail: william.pettersson@glasgow.ac.uk

Ana Viana
INESC TEC,
School of Engineering
Polytechnic of Porto
4200-072 Porto, Portugal
E-mail: ana.viana@inesctec.pt

KEYWORDS

Kidney exchange programmes, integer programming, simulations, compensation scheme, Shapley value.

ABSTRACT

Following up the proposal of (Klimentova, Viana, Pedroso and Santos 2019), we consider the usage of a compensation scheme for multi-country kidney exchange programmes to balance out the benefits of cooperation. The novelty of our study is to base the target solution on the Shapley value of the corresponding TU-game, rather than on marginal contributions. We compare the long term performances of the above two fairness concepts by conducting simulations on realistically generated kidney exchange pools.

INTRODUCTION

Currently, no cure exists for Chronic Kidney Disease (CKD). Dialysis is a common treatment for CKD, but is costly, and has poor life expectancy and quality of life for patients. Transplantation is the preferred treatment of choice, but there is a shortage of deceased donors, so the waiting lists for a deceased donor kidney are long. The other alternative for transplantation is living donation. However even if there is a willing donor for the patient, the donor and recipient may be immunologically incompatible.

A Kidney Exchange Programme (KEP) is a platform that allows incompatible *recipient-donor pairs* (RDPs) to exchange their incompatible donor organ for a compatible donor organ. In Europe, as surveyed in 2016 (Biró, Haase-Kromwijk, van de Klundert et al. 2019), ten countries have operated a KEP. The first was the programme of the Netherlands in 2004. In Europe, the UK programme is the largest, with an average of 135 transplants per year.

In the European KEPs matching runs are conducted at regular time intervals, typically every 3 months. Virtual crossmatch tests are used to estimate compatibility between all donors and recipients, and the results of said tests are used to determine optimal exchanges. Compatibility is represented by a directed graph, where the vertices denote the incompatible recipient-donor pairs and the directed arcs denote the potential compatible transplants between the pairs. The objective of the model is usually to maximise the number of transplants. In advanced KEPs, IP models are used to solve this problem (Abraham, Blum and Sandholm 2007).

For ethical reasons, the operations in an exchange cycle should happen simultaneously. Hence in practice logistical constraints create an upper bound on viable cycles. In Europe, the most common such bound in KEPs is 3 (e.g. the UK, Spain and Belgium), but there are examples for bound of 2 (e.g. France) and 4 (Netherlands) as well (Biró et al. 2019). Besides the number of the transplants, further considerations are given for improving the quality of the transplants or the long term success of the scheme by prioritising hard-to-match pairs, e.g. sensitive patients or patients with blood type O. A complete list of objectives in European KEPs and their possible implementations with IP techniques can be found in (Biró, van de Klundert, Manlove et al. 2020).

In some programmes, the usage of non-directed donors (NDDs) is also allowed. In most countries an NDD is an altruistic donor, who has no paired recipient, but they would offer their kidney to any recipient in the pool. Another possible NDD can be a deceased donor. In these cases, the operations can be performed consecutively. Hence these donors may initiate longer chains involving numerous pairs (Ashlagi and Roth 2014).

Larger KEP pools may result in a higher number of transplants, and merging pools can bring extra ben-

efits due to the increased number of options. In a multi-country Kidney Exchange Programmes (mKEP) the national KEPs merge their pools. In the last few years, there several such collaborations have been established in Europe. Since 2016 the Czech Republic and Austria started a collaboration by joining the pools of two transplant centres (Böhmg, Fronek, Slavcev, Fischer, Berlakovich and Viklicky 2017). In 2018 a collaboration started between Italy, Portugal and Spain (Valentín, Garcia., Costa, Bolotinha, Guirado, Vistoli, Breda, Fiaschetti and Dominguez-Gil 2019), and Scandiatransplant has started to coordinate a joint scheme in 2019 for Sweden, Norway and Denmark (STEP Documentation 2016).

There are two types of collaboration possible for joining the pool in an mKEP. In the so-called *Consecutive runs* (CR) the countries first find the optimal solutions separately on their national pools; then they consider an international pool on the remaining pairs. An example for this is the Spanish-Italian-Portuguese collaboration (Valentín et al. 2019). The other possibility is the *Merged pool* (MP) collaboration, where the countries merge their national pools from the beginning of the optimisation process. This concept is present in the Austro-Czech and the Scandinavian programmes.

From the point of view of optimisation, an mKEP can be modelled with a similar compatibility graph, where the objective is to find the maximal number of transplants. However, in an mKEP, there are multiple countries involved, so the pairs of a cycle may belong to different countries, which may increase the logistical difficulty of the programme. Furthermore, the countries may have different constraints and objectives.

When multiple transplant centres collaborate in a regional or national KEP, it is crucial that all participants receive fair benefits. In (Ashlagi and Roth 2011), (Ashlagi and Roth 2012) the authors suggest that individual rationality (IR) of the centres should be guaranteed as a constraint in the optimisation in order to incentivise centres to reveal all of their RDPs in the programme. The constraint of iIR would guarantee that each centre gets as many transplants as they would make without the collaboration.

European mKEPs differ to US KEPs in a number of ways. In Europe the national KEPs conduct their matching runs at regular intervals, whilst the US KEPs operate on a daily basis due to the competition among the alternative programmes (Agarwal, Ashlagi, Azevedo, Featherstone and Karaduman 2018). In Europe the countries have different health care systems and various performance of their deceased programmes, so their KEPs are also different in pool sizes, their regulations and also in their distributions of RDPs, whilst in the US the pools of the transplant centres are more similar to each other. The largest national programme in the US, the National Kidney Registry, uses a scoring system to incentivise the transplant centres to fully register all of their pairs, and the registration of easy-to-match pairs and altruistic donors are awarded bonus scores to ensure fairness among centres.

When the objective of the mKEP is to obtain the maximal number of transplants, the benefits from the collaboration can be unbalanced across countries. It means that some countries would proportionally increase their number of transplants less than other countries, if one considered the benefit they brought to the pool. The paper by (Klimentova et al. 2019) investigated the fairness of an mKEP by proposing the usage of a compensation scheme. The authors presented an extended MILP model and algorithm to make an mKEP fairer when running for a long period of time. The process is based on a credit system that tracks and balances out the benefits for the contributors. The authors use two fair-values for the determination of the contribution of each participant.

The Shapley value is a well-known solution concept from the field of cooperative game theory (Shapley 1953), that satisfies four important fairness axioms. An mKEP can be considered as a game with transferable utility (TU-game), where the players are the participating countries, and the value of a coalition is the number of transplants these countries can achieve together in the collaboration. Several recent articles investigated an mKEP as a cooperative game, studying the Nash-equilibria (Carvalho, Lodi, Pedroso and Viana 2017), (Carvalho and Lodi 2019) and the core (Biró, Kern, Pálvölgyi and Paulusma 2019) of the corresponding games. In this paper, we compare one of the fair-values, the Benefit value, used in (Klimentova et al. 2019) to the Shapley value. Our results show that the Shapley value provides similar effects as the Benefit value, essentially giving better outcomes for the larger countries and worse for the smaller countries, that can be seen as fairer from the point of view of their value-adding contributions. However, we found the compensation scheme with the Shapley value more stable than with the Benefit value respecting the dynamic biases from the fair values.

FAIRNESS MODELS FOR MKEP

First we describe the standard KEP model and the basic IP formulations, and then we introduce the concept of dynamic compensation schemes.

Notation

Let $D(V, A)$ denote a directed *compatibility graph* D , where V corresponds to the recipient-donor pairs and A corresponds to the compatible transplants. The arc (i, j) represent the compatibility between the donor of pair i and the recipient of pair j , hence the arc only exists if donor i is compatible with recipient j . Let \mathcal{C} denote the set of cycles up to length K . Let each cycle c be a set of arcs and let $V(c)$ denote the set of vertices covered by cycle c .

We can include NDDs in the model as follows. For each NDD we add a vertex, which has outgoing arcs as any other vertex in the graph representing possible donations from the NDD. We also add incoming arcs to the NDD vertex from all the vertices representing RDPs. The latter dummy arcs will correspond to do-

nations at the end of the NDD-chain that are either not performed or the donation is from the last living donor to the deceased waiting list.

In the international collaboration, let N denote the set of participating countries, where the set of RDPs is partitioned according to countries as follows: $V = V^1 \cup V^2 \cup \dots \cup V^{|N|}$. The set of arcs A^n denotes the arcs pointing to the pool of the country n , denoted by V^n . Note that $A = A^1 \cup A^2 \cup \dots \cup A^{|N|}$.

Model with the individual rationality constraint

There are two types of basic IP models known in the literature for maximising the number of transplants in a KEP. One is the so-called edge-formulation, and the other one is the cycle-formulation (Abraham, Blum and Sandholm 2007). The edge-formulation considers the arcs of the graph as the variables of the model. In the cycle formulation, all of the possible cycles are enumerated, and each cycle is represented by a binary variable.

Both formulations are usable in the multi-country situation, however the cycle-formulation is usually faster to solve (Abraham, Blum and Sandholm 2007). Hence we used this formulation in our computational analysis.

As in (Klimentova et al. 2019), we also study a cycle-formulation model with the IR constraint. Let Z^n be the number of transplant that country n would accomplish without any collaboration. The model $\mathcal{M}(\mathcal{C})$ considers the IR constraints in the maximisation of the number of transplants:

$$\mathcal{M}(\mathcal{C}) : \quad \max \sum_{c \in \mathcal{C}} |c|x_c \quad (\text{IR:obj})$$

subject to

$$\sum_{c: i \in V(c)} x_c \leq 1 \quad , \forall i \in V, \quad (\text{IR:1})$$

$$\sum_{c \in \mathcal{C}} |c|^n x_c \geq Z^n \quad , \forall n \in N, \quad (\text{IR:2})$$

$$x_c \in \{0, 1\} \quad \forall c \in \mathcal{C}.$$

Where $|c|^n$ denotes the number of recipients of country n in cycle c . Constraints (IR:1) enforce each RDP to participate in at most one cycle. Constraints (IR:2) is the IR constraint; they ensure that each country receives as many transplants in the mKEP as it could have achieved alone.

Static fair model

Let \mathcal{M}^* denote the optimum of model $\mathcal{M}(\mathcal{C})$. When multiple countries start a collaboration, the model may result in an unfair solution. Some of the participants may increase the number of transplants significantly, but the increase for others may not be that large. The next model, introduced by (Klimentova et al. 2019) minimises the difference between the number of transplants and the fair-values of the countries. The fair-value is computed with respect to the contribution of

the participants. Let ψ^n denote the fair-value of country n . And let δ^n be the absolute deviation variable, which shows the difference between the number of possible transplants and the fair-value of country n . Under this notation, the model is written as follows:

$$\mathcal{F}(\mathcal{C}, \psi) : \quad \min \sum_{n \in N} \delta^n \quad (1)$$

subject to

$$(\text{IR:1}), (\text{IR:2}),$$

$$\sum_{c \in \mathcal{C}} |c|x_c \geq \mathcal{M}^*, \quad (2)$$

$$\delta^n \geq \sum_{c \in \mathcal{C}} |c|^n x_c - \psi^n \quad \forall n \in N, \quad (3)$$

$$\delta^n \geq \psi^n - \sum_{c \in \mathcal{C}} |c|^n x_c \quad \forall n \in N, \quad (4)$$

$$x_c \in \{0, 1\}, \delta^n \geq 0 \quad \forall c \in \mathcal{C}, \forall n \in N.$$

The objective function (1) is to minimise the absolute deviation between the real and the fair-values of each country. Constraints (3)-(4) define the absolute deviations, which is $\delta^n = |\psi^n - \sum_{c \in \mathcal{C}} |c|^n x_c|$. In this model, the objective is to find a solution which is as close to the fair-values as possible, without the reduction of the total number of transplants. Hence the constraint (2) guarantees that introducing fairness will not reduce the number of transplants.

Dynamic fairness model

A fair solution may not exist in a single period of an mKEP. Therefore (Klimentova et al. 2019) proposed an algorithm to guarantee the fairness of the collaboration in the long run. Let t denote a period when a matching is performed in the mKEP. The programme is fair if the differences between the fair-values and the actual transplants are small in general. For the consideration of the past runs, we introduce credits for each country. When one country got less transplants in previous periods than it should, as per fair-value indications, it should get more transplant in the current period, indicated by a positive *credit*.

Let τ_t^n denote the *target value* of country n in the period t . The target value of a country depends on the previous periods' credits (c_t^n), and the fair value of the current period (ψ_t^n). We call the credit as the difference between the previous periods' target value and the number of transplants obtained:

$$c_t^n = \tau_{t-1}^n - s_{t-1}^n \quad (5)$$

Where the s_{t-1}^n denotes the number of transplants that country n obtained in the previous period. Therefore the target value of the country n in period t can be written as:

$$\tau_t^n = \psi_t^n + c_t^n \quad (6)$$

At the beginning of an mKEP $\tau_0^n = s_0^n = 0$ for all country. The algorithm differs a bit in the two types of collaborations. Let $\{\text{MP}\}$ denote of the merged pools case and $\{\text{CR}\}$ the case of the consecutive runs.

Algorithm 1 Dynamic fairness algorithm

- Require:** \mathcal{C} and \mathcal{C}^\setminus for $n \in N$, and the credit, c_t^n .
- 1: $\{\text{MP}; \text{CR}\}$ Find the maximal number of transplants that the countries would conclude without the collaboration: Z^n .
 - 2: $\{\text{CR}\}$ Remove the pairs of the optimal solution of the national matching from the pools. Then create the graph for the remaining pairs.
 - 3: $\{\text{MP}\}$ Solve the $\mathcal{M}(\mathcal{C})$ for \mathcal{M}^* . In the case of $\{\text{CR}\}$ it is enough to solve the model $\mathcal{M}(\mathcal{C})$ without constraint (IR:2).
 - 4: $\{\text{MP}; \text{CR}\}$ Calculate the fair-value ψ^n .
 - 5: $\{\text{MP}; \text{CR}\}$ Determine the target value: $\tau_t^n = \psi_t^n + c_t^n$.
 - 6: $\{\text{MP}\}$ Solve $\mathcal{F}(\mathcal{C}, \tau_\perp)$, in the $\{\text{CR}\}$ case the (IR:2) constraint is redundant.
 - 7: $\{\text{MP}; \text{CR}\}$ Save result, calculate the transplant (s_t), determine the remaining pairs and compute the credit c_t^n .
-

THE FAIR VALUES

In the previous section, the algorithm is based on a fair allocation, with fair share ψ^n for each country n . In (Klimentova et al. 2019) the authors used two different fair values, the *Benefit* and the *Potential*.

The Potential value is the number of additional transplants in an optimal mKEP that includes this country compared to the number of transplants in the mKEP if said country was excluded. The Benefit value is the result of dividing the potential value by the number of recipients within that country's individual KEP. This helps avoid situations wherein countries with larger KEPs dominate an mKEP. According to results of experiment the Benefit value provide more balanced results, so we will only consider Benefit value in this paper.

The Benefit value is similar to the Shapley value, that is well-known in the field of cooperative game theory (Shapley 1953). The Shapley value considers the payoffs of the participants by the concept of marginal contributions. The contribution of this paper is to make a comparison between the Shapley value to the Benefit value.

mKEP as a game and Shapley value

Several articles consider an mKEP as a game. The players of this game are the participating countries. The value that the players can make in any coalition S is denoted by $v(S)$, and it is equal to the optimum of the model \mathcal{M} , when the countries of S collaborate. The coalition involving all the players is called as the grand coalition. Hence it means that the value of the

coalition is the maximal number of transplants with the consideration of individual rationality.

The Shapley value, introduced in (Shapley 1953), is based on the concept of the marginal contribution of the player. The improvement that the player makes in a particular ordering for the players joining the game is the marginal contribution. The Shapley value is the average of these improvements over all possible coalitions.

Let Θ denote all of the different orders in which the players can enter the game. For a $\theta \in \Theta$ order q_θ^n indicates the contribution that country n adds to the coalition when she joins. Then the Shapley value can be calculated as:

$$\psi^n = \frac{1}{|N|!} \sum_{\theta \in \Theta} q_\theta^n \quad (7)$$

In an mKEP, q_θ^n is the increase of the transplants when the country n joins to the θ ordered collaboration. In an order, θ , the first country's marginal contribution is $q_\theta^{n_1} = v(n_1) - v(\emptyset)$. The second one's is $q_\theta^{n_2} = v(n_2 \cup n_1) - v(n_1)$ and the third one's is $q_\theta^{n_3} = v(n_3 \cup n_2 \cup n_1) - v(n_2 \cup n_1)$, and so on. When all of the θ orders are considered, the (7) can be simplify to the next formula:

$$\psi^n = \sum_{S \subseteq (N \setminus n)} \gamma_S [v(S \cup n) - v(S)] \quad (8)$$

The $\gamma_S = \frac{|S|!(N-|S|-1)!}{N!}$ denotes the probability of the occurrence of the S coalition. Then (8) considers every possible coalition without country n (even the zero-coalition, when there is no one to collaborate), and sums the effects of the inclusion of country n with the weight of γ_S .

Benefit value

The Benefit value of the article (Klimentova et al. 2019) distributes the surplus of the coalition over the sum of the individual solutions. Hence this excess (σ) is the value of the grand coalition minus all of the result of the separated KEPs:

$$\sigma = v(N) - \sum_{i \in N} v(i) \quad (9)$$

The ratio that the players can get from the surplus depends on their contribution to the grand coalition value. This contribution is calculated by the total improvement in the number of transplants if the country joins the others:

$$a^n = v(N) - v(N \setminus n) - v(n) \quad (10)$$

Therefore the Benefit payoff for the country n is:

$$\psi_n = v(n) + \sigma \frac{a^n}{\sum_{i \in N} a^i}$$

In the case of $\sum_{i \in N} a^i = 0$, the surplus is distributed equally for the players:

$$\psi_n = v(n) + \frac{\sigma}{|N|} \quad \forall i.$$

SIMULATION

We used the simulator form (Santos, Tubertini, Viana and Pedroso 2017, Klimentova et al. 2019) to generate a pool of mKEP that dynamically changes over time. We investigated seven consecutive years of the mKEP, with the matching run every 90 days (4 times a year). Over the considered years the incompatible RDPs arrive, depending on the country, but stays there minimum 60 to the maximum of 2190 days, on average of 365 days uniformly. For the analysis of the data, we considered the first year as a warm-up period, to populate the pool. Hence in the statistical results, we analysed only the last six years.

We compared the results when the countries do not collaborate with the MP and CR type of collaborations. For each collaboration levels, we considered the maximisation of the number of transplants. For the latter two, we calculated the result of the fair model as well, with the Shapley and the Benefit value. We restricted the number of transplant per cycle to 3 and the number of transplants per chains initiated by NDDs to 2.

Collaboration of same sized countries.

First, we considered a case, when three countries with similar attributes of the pools start the coalition. In this case, because the pools are identical, the contributions also should be similar. Therefore it would make the same effect for each country. Hence, in the long run, the ratio of a participant gets on average should be around $\frac{1}{3}$. We investigated this case because in this setup, with the fair models, the effect of an additional random contribution of a country can be more noticeable.

In this setup, we excluded the possibilities of NDDs. We investigated three different cases: *Small*, *Medium* and *Large* which refers to the sizes of the pools. In the *Small* case the pairs arrive on an average of 20 days, in the *Medium* on every ten days and for the last case, on every five days. In each case, we considered 100 randomly generated pools.

TABLE 1: Average improvement to the No-collaboration

None	Small		Medium		Large	
	CR	MP	CR	MP	CR	MP
	45.37		104.76		231.43	
Max.Numb.	116.66%	124.06%	108.40%	114.75%	103.37%	108.13%
Benefit	116.78%	123.38%	108.46%	114.57%	103.34%	108.20%
Shapley	116.85%	123.33%	108.47%	114.54%	103.37%	108.15%

Table 1 presents the average results of the no-collaboration and the improvement with the different levels of collaborations and models. In every case, the collaboration resulted in an improvement in the number of transplants. Generally, the MP resulted in much higher transplants than the CR. The fair-models had a similar result in the average selected pairs, as expected.

Although the result does not change too much with the fair model, because of constraint (2), there is a tiny difference. This difference is because the alternative solution with the Benefit or Shapley value results in a different pool of remainders. As the results show, the

different remaining pairs did not change the number of the selected pairs significantly. Furthermore, interestingly as the sizes of the pools increase the effect of the collaboration on the number of transplants decreases.

Comparing the bias between the target value and the concluded transplants

The result of the number of transplants was not significantly different for the Shapley and the Benefit value. To examine the differences, we investigated the bias between the target value and the concluded transplants, $bias = \tau_t^n - s_{t+1}^n$. The bias shows the average fairness of the model. If the average of the bias is close to zero, the participants received results that are close to their contribution. When a participant has a large bias, it means that her result from the model was far from her fair solution. When the bias is positive, the country did not get as much transplant as she has contributed since the beginning of the collaboration.

Fig. 1: biases of the models

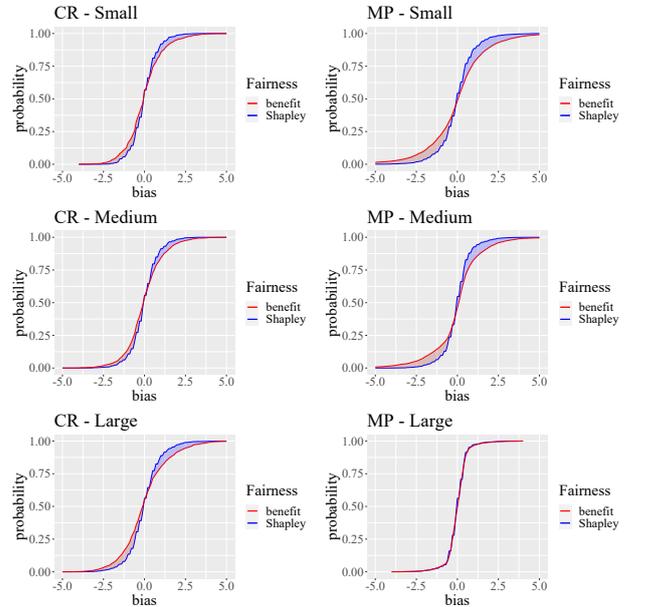


Figure 1 presents the cumulative distribution functions of the biases. The line shows the probability of the bias was less than or equal to the value of the x-axis. The larger probabilities define the colour of the area between the two lines. In general, the Shapley value resulted in a steeper line. Therefore the biases were more concentrated near zero, than the biases of the Benefit value. The descriptive statistics of the biases can be seen in Table 2.

In general, the bias of the model with the Shapley value resulted in a much smaller standard deviation. According to Levene's test, the difference is significant with 1% significance level for all cases, except for the Large MP case, which significantly differs only with 5% significance level. Also, the distance between the minimum and maximum were higher in almost every case with the Benefit value. Furthermore, both models resulted in a more peaked distribution, than the stan-

TABLE 2: Descriptive statistics of the biases

Small	CR			MP		
		Benefit	Shapley		Benefit	Shapley
	min	-4.15	-2.67	min	-9.25	-5.00
	max	4.86	4.00	max	8.55	5.00
	sd	1.11	0.78	sd	1.81	1.05
	skewness	0.44	0.29	skewness	-0.17	0.04
	kurtosis	1.22	1.04	kurtosis	2.39	2.44
Medium		Benefit	Shapley		Benefit	Shapley
	min	-5.11	-4.00	min	-11.79	-4.17
	max	5.69	3.17	max	11.76	4.67
	sd	1.13	0.84	sd	1.54	0.84
	skewness	0.27	0.03	skewness	-0.56	0.04
	kurtosis	1.61	1.45	kurtosis	5.77	2.94
Large		Benefit	Shapley		Benefit	Shapley
	min	-4.53	-3.83	min	-4.50	-3.83
	max	5.33	5.67	max	4.07	4.33
	sd	1.40	0.97	sd	0.61	0.59
	skewness	0.43	0.31	skewness	-0.11	-0.09
	kurtosis	0.83	2.19	kurtosis	7.48	6.67

dard normal distribution (when kurtosis and skewness equals to 0).

In summary, when the participants had similar characteristics, the results on the number of transplants were very similar for both models. However, the Shapley value resulted in a less spread result on the bias. The smaller deviation is important because it can result in a more stable mKEP.

Collaboration of different sized countries.

When the pools of the countries are not identical, their contributions to the collaboration may be quite different. In this section, we investigated the case when the size of the pools differ across countries. In this case, using a fair model is important because it would make the mKEP more beneficial to the participants with a high contribution to the shared pool.

To consider all possibilities (collaboration with larger or with smaller country), we considered an instance when three countries with different pool-sizes collaborate. We analysed a case when there is a small country with incoming pairs on 20 days average, a medium-sized country with 10 days on average and a country with a large pool where the RDPs register into the KEP on average every 5 days.

The fair models change the allocation of the results, compared to the model maximisation of the number of transplants. To investigate which participant would benefit from considering the fair values, we compared the result of the fair model to the original maximisation model without the IR constraint. For the sake of comparison of the different sized countries, we used relative changes. Therefore we calculated a *relative Price of Fairness* (RPoF) indicator, which shows the relative loss of the participant when the fair model is used.

Let v the number of transplants that a country made by the simple maximisation model. And let the w be the total number of transplants with the fair-model. Then

$$\text{RPoF} = \frac{v}{w} - 1. \quad (11)$$

Therefore if the RPoF is positive, then the participant lost transplants when the fair model was used. Table 3 present the *RPoF* results of each country.

TABLE 3: RPoF of the three different sized countries.

	Small		Medium		Large	
	MP	CR	MP	CR	MP	CR
Benefit	10.28%	0.04%	2.05%	-0.23%	-2.65%	0.14%
Shapley	9.40%	0.31%	2.37%	-0.50%	-2.70%	0.10%

When the collaboration was CR, the result of the fair-model slightly differs from the maximal number of transplants. In this case, in the first run, the set of remaining pairs the participants share, is usually much smaller than the original pool. Hence the effect of fairness is much smaller. However, in case of the collaboration is MP, then there is a much higher effect across countries. For the small-pooled country, the RPoF is higher. Therefore this country gets fewer transplants with the consideration of the Fair-value. The effect is similar but smaller for the medium-pooled country. However, the country with the largest pool got a negative RPoF. That means with the consideration of the fair-model, the large-pooled country gets more transplant than in the case of the simple maximisation of the number of transplants. In total, the results indicate that the simple maximisation of the number of transplants is not fair for the countries with larger pools. Therefore in the original model, the larger pools contribute excessively to the collaboration, but their advantage from the collaboration is smaller.

The results of the Benefit and the Shapley values slightly differ. In the case of the Shapley value, the medium-pooled country gets a somewhat fewer transplant, and the smaller one gets more.

Effects of NDDs

In the previous tests, we did not consider any NDDs. In this section, we investigate the effects when one of the participants allows NDDs in her pool. We investigated the cases where the NDDs could arrive at the pool in 180 days on average.

We again considered three country cases. Among the three countries, two have equal-sized pools, and the third one has a Large-pool. We chose this setup, to investigate the effect of the same sized pools with NDDs and the different sized pools with NDDs at the same time. With the principle of *ceteris paribus*, we compared three different instances: There are no NDDs; One of the small pooled countries has NDDs; the Large country has NDDs. With these instances, we would investigate the effect when a similar country, a larger one, or a smaller country, has NDDs. However, for the comparison, we calculated the results without any NDDs as well.

Table 4 presents the RPoF when no country has NDDs. The fair models in the collaboration level of CR has almost zero effect compared to the simple maximisation of the number of transplants. In the case of MP, the small countries benefited more from the

TABLE 4: RPoF:No country has NDDs

	MP			CR		
	Large	Small	Small-2	Large	Small	Small-2
Beneit	-2.39%	7.94%	5.25%	0.32%	0.23%	-1.14%
Shapley	-2.40%	8.39%	5.50%	0.31%	0.26%	-0.87%

model without fair-values. When a small pool with a larger one merges, the pairs from the smaller one has a much higher chance for transplantation. However, the pairs from the larger pool do not have that much increase. Hence the smaller countries contribution was much smaller. The large country, on the other hand, had a more significant effect on the transplants number. Therefore the country with the largest pool size would get more transplant with the fair model.

TABLE 5: RPoF: One small country has NDDs

	MP			CR		
	Large	Small	Small-NDD	Large	Small	Small-NDD
Beneit	1.14%	11.12%	-11.69%	0.62%	0.63%	-1.97%
Shapley	0.99%	10.90%	-11.05%	0.57%	0.43%	-2.02%

In the case, when one small country can have NDDs, with the fair model, this country gets an improvement in the number of transplants. Furthermore, in this case, the large-pooled country has a slightly positive RPoF, which means it loses transplants in the fair instances. Interestingly it leads to a conclusion that even a few NDDs (registered on every half-year on average) can have more effects on the contribution to the collaboration than the size of the pool. Therefore the RPoF of the small country with NDDs became negative while the Large country's RPoF shifted positively. The other small-pooled country has even more reduction in the number of transplants compared to the original case without NDDs.

TABLE 6: RPoF:Large pooled country has NDDs

	MP			CR		
	Large-NDD	Small	Small-2	Large-NDD	Small	Small-2
Beneit	-4.27%	13.44%	8.32%	0.24%	1.17%	-0.86%
Shapley	-4.38%	13.11%	8.94%	0.13%	0.98%	-0.59%

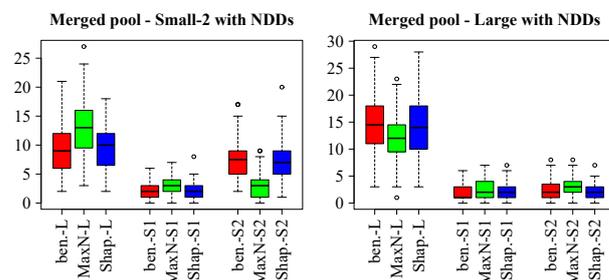
As in Table 4, the simple maximisation model is unfair against the larger-pooled countries. Furthermore, Table 5 shows that the NDDs made more effect on the level of contribution to the collaboration than the size of the large country. Moreover, we investigated when only the largest-pooled country has NDDs. Unsurprisingly the large-pooled country with NDDs got an even more significant improvement on the number of transplants when the fair model was applied. The RPoF of this situation is shown in table 6. While in the no-NDDs case, the increase of the fair model was around 2.4%, with NDDs the improvement increased to around 4.3%.

Effects of NDDs on transplants

Permitting NDDs in the multinational case may have a higher effect on other countries. For this, we considered every chain, initiated by a NDD and calculated

the number of recipients of each country. Hence, figure 2 presents the distribution of those transplants, which concluded because of a NDD.

Fig. 2: Effects of the NDDs



In the left side of figure 2, there is the effect of the small country's NDDs. Interestingly the NDDs resulted in more improvement for the Large country than the country of the NDDs. The fair-models somewhat reduce the difference, but still, the Large country remains the most affected. The other small country gets a bit of improvement because of the NDDs, but she also lost some transplant because of the fair values.

On the right side, there is the effect when the larger country permits NDDs. The Large country gets the most improvement in this case, similarly to the other case. Due to the provider of these donors are also the Large country, the fair values improve her results.

CONCLUSION

We concluded some tests over the Shapley value and the Benefit value. We used a multi-period simulation of an mKEP with three countries. The Shapley- and Benefit value resulted in a similar result in the number of transplants. However, the Shapley value has a slightly better result, on the bias, a.k.a the temporary unfairness of the mKEP. The bias of the Shapley value had a much smaller standard deviation in most of the cases. In general, the fair model gives an improvement of the countries with a larger pool. Hence their contribution is more relevant in the collaboration. The inclusion of NDDs mostly affected the country with larger pools. Hence when one smaller country has NDDs, they get more benefit from the fair-model.

ACKNOWLEDGEMENTS

Supported by COST Action CA15210 ENCKEP, supported by COST (European Cooperation in Science and Technology) – <http://www.cost.eu/> Biró and Gyetvai were supported by the Hungarian Academy of Sciences under its Momentum Programme (LP2016-3/2018) and Cooperation of Excellences Grant (KEP-6/2018), and by the Hungarian Scientific Research Fund – OTKA (no. K129086)

REFERENCES

Abraham, D.J., A. Blum and T. Sandholm. 2007. "Clearing algorithms for Barter exchange markets: Enabling nationwide kidney exchanges." *Proceedings of the 8th ACM conference on Electronic commerce* pp. 295–304.

- Agarwal, Nikhil, Itai Ashlagi, Eduardo Azevedo, Clayton R. Featherstone and Ömer Karaduman. 2018. Market failure in kidney exchange. Technical report National Bureau of Economic Research.
- Ashlagi, I. and A.E. Roth. 2011. Individual rationality and participation in large scale, multi-hospital kidney exchange. In *EC '11 Proceedings of the 12th ACM conference on Electronic commerce*. pp. 321–322.
- Ashlagi, I. and A.E. Roth. 2012. “New challenges in multihospital kidney exchange.” *American Economic Review* 102(3):354–359.
- Ashlagi, Itai and Alvin Roth. 2014. “Free riding and participation in large scale, multi-hospital kidney exchange.” *Theoretical Economics* 9(3).
- Biró, Péter, Bernadette Haase-Kromwijk, Joris van de Klundert et al. 2019. “Building kidney exchange programmes in Europe – an overview of exchange practice and activities.” *Transplantation* 103(7):1514–1522.
- Biró, Péter, Joris van de Klundert, David Manlove et al. 2020. “Modelling and optimisation in European Kidney Exchange Programmes.” *European Journal of Operational Research*.
- Biró, P., W. Kern, D. Pálvölgyi and D. Paulusma. 2019. Generalized Matching Games for International Kidney Exchange. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*. pp. 413–421.
- Böhmig, Georg A, Jiří Fronek, Antonij Slavcev, Gottfried F Fischer, Gabriela Berlakovich and Ondrej Viklicky. 2017. “Czech-Austrian kidney paired donation: first European cross-border living donor kidney exchange.” *Transplant International* 30(6):638–639.
- Carvalho, Margarida and Andrea Lodi. 2019. “Game theoretical analysis of Kidney Exchange Programs.” *arXiv preprint arXiv:1911.09207*.
- Carvalho, Margarida, Andrea Lodi, Joao Pedro Pedroso and Ana Viana. 2017. “Nash equilibria in the two-player kidney exchange game.” *Mathematical Programming* 161(1-2):389–417.
- Klimentova, X., A. Viana, J.P. Pedroso and N. Santos. 2019. “Fairness models for multi-agent kidney exchange programmes.” Submitted to Omega: The International Journal of Management Science.
- Santos, N., P. Tubertini, A. Viana and J.P. Pedroso. 2017. “Kidney exchange simulation and optimization.” *Journal of the Operational Research Society* pp. 1–12.
- Shapley, L. S. 1953. A Value for n-person Games. In *Contributions to the Theory of Games. Annals of Mathematical Studies 28.*, ed. H. W. Kuhn and A. W. Tucker. Princeton, New Jersey: Princeton University Press p. 307–317.
- STEP Documentation. 2016. “Scandiatransplant kidney exchange programme (STEP), Internal working document, Sahlgrenska Universitetssjukhuset.” Version 1.7, January 26, 2016.
- Valentín, M.O., M. Garcia., A.N. Costa, C. Bolotinha, L. Guirado, F. Vistoli, A. Breda, P. Fiaschetti and B. Dominguez-Gil. 2019. “International Cooperation for Kidney Exchange Success.” *Transplantation* 103(6):e180–e181.

AUTHOR BIOGRAPHIES

PÉTER BIRÓ has received his PhD in mathematics and computer science at Budapest University of Technology in 2007 and then he was a postdoc at the Computer Science Department of Glasgow University for three years. In 2010 he joined the Institute of Economics of the Hungarian Academy of Sciences, and currently he is the head of the Momentum research group on Mechanism Design. He is also a part-time lecturer at the Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest since 2013. His email address is peter.biro@krtk.mta.hu

MÁRTON GYETVAI graduated as an Actuary from the Corvinus University of Budapest in 2017 and started his PhD studies at the same institute

in the same year, at the Doctoral School of General and Quantitative Economics. In November 2017 he joined the Mechanism Design Research Group at the Research Centre for Economic and Regional Studies. His research interest lies in the fields of Combinatorial optimization. His e-mail address is gyetvai.marton@krtk.mta.hu.

XENIA KLIMENTOVA holds a PhD in Operations research from Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the RAS (2010). She is a senior researcher at the Center for Industrial Engineering and Management, IN-ESC TEC. Her research interests lie in the field of Combinatorial Optimization. Her e-mail address is xenia.klimentova@inesctec.pt.

JOÃO PEDRO PEDROSO holds a PhD in mathematical engineering from the Université catholique de Louvain, Belgium and is presently auxiliary professor at the Faculty of Sciences, University of Porto. His research interests are on computational methods for combinatorial optimization, involving simulation and machine learning. His e-mail address is jpp@fc.up.pt.

WILLIAM PETERSSON holds a PhD in Computational Graph Theory from the University of Queensland (2014). He is currently a Research Associate in the School of Computing Science at the University of Glasgow, and he studies Algorithmic Complexity, with a focus on Parameterised Complexity and Matching Problems. His email address is william.petersson@glasgow.ac.uk.

ANA VIANA holds a PhD in Electrical and Computers Engineering from the University of Porto (2004). She is Coordinator Professor at the Polytechnic of Porto and Head of Research of the Center for Industrial Engineering and Management. Her research interests lie in the field of Combinatorial Optimization. Her e-mail address is ana.viana@inesctec.pt.

Modelling, Simulation and Control of Technological Processes

RETROFIT OPTIMIZATION OF BATTERY AIR COOLING BY CFD AND MACHINE LEARNING

Eero Immonen*, Janne Sovela and Samuli Ranta
Engineering and Business, Technology Industry
Turku University of Applied Sciences
Joukahaisenkatu 3, 20520 Turku, C-siipi, Finland
Email: *Eero.Immonen@turkuamk.fi

Kirill Murashko and Paula Immonen
Laboratory of Electrical Engineering
LUT University
Yliopistonkatu 34, 53850 Lappeenranta, Finland

KEYWORDS

Battery; Air cooling; Design optimization; Retrofit; Computational fluid dynamics; Random forest; Feature importance

ABSTRACT

We investigate a simulation methodology for systematically optimizing air cooling in an *existing* battery system by placement of passive components. The goal in such *retrofit* optimization is to achieve design improvement by making as few and cheap changes in the original system as possible. Our methodology utilizes CFD for fluid flow and heat transfer modeling and machine learning for cause-effect assessment across binary design variables, such as wall placement for passive flow control. As an application, we consider computational optimization of air cooling in a scaled-down electric bus charging station battery system.

I INTRODUCTION

I-A Background

Battery systems are seen today as a promising alternative for fossil fuels in mobile machinery, for reducing carbon dioxide emissions. Indeed, the electric vehicle technology in consumer use is already relatively mature, and progress has also been reported in electrification of public transport (Valenti et al., 2017), as well as heavy machinery (Moreda, Muñoz-García, & Barreiro, 2016; Valenzuela Cruzat & Anibal Valenzuela, 2018). It is important to note that batteries are used in their charging stations, too, for power peak control among others (Li, Huang, Zhang, & Bao, 2017).

An industrial-scale battery often requires a separate cooling system that aims to keep the temperature across the battery cells below a threshold level, typically 40°C-60°C, specified by the cell manufacturer. Excess heat is known to deteriorate the cell and reduce its lifetime (K. Xu, Zhang, Jow, Xu, & Angell, 2002). Consequently, the cooling system should also ensure that the temperature *variation between* the cells is minimal, in order to ensure that the cells wear out evenly within a battery (Wang, Tseng, Zhao, & Wei, 2014).

The design of a battery cooling system is non-trivial, because the design objectives are complex and contradictory. On the one hand, the cells should be efficiently and uniformly cooled. On the other hand, the cooling system should be cheap and easy to install and maintain

in practical use. In mobile applications, also a low weight for the cooling system is an obvious pre-requisite. Battery cooling systems are typically designed by using Computational Fluid Dynamics (CFD), whereby designers attempt to find an optimum cooling medium as well as a means to distribute it around the cells; see e.g. (An, Chen, Zhao, & Gao, 2019; Xun, Liu, & Jiao, 2013) and the references therein. For a survey of modern battery cooling methodologies, we refer the reader to (Peng, Garg, Zhang, & Shui, 2017).

Air cooling is often preferred in practice, because it yields a simple-to-build system which is also lightweight and low-cost. Furthermore, the cooling medium (cold air) can, at least in principle, be distributed around the cells without additional ductwork. However, the heat capacity rates in air cooling are also typically quite low. This implies that flow control, i.e. guiding the cooling air efficiently, plays a pivotal role in battery air cooling system design. The fact that the heat transfer coefficient for *forced* convection is up to an order of magnitude larger than that for *natural* convection further necessitates air *motion* around the cells in an air-cooled battery.

I-B Contribution of the article

Several computational approaches for optimizing battery air cooling have been proposed in the academic literature, e.g. (Li, He, & Ma, 2013; Na, Kang, Wang, & Wang, 2018; Wang et al., 2014; X. Xu & He, 2013). However, to the authors' knowledge, little research has been reported on optimizing *existing* battery systems. In such "retrofit" optimization scenarios, one attempts to improve an existing design by making minimal modifications, as the equipment may be in daily use or difficult to access in practice.

The purpose of this article is to investigate the degree to which retrofit battery air cooling system optimization can be carried out by using passive components alone, and whether this optimization can be systematically carried out by utilizing CFD and machine learning. This optimization problem is not an easy one, because one should be able to draw conclusions robustly from a relatively small subset of the vast number of all conceivable design alternatives. In addition, as illustrated in Subsection I-C below, even simple geometry optimization tasks that involve fluid flow are often nonlinear mathematical problems. In this article, we present a generic methodology attempting to address these challenges and, as an application, we consider a down-scaled electric

bus charging station battery built at Turku University of Applied Sciences.

I-C Motivating example

In a simple retrofit flow optimization case, an existing design layout is modified to locally enhance or balance air flow by *passive* equipment, such as plate blockages and guide vanes. However, even this simple scenario is a nonlinear mathematical problem, with complex cause-effect interactions. This is illustrated in Figure 1, which displays the velocity field in a turbulent air flow CFD simulation (see Subsection II-A for details) across an array of 4 square objects. In both cases, the flow is specified as 1 m/s from left to right, extending to a virtual wind tunnel geometry with only the zone around the squares shown. The only difference between the two geometries is the baffle blockage connecting the squares on the right. Yet this small difference affects the entire flow field: Notice that not only downstream, but also flow conditions *upstream* from the blockage are affected, and that effects are seen away from the baffle.

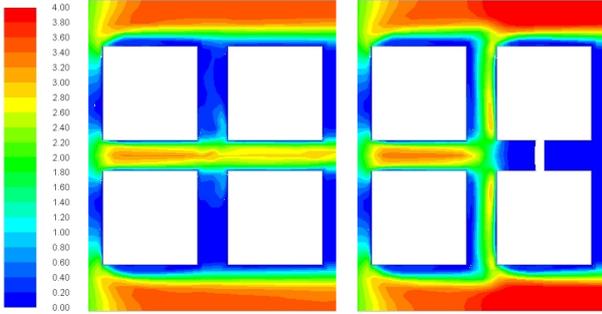


Fig. 1: Effect of Blockage on Flow Velocity Magnitude

II METHODOLOGY

The retrofit design optimization methodology we consider in this article is based on fluid flow and heat transfer modeling by CFD and a systematic cause-effect assessment across binary design variables by machine learning. The details of this approach are given in this section, and the methodology is then applied for an existing battery system in the remainder of this article.

II-A CFD modeling

1) Fluid flow

We assume that air motion in the battery system is incompressible turbulent fluid flow, which is described by the Reynolds Averaged Navier-Stokes (RANS) equations in the steady state. In tensor notation, they read:

$$\frac{\partial \bar{u}_i}{\partial x_i} = 0 \quad (1)$$

$$\rho \frac{\partial (\bar{u}_i \bar{u}_j)}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \left[2\mu \bar{S}_{ij} - \rho \overline{u'_i u'_j} \right] \quad (2)$$

with $\bar{S}_{ij} = \frac{1}{2} \left(\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i} \right)$. Here $\mathbf{u} = (u_i, u_j, u_k)^T$ denotes the fluid velocity, ρ denotes the (constant) fluid density, p denotes pressure, and μ denotes dynamic viscosity. In addition, subscripts denote coordinates, and for any scalar

quantity a , \bar{a} denotes its time average and a' denotes the fluctuating component in the Reynolds decomposition $a = \bar{a} + a'$.

The above equations require modification for any fan sections, as for design optimization it is not feasible to model fan blade motion directly. Instead, each fan zone constitutes a rotating reference frame, which yields a time-averaged (steady state) flow solution through the blades via the absolute velocity formulation.

To model turbulence, similar to the approach of (X. Xu & He, 2013), we assume that the Reynolds stresses $\rho \overline{u'_i u'_j}$ can be described via turbulence kinetic energy k and its dissipation rate ϵ , as in the $k-\epsilon$ turbulence model:

$$\rho \frac{\partial (k u_i)}{\partial x_i} = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + G - \rho \epsilon \quad (3)$$

$$\rho \frac{\partial (\epsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_\epsilon} \right) \frac{\partial \epsilon}{\partial x_j} \right] + C_{1\epsilon} \frac{\epsilon}{k} G - C_{2\epsilon} \rho \frac{\epsilon^2}{k} \quad (4)$$

where G represents the effect of turbulence generation by velocity gradients, and $\mu_t = \rho C_\mu \frac{k^2}{\epsilon}$ represents eddy viscosity, which yields the Reynolds stresses via the Boussinesq hypothesis. We use the default constant values $C_{1\epsilon} = 1.44$, $C_{2\epsilon} = 1.92$, $C_\mu = 0.09$, $\sigma_k = 1.0$, and $\sigma_\epsilon = 1.3$ reported in the academic literature.

2) Heat transfer

Heat transfer is described by the equation for conservation of energy E , for which each battery cell c_k contributes as a local volumetric heat source S_{c_k} (W/m^3):

$$\rho \nabla \cdot (\mathbf{u} E) = \nabla \cdot (\lambda \nabla T + \tau_{eff} \cdot \mathbf{u}) + \Sigma S_{c_k} \quad (5)$$

where λ denotes thermal conductivity and $\tau_{eff} \cdot \mathbf{u}$ represents energy transfer by viscous dissipation.

II-B Design optimization

The approach for CFD-based design optimization considered herein has two stages. In the first stage, a large batch of candidate designs, generated via statistical design of experiments, is simulated. In the second stage, the effect of each individual geometry modification on the design objective is estimated from this simulation data. The degree to which any given design variable improves or worsens the objective is estimated by employing the feature importance metric in the *random forest* machine learning scheme. Monte Carlo simulations help address the inherent stochasticity in the approach, as explained below.

1) Design variables

We assume that there are N locations in the air-cooled battery geometry model which can be blocked by introducing a *baffle plate*. As planar walls, they are easy to install, and their presence can be modeled as N binary (on/off) design variables. In CFD simulations, such baffle plates can be modeled as infinitely thin walls. This facilitates using precisely the same computational mesh for simulating baffled design candidates as for simulating the present situation without baffles, i.e. the *baseline case*. It is well-known, see e.g. (Immonen, 2017), that this is important for drawing conclusions from the simulation

results: Even if the numerical values obtained from CFD calculations may have limited accuracy, one can more robustly infer which one of any two competing geometries performs better than the other.

2) Fractional factorial designs

For N binary design variables, there are 2^N different combinations (the full factorial design), and evaluating them all becomes quickly impossible as N increases. Fractional factorial designs (FFDs) are statistical design of experiments that are widely used for statistical cause-effect analyses (Pham, 2006). They are well known to efficiently exploit the fact that often many elements of the full factorial design are redundant (the sparsity-of-effects principle). Moreover, they facilitate controlling the degree of variable confounding, i.e. attributing an effect to some (combination of) design variables when in fact it is due to others. By definition, an FFD of resolution R is one in which no n -factor interaction is confounded with any other effect containing less than $R-n$ factors.

3) Feature importance by random forests

Random forests are an ensemble-based machine learning method for classification and regression that operate by constructing a number of decision trees. Each tree attempts to model a subset of the full input-output relationship, and ensemble averaging aims to correct for decision trees' typical overfitting of the training set (Hastie, Tibshirani, & Friedman, 2009, pp. 587–588). For the retrofit battery air cooling system optimization problem, the inputs (or predictor variables) are the N binary design variables, and the output (or response) is the global maximum cell temperature within the battery. Here, identification of a random forest model is carried out on the FFD batch data described in Subsection II-B2. For any given random forest, the predictor variable importances (i.e. their relative significance for the response) can be estimated by permuting out-of-bag observations among the trees (Loh, 2002).

4) Monte Carlo simulation

The outcome of a feature importance analysis for a given random forest, described in Subsection II-B3, not only depends on tree-forest structure (e.g. the number of trees, leaves etc.) but also on how the underlying input-output relation was partitioned across the trees in the random forest. Therefore, this outcome is stochastic. To address this, we consider a Monte Carlo simulation in which the feature importance analysis of Subsection II-B3 is carried out for several different random forest specifications. This yields a statistical description of the feature importances, and one can more robustly infer which (if any) of the design variables have the desired effect on the output.

III APPLICATION: A DOWN-SCALED ELECTRIC BUS CHARGING STATION BATTERY SYSTEM

III-A The SeBNet battery system

During the SeBNet project (*Smart Electric Bus Network Integration*, from 1.7.2017 to 31.12.2019), a student team at Turku University of Applied Sciences developed

and built the initial down-scaled (1:10) version of an electric bus charging station battery system shown in Figure 2. It consists of a series connection of 16 LiFePO4 cells (GWL Power, 3.2 V, 20 Ah, 64 Wh, 0.65 kg), which must be kept below 45°C during charge and below 55°C during discharge, according to the manufacturer's specification. Further, the maximal continuous charge and discharge currents for the cells are 20A (1C) and 60A (3C), respectively. The cell's internal resistance is stated by the manufacturer to be less than 0.002 Ω. These yield an estimate of the maximum heat generation rate: $16 \times 7.2 \text{ W} = 115.2 \text{ W}$ for the whole battery system. The two fans (Arctic F14PWM, each rated at 126 m³/h) on the back panel, operated at the maximum rpm, attempt to remove this heat from the casing. In practical use, the air cooling system utilizes dry air at room temperature 20°C drawn through the front panel only; the top and bottom sections of the casing are sealed closed.

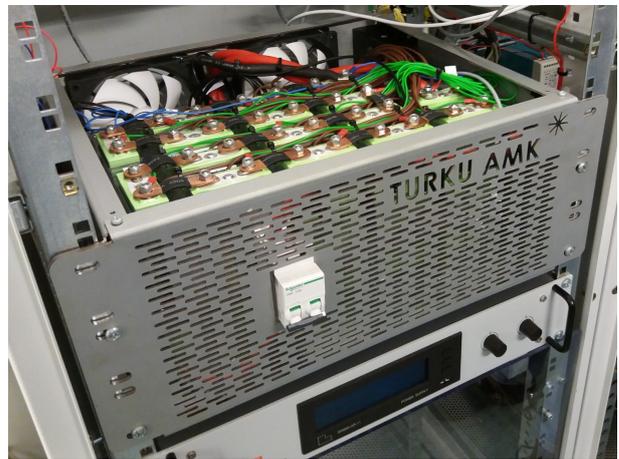


Fig. 2: The SeBNet Battery System

III-B Baseline CFD model development and simplifications

1) Geometry

The CFD geometry model for the SeBNet battery system, adapted from the original SolidWorks design drawings, is displayed in Figure 3.

The 16 cells (shown in blue color in Figure 3) were modeled as 71 mm × 178 mm × 28 mm solid rectangular regions, as in the cell specification document. The cells are connected by rectangular solid busbars (shown in brown color in Figure 3). The CFD geometry model includes the battery management system, capacitors, relays and support structures as flow blocking objects. No wire connections were included in the CFD model because they only fill a relatively small part of the total battery volume and their exact locations is may change or even be unknown in practice.

2) Mesh

The computational mesh (i.e. spatial discretization) chosen for the model consists of approximately 601000 cells. A mesh sensitivity analysis was carried out, for meshes up to 5.2 million cells to ensure independence of the results on mesh resolution.

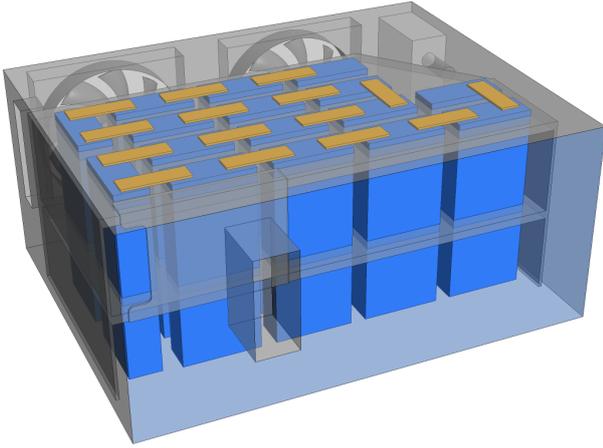


Fig. 3: The SeBNet Battery System CFD Geometry

3) Casing

The casing walls were assumed to be adiabatic.

4) Fans

The fan blades were modeled by hand as accurately as possible. To account for modeling errors, the fan speed in the simulations was calibrated to match the specified maximum volume flow rate of 126 m³/h in virtual wind tunnel simulations (see Figure 4 for an example).

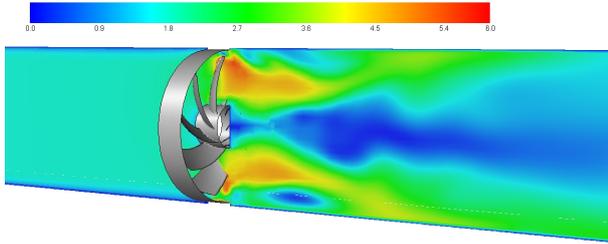


Fig. 4: Simulated Air Velocity in a Virtual Wind Tunnel (Fan Speed 1200 RPM)

5) Front panel

The front panel with TURKU AMK branding (cf. Figure 2) was modeled as a homogeneous inlet vent boundary yielding a pressure drop $\Delta p = k_L \frac{1}{2} \rho v^2$. The loss coefficient value $k_L = 21.5$ was obtained from virtual wind tunnel simulations, similar to the fans described in Subsection III-B4 above.

6) Battery cells

The internal structure and thermo-electric behavior of the cells was not modeled in detail. Instead, each battery is treated as a homogeneous volumetric heat source of 7.2 W. This corresponds to the maximum continuous discharge conditions with an infinite battery capacity, i.e. extremal use conditions from the practical point of view. The battery terminals were not included separately in the simulation model, as they are covered by the busbars.

7) Materials

The incoming air is treated as dry at 20°C, the busbars are made of copper and the LiFePO₄ cells' thermal

properties were obtained from the academic literature (Mathewson, 2014).

III-C Simulation environment and solver settings

For CFD simulations, we utilized ANSYS Fluent 2019 R3 with the Coupled pressure-velocity scheme and second-order discretization everywhere except for turbulence, which was specified as the First Order Upwind scheme. All simulations were carried out on the “Puhti” supercomputer at CSC – IT Center for Science Ltd, Finland. Each case was simulated using 40 CPU cores and 10000 iterations. Such a high iteration count is necessary for reaching a steady state in the flow field, as especially convergence of the energy equation solution turned out to be slow. The simulation time for one case was approximately 50 minutes.

III-D Baseline model simulation results

Figure 5 displays the simulated velocity magnitude profile within the battery system on two artificial perpendicular planes. Note that there is significant air flow through the opening above the cells, where there is little cell area exposed to cooling. Also note that there are higher and lower velocity regions between the cells, which causes some cells to be less effectively cooled than others.

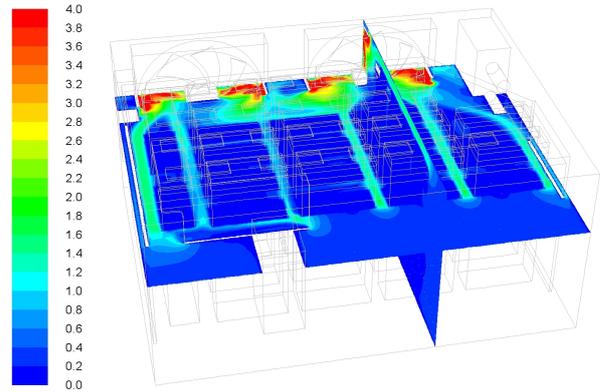


Fig. 5: Flow Velocity Magnitude (Baseline CFD)

Figure 6 displays the simulated temperature profile at an artificial midsection plane. Clearly there is significant variation between the cell temperatures. Moreover, the hottest cells are close to the low air velocity regions (cf. Figure 5).

III-E Model validation

In order to validate the simulation model by measurements, the SeBNet battery system was first charged from 0% to 100% in 60 minutes at 1C (20A), then left to settle for 15 minutes, and finally discharged from 100% to 0% in 20 minutes at 3C (60A). Cell temperatures were continuously measured, with the temperature probes placed at the center of the large side on each cell.

Table I displays the simulated maximum cell temperatures (middle column) and the corresponding measured cell temperatures (right column). The average error

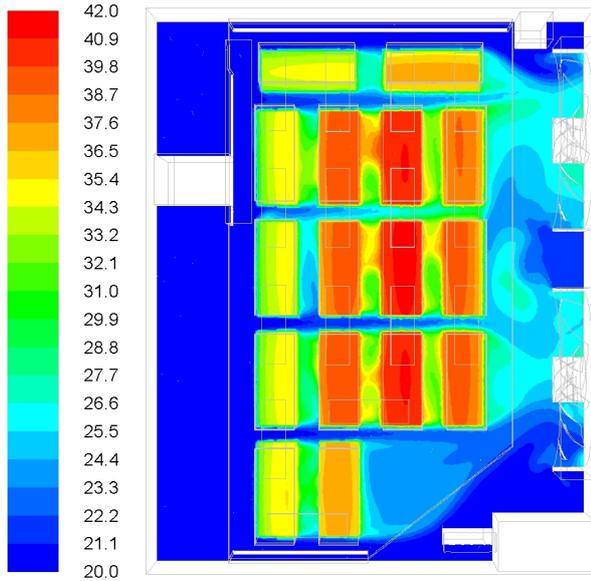


Fig. 6: Temperature (Baseline CFD)

between simulation results and measurements is 6%. We consider this accuracy reasonably good especially bearing in mind the differences between the simulation and measured cases: First of all, the measurements were not obtained from a steady state situation, but after a transient cycle. Second, the cells' internal resistances were measured, and they ranged from 1.81 m Ω to 2.35 m Ω , which is grossly outside of the manufacturer's specification (below 2 m Ω). Finally, a comparison is made between the simulated *maximum* cell temperature and temperature measurement *at a specific location*.

IV CFD OPTIMIZATION OF THE AIR COOLING SYSTEM

IV-A Retrofit design optimization

1) Baffle plate placement options

In total 24 possible baffle plate placement locations, shown in green color in Figure 7, were identified for the SeBNet model. This set of possible baffle locations is a compromise between computational simplicity and ability to guide the flow in complex patterns through the casing. In the optimum design, some of the 24 zones are walls and the others are interior zones that are transparent to flow and heat transfer in CFD simulations.

2) Design of experiments

We used the Franklin-Bailey algorithm (Franklin & Bailey, 1977) implemented in MATLAB 2019b for constructing an FFD of resolution 4. The resulting design-of-experiments was a binary 256×24 matrix, for which the elements of each row determined the subset of the 24 baffles that was to be included as walls in the CFD simulation. This FFD is a compromise of simulation time and case comprehension: At resolution 4, the design does not confound either main effects or two-way interactions but may confound two-way interactions (and beyond) with each other.

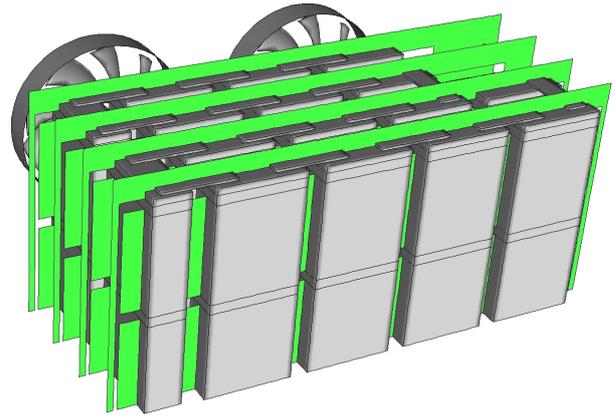


Fig. 7: Possible Baffle Plate Locations (Green).

3) Effect of baffles 1-24 on cell temperatures

Figure 8 shows the individual effects of the baffle plates 1 – 24 on the global maximum cell temperature observed in the CFD simulations in the steady state. The underlying data is from the 256 CFD simulation cases of the FFD. The boxplot shows the importance of the features across a Monte Carlo simulation of 2000 iterations, with the random forest tree count ranging between 30 – 100 (uniform distribution). Each random forest model was fitted using bootstrap aggregation in MATLAB 2019b (Breiman, 2001).

By Figure 8, only baffles 7 and 8 contribute to lowering the cell temperatures. These baffle plates are shown in Figure 9. For all other baffle plates, the boxplot maxima in Figure 8 extend above zero, indicating that they may not help reducing the maximum cell temperatures.

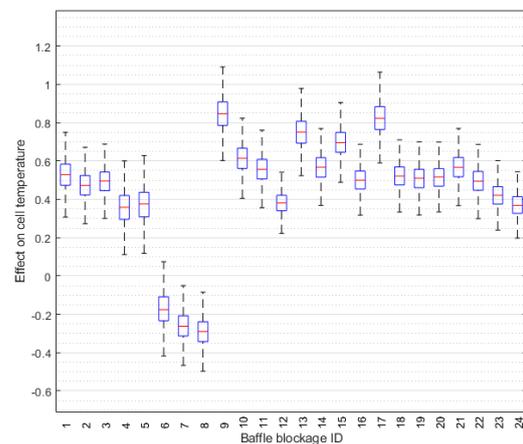


Fig. 8: Effect of Baffle Plates 1 – 24 on Global Maximum Cell Temperature

IV-B Optimal design: Results and analysis

1) Optimal design

In Subsection IV-A3, we concluded that only the baffles 7 and 8 could be expected to reduce the maximum cell temperature in the battery system. The *optimal design*

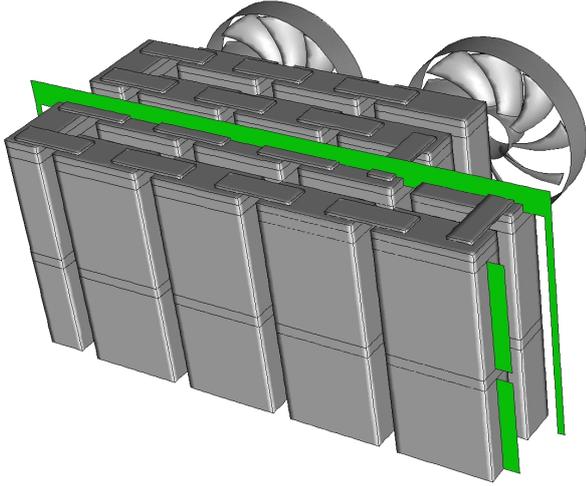


Fig. 9: Baffle Plate 7 (Two Vertical Sections on the First Row of Cells) and Baffle Plate 8 (Single Section Extending to the Top of the Second Row of Cells) Shown in Green

is thus that presented in Figure 9. In the remainder of this section, we show that it indeed displays a lower simulated global maximum temperature than in the baseline model. We also analytically study the fluid mechanics properties of the CFD simulation results for the optimal case.

2) Optimal design simulation results

Figure 10 displays the velocity magnitude profile within the battery system on the same two artificial perpendicular planes as in Figure 5, with baffles 7 and 8 now included as walls. In comparison to the baseline model (cf. Figure 5), there is no longer air flow through the compartment above the cells. The modifications yield higher flow speeds between the cells and potentially better cooling of the hottest cells of the baseline model.

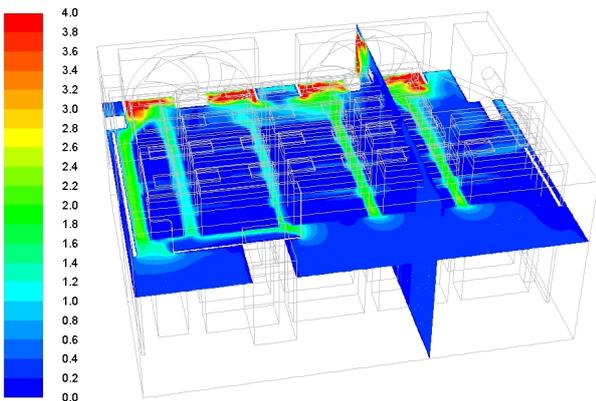


Fig. 10: Flow velocity magnitude (Optimal CFD)

The simulated maximum cell temperatures for the optimal design are shown in Table I (left column). In addition, Figure 11 displays the simulated temperature profile at the same artificial midsection plane as in Figure 6. In comparison to the baseline model (cf. Figure, 6), the highest simulated cell temperatures are indeed lower in the optimal design: The maximum cell temperature is

reduced from 42.5°C to 41.5°C. The difference is clearly not a large one, and, moreover, some cells are hotter in the optimum design than in the baseline case. However, the design objective utilised in Subsection IV-A3 was specifically the reduction of the global maximum cell temperature, which was thus achieved.

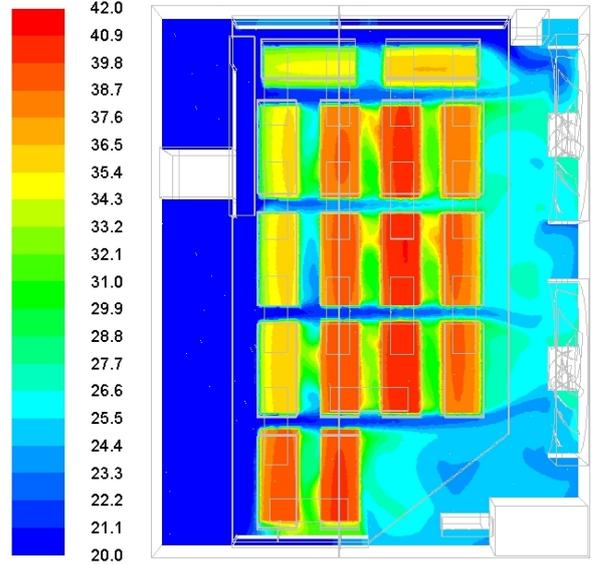


Fig. 11: Temperature (Optimal CFD)

TABLE I: Cell Temperatures in Simulations and Measurements

Cell id	Optimal (CFD)	Baseline (CFD)	Baseline (measured)
#1	39.3	39.6	41.8
#2	39.4	40.1	42.1
#3	38.9	39.5	42.4
#4	36.7	37.2	37.9
#5	40.9	42.2	43.5
#6	41.5	42.5	41.9
#7	41.2	41.8	40.8
#8	39.9	40.4	41.6
#9	39.6	40.2	37.8
#10	39.7	40.8	37.5
#11	35.3	35.0	37.7
#12	36.2	36.5	37.2
#13	36.2	35.8	37.6
#14	36.9	36.1	34.8
#15	39.9	36.2	40.3
#16	40.1	38.2	40.5

3) Evaluation

From fundamental thermodynamics, the heat transfer rate for an air-cooled battery system can be increased by increasing the rate of cooling air. However, introduction of baffles in the SeBNet battery system in fact *reduces* the total mass flow rate through the front panel, because they result in an additional pressure loss. This is seen in Figure 8: Most of the modifications yield an increase in the maximum cell temperature. Consequently, optimization of the SeBNet battery system by any baffle plate arrangement is challenging, which is reflected in the modest improvement achieved here relative to the baseline model.

V CONCLUSIONS

In the present article, we have investigated a simulation-based methodology for optimizing air cooling in an *existing* battery system by passive components, such as baffle plates. We utilized CFD for fluid flow and heat transfer modeling and machine learning for cause-effect assessment across binary design variables. We demonstrated use of the methodology for a down-scaled electric bus charging station battery system. In this application, a simple combination of baffle plates resulting in a lower global maximum cell temperature was systematically extracted out of a vast set of 2^{24} design alternatives. Although the temperature reduction achieved was relatively small, the results indicate that the approach considered herein may be beneficial for retrofit design optimization. In future applications, one should not only consider introduction of blockages, but removing existing ones, for improved air circulation and better cooling efficiency. Moreover, future research on the methodology outlined in this article should focus on accounting for both the global maximum cell temperature and inter-cell temperature variation. The latter was not considered here.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from Business Finland (e3Power project). The ability to use CSC's scientific computing facilities is also gratefully acknowledged.

REFERENCES

- An, Z., Chen, X., Zhao, L., & Gao, Z. (2019). Numerical investigation on integrated thermal management for a lithium-ion battery module with a composite phase change material and liquid cooling. *Applied Thermal Engineering*, 163, 114345.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Franklin, M., & Bailey, R. (1977). Selection of defining contrasts and confounded effects in two-level experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3), 321–326.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Immonen, E. (2017). 2D shape optimization under proximity constraints by CFD and response surface methodology. *Applied Mathematical Modelling*, 41, 508 - 529.
- Li, Q., Huang, M., Zhang, W., & Bao, Y. (2017, Aug). Economic study on bus fast charging station with battery energy storage system. In *2017 IEEE ITEC Asia-Pacific Conference* (p. 1-6).
- Li, X., He, F., & Ma, L. (2013). Thermal management of cylindrical batteries investigated using wind tunnel testing and computational fluid dynamics simulation. *Journal of Power Sources*, 238, 395 - 402.
- Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica sinica*, 361–386.
- Mathewson, S. (2014). *Experimental measurements of LiFePO4 battery thermal characteristics* (Unpublished master's thesis). University of Waterloo.
- Moreda, G., Muñoz-García, M., & Barreiro, P. (2016). High voltage electrification of tractor and agricultural machinery—a review. *Energy Conversion and Management*, 115, 117–131.
- Na, X., Kang, H., Wang, T., & Wang, Y. (2018). Reverse layered air flow for li-ion battery thermal management. *Applied Thermal Engineering*, 143, 257 - 262.
- Peng, X., Garg, A., Zhang, J., & Shui, L. (2017, Oct). Thermal management system design for batteries packs of electric vehicles: A survey. In *2017 asian conference on energy, power and transportation electrification (ACEPT)* (p. 1-5).
- Pham, H. (2006). *Springer handbook of engineering statistics*. Springer Science & Business Media.
- Valenti, G., Liberto, C., Lelli, M., Ferrara, M., Nigro, M., & Villante, C. (2017, June). The impact of battery electric buses in public transport. In *2017 IEEE EEEIC / I CPS Europe conference* (p. 1-5).
- Valenzuela Cruzat, J., & Anibal Valenzuela, M. (2018, Nov). Integrated modeling and evaluation of electric mining trucks during propel and retarding modes. *IEEE Transactions on Industry Applications*, 54(6), 6586-6597.
- Wang, T., Tseng, K., Zhao, J., & Wei, Z. (2014). Thermal investigation of lithium-ion battery module with different cell arrangement structures and forced air-cooling strategies. *Applied Energy*, 134, 229 - 238.
- Xu, K., Zhang, S., Jow, T. R., Xu, W., & Angell, C. A. (2002). Libob as salt for lithium-ion batteries: a possible solution for high temperature operation. *Electrochemical and Solid-State Letters*, 5(1), A26–A29.
- Xu, X., & He, R. (2013). Research on the heat dissipation performance of battery pack based on forced air cooling. *Journal of Power Sources*, 240, 33 - 41.
- Xun, J., Liu, R., & Jiao, K. (2013). Numerical and analytical modeling of lithium ion battery thermal behaviors with different cooling designs. *Journal of Power Sources*, 233, 47 - 61.

AUTHOR BIOGRAPHIES

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.

JANNE SOVELA works as Senior Advisor at New Energy Research Group at Turku University of Applied Sciences, Finland.

SAMULI RANTA works as Senior Lecturer and New Energy Research Group Leader at Turku University of Applied Sciences, Finland.

KIRILL MURASHKO works as Post-Doctoral Researcher at LUT University, Finland.

PAULA IMMONEN works as Post-Doctoral Researcher at LUT University, Finland.

ANALYTICAL APPROACHES FOR DETERMINING THE EFFECTS OF WORT EXTRACT ON THE SPECIFIC GROWTH RATE OF THE YEAST POPULATION

Georgi Kostov*, Rositsa Denkova-Kostova**, Vesela Shopska*,
*Department of Wine and Beer ** Department of Biochemistry and Molecular Biology”
University of Food Technologies, 4002, 26 Maritza Blvd., Plovdiv, Bulgaria
E-mail: george_kostov2@abv.bg; rositsa_denkova@mail.bg; vesi_nevelinova@abv.bg
Bogdan Goranov
LBLact, Plovdiv, Bulgaria, E-mail: goranov_chemistry@abv.bg
Kristina Ivanova
Institute of canning and food quality, Bulgaria, E-mail: kriss_k@abv.bg

KEYWORDS

brewing, kinetics of biomass growth, high gravity fermentation

ABSTRACT

The effect of the original wort extract on the biomass growth rate was examined in the present work. For this purpose, analytical dependencies to determine the kinetic parameters that reveal different sides of the fermentation process, have been used. The joint influence of the wort extract and the immobilization process on the growth of the microbial population as well as on the accuracy of the analytical dependences used to describe the microbial growth kinetics in the yeast population was established by analyzing the kinetics of the fermentation process.

INTRODUCTION

Basic processes in beer production

The main stages of the brewing process are wort production (mashing, filtering, boiling, clarification and cooling), fermentation (consisting of main fermentation and beer maturation processes), and processes related to the processing and stabilization of the finished beer (Kunze 2003). The main stage in beer production is the process of alcohol fermentation. Ethanol fermentation occurs as a result of the yeast enzymatic activity in the Embden-Meyerhof Parnas pathway, which leads to glucose conversion to pyruvate. Under anaerobic conditions, the yeasts convert pyruvate to ethanol and CO₂. In aerobic conditions, yeasts consume sugars, mainly for biomass accumulation and CO₂ production (Boulton and Quain, 2001).

Fermentation and maturation are the longest processes in brewing. The main fermentation lasts between 3-6 days and the maturation lasts up to 2 weeks depending on the fermentation type and the equipment used. On such a competitive market, the potential time savings offered by immobilized cell technology (ICT) have to be taken into account. Immobilized cells are physically limited or localized in a specific space while preserving their catalytic activity, and if possible, and even necessary, viability, and which can be used repeatedly and continuously (Godia et al., 1987; Shopska et al., 2019). The use of immobilized cell systems offers a number of advantages (Boulton and Quain, 2001; Hayes et al., 1991; Shopska et al., 2019): increased cell concentration in the reactor operating volume and

increased reaction rate; smaller bioreactor sizes, and in continuous processes - lower reaction times; easy separation and regeneration of biomass; possibility for use in both batch and continuous alcohol fermentation. Immobilization is characterized by the following disadvantages (Mensour et al., 1996; Shopska et al., 2019): limited mass exchange due to the presence of diffusion resistances; possibility of destroying the matrix at a high rate of the fermentation process and/or the formation of gaseous metabolites; changes in the physiological state of the immobilized culture as well as in the growth rate and the overall stoichiometry of the reactions in the immobilized system. Immobilized yeast cell technology allows beer production to be accomplished in as little as 2-3 days (Branyik et al., 2005). Immobilized cell systems are heterogeneous systems in which considerable mass transfer limitations can occur, resulting in a changed yeast metabolism (Willaert, 2007). Consequently, the main challenge for ICT is to reproduce the traditional beer flavor.

The fermentation rate depends largely on the original wort extract. The increase in the wort extract leads not only to prolongation of the fermentation time, but also to changes in both primary and secondary yeast metabolism. In recent years, the so-called. high gravity brewing was introduced. The wort has a high original extract (15-17.5 °P). The final value of the extract depends on biochemical and microbiological limitations. The increase in the original extract has a negative effect on the fermentation process and the quality of the finished beer (Kunze 2003).

Equation of fermentation

The process of alcohol fermentation in beer production can be described with the following system of differential equations (eq. 1). The Monod equation (eq. 2 and eq.3) can be used to describe the kinetics of microbial growth in beer production. The parameter determination in the equation can be done analytically or by solving the differential growth equation by numerical methods. In our previous work (Kostov et al., 2019), we described analytical approaches to determine the yeast specific growth rate. Out of the methods discussed in our previous work, the methods of Warpholomeew-Gurevich Linearization and Linearization by Substrate Consumption and Product Accumulation were most suitable for describing the fermentation process in the brewing industry (Kostov et al., 2019).

$$\begin{aligned}
\frac{dX}{d\tau} &= \mu X \\
\frac{dP}{d\tau} &= qX \\
\frac{dS}{d\tau} &= -\frac{1}{Y_{X/S}} \frac{dX}{d\tau} - \frac{1}{Y_{P/S}} \frac{dP}{d\tau} \\
\frac{dE}{d\tau} &= Y_E \mu X \\
\frac{dHA}{d\tau} &= Y_{HA} \mu X \\
\frac{dA}{d\tau} &= Y_A \mu X - k_A X A \\
\frac{dVDK}{d\tau} &= Y_{VDK} \mu X - k_{VDK} X VDK
\end{aligned} \quad (1)$$

where: X – biomass concentration, g/dm³; P – ethanol concentration, g/dm³; S – real extract, g/dm³; Y_{P/S}, Y_{X/S} – yield coefficients; μ – specific growth rate, h⁻¹; q – specific ethanol accumulation rate, g/(g.h); E – ester concentration, mg/dm³; HA – higher alcohol concentration, mg/dm³; A – aldehyde concentration, mg/dm³; VDK – vicinal diketone concentration, mg/dm³; Y_{HA}, Y_E, Y_A, Y_{VDK} – yield coefficients of the corresponding metabolites, mg/(g.h); k_A, k_{VDK} – reduction coefficients for aldehydes and vicinal diketones, mg/(g.h); K_{SX}, K_{SP} – Monod constants, g/dm³; K_{SXi}, K_{SPi} – inhibition constants, g/dm³; P_M, P_{MP} – maximal ethanol concentration for full inhibition of the process, g/dm³

$$\frac{dX}{d\tau} = \mu X \quad (2)$$

$$\mu = \mu_{\max} \frac{S}{K_S + S} \quad (3)$$

where: μ_{max} – maximum specific biomass growth rate, h⁻¹; K_S – saturation constant, g/dm³.

Warpholomeew-Gurevich Linearization

For the determination of the kinetic parameters in equation (3), the approach suggested by Warpholomeew-Gurevich can also be used. In this case, the experimental parameters are represented by modifications of equations (2) and (3) as follows (Warpholomeew and Gurevich, 1999):

$$\frac{\ln\left(\frac{X}{X_0}\right)}{\tau} = \frac{\mu_{\max} S_0}{K_S + S_0} + \frac{K_S}{K_S + S_0} \frac{\ln\left(\frac{X_F - X}{X_F - X_0}\right)}{\tau} \quad (4)$$

where: X_F – final biomass concentration, g/dm³; X – current biomass concentration, g/dm³; X₀ – initial biomass concentration, g/dm³; S₀ and S – initial and current substrate concentration, g/dm³.

To determine the kinetic parameters in the Monod equation, the dependence $\frac{\ln\left(\frac{X}{X_0}\right)}{\tau} = f\left(\frac{\ln\left(\frac{X_F - X}{X_F - X_0}\right)}{\tau}\right)$ is used,

and the kinetic parameters in the Monod model are determined from the straight line equation.

Linearization by Substrate Consumption and Product Accumulation

Warpholomeew-Gurevich linearization approach also allows the growth rate to be determined using the changes in the substrate and the product concentration. For this purpose, equation (4) would acquire the form (Warpholomeew and Gurevich, 1999):

$$\frac{\ln\left(\frac{S_0 - S_j}{S_0 - S_i}\right)}{\tau_j - \tau_i} = \frac{\mu_{\max} S_0}{K_S + S_0} + \frac{K_S}{K_S + S_0} \frac{\ln\left(\frac{S_j}{S_i}\right)}{\tau_j - \tau_i} \quad (5a)$$

$$\frac{\ln\left(\frac{P_j}{P_i}\right)}{\tau_j - \tau_i} = \frac{\mu_{\max} S_0}{K_S + S_0} + \frac{K_S}{K_S + S_0} \frac{\ln\left(\frac{P_m - P_j}{P_m - P_i}\right)}{\tau_j - \tau_i} \quad (5b)$$

where S₀; S_j; S_i – initial substrate concentration, g/dm³; substrate concentration corresponding to t_j, g/dm³; substrate concentration corresponding to t_i, g/dm³; τ_j-t_i – time interval equal to the difference between the final and the current process time, h; P_j; P_i – product concentration corresponding to t_j, g/dm³; product concentration corresponding to t_i, g/dm³; P_m – maximum concentration of the product, g/dm³.

The equations are plotted in the corresponding coordinates $\frac{\ln\left(\frac{S_0 - S_j}{S_0 - S_i}\right)}{\tau_j - \tau_i}$ of $\frac{\ln\left(\frac{S_j}{S_i}\right)}{\tau_j - \tau_i}$ and $\frac{\ln\left(\frac{P_j}{P_i}\right)}{\tau_j - \tau_i}$ of $\frac{\ln\left(\frac{P_m - P_j}{P_m - P_i}\right)}{\tau_j - \tau_i}$.

The parameter K_S is determined from the angular coefficient of the straight lines, while μ_{max} is determined from the cut-off.

The purpose of the present work was to investigate the influence of the original wort extract on the accuracy of the analytical models for describing the fermentation process kinetics. The aim was to define an analytical approach making it is easy, accurate and quick to look for the influence of a parameter on the fermentation process kinetics and to make comparisons between different fermentation regimes. In addition, the effect of the immobilization process on the yeast primary metabolism kinetics was investigated.

MICROORGANISMS AND FERMENTATION CONDITIONS

The fermentations were carried out with top-fermenting yeast strain *Saccharomyces pastorianus* (*carlsbergensis*) S-23. Wort with 5 different original extracts – 9, 11, 13, 15 and 17 % was used for fermentations. All media were sterilized at 121 °C for 20 min before fermentations. The immobilization procedure and the fermentation conditions were previously reported in (Parcunev et.al. 2012). In this study the fermentations with free and immobilized cells were investigated for the same time intervals to determine the impact of immobilization on the yeast metabolism. Biomass concentration of immobilized cells was determined according Parcunev et.al. 2012.

The fermentation of all variants was carried out in plastic bottles, with a volume of 500 cm³, equipped with an airlock system. The 400 cm³ of wort was placed into the bottles and inoculated with yeast suspension at a concentration of 10⁷ cfu/cm³. For the variants with immobilized cells, the microcapsule mass was 15 g for 400 cm³ wort.

All fermentation processes were carried out at a constant temperature (main fermentation and maturation) of 15 °C in order to avoid the effect of temperature on biochemical reactions during fermentation.

The characterization of wort, green beer and beer (OE, degree of attenuation, extract, and alcohol) was

conducted according to the current methods recommended by the European Brewery Convention (Analytica-EBC, 2004).

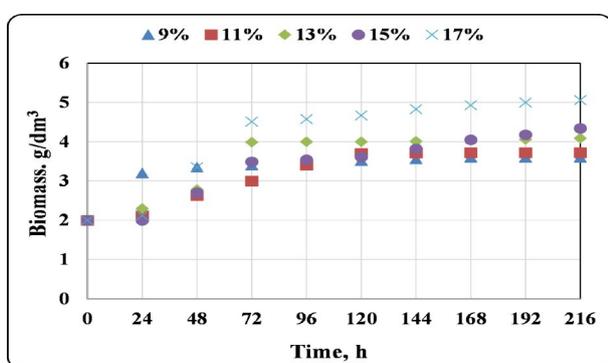
The efficiency coefficient was used to determine the impact of the immobilization process on the yeast biomass growth:

$$\eta = \frac{\mu_{IM}}{\mu_F} \quad (6)$$

where: η - efficiency coefficient; μ_{IM} - maximum specific biomass growth rate for immobilized cells, h^{-1} ; μ_F - maximum specific biomass growth rate for free cells, h^{-1} ;

RESULTS AND DISCUSSION

Before considering the possibilities for modeling the biomass specific growth rate at different initial values of the wort extract, it is necessary to consider the dynamics of biomass formation. The results for biomass accumulation in free cell fermentation and immobilized cell fermentation are presented in Fig. 1.



a) free cells

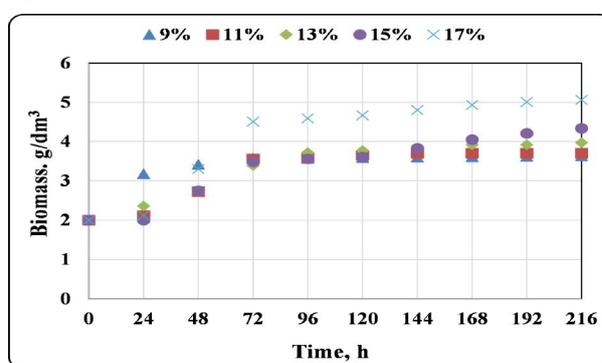
Figure 1: Biomass concentration dynamics

The increase in the extract also provoked longer biomass growth. At low original extracts, the cells entered the stationary phase rather quickly, while prolongation of biomass growth in the high gravity wort was observed. The immobilization process did not generally change the observed trends, but there was a slight extension of the lag phase. Due to the increased cell volume in the fermentation volume, the main fermentation ended faster, i.e. a decrease in the total fermentation time was observed. Another important observation that is relevant to the accuracy of the models used is the fact that smaller biomass amounts accumulated in the capsules than in free cell fermentation for the same fermentation time. This was due to the physical limitations of the cell growth inside the capsules. This, in turn, would affect both the specific growth rate and the accuracy of the models.

Warpholomeew-Gurevich Linearization

The results for the linearization by the Warpholomeew-Gurevich method are presented in Fig. 2 and Table 1. Fig. 2 shows the linear dependencies on the basis of which the kinetic parameters are determined (Table 2). The data show that the original extract directly affected the kinetic parameters determined according to the proposed method. The data in Table 1 confirm the fact

Data on free cell accumulation indicate that the biomass concentration increased with the increase in the original wort extract. This is logical since the wort with higher original extract contains more fermentable carbohydrates. On the other hand, the process lag phase increased with the increase in the original wort extract, and although a relatively close biomass concentration was found at the end of the fermentation process, it is likely that the increase in the lag phase would be reflected in the accuracy of the the models. The results show that the cells adapted relatively quickly to the fermentation conditions at low original extracts - 9% and 11%. Under these conditions, the lag phase was about 24 hours, after which the cells began active fermentation. In this case, the main fermentation ended in about 3-4 days. With the increase in the extract, the duration of the lag phase increased up to 48 hours, and it was the longest in the high gravity wort - 15% and 17%. This in turn provoked an extension of the fermentation process and the main fermentation took 4-6 days.



b) immobilized cells

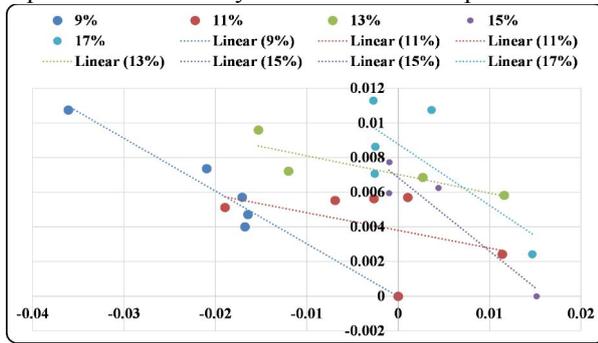
that the kinetic parameters could be determined according to the proposed method, since the value of the saturation constant K_S is in the same order as the wort extract. This, in turn, confirms the observations of Warpholomeew-Gurevich (Warpholomeew and Gurevich, 1999) that constants in the Monod equation can only be determined if the free member in the linear equation is less than 1, i.e. the saturation constant is in the order of the substrate concentration. An interesting fact is that there was obviously an optimum zone of saturation constant values where the cells grew at the highest specific growth rate. This zone is in the range of 11-13% original wort extract. This was probably due to the presence of more glucose in the wort. The accuracy of the model is greatly improved with the decrease in the saturation constant, i.e. the cells generally grow at a higher specific growth rate.

This trend was maintained in immobilized cell fermentation, but there was a shift in accuracy to wort with 13% and 15% wort extract. This could be explained by the presence of diffusion resistances and the need for higher substrate concentrations to overcome the diffusion barrier. In this case, the presence of greater amount of fermentable extract in the wort caused the higher concentration difference.

In general, the process of immobilization had no major adverse effect on the microbial population. With the exception of the variant at 9% original wort extract, in the other variants the efficiency coefficient was below 1, i.e. the immobilized cells were affected by the immobilization process. The expected tendency to increase the effect of immobilization on the yeast population was not confirmed. In general, however, the average biomass specific growth rate was affected negatively and decreased with the increase in the extract. This was associated with an increase in the saturation constant K_S , especially in the high gravity wort. K_S was in the range of 4-5% in the high gravity wort, which was very close to the non-fermentable extract of the wort. The saturation constant was lower and the cells grew at a higher average specific growth rate at low original extracts of up to 11-13%. These observations suggest that main fermentation might occur in a shorter time, while the high extract quite expectedly led to an increase in the main fermentation time.

Linearization by Substrate Consumption and Product Accumulation

Linearization by substrate and product is a modification of the Warpholomeew-Gurevich Linearization, which reflect the effect of substrate consumption and product accumulation during fermentation. It reveals whether the process is limited by the substrate or the product.



$$y_9 = -0.3042x - 2E-05; R^2 = 0.9595$$

$$y_{11} = -0.2181x + 0.0037; R^2 = 0.4179$$

$$y_{13} = -0.1077x + 0.007; R^2 = 0.7284$$

$$y_{15} = 0.3879x + 0.0099; R^2 = 0.964$$

$$y_{17} = -0.1579x + 0.0096; R^2 = 0.2873$$

a) free cells

on graph – by axis X - $\ln\left(\frac{X_F - X}{X_F - X_0}\right)$; by axis Y - $\ln\left(\frac{X}{X_0}\right)$

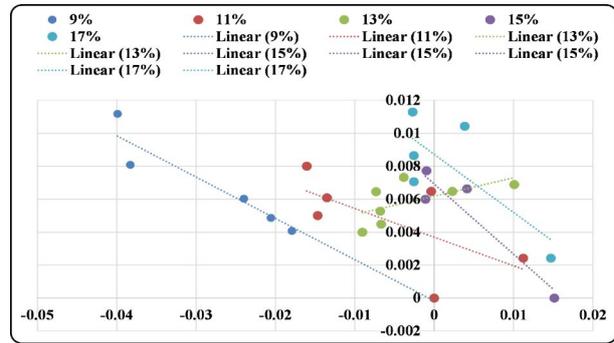
Figure 2: Warpholomeew-Gurevich Linearization

The results of the two linearization methods are presented in Fig. 3 and Fig. 4, and the kinetic parameters obtained are presented in Table 2 and Table 3. In free cells, the linearization accuracy was extremely high. This indicated that the fermentation process was substrate dependent. The data in Table 2 show a decrease in the average specific growth rate with an increase in the wort extract. Once again, two zones can be highlighted - up to 13-15% wort extract and over 15% wort extract. Cells grew at a high average growth rate in the first zone, while the average rate decreased in

Table 1: Kinetic characteristics of biomass growth determined by Warpholomeew-Gurevich Linearization

μ_{max}, h^{-1}	$K_S, g/dm^3$	μ_{max}, h^{-1}	$K_S, g/dm^3$
Free cells		Immobilized cells	
9%			
0.000029	39.35	0.00025	23.31
$R^2=0.9669$		$R^2=0.9231$	
$\eta=8.621$			
11%			
0.0047	30.66	0.0042	12.51
$R^2=0.9521$		$R^2=0.8932$	
$\eta=0.894$			
13%			
0.0078	15.58	0.0068	12.98
$R^2=0.9123$		$R^2=0.9523$	
$\eta=0.872$			
15%			
0.0127	41.92	0.0089	44.81
$R^2=0.9359$		$R^2=0.9612$	
$\eta=0.700$			
17%			
0.0114	31.87	0.0109	44.30
$R^2=0.8232$		$R^2=0.9326$	
$\eta=0.956$			

* correlation coefficients between the experimental biomass concentration and the calculated biomass concentration were calculated using the kinetic parameters obtained.



$$y_9 = -0.2507x - 0.0002; R^2 = 0.9476$$

$$y_{11} = -0.1021x + 0.0038; R^2 = 0.1894$$

$$y_{13} = 0.1111x + 0.0062; R^2 = 0.3630$$

$$y_{15} = -0.4259x + 0.0069; R^2 = 0.8762$$

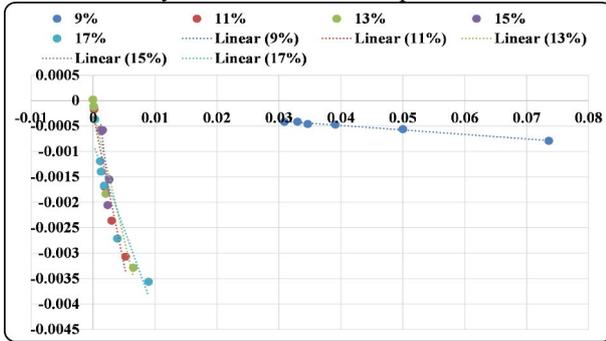
$$y_{17} = 0.3525x + 0.0087; R^2 = 0.5771$$

b) immobilized cells

the second zone. This was due to an increase in the saturation constant, which was associated with an increase in the wort extract. The effect of the substrate consumption should be explained by the process of catabolic repression in yeast. In this process, cells can consume complex carbohydrates - maltose and maltotriose only after glucose is consumed (its concentration in the wort must decrease below 2%). Therefore, the substrate uptake model can be considered as a function reflecting this process.

The model using substrate consumption more clearly emphasizes the importance of immobilization. In this case, all efficiency ratios were higher than 1, reflecting the fact that the fermentation process was completed faster in the immobilized cell fermentation. The increased amount of cells in the fermentation volume decreased the importance of the catabolite repression, i.e. the cells grew at a higher specific growth rate (Table 2). Another important feature of the model by the linearization by substrate consumption is that it

incorporates all experimental data from the end of the lag phase to the stationary phase, which increases its accuracy. The model accuracy is only affected in the zone of transition to fermentation of high gravity wort with an original extract of 15%. This is due to the fact that at that value the process begins to slow down due to the influence of substrate inhibition and the delayed maltose absorption in this case. This peculiarity should be taken into account when using the model to describe the fermentation process.



$$y_9 = -0.0088x - 0.0001; R^2 = 0.9913$$

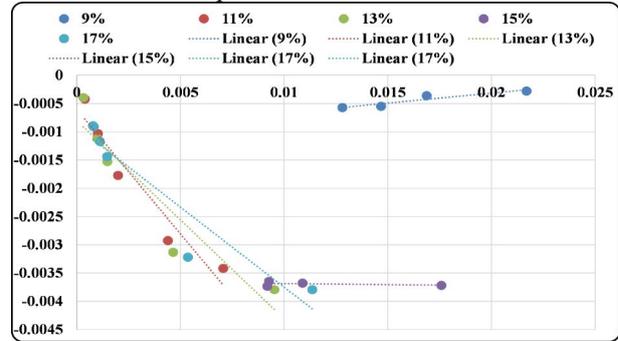
$$y_{11} = -0.5911x - 0.0003; R^2 = 0.9434$$

$$y_{15} = -0.5069x - 0.0002; R^2 = 0.9447$$

$$y_{17} = -1.0262x + 0.0008; R^2 = 0.8223$$

$$y_{13} = -0.3347x - 0.0008; R^2 = 0.8786$$

a) free cells



$$y_9 = 0.0352x - 0.001; R^2 = 0.8922$$

$$y_{11} = -0.4401x - 0.0006; R^2 = 0.9292$$

$$y_{13} = -0.351x - 0.0008; R^2 = 0.8859$$

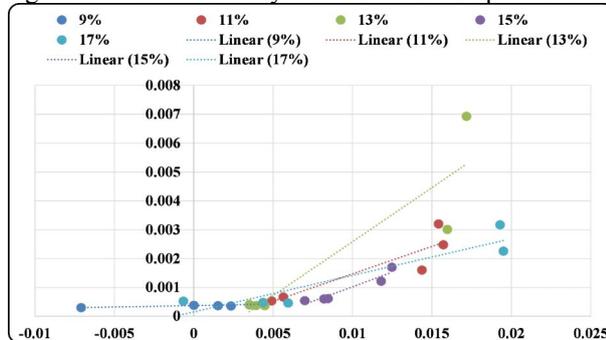
$$y_{15} = -0.0035x - 0.0037; R^2 = 0.1214$$

$$y_{17} = -0.2835x + 0.0009; R^2 = 0.8930$$

b) immobilized cells

on graph – by axis X - $\frac{\ln\left(\frac{S_j}{S_i}\right)}{\tau_j - \tau_i}$; by axis Y - $\frac{\ln\left(\frac{S_0 - S_j}{S_0 - S_i}\right)}{\tau_j - \tau_i}$

Figure 3: Linearization by Substrate Consumption



$$y_9 = 0.0101x + 0.0004; R^2 = 0.742$$

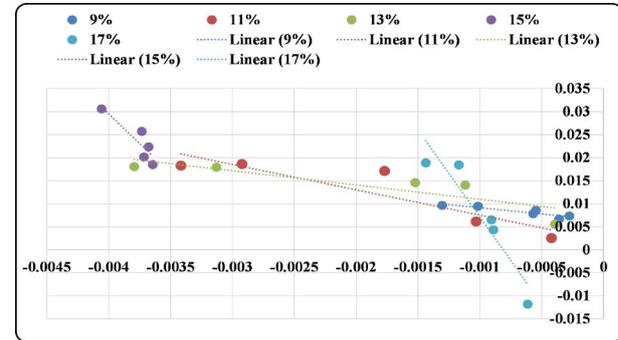
$$y_{11} = 0.1911x - 0.0004; R^2 = 0.8139$$

$$y_{13} = 0.3744x - 0.0012; R^2 = 0.8156$$

$$y_{15} = 0.2020x - 0.001; R^2 = 0.9138$$

$$y_{17} = -0.1274x + 0.0001; R^2 = 0.8555$$

a) free cells



$$y_9 = -2.691x + 0.0064; R^2 = 0.8303$$

$$y_{11} = -5.5013x + 0.002; R^2 = 0.8223$$

$$y_{13} = -3.0981x + 0.0079; R^2 = 0.7435$$

$$y_{15} = -25.81x + 0.0737; R^2 = 0.8003$$

$$y_{17} = -37.802x + 0.0306; R^2 = 0.8802$$

b) immobilized cells

on graph – by axis X - $\frac{\ln\left(\frac{P_m - P_j}{P_m - P_i}\right)}{\tau_j - \tau_i}$; by axis Y - $\frac{\ln\left(\frac{P_j}{P_i}\right)}{\tau_j - \tau_i}$

Figure 4: Linearization by Product Accumulation

The kinetic parameters determined by the linearization by product accumulation largely confirm the previously described trends. This indicates that ethanol accumulation in the medium, which is associated with primary yeast metabolism, is a paramount process during fermentation. It is interesting that in free cell fermentation the specific growth rates determined were

one order of magnitude higher than those determined by linearization by substrate consumption.

This proves again that the beer production process is substrate dependent and strongly depends on the order of consumption of carbohydrates. The accuracy of the linearization by product accumulation also depends on the inclusion of data mainly on the main fermentation process. This means that prolonged fermentation

processes using high-gravity wort led to the reduction of the model accuracy. One of the conclusions drawn from these observations and from previous studies (Kostov et al., 2019) is the following: if modeling the effect of the wort extract, it is good to use linearization by substrate consumption, but if comparing fermentation conditions (influence of temperature, pH, oxygen amount, etc.), it is good to apply linearization by product accumulation, but only on condition that the wort has the same qualitative and quantitative composition.

Table 2: Kinetic characteristics of biomass growth determined by Substrate consumption linearization

μ_{max}, h^{-1}	$K_S, g/dm^3$	μ_{max}, h^{-1}	$K_S, g/dm^3$
Free cells		Immobilized cells	
9%			
0.0001	0.96	0.00100	0.31
R ² =0.9921		R ² =0.9231	
$\eta=9.947$			
11%			
0.00041	40.87	0.0007	33.62
R ² =0.9236		R ² =0.8962	
$\eta=1.904$			
13%			
0.000267	43.73	0.0010	33.77
R ² =0.8932		R ² =0.8859	
$\eta=3.771$			
15%			
0.00121	75.95	0.0037	0.52
R ² =0.8631		R ² =0.214	
$\eta=3.080$			
17%			
0.00100	42.63	0.00101	37.55
R ² =0.8232		R ² =0.8621	
$\eta=1.098$			

* correlation coefficients between the experimental biomass concentration and the calculated biomass concentration were calculated using the kinetic parameters obtained.

Perhaps the most unexpected results are obtained by applying linearization by product accumulation to the immobilized cell fermentation process (Table 3). The specific growth rates obtained are between 7 and 527 times higher than those in the free cell fermentation. The results obtained more closely approximate the specific ethanol accumulation rate determined in Parcunev et al., 2012. Analogously, the data obtained for free cell fermentation also approximate those results. We can assume that actually the so-called specific growth rate is the specific rate of product accumulation. Furthermore, the analytical solution of the differential equation does not account for the diffusion of ethanol from the capsules to the fermentation medium. Probably the values obtained somehow include the diffusion coefficients of the product in the fermenting liquid, but this would be a subject to future studies.

The results in Table 1 to Table 3 show that some of the models were characterized by low accuracy of the description of the fermentation process. This was due to the inability of these models to evaluate the influence of some parameters, such as substrate and product diffusion to and from the yeast cells through the immobilization matrix, the effect of some of the

metabolic products on the yeast cell growth, etc. Usually, the change in a model accuracy is on the verge of whether the beer wort is considered highly extractive or not. Despite these problems, the models could be used to quickly and accurately determine kinetic parameters and to make comparisons between different fermentation regimes with both free and immobilized cells.

Table 3: Kinetic characteristics of biomass growth determined by Product accumulation linearization

μ_{max}, h^{-1}	$K_S, g/dm^3$	μ_{max}, h^{-1}	$K_S, g/dm^3$
Free cells		Immobilized cells	
9%			
0.0004	1.01	0.011	65.62
R ² =0.9556		R ² =0.9264	
$\eta=27.36$			
11%			
0.00049	25.99	0.0037	93.08
R ² =0.9432		R ² =0.8264	
$\eta=7.47$			
13%			
0.00186	71.36	0.0139	98.28
R ² =0.8531		R ² =0.8146	
$\eta=7.46$			
15%			
0.0011	25.23	0.147	144.41
R ² =0.8142		R ² =0.8631	
$\eta=123.85$			
17%			
0.00011	24.82	0.0604	165.61
R ² =0.9126		R ² =0.9326	
$\eta=527.14$			

* correlation coefficients between the experimental biomass concentration and the calculated biomass concentration were calculated using the kinetic parameters obtained.

Another important feature was that the models gave up to 29 times difference in the specific growth rate (Table 1 to Table 3). This was due to the fact that they described different sides of the fermentation process. The first model described the overall biomass growth, and the second and third one provided the connection between substrate consumption, ethanol accumulation, and biomass growth. This also implied differences in the specific growth rate obtained as the models took into account the biochemical side of the process. Finally, it should be noted that the models allowed clear distinction to be made between the processes that take place in the wide range of original extracts (Table 1 to Table 3). The data shows that beer wort can be divided into three conventional groups - with "low" extract (9%), with "normal extract" (11% -13%) and with "high" extract (15% -17%). The observed difference between the kinetic parameters was due to the different composition of fermentable sugars in each of the three beer wort groups. As a result, the accuracy of the models differed in the area of transition from one group to another.

CONCLUSION

The effect of the original wort extract on three analytical methods for determining the specific growth

rate of yeast biomass was investigated. The data show that as a substrate dependent process, the kinetics of the fermentation process is best described by linearization by substrate consumption. In this case, the model takes into account in a certain way the catabolic repression that is characteristic of alcohol fermentation in brewing. Besides, it is found that the model using linearization by product accumulation can detect not the specific growth rate of the microbial population but the specific rate of product accumulation. Moreover, it is found that the linearization by metabolic product accumulation also takes into account to some extent the product diffusion from and to the immobilized preparations.

It was also found that the immobilization process was irrelevant to the accuracy of the models, and the data obtained on the efficiency coefficients confirm previous observations that the strain used was not significantly affected by the immobilization procedure.

Lastly, the accuracy of the models was found to be highly dependent on the original wort extract, with the change in the accuracy being related to the transition zone between the wort with normal extracts and the high gravity wort (15% -17% extract).

REFERENCES

- Analytica EBC. 2004, Fachverlag Hans Carl, Nurnberg.
- Branyik, T.; A. Vicente; P. Dostalek; and J.A. Teixeira. 2005 "Continuous beer fermentation using immobilized yeast cell bioreactor systems", *Biotechnol. Prog.*, 21, 653-663.
- Boulton C. and D. Quain 2001. "Brewing yeast and fermentation." *Blackwell Science*, ISBN 0-632-05475-1.
- Godia, F., Casas, C., Sola, C., 1987. "A survey of continuous ethanol fermentation systems using immobilized cell". *Process Biochem.* 22, 43-48.
- Hayes, S.; Power, J.; Ryder, D. 1991. "Immobilized cell technology for brewing: a progress report Part 1: physiology of immobilized cells and the application to brewing". *Brewer's Digest* 9, 14-22.
- Kostov, G.; R. Denkova-Kostova; V. Shopska; B. Goranov. 2019. "Analytical approaches to determine the specific biomass growth rate in brewing" In: Iacono, M., Palmieri, F., Gribaudo, M., Ficco, M. (Editors), ECMS 2019 Proceedings, 125-131. doi: 10.7148/2019-0125
- Kunze W. 2003. "Technology of brewing and malting, 3rd edition", ISBN 3-921690-49-8, *VLB-Berlin*
- Mensour, N.; Margartitis, A.; Briens, C.L.; Pilkington, H.; Russell, I. 1996. "Application of immobilised cell fermentations". In: Wijfels, R.H., Buitelaar, R.M., Bucke, C., Tramper, J. (Eds.), *Immobilised Cells: Basics and Applications*. Elsevier Science, Amsterdam, pp. 661-671.
- Parcunev, I.; V. Naydenova; G. Kostov; Y. Yanakiev, Zh. Popova; M Kaneva; I. Ignatov, 2012. "Modelling of alcoholic fermentation in brewing – some practical approaches". In: Troitzsch, K. G, Möhring., M. and Lotzmann, U. (Editors), *Proceedings 26th European Conference on Modelling and Simulation*, ISBN: 978-0-9564944-4-3, pp. 434-440.
- Shopska, V.; R. Denkova; V. Lubenova; G. Kostov. 2019. "Kinetic characteristics of alcohol fermentation in brewing: state of art and control of fermentation process", In: Grumezescu, A. and Holban, A.-M., *Fermented beverages* ISBN: 9780-1281-5271-3, Woodhead Publishing, in press.
- Warpholomeew, S. D and K. G. Gurevich. 1999. "*Biokinetic – practical course*". ISBN: 5-8183-0050-1, Fair-Press, Moscow, pp.720 (in Russian)
- Willaert R., 2007." The beer brewing process: wort production and beer fermentation". In: Y.H. Hui (Editor) *Handbook of food products manufacturing*, John Wiley & Sons, Inc., Hoboken, New Jersey, 443-507

ACKNOWLEDGEMENTS

This research has been funded under Project No KII-06-M27/3 "Technological and microbiological approaches for the production of new types of low-alcohol and non-alcoholic drinks with increased biological value" of the National Science Fund, Bulgaria.

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is Associate 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. 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 Head Assistant 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 Head Assistant 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, identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is a researcher in the LBLact Company, 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.

KRISTINA IVANOVA is assistant professor at the department of "Food Technologies" at Food Research and Development Institute, Plovdiv. She received her MSc in "Food Safety - University of Food Technologies – Plovdiv" in 2014 and PhD in "Food technologies" in 2018. Her research interests are in the modeling of food technologies processes and use of food and beverages wastes for functional food development.

KINETICS OF MICROBIAL INACTIVATION OF HUMAN PATHOGENS BY BIOLOGICAL FACTORS

Georgi Kostov*, Rositsa Denkova-Kostova**, Vesela Shopska*, 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; zdenkova@abv.bg

Bogdan Goranov

LBLact, Plovdiv, Bulgaria, E-mail: goranov_chemistry@abv.bg

Desislava Teneva

Institute of organic chemistry, Bulgarian academy of sciences, Bulgaria, E-mail: desi_gerinska@yahoo.com

KEYWORDS

antimicrobial activity, *Lactobacillus*, *Staphylococcus aureus*, *Salmonella abony*, co-culturing, modelling, kinetics

ABSTRACT

The antimicrobial activity of various lactic acid bacteria is an important characteristic for their incorporation in the composition of probiotic preparations and functional foods. The purpose of the present work was to present a mathematical approach to determine the kinetics of antimicrobial action of probiotic lactic acid bacteria *Lactobacillus plantarum* BZ1, *Lactobacillus plantarum* BZ2 and *Lactobacillus plantarum* BZ3 when co-cultured with the pathogenic microorganisms *Staphylococcus aureus* ATCC 25093; *Staphylococcus aureus* ATCC 6538P; *Salmonella* sp. (clinical isolate), *Salmonella abony* NTCC 6017. The pathogen inactivation was achieved by the antagonistic action of the lactic acid bacteria strains, which is a biological factor of inactivation. Three kinetic models to reveal different sides of the antagonism between beneficial lactic acid bacteria and pathogenic microorganisms were used in the present work. Only probiotic strains with good antimicrobial activity against pathogenic microorganisms can be included in the composition of starters for functional foods and beverages and probiotic formulations so that upon consumption the selected lactobacilli strains could execute their inherent role to restore and maintain the microbial balance in the gastrointestinal tract.

INTRODUCTION

A. Theoretical foundations of the kinetics of dying of microorganisms

Mathematically, microbial dying follows one and the same relationship, regardless of the factors that lead to inactivation. Creating inactivation conditions does not lead to the immediate death of the whole cell population. The cells to be destroyed are reduced in number in time under the action of the respective factor. The factors can be chemical, physical and biological. The action of chemical factors (various preservatives and disinfectants) and physical factors (mainly heat generated by various means) is at the heart of sterilization processes in the microbiological industry (Chen et al., 2013).

The biological factors causing a decrease in the number of a group of microorganisms are due to the antagonistic action of beneficial over harmful microorganisms and is expressed in the competitive absorption of the substrate and the production of organic acids, bacteriocins and BLIS and other components causing the inhibitory action against the pathogenic microflora (Denkova et al., 2017).

Microorganisms do not die simultaneously after a certain effect of the inhibitory factor, but by gradually reducing the number of surviving microorganism cells due to their different resistance. If the microbial culture is homogeneous, then the dying rate related to the number of living microorganisms is constant (Stanbury et al., 2003):

$$-\frac{dX}{d\tau} = kX \quad (1)$$

where: X is the concentration of viable microorganisms (spores of the inactivated microorganism) at the moment τ ; k - specific dying rate of the microorganisms, s^{-1} .

Integrating this equation within the limits from N_0 to N and from 0 to τ , the following equation is obtained:

$$\ln \frac{X}{X_0} = -k\tau \quad (2)$$

where: X_0 - concentration of viable microorganisms to be inactivated in the fermentation volume

This equation in coordinate's $\ln X$ - τ is a straight line (Fig. 1). The rate constant is the angular coefficient of the straight line with a negative sign. It is independent of the microorganisms' concentration $\ln X_0$ and the duration of the process and is numerically equal to the proportion of organisms dying per unit time. The physical meaning of the parameter that is inverse of the constant, is the average life span of the individual microorganism during the dying period and characterizes its resistance to the inhibitory factor.

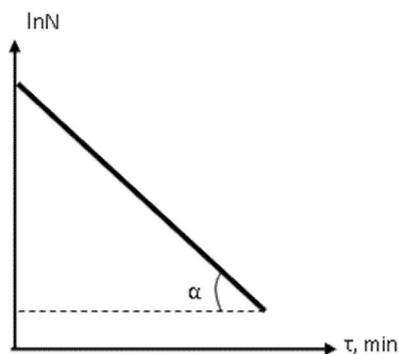


Figure 1: Graph of the kinetics of dying of microorganisms

When studying the influence of various parameters (temperature, pH, etc.) on the destruction of microorganisms, the target function is the rate constant, which characterizes the behavior of microorganisms with average properties, not the number of dead microorganisms.

The specific dying rate is a characteristic of the individual microbial species. Physical and chemical effects (Stanbury et al., 2003) have great influence on the constant in addition to the nature of the organism and the conditions for culture growth.

B. Mathematical models for describing the kinetics of dying of pathogenic microorganisms

The following three models were used to model the kinetics of dying of pathogenic microorganisms in the presence of a biological factor:

$$\frac{dX}{d\tau} = \mu X - \beta X^2 \quad (3)$$

$$\frac{dX}{d\tau} = (\mu X - \beta X^2)^n \quad (4)$$

$$\frac{dX}{d\tau} = -kX \quad (5)$$

The logistic curve equation (equation 3) describes in general terms the growth of a microbial population in a limited volume. It expresses the effect of the increasing biomass concentration on the maximum specific growth rate. The model has two parameters - the maximum specific growth rate μ and the internal population competition coefficient β . The coefficient β characterizes the effect of the interaction of cells in the microbial population on one another as a result of substrate deficiency and the inhibitory effect of the accumulating metabolic products and shows both the amount of cells killed per unit volume of culture medium per unit time and the inhibition degree of the potential maximum growth rate of the microbial population. This parameter indirectly indicates the influence of the growth conditions on the microbial population. The modified logistic curve model (equation 4) contains the parameter n , which shows the influence of the culture medium composition (local substrate concentrations and metabolic products) on the microbial population. The parameter n indicates the sensitivity (resistance) of the pathogenic cells to the presence of lactobacilli and the acids and antimicrobial substances produced by the lactobacilli, as well as the sensitivity (resistance) of the lactobacilli cells to the presence of pathogens and their metabolites. Equation (5) is used to describe the kinetics of pathogen cell death. It describes first-order kinetics of chemical reactions. The models presented are generally accepted to describe the kinetics of microbial growth and the inactivation of the microbial population by physical, biological and chemical factors (Denkova et al., 2017; Stanbury et al., 2003)

C. Antimicrobial activity of lactic acid bacteria

Probiotics are „live microorganisms which when administered in adequate amounts confer a health benefit on the host“. Lactic acid bacteria are the major bacterial species used for the production of probiotics and probiotic foods. They are traditional cultures in the production of fermented foods. Probiotic microorganisms contribute to the restoration of the intestinal balance, play an important role in maintaining health and improve the quality of certain foods with

their inclusion (Charalampopoulos et al., 2002; Charalampopoulos et al., 2003; Stanton et al., 2005; Siro et al., 2008; López de Lacey et al., 2014; Soccol et al., 2010; Kociubinski and Salminen, 2006, Denkova-Kostova et al., 2018).

The suppression of conditionally pathogenic, carcinogenic and pathogenic microorganisms is associated with the inactivation of their enzyme systems, inhibition of their adhesion and growth by expelling them from the gastrointestinal tract and normalizing the gastrointestinal microflora. The antimicrobial activity of lactic acid bacteria is mainly related to the production of lactic acid and acetic acid but also to the production of propionic acid, sorbic acid, benzoic acid, hydrogen peroxide, diacetyl, ethanol, phenolic and protein compounds as well as bacteriocins. The produced organic acids alter the medium pH and inhibit the growth of putrefactive, pathogenic and toxigenic microorganisms, while antibacterial substances of peptide nature (bacteriocins) act directly on the microbial cells (Dalié et al., 2010; Eswaranandam et al., 2004; Denkova-Kostova et al., 2018).

The purpose of the present work was to study the kinetics of dying of pathogenic microorganisms when co-cultured with lactic acid bacteria. Three mathematical dependencies, which reveal different sides of the process of pathogen inactivation in the presence of the biological factor - the lactic acid bacteria cells, were used to accomplish this purpose.

MATERIAL AND METHODS

A. Microorganisms

- The research was carried out with 3 *Lactobacillus plantarum* strains, isolated from spontaneously fermented vegetables/fruits - *Lactobacillus plantarum* BZ1, *Lactobacillus plantarum* BZ2, *Lactobacillus plantarum* BZ3
- Pathogenic microorganisms: *Staphylococcus aureus* ATCC 25093; *Staphylococcus aureus* ATCC 6538P; *Salmonella* sp. (clinical isolate), *Salmonella abony* NTCC 6017.

B. Growth media:

- LAPTg10-broth (g/dm³): peptone - 15; yeast extract - 10; tryptone - 10; glucose - 10. pH was adjusted to 6.6-6.8 and Tween 80 - 1cm³/dm³ was added. Sterilization - 20 minutes at 121 °C.
- LAPTg10-agar (g/dm³): peptone - 15; yeast extract - 10; tryptone - 10; glucose - 10. pH was adjusted to 6.6-6.8 and Tween 80 - 1cm³/dm³ and agar - 20 g were added. Sterilization - 20 minutes at 121 °C.
- LBG-agar (g/dm³): tryptone - 10, yeast extract - 5, NaCl - 10, glucose - 10, agar - 20. Sterilization - 20 minutes at 121 °C.

C. Determination of the antimicrobial activity of *Lactobacillus plantarum* strains against pathogenic microorganisms - by co-cultivation

To determine the antimicrobial activity of the studied *Lactobacillus plantarum* strains against the test-pathogenic microorganisms, the following variants were prepared:

Variant	LAPTg10-broth	<i>Lactobacillus plantarum</i> cm ³	Pathogen
LAB C	9.5	0.5	-
Pathogen C	9.5	-	0.5
Mixture	9.0	0.5	0.5

Co-cultivation of each *Lactobacillus plantarum* strain and each pathogen under static conditions in a thermostat at 37±1°C for 60 to 72 hours, taking samples at 0, 12, 24, 36, 48, 60 and 72 h and monitoring the changes in the titratable acidity and the concentration of viable cells of both the pathogen and the *Lactobacillus plantarum* strain, was performed. The number of viable cells was determined through appropriate tenfold dilutions of the samples and spread plating on LBG-agar medium (to determine the number of viable pathogen cells) and on LAPTg10 – agar medium (to determine the number of viable *Lactobacillus plantarum* cells). The Petri dishes were cultured for 72 hours at 37±1°C until the appearance of countable single colonies. The titratable acidity was determined after sterilization of the samples (to kill the pathogen) using 0.1N NaOH. 5 cm³ of each sample were mixed with 10 cm³ dH₂O and titrated with 0.1N NaOH, using phenolphthalein as an indicator, until the appearance of pale pink colour, which retained for 1 minute. The value for the titratable acidity was obtained by multiplying the millilitres 0.1N NaOH by the factor of the 0.1N NaOH and the number 20. Bacterial counts were transformed to log values. Results are shown as the average values and standard deviations obtained from three independent experiments (Denkova et al., 2013).

D. Modeling of antimicrobial activity and determination of process kinetic parameters

Equations (3) to (5) have been used to model the process of inactivation of the microbial population of the pathogenic microorganisms. The modeling was performed in an Excel environment, and the accuracy of the models was determined based on the algorithms contained in the software.

RESULTS AND DISCUSSION

Figures 2 to 4 show the growth dynamics of one of the lactobacilli strains (*Lactobacillus plantarum* BZ1) when co-cultured with each of the pathogenic strains. The rest of the figures are of a similar nature and are therefore not presented in the present publication. The results of the identification of the kinetic parameters of the three models are presented in Table 1.

The kinetic parameters presented in the table show that both lactobacilli and pathogens cultured as pure cultures had relatively high maximum specific growth rates. In co-cultivation, the maximum specific growth rates of both lactobacilli and pathogens were reduced. It is noteworthy that the values of the coefficients of internal population competition predicted by the logistic curve model were comparable both in the separate cultivation of the studied strains and in their co-cultivation. The value of β varied from 0.0049 to 0.02833 cfu/cm³.h. Therefore, for a more detailed study of the growth kinetics of the studied strains and the antimicrobial

activity, a modified logistic curve model containing a power factor n reflecting the effect of the medium composition and the released metabolic products of the tested microorganisms (lactobacilli and pathogens) on the cell growth, was used. The values of the maximum specific growth rates predicted by the two logistic curve models were very close in the separate cultivation of *Lactobacillus plantarum* BZ1, *Lactobacillus plantarum* BZ2 and *Lactobacillus plantarum* BZ3. μ of the strains tested varied in the range of 0.112 to 0.112 h⁻¹ according to the classical logistic curve model (model 1); and in the range of 0.100 to 0.103 h⁻¹ according to the modified logistic curve model (model 2). Similar values were also observed for the coefficient of internal population competition (β), which, according to models 1 and 2, varied in the range from 0.0078 to 0.0093 cfu/cm³.h. The parameter n ranged from 0.8850 to 0.8874. These values indicate that the influence of the medium and the accumulating metabolites (mainly lactic acid) had little effect on the growth of the strains. This was also supported by the high values of the maximum biomass concentration for the three strains predicted by the models (X_k varied from 12.86 to 13.11 log N).

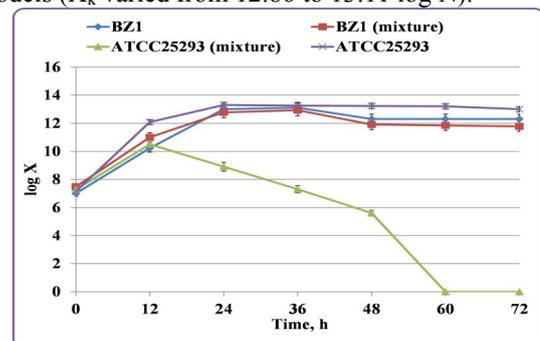


Figure 2: Dynamics of growth of *Lactobacillus plantarum* BZ1 and *Staphylococcus aureus* ATCC 25093 in separate cultivation and in co-cultivation

In separate cultivation, the strains *Staphylococcus aureus* ATCC 25093 and *Staphylococcus aureus* ATCC 6538P were characterized by relatively high values of μ . According to the classical logistic curve model, *Staphylococcus aureus* ATCC 6538P had higher growth rate (0.359 h⁻¹) than *Staphylococcus aureus* ATCC 25093 ($\mu = 0.144$ h⁻¹), whereas, according to Model 2, both strains had close maximum growth rates of 0.105 h⁻¹ and 0.104 h⁻¹, respectively. The same trend was observed in the values of the coefficient of internal population competition (0.0078 cfu/cm³.h for *Staphylococcus aureus* ATCC 6538P and 0.0076 cfu/cm³.h for *Staphylococcus aureus* ATCC 25093), while higher values of β (0.0283 cfu/cm³.h and 0.0107 cfu/cm³.h, respectively) were observed in the classical logistic curve model. Both logistic curve models predicted high concentration of active staphylococci cells varying in the range of 13.30 log N to 13.70 log N. In both *Staphylococcus aureus* strains cultivated individually, the values of the parameter n were less than 1 (0.8779 and 0.8774, respectively). Relatively high maximum specific growth rates were also observed in the separate cultivation of *Salmonella*

abony ATCC 6017 and *Salmonella* sp. As a classic model, the logistic curve model predicted higher maximum growth rate for *Salmonella abony* ATCC 6017 (0.295 h⁻¹) than for *Salmonella* sp. (0.119 h⁻¹). A similar trend was observed in the values of β , which was 0.0252 cfu/cm³.h for *Salmonella abony* ATCC 6017 and 0.0097 cfu/cm³.h for *Salmonella* sp. According to the modified logistic curve model (model 2), the two pathogens were characterized by close maximum specific growth rates (0.113 h⁻¹ for *Salmonella abony* ATCC 6017 and 0.111 h⁻¹ for *Salmonella* sp.). The same trend was observed for the values of the parameter β (0.0095 cfu/cm³.h and 0.0092 cfu/cm³.h, respectively). For *Salmonella* sp. both logistic curve models predicted higher maximum concentrations of active cells (12.21

log N by the classical logistic curve model and 12.10 log N by the modified logistic curve model) compared to *Salmonella abony* ATCC 6017 (11.70 log N and 11.82 log N, respectively). The co-cultivation of *Lactobacillus plantarum* BZ1 and *Staphylococcus aureus* ATCC 25093 (Figure 1), *Staphylococcus aureus* ATCC 6538P, *Salmonella abony* ATCC 6017 and *Salmonella* sp. showed a slight decrease in the maximum specific growth rate of *Lactobacillus plantarum* BZ1, which according to the two logistic curve models ranged from 0.074 h⁻¹ to 0.098 h⁻¹ (μ ranged between 0.101 h⁻¹ and 0.121 h⁻¹ in separate cultivation). Comparable values were also observed for β , which varied in the range of 0.0022 cfu/cm³.h to 0.0077 cfu/cm³.h.

Table 1: Kinetic characteristics of pathogen inactivation upon co-cultivation with *Lactobacillus plantarum* strains

Variant	Kinetic parameters							
	Equation 3			Equation 4				Equation 5
	μ h ⁻¹	β cfu/cm ³ .h	X _k cfu/cm ³	μ h ⁻¹	β cfu/cm ³ .h	X _k cfu/cm ³	n	k _i h ⁻¹
<i>L. plantarum</i> BZ1 control	0.121	0.0093	13.11	0.101	0.0078	13.00	0.8850	-
<i>L. plantarum</i> BZ2 control	0.114	0.0088	12.94	0.100	0.0078	12.86	0.8866	-
<i>L. plantarum</i> BZ3 control	0.112	0.0085	13.07	0.103	0.0079	12.95	0.8874	-
<i>St. aureus</i> ATCC 25093 control	0.144	0.0107	13.52	0.104	0.0076	13.70	0.8779	-
<i>St. aureus</i> ATCC 6538P control	0.359	0.0283	13.30	0.105	0.0078	13.35	0.8774	-
<i>S. abony</i> ATCC 6017 control	0.295	0.0252	11.70	0.113	0.0095	11.82	0.8255	-
<i>Salmonella</i> sp. control	0.119	0.0097	12.21	0.111	0.0092	12.10	0.8841	-
<i>L. plantarum</i> BZ1+<i>St. aureus</i> ATCC 25093								
<i>L. plantarum</i> BZ1 (in mixture)	0.087	0.0060	12.95	0.074	0.0057	12.92	0.9056	-
<i>St. aureus</i> ATCC 25093 (in mixture)	0.078	0.0067	11.55	0.119	0.0111	10.43	1.1232	0.313
<i>L. plantarum</i> BZ2 + <i>St. aureus</i> ATCC 25093								
<i>L. plantarum</i> BZ2 (in mixture)	0.092	0.0070	12.78	0.084	0.0066	12.75	0.9490	-
<i>St. aureus</i> ATCC 25093 (in mixture)	0.088	0.0080	11.04	0.086	0.0065	13.20	1.1948	0.318
<i>L. plantarum</i> BZ3 + <i>St. aureus</i> ATCC 25093								
<i>L. plantarum</i> BZ3 (in mixture)	0.090	0.0071	12.72	0.084	0.0066	12.66	0.9538	-
<i>St. aureus</i> ATCC 25093 (in mixture)	0.081	0.0072	11.31	0.097	0.0066	10.51	1.2000	0.325
<i>L. plantarum</i> BZ1 + <i>St. aureus</i> ATCC 6538 P								
<i>L. plantarum</i> BZ1 (in mixture)	0.092	0.0072	12.78	0.080	0.0066	12.74	0.9137	-
<i>St. aureus</i> ATCC 6538 P (in mixture)	0.063	0.0049	12.80	0.022	0.0019	10.41	1.3523	0.317
<i>L. plantarum</i> BZ2 + <i>St. aureus</i> ATCC 6538 P								
<i>L. plantarum</i> BZ2 (in mixture)	0.087	0.0067	12.96	0.085	0.0065	12.91	0.9932	-
<i>St. aureus</i> ATCC 6538 P (in mixture)	0.074	0.0065	11.56	0.022	0.0021	10.48	1.5623	0.307
<i>L. plantarum</i> BZ3 + <i>St. aureus</i> ATCC 6538 P								
<i>L. plantarum</i> BZ3 (in mixture)	0.079	0.0062	12.87	0.084	0.0067	12.52	0.9069	-
<i>St. aureus</i> ATCC 6538 P (in mixture)	0.077	0.0065	11.83	0.019	0.0019	10.50	1.7424	0.307
<i>L. plantarum</i> BZ1 + <i>S. abony</i> ATCC 6017								
<i>L. plantarum</i> BZ1 (in mixture)	0.082	0.0060	13.29	0.081	0.0022	13.30	0.9075	-
<i>S. abony</i> ATCC 6017 (in mixture)	0.107	0.0090	11.84	0.115	0.0102	11.27	0.9032	0.449
<i>L. plantarum</i> BZ2 + <i>S. abony</i> ATCC 6017								
<i>L. plantarum</i> BZ2 (in mixture)	0.099	0.0076	13.02	0.098	0.0022	12.84	0.9088	-
<i>S. abony</i> ATCC 6017 (in mixture)	0.099	0.0084	11.70	0.115	0.0101	11.05	0.9976	0.462
<i>L. plantarum</i> BZ3 + <i>S. abony</i> ATCC 6017								
<i>L. plantarum</i> BZ3 (in mixture)	0.080	0.0060	13.30	0.081	0.0022	13.50	0.9054	-
<i>S. abony</i> ATCC 6017 (in mixture)	0.098	0.0085	11.55	0.098	0.0086	11.43	0.9753	0.462
<i>L. plantarum</i> BZ1 + <i>Salmonella</i> sp.								
<i>L. plantarum</i> BZ1 (in mixture)	0.095	0.0077	12.95	0.098	0.0077	13.17	0.8909	-
<i>Salmonella</i> sp. (in mixture)	-	-	-	-	-	-	-	0.587
<i>L. plantarum</i> BZ2 + <i>Salmonella</i> sp.								
<i>L. plantarum</i> BZ2 (in mixture)	0.089	0.0068	13.15	0.076	0.0057	13.23	1.1179	-
<i>Salmonella</i> sp. (in mixture)	-	-	-	-	-	-	-	0.628
<i>L. plantarum</i> BZ3 + <i>Salmonella</i> sp.								
<i>L. plantarum</i> BZ3 (in mixture)	0.089	0.0069	12.89	0.076	0.0056	13.56	1.0928	-
<i>Salmonella</i> sp. (in mixture)	-	-	-	-	-	-	-	0.394

A slight increase in the parameter n , which varied in the range from 0.8909 to 0.9137, were observed in the co-cultivation of *Lactobacillus plantarum* BZ1 and the pathogens studied. This indicates that *Lactobacillus*

plantarum BZ1 was very slightly affected by the presence of the pathogens and the metabolites secreted during their growth. The high values of the maximum active cell concentration of *Lactobacillus plantarum*

BZ1 predicted by both models also serve as a confirmation of this conclusion. The value for X_k was close to that of the control (separate cultivation of the strain) - from 12.74 log N to 13.30 log N. In the co-cultivation of *Staphylococcus aureus* ATCC 25093 or *Staphylococcus aureus* ATCC 6538P and *Lactobacillus plantarum* BZ1, a reduction in the maximum specific growth rate of the pathogens, especially for *Staphylococcus aureus* ATCC 6538P, in which μ decreased from 0.359 h⁻¹ to 0.063 h⁻¹, according to the classical logistic curve model, and to 0.019 h⁻¹, according to model 2, was observed. According to both models for this strain, β ranged from 0.0019 cfu/cm³.h to 0.0077cfu/cm³.h.

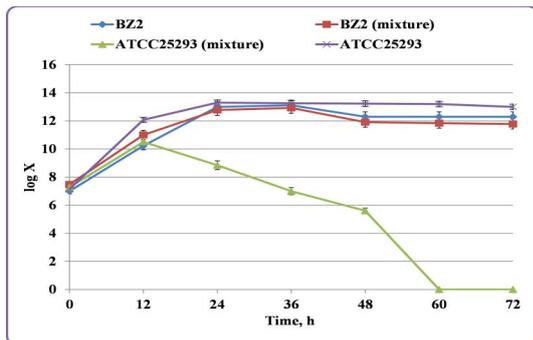


Figure 3: Dynamics of growth of *Lactobacillus plantarum* BZ2 and *Staphylococcus aureus* ATCC 25093 in separate cultivation and in co-cultivation

In *Staphylococcus aureus* ATCC 25093, a reduction in the maximum specific growth rate to 0.078 h⁻¹ and 0.119 h⁻¹ and in the internal population competition β to 0.0067 cfu/cm³.h and 0.0111 cfu/cm³.h, according to the two logistic curve models used was observed. What is striking are the high values of the parameter n , which was 1.1232 for *Staphylococcus aureus* ATCC 25093 and 1.3523 for *Staphylococcus aureus* ATCC 6538P. This indicated that the pathogenic microorganisms were strongly influenced by the presence of the lactobacilli and the released substances with antimicrobial activity (organic acids, bacteriocins, etc.). This was also confirmed by the fact that, according to the mathematical models, both pathogens were characterized by a significantly lower maximum concentration of active pathogen cells in the mixed population, which varied in the range from 10.41 log N to 12.80 log N for *Staphylococcus aureus* ATCC 6538P and from 10.43 log N to 11.55 log N for *Staphylococcus aureus* ATCC 25093. In the separate cultivation both pathogens showed a maximum final concentration of active cells in the range from 13.52 log N to 13.70 log N for *Staphylococcus aureus* ATCC 25093 and from 13.30 log N to 13.35 log N for *Staphylococcus aureus* ATCC 6538P. Comparable values of the dying rate constant were observed in the conducted modelling of the kinetics of dying of the pathogenic strains of *Staphylococcus aureus* - 0.313 h⁻¹ for *Staphylococcus aureus* ATCC 25093 and 0.317 h⁻¹ for *Staphylococcus aureus* ATCC 6538P.

The co-cultivation of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ1 resulted in a

reduction in the maximum specific growth rate of the pathogen, but to a lesser extent than that of *Staphylococcus aureus*. In this strain, μ varied in the range from 0.107 h⁻¹ and 0.115 h⁻¹, with the parameter β varying from 0.0090 cfu/cm³.h to 0.0102 cfu/cm³.h. A lower value of the parameter n ($n=0.9032$) was also observed in this strain compared to the two representatives of *Staphylococcus aureus*. This indicated that *Salmonella abony* ATCC 6017 would exhibit resistance to the presence of lactobacilli and their metabolites in comparison with the two strains of *Staphylococcus aureus*. This was further confirmed by the fact that the values of the maximum active cell concentration of *Salmonella abony* ATCC 6017 in the mixed population were commensurable with that of the control (pathogen separate cultivation), namely 11.84 log N and 11.27 log N. However, the rate constant of dying of the pathogen was 0.449 h⁻¹ and it was higher than that of *Staphylococcus aureus*.

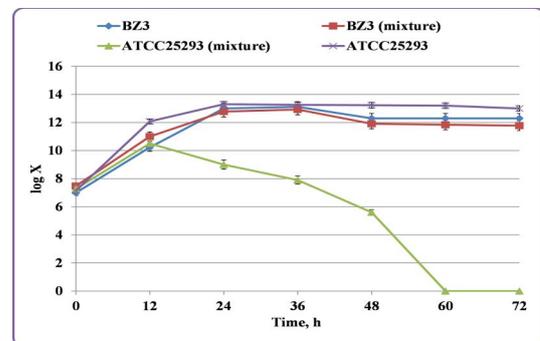


Figure 4: Dynamics of growth of *Lactobacillus plantarum* BZ3 and *Staphylococcus aureus* ATCC 25093 in separate cultivation and in co-cultivation

The co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ1 resulted in a complete reduction of the maximum specific growth rate compared to that in the separate cultivation of the pathogen alone. Since the beginning of co-cultivation, there had been continuous death of the pathogen cells. The rate constant of dying was 0.587 h⁻¹ in the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ1 and it was the highest compared to that of the other pathogenic microorganisms.

The co-cultivation of *Lactobacillus plantarum* BZ2 (Figure 3) with the pathogens examined showed a similar trend as in the previous strain *Lactobacillus plantarum* BZ1. A slight reduction in the maximum specific growth rate was observed, which varied from 0.087 h⁻¹ to 0.099 h⁻¹, and β ranged from 0.0067 cfu/cm³.h to 0.0076 cfu/cm³.h, according to model 1 and, μ varied from 0.076 h⁻¹ to 0.098 h⁻¹, and β ranged from 0.0022 cfu/cm³.h to 0.0066 cfu/cm³.h, according to model 2. Again, a slight increase in the parameter n was observed in this strain, whose values ranged from 0.9088 to 1.1179. This indicated that this strain was also poorly affected by the presence of the studied pathogens and their metabolites. As a confirmation of this conclusion was the high value of the maximum concentration of active cells - X_k varied in the range from 12.78 log N to 13.15 log N according to model 1

and from 12.75 log N to 13.23 log N according to model 2 and these values were close to those of the control.

The co-cultivation of *Staphylococcus aureus* ATCC 25093 or *Staphylococcus aureus* ATCC 6538P and *Lactobacillus plantarum* BZ2 resulted in a reduction in the pathogen maximum specific growth rate. μ for *Staphylococcus aureus* ATCC 25093 changed from 0.086 h⁻¹ to 0.088 h⁻¹, and β ranged from 0.0065 cfu/cm³.h to 0.0080 cfu/cm³.h, according to the mathematical models used. A maximum reduction in the maximum specific growth rate of *Staphylococcus aureus* ATCC 6538P - between 0.022 h⁻¹ and 0.074 h⁻¹ was observed, while β varied between 0.0021 cfu/cm³.h and 0.0065 cfu/cm³.h. The parameter n had a higher value ($n=1.5623$) in the co-cultivation of *Staphylococcus aureus* ATCC 6538P and *Lactobacillus plantarum* BZ2 compared to *Staphylococcus aureus* ATCC 25093 ($n=1.1948$). This indicated that *Staphylococcus aureus* ATCC 6538P was more strongly influenced by the presence of *Lactobacillus plantarum* BZ2 and the secreted metabolites with antimicrobial activity, which was also evidenced by the lower maximum growth rates of this strain in the mixed population. The higher sensitivity of this pathogen to lactic acid bacteria was also confirmed by the lower values of the maximum concentration of active cells in the mixed population for *Staphylococcus aureus* ATCC 6538P, which, according to the models, varies from 10.48 log N and 11.56 log N, compared with that of *Staphylococcus aureus* ATCC 25093, which, according to the mathematical models, varied in the range from 11.04 log N to 13.20 log N.

Staphylococcus aureus ATCC 25093 or *Staphylococcus aureus* ATCC 6538P cultured in a mixed population with *Lactobacillus plantarum* BZ2 were characterized by compatible and relatively high dying rates - 0.318 h⁻¹ for *Staphylococcus aureus* ATCC and 0.307 h⁻¹ for *Staphylococcus aureus* ATCC 6538P.

The co-cultivation of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ2 resulted in a reduction in the maximum specific growth rate of the pathogen, once again to a lesser extent than that of the two *Staphylococcus aureus* strains. In the co-cultivation of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ2, μ for *Salmonella abony* ATCC 6017 changed between 0.099 h⁻¹ and 0.115 h⁻¹, and β ranged from 0.0084 cfu/cm³.h to 0.0101 cfu/cm³.h. The parameter n value was lower ($n=0.9776$) compared to the same parameter in the co-cultivation of *Lactobacillus plantarum* BZ2 with the two representatives of *Staphylococcus aureus*, indicating resistance of the pathogen to the presence of *Lactobacillus plantarum* BZ2 and its metabolic products. To support this, the maximum concentration of pathogen active cells in the mixed population was 11.05 log N and 11.70 log N, which was close to that of the control (separate cultivation of *Salmonella abony* ATCC 6017) - 11.70 log N and 11.82 log N. Nevertheless, the dying rate of *Salmonella abony* ATCC 6017 was significantly higher than that of

Staphylococcus aureus ATCC 25093 and *Staphylococcus aureus* ATCC 6538P. For *Salmonella abony* ATCC 6017, the dying constant was 0.462 h⁻¹.

A complete reduction of the maximum specific growth rate in comparison with the separate cultivation of *Salmonella* sp. alone was observed in the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ2. From the beginning of the co-cultivation, there had been determined continuous death of the pathogen cells. In the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ2, the dying rate constant for *Salmonella* sp. was the highest (0.628 h⁻¹) compared to the same parameter for the other pathogens. This value of the dying rate constant was higher but close to the dying rate constant value of *Salmonella* sp. in the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ1 (0.587 h⁻¹). This in turn indicated that *Salmonella* sp. was more sensitive to the presence of *Lactobacillus plantarum* BZ2 and its metabolites secreted in the medium.

In co-cultivation of *Lactobacillus plantarum* BZ3 with the pathogens examined, a slight reduction in the maximum specific growth rate of *Lactobacillus plantarum* BZ3 was observed, varying from 0.076 h⁻¹ to 0.089 h⁻¹, and β ranging from 0.0060 cfu/cm³.h to 0.0071 cfu/cm³.h according to model 1; μ varied from 0.081 h⁻¹ to 0.084 h⁻¹, and β ranged from 0.0022 cfu/cm³.h to 0.0066 cfu/cm³.h according to model 2. Once again, a slight increase in the parameter n for *Staphylococcus aureus* ATCC 25093, *Staphylococcus aureus* ATCC 6538P and *Salmonella abony* ATCC 6017 was observed. Its values were 0.9538, 0.9069 and 0.9054, respectively, which indicated that *Lactobacillus plantarum* BZ3 was also affected by the presence of these pathogenic strains and their metabolites. This was evidenced by the high values of the maximum concentration of active lactobacilli cells in the mixed population, which varied for the respective pathogens - between 12.72 log N and 12.66 log N in the co-cultivation with *Staphylococcus aureus* ATCC 25093; between 12.87 log N and 12.52 log N in the co-cultivation with *Staphylococcus aureus* ATCC 6538P; between 13.30 log N and 13.50 log N in the co-cultivation with *Salmonella abony* ATCC 6017. These values were close to those of the control (*Lactobacillus plantarum* BZ3 cultivated alone). The co-cultivation of *Lactobacillus plantarum* BZ3 and *Salmonella* sp. resulted in a higher value of n ($n=1.0918$), compared to the other *Lactobacillus plantarum* strains tested. However, *Lactobacillus plantarum* BZ3 also achieved high maximum final concentration of active cells in the mixed population of 12.89 log N and 13.56 log N, indicating a negligible effect of the pathogen and its metabolites on the lactobacilli cells.

In the co-cultivation of *Staphylococcus aureus* ATCC 25093 or *Staphylococcus aureus* ATCC 6538P and *Lactobacillus plantarum* BZ3, a reduction in the maximum specific growth rate of the pathogens was observed. For *Staphylococcus aureus* ATCC 25093 μ varied from 0.081 h⁻¹ to 0.097 h⁻¹, and β ranged from

0.0066 cfu/cm³.h to 0.0072 cfu/cm³.h, according to the mathematical models used. Again, *Staphylococcus aureus* ATCC 6538P showed greater reduction in the maximum specific growth rate - between 0.019 h⁻¹ and 0.077 h⁻¹, while β varied between 0.0019 cfu/cm³.h and 0.0065 cfu/cm³.h. In the co-cultivation of *Staphylococcus aureus* ATCC 6538P and *Lactobacillus plantarum* BZ3, the parameter n had higher value ($n=1.7424$) compared to the co-cultivation of the same lactobacilli strain and *Staphylococcus aureus* ATCC 25093 ($n=1.2000$). The value of n in *Staphylococcus aureus* ATCC 6538P was the highest compared to the values in co-cultivation of the same pathogenic strain with the other lactobacilli strains, indicating that this pathogen was most sensitive to the presence of *Lactobacillus plantarum* BZ3, compared to the other *Lactobacillus plantarum* strains. The same trend was observed for *Staphylococcus aureus* ATCC 25093. The high impact of *Lactobacillus plantarum* BZ3 on these two pathogens can also be seen in the significantly lower values of the maximum final concentrations of active pathogen cells in the mixed populations, compared to the controls. The maximum final active cell concentration varied from 10.51 log N to 11.31 log N for *Staphylococcus aureus* ATCC 25093 and from 11.38 log N to 10.50 log N for *Staphylococcus aureus* ATCC 6538P.

Staphylococcus aureus ATCC 25093 and *Staphylococcus aureus* ATCC 6538P co-cultured with *Lactobacillus plantarum* BZ3 were again characterized by consistent and relatively high dying rates - 0.325 h⁻¹ for *Staphylococcus aureus* ATCC 25093 and 0.307 h⁻¹ for *Staphylococcus aureus* ATCC 6538P.

In the co-cultivation of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ3, a reduction in the maximum specific growth rate of the pathogen was observed, with both models predicting an equal reduction in the maximum specific growth rate to 0.098 h⁻¹, as well as close β values of 0.0085 cfu/cm³.h and 0.0086 cfu/cm³.h. The parameter n (0.9753) was lower in the co-cultivation of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ3 than in the co-culturing of the same lactobacilli strain and the representatives of *Staphylococcus aureus*, indicating resistance of *Salmonella abony* ATCC 6017 to the presence of *Lactobacillus plantarum* BZ3 and its metabolites. The maximum concentration of pathogen active cells in the mixed population can serve as evidence - 11.43 log N and 11.55 log N, which was close to the values of the control (separate cultivation of *Salmonella abony* ATCC 6017 alone) - 11.70 log N and 11.82 log N.

Salmonella abony ATCC 6017 dying rate in the co-culturing of *Salmonella abony* ATCC 6017 and *Lactobacillus plantarum* BZ3 was significantly higher than that of *Staphylococcus aureus* ATCC 25093 and *Staphylococcus aureus* ATCC 6538P. The *Salmonella abony* ATCC 6017 dying rate value was equal to that in the co-cultivation of the same pathogen with *Lactobacillus plantarum* BZ2 - 0.462 h⁻¹.

A complete reduction of the maximum specific growth rate of *Salmonella* sp. in the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ3 compared to the cultivation of the pathogen alone was observed. From the beginning of the co-cultivation, there had been continuous dying of the pathogen cells. In the co-cultivation of *Salmonella* sp. and *Lactobacillus plantarum* BZ3, a lower value of the dying rate constant (0.394 h⁻¹) compared to the co-cultivation of this pathogen with the other lactobacilli strains was observed. This lower value of the dying rate constant indicated that *Salmonella* sp. was resistant to the presence of *Lactobacillus plantarum* BZ3 and its metabolites.

The models used had high accuracy, ranging from 0.85 to 0.99 (evaluated by the R²-value). They were distinguished by their simple and high appreciation of the inactivation of the pathogens. The data in Table 1 show that the three strains tested had similar values with respect to the kinetic parameters of pathogen inactivation. This was due to the fact that they had been isolated from similar sources, suggesting similarities in their specific metabolism, including the principles and mechanisms of inactivation of pathogenic microorganisms.

CONCLUSION

The antimicrobial activity of lactic acid bacteria against pathogens is a paramount prerequisite for their selection for inclusion in the composition of probiotic preparations and different functional foods. The kinetics of the antimicrobial activity of three *Lactobacillus plantarum* strains against 2 *Staphylococcus aureus* strains and 2 *Salmonella* strains was determined using 3 kinetic models. The classical logistic curve equation and the modified logistic curve equation revealed different sides of the antagonism between beneficial *Lactobacillus plantarum* strains and the pathogenic microorganisms and the very inactivation of the pathogens under the action of this biological factor. The kinetic parameters showed that *Salmonella* sp. was the most sensitive pathogen to the presence of the *Lactobacillus plantarum* strains and their metabolites, followed by *Staphylococcus aureus* ATCC 6538P, *Staphylococcus aureus* ATCC 25093 and *Salmonella abony* ATCC 6017. The applied kinetic models were adequate and appropriate for examination of the antagonism kinetics between the lactic acid bacteria strains and the pathogen strains.

REFERENCES

- Charalampopoulos, D.; R. Wang; S. Pandiella; C. Webb. 2002 "Application of cereals and cereal components in functional foods: A review", *International Journal of Food Microbiology*, 79, 131-141;
- Charalampopoulos, D.; S. Pandiella; C. Webb. 2003 "Evaluation of the effect of malt, wheat and barley extracts on the viability of potentially probiotic lactic acid bacteria under acidic conditions", *International Journal of Food Microbiology*, 82, 133-140.
- Chen, Y.; L. J. Yu; H. P. V. Rupasinghe. 2013 "Effect of thermal and non-thermal pasteurisation on the microbial

- inactivation and phenolic degradation in fruit juice: a mini-review”, *Journal of the Science of Food and Agriculture*, 93 (5), 981 - 986.
- Dalić, D.K.D.; A. M. Deschamps; F. Richard-Forget. 2010 “Lactic acid bacteria - potential for control of mould growth and mycotoxins: a review”, *Food Control*, 21, 370 – 380.
- Denkova, R.; B. Goranov; D. Teneva; Z. Denkova; G. Kostov. 2017 “Antimicrobial activity of probiotic microorganisms: mechanisms of interaction and methods of examination”. In: *Antimicrobial Research: Novel bioknowledge and educational programs* (Microbiology Book Series - Volume #6) 201 - 212.
- Denkova, R.; S. Ilieva; D. Nikolova; Y. Evstatieva; Z. Denkova; M. Yordanova; V. Yanakieva. 2013 “Antimicrobial activity of *Lactobacillus plantarum* X2 against pathogenic microorganisms”, *Bulgarian Journal of Agricultural Science*, 19 (2), 108–111.
- Denkova-Kostova, R.; B. Goranov; D. Teneva; Z. Denkova; G. Kostov. 2018 “Antimicrobial activity of *Lactobacillus* strains against *Escherichia coli*: a multimethod approach to explore the mechanisms and factors determining the antimicrobial action”. In: *Antimicrobial Research: Novel bioknowledge and educational programs* (Microbiology Book Series - Volume #7) (in press)
- Eswaranandam, S.; N. S. Hettlarachy; M. G. Johnson. 2004 “Antimicrobial activity of citric, lactic, malic or tartaric acids and nisin-incorporated soy protein film against *L. monocytogenes*, *E.coli* 0157:H7, and *Salmonella gaminara*”, *J. of Food Science*, 69(3), FMS 79.
- Kociubinski, G.; S. Salminen. 2006 “Probiotics: Basis, state of the art and future perspectives”, *Functional food network general meeting*.
- López de Lacey, A.M.; E. Pérez-Santín; M. E. López-Caballero; P. Montero. 2014 “Survival and metabolic activity of probiotic bacteria in green tea”, *LWT - Food Science and Technology*, 55, 314-322.
- Siro, I.; E. Kapolna; B. Kapolna; A. Lugasi. 2008 “Functional food. Product development, marketing and consumer acceptance —A review”, *Appetite*, 51, 456–467.
- Socol, C.R.; L. P. S. Vandenberghe; M. R. Spier; A. B. P. Medeiros; C. T. Yamagishi; J. De Dea Lindner. 2010 “The potential of probiotics: a review”, *Food Technology and Biotechnology*, 48, 413-434.
- Stanbury, P. F.; A. Whitaker; S. Hall. 2003 “Principles of fermentation technology”, *Elsevier Ltd*, pp. 123-144.
- Stanton, C.; R. Ross; G. Fitzgerald; D. Van Sinderen. 2005 “Fermented functional foods based on probiotics and their biogenic metabolites”, *Current Opinion in Biotechnology*, 16, 198–203.

ACKNOWLEDGEMENTS

This work was 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 Bulgarian National Fund "Scientific Research" under the project КП-06-Rila/2 from 20.12.2018 "Bio-preservation by the Synergistic Action of Probiotics and plant Extracts (ESCAPE)". Cooperation between University of Food Technologies Plovdiv (Bulgaria), BioDyMIA research unit (EA n°3733, Université Claude Bernard Lyon 1 - ISARA Lyon, France), and PAM (Université de Bourgogne - AgroSup Dijon Joined Research Unit) was supported by Franco-Bulgarian cooperation Hubert Curien Programme (PHC Rila) (ESCAPE project). The authors wish to express their gratitude for the supports of the Ministry of Education and Science

(Bulgaria), the Embassy of France to Bulgaria in Sofia (“Ministère des Affaires Etrangères”, France), Campus France and “Ministère de l’Enseignement Supérieur, de la Recherche et de l’Innovation” (France).

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is Associate 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. 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 Head Assistant 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 Head Assistant 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, identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is a researcher in the LBLact Company, 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.

DESI SLAVA TENEVA is biologist, PhD of the Laboratory of Biologically Active Substances, Plovdiv – Institute of Organic Chemistry with Centre of Phytochemistry – Bulgarian Academy of Sciences. She received her MSc degree in Analysis and Control of Food Products in 2008 and a PhD degree in Biological sciences (Microbiology) in 2017. Her research interests are in the area of Chemical Sciences: Phytochemistry, biologically active phyto-substances (extraction, purification, isolation and characterization); polyphenol compounds; flavonoids; antioxidants; antioxidant activity determination; nutraceuticals and functional foods development; evaluation of biological activity. Additional research areas: Biological sciences: Antimicrobial activity of potentially probiotic strains, essential oils, plant extracts against pathogenic and saprophytic microorganisms.

AUTOMATIC PRODUCTION OF PATIENT ADAPTED ORTHOPAEDIC BRACES USING 3D -MODELLING TECHNOLOGY

Paul Steffen Kleppe
Department of Ocean Operations and Civil Engineering
Faculty of Engineering
NTNU, Norwegian University of Science and Technology
N-6025 Aalesund, Norway
E-mail: paul.s.kleppe@ntnu.no

Webjørn Rekdalsbakken
Department of ICT and Natural Sciences
Faculty of Information Technology and Electrical
Engineering
NTNU, Norwegian University of Science and Technology
N-6025 Aalesund, Norway

KEYWORDS

Surface photogrammetry, anatomic surface modelling, 3D printing, orthopaedics, prosthetics.

ABSTRACT

The research group in biomechanics at NTNU Aalesund works in close cooperation with the orthopaedic surgeons at Aalesund Hospital. One of the research activities has been to develop an automatic procedure for producing individual patient adapted orthopaedic braces for hand fractures. This paper is a result of this cooperation. The work has so far resulted in the design and 3D printing of individually adapted orthopaedic braces for simple fractures in the hand and arm. However, much of the production of these has been manual and time consuming. Now, a practical procedure for producing such braces in the clinic is about to be realized. This paper presents the development of the production procedure and testing of the resulting braces. The results of this research are then discussed regarding the challenges involved and benefits of introducing this procedure into the orthopaedic clinic.

INTRODUCTION

For many years NTNU Aalesund has worked in close cooperation with the local Department of Health in Møre og Romsdal. In the field of biomechanics there are several interesting ongoing research projects between technology staff at NTNU and medical doctors. These projects represent a broader research context than most internal projects at NTNU and are very inspiring and stimulating for the researchers both at NTNU and Aalesund Hospital (Mork, Hansen, Strand, Giske, Kleppe, 2016). This ongoing activity has already resulted in several research articles, in the combined field of technology and medicine. Related to the work discussed here, a first paper on orthopaedic braces was presented in 2018 at the first International Industrial Conference on Cyber-physical Systems (IICPS2018) in St. Petersburg (Kleppe et. al, 2018).

In this research, we have combined the competence of the senior surgeons at the hospital with NTNU's considerable competence in the fields of mechatronics and 3D modelling. The combination of these two disciplines provides a strong basis for research in the field of biomechanics and we have managed to produce patient adapted orthopaedic braces. However, the process in the

beginning was quite time-consuming involving many manual operations. This paper presents the work done to improve this process, looking closely at the different activities in order to find areas for improvements. The basis for such a project is a thorough competence in the field of 3D technology, including 3D scanning, data formatting, concept solution, surface processing, and 3D printing. Last, it is the orthopaedic surgeon's detailed knowledge of the anatomy and treatment of fractures that guides us towards a functional procedure for producing patient adapted orthopaedic braces in the clinic.

At the current stage of development, there are some shortcomings regarding the quality of the braces, but more seriously is the production time. The process is too slow within reasonable product costs. This problem involves both the scanning and the printing stages of the process. There are solutions, but these are still too expensive. However, there are many advantages of introducing such an automated process in the clinic, primarily for the patient, but also for the hospital. Both working time and material costs will decrease substantially once the equipment is installed in the clinic. This research has concentrated on the improvement and optimization of the methods and technology that are already included in the process. The general development in today's world of 3D technology is enormous, and our group has spent much time finding the correct technology and most suitable software for the implementation of a functional production procedure, (Gya and Thorsen, 2017). We regard it as only a matter of time until the obstacles to the realization of this procedure are removed. This novel way of producing orthopaedic braces is tailored to each patient, and will be both faster and cheaper than previous methods.

ORTHOPAEDIC BASIS

The motivation and benefits for the orthopaedic clinic are well described by Kleppe et al. (2018). The traditional way of making plasters is both time consuming and labour intensive, and represents a past technology. See Figure 1 for a traditional plaster cast. It is high time for this to be replaced by modern technology and an efficient new procedure. Wrist fractures (distal radius) are the most common types of fractures in Norway; it is estimated that there are more than 15,000 fractures annually among adults. After the introduction and spreading of electric scooters in Norway and most other countries, the notion of "The last mile" has become

synonymous with a substantial increase in hand fractures. For a long time, the orthopaedic doctors At Aalesund Hospital been looking for a more efficient and accurate way to make casts. A customized cast fitted to each patient and produced by 3D scanning and 3D printing is now a realistic option that could meet their needs. The production process will be highly automated and therefore save valuable time for the medical staff, and the quality of the product will be independent of the staff member's experience. The implementation of this 3D scanning and printing technology in the conservative treatment of fractures will make the process faster, more reliable and more cost effective compared to traditional manual work.

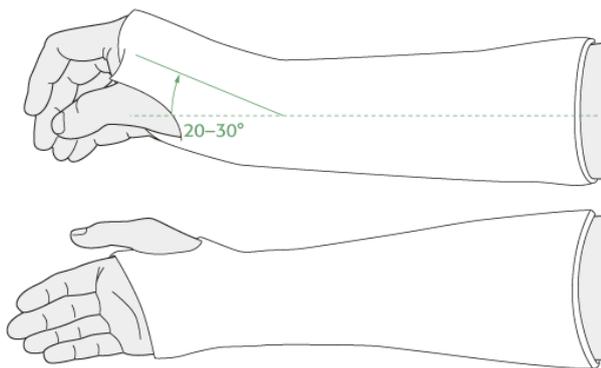


Figure 1: Illustration of a Traditional Plaster Cast

TECHNOLOGICAL BASIS

When it comes to technology, the necessary equipment already exists in the market. However, 3D technology is still an immature field regarding commercial applications and there are many challenges in cost, processing speed, software development and software interfacing. However, the current speed of development of this technology means that these obstacles will soon be overcome. The main focus will therefore be on system integration: choosing the correct equipment, finding the correct software and adapting it to the particular functioning product. In the process of making patient adapted braces there are four steps. First, the hand is scanned with the appropriate equipment to obtain a point cloud of the object. Second, the point cloud is imported in a suitable 3D modelling program, and a 3D surface model of the hand is made based on the data in the point cloud. Third, the orthopaedic brace is modelled based on the anatomy of the hand using special techniques for surface modelling. The last step is the 3D printing of the brace with a printer that fulfils the requirements of product quality and processing time at an affordable price. See as example Summitid (2014).

OBJECT SCANNING

The scanning units in this project are ordinary photographic cameras. The reason for this is that very good cameras are available at reasonable prices, and the 3D scanning technology and associated software using cameras are developing very fast. There are products in the market well suited for this kind of research and an

increasing number of sources for free public domain software.

Photogrammetry

The basic platform for the reconstruction of 3D surfaces is the field of photogrammetry. With the introduction of 3D surface scanners, this technology is used in an increasing number of applications in many fields. Today there are many software programs available on the market that can generate dense point clouds and 3D surfaces from still images. There is also an increasing number of software tools available in the public domain for free use.

Photogrammetric methods

We use photogrammetric techniques in the process of scanning and reconstructing the surface of the patient's hand. This is based on the recording of synchronized images taken with several cameras at different angles surrounding the patient's hand. The reconstruction is based on the methods called Scale-Invariant Feature transform (SIFT) and Bundle Adjustment. The position and orientation of each camera must be determined exactly, and the necessary overlap of images decided upon.

SIFT-algorithm

Scale-Invariant Feature transform is a patented algorithm that seeks to find hallmarks of images, although there are several other algorithms that give similar results. The algorithm starts by searching for a set of reference images, so-called key points. These points have a unique identification and are found by searching for local maximums and minimums by the Difference of Gaussian (DoG) method. DoG uses a Gaussian blur filter to calculate new values of the image pixels. The filter consists of a Gaussian function to obtain these new values. New images are created by subtracting a blurred image from the original image. This process is repeated several times. In this way, DoG enhances the details of the image. The Gaussian function in two dimensions is given as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

where σ is the standard deviation of the Gaussian distribution. When the key points are derived, they are compared to distinctive marks and details on the images that are manually defined in advance. The key points are used to recognize specific points in many images, and thus connect the images. The SIFT algorithm is very robust regarding rotation, stretching, distortion and light changes of the images. However, the algorithm may be sensitive to repetitive patterns by defining one key point at more places. The SIFT algorithm needs as much information as possible to derive the key points. The pictures must imitate the reality as closely as possible. Since the DoG principle creates key points in areas with much variation, images with high contrast and sharpness are valuable in photogrammetry. High spatial resolution

images are necessary to obtain the required details of the object, while accurate time resolution and synchronization between images is important to achieve sharp images. All details in an image are important, also in the background of the object. The best result is obtained when the object is close and covers most of the image size. Objects with big shiny surface areas are also a problem because they have too few details to create good key points. In such cases, it may be necessary to change the texture of the surface by placing suitable patterns on it.

Bundle Adjustment

Bundle Adjustment is a method used to estimate structure in 3D from 2D images. By recognizing two-dimensional bundles of key points in different images, the method will calculate the position and direction from which each image is taken. The calculations are based on the distances and directions between the points of the bundle in each image. Changes in the orientation of points from image to image are due to either rotation or scaling, in addition to image distortion. Combined with knowledge about the characteristics of the cameras and their positions, this method can be used to find the orientation of points on the object.

Dense point cloud and surface reconstruction

A dense point cloud is the basis for the reconstruction of the 3D model of the scanned object. The point cloud consists of the key points found by SIFT, correctly oriented relative to each other in space by use of the Bundle Adjustment algorithm. From the information from the dense point cloud, the surface of the 3D model can be reconstructed. This reconstruction is based on algorithms using nearby points in the cloud to build small patches of the surface. These patches constitute a mesh of triangles representing the surface. Triangulation is a method much used in these algorithms. The size of the triangles determines the smoothness of the surface.

Smartphone as a photogrammetric device

In this research, we have tested the use of an iPhone for scanning. There has been great improvements in such software for the latest of these models. Scanning with a smart phone is very easy, and the software gives a point cloud in a format that is suitable for further processing in 3D modelling programs, such as Siemens NX.

SCANNER CONFIGURATIONS

One important issue is the design of the scanner. Depending on the scanning technique chosen, the scanner may have different configurations, and it is a question of choosing the optimum solution. The following presents the designs that have been tested.

Tube configuration with fixed cameras.

To ensure sufficient image overlap in radial and axial directions around the object, a concept for a photogrammetric device with eight cameras mounted on a ring was developed. The rings were stackable with four

rings needed to cover one forearm. The outer diameter of the machine was carefully selected to be able to print parts on a standard 3D printer. The rings were divided into one segment per camera to modularise and simplify assembly. The camera stack, segments and rings, were kept together by truss rods that were slightly pretensioned. See Figures 2 and 3 for camera configuration and ring design. In this cylinder, each of the segments was equipped with a Raspberry PI camera, and the operation of all the cameras were synchronized through an Arduino controller. In this way, all cameras were remotely triggered from one master computer. Figure 4 shows the final design of this scanning device including a support for the patient's hand (Aarsæther, Dale, 2019 and Alvestad, Nedreliid, Sjøstad, 2019).

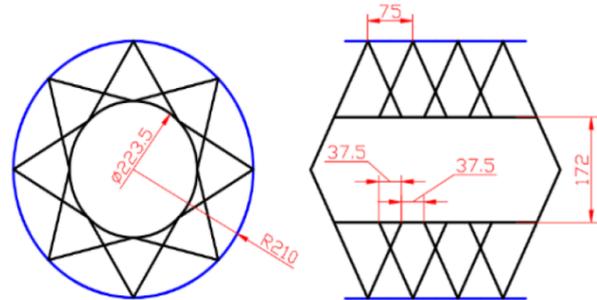


Figure 2: 32 (8x4) camera configuration

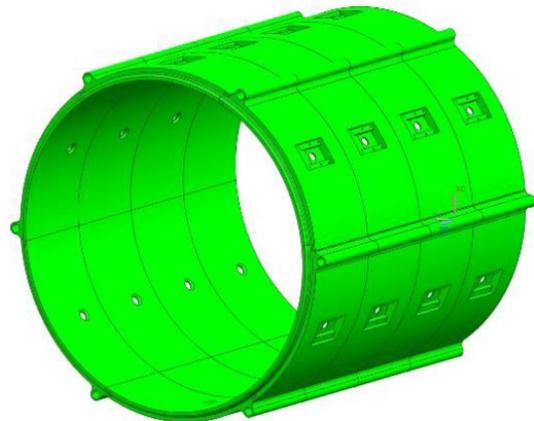


Figure 3: Scanning cylinder with camera segments



Figure 4: Scanner with fixed cameras

Rotating scanner.

Though some flexibility was built into the design of the fixed scanner by installing additional rings, the design became too rigid. In addition, the availability of new types of scanners, such as lidars, and the fast development of photogrammetry software on smart phones, led our thinking towards a more flexible solution. We decided to build a simple device that could be used in testing different kinds of scanners. The result became a rotating arm controlled by a motor moving at the desired speed. The arm was supplied with moveable brackets to hold the scanning element (either a lidar or a mobile phone) in the desired position. The distance to the object could also easily be changed. With this device, we were able to test different scanner configurations to obtain the best result. See Figure 5.

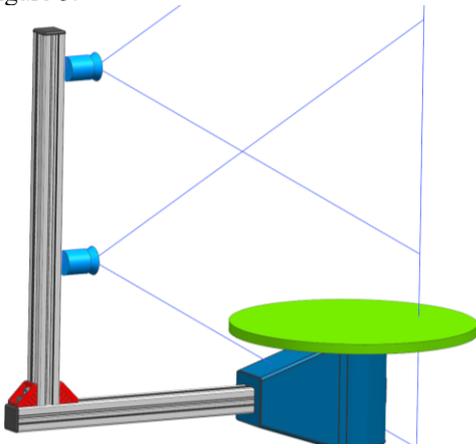


Figure 5: Design of the rotating scanner

GENERATING 3D-GEOMETRY

3D models and modelling tools

As described in Kleppe et al. (2018), several specialised 3D-applications were needed to efficiently create a 3D-supporting brace model of an adequate quality. Scan-data were imported into Geomagic X for clean-up, post-processing and to create anatomic surfaces. These surfaces were imported into Siemens to create the geometry of the brace. New and improved features in Siemens NX enables the CAD-operator (designer) to efficiently work with point clouds, facet data, b-rep surfaces, product configurators and manufacturing preparations in one single application.

Polygon modelling and polygon mesh is an approximate method to describe surfaces, while a vertex is a point in three-dimensional space. Two vertices connected by a straight line become an edge, and three edges connected to each other becomes a triangle (Hahnmann, Brunett, Farin, Goldman, 2002). Polygon modelling is also referred to as facet topology in Siemens NX. When capturing and post-processing 3D-scan data, the polygon approach is a common way of visualising the object. Scan resolution and number of points affect the accuracy of the model and the computer power needed to process the data.

NURBS or B-splines is a mathematical description of curves and surfaces. NURBS are quite common in computer aided design and engineering (CAD, CEA) software. Because of the mathematical nature of NURBS, they are quite efficiently handled by computer software. NURBS and B-spline are also referred to as B-rep topology in Siemens NX.

Both polygons and NURBS have their benefits in computer science and visualisation, but until recently combining both data formats is still a cumbersome process. However, some advances have been made, and with Siemens NX 11 convergent modelling has been introduced as a new feature combining polygon and NURBS as a modelling tool.

Convergent modelling in NX enables the designer to process scan-data, combining polygon and NURBS into one robust and efficient data processing CAD-model. Furthermore, this enables the use of product configurators to automate the design process and connect to manufacturing applications and software. These are the 3D modelling methods that were used to manage the reconstruction of the hand surface. See the following references.

Post-processing and model surface optimization

Based on the scanner setup and scan quality, several steps must be taken during the post-processing to achieve anatomic surfaces of the desired quality. A similar procedure was described in Kleppe (2018) and Scarano, Chiara and Erra (2008), but now it is possible to post-process scan data and design the brace in one single application. Based on the scanner setup and scan quality several of the steps may be skipped or automated by scripting during import from the scanner to Siemens NX. This procedure is described below:

First, remove background noise. Usually the scanner captures more than just the hand, but by using a bounding box, only the geometry required is selected for further processing. Second, heal the mesh. This procedure fills small holes and gaps in the mesh and aligns small surfaces with each other. Third, perform a global re-mesh. Based on modifications in the second step, further optimizations are carried out automatically by moving and aligning points to improve quality and reduce the number of triangles while keeping the initial shape of the geometry. Fifth, fill holes or replace rough surface areas by manual patching the mesh. Shown in Figure 6 and Figure 7. Sixth, optimize and smooth the mesh.

Defining cut boundaries

A concept for plaster design was developed by Dale, Thorsen et al. (2017). The procedure and 3D-modelling technique was further developed and integrated into one application (Siemens NX) by Alvestad, Nedreliid and Sjästad (2019).

To create the cast geometry, the required surfaces are extracted from the scanned model by defining cut-boundaries. Three main cut planes are defined, on the fingers (1), thumb (2) and forearm near the elbow (3) as shown in Figure 8. Figure 9 shows anatomic surfaces of the arm to be used as a template for designing the customized cast. In addition, a fourth cut plane is added later in the design stage along the arm-axis to open up the plaster and enable it to be slid onto the arm. Slots for Velcro straps are added next. See Figure 10.

The end-result after all this processing is a brace geometry ready for 3D production, as shown in Figure 11.

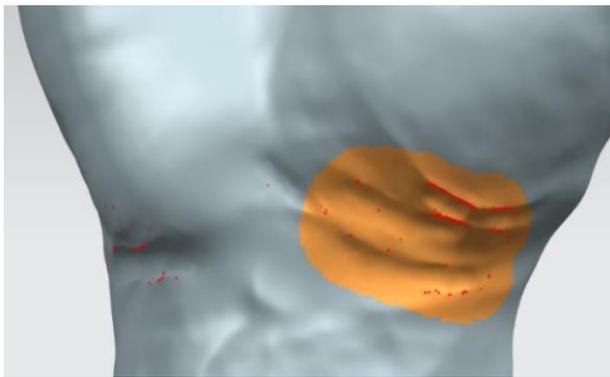


Figure 6: Before manual patching

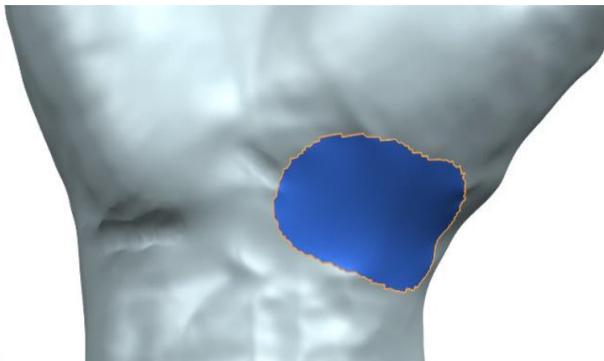


Figure 7: After manual patching

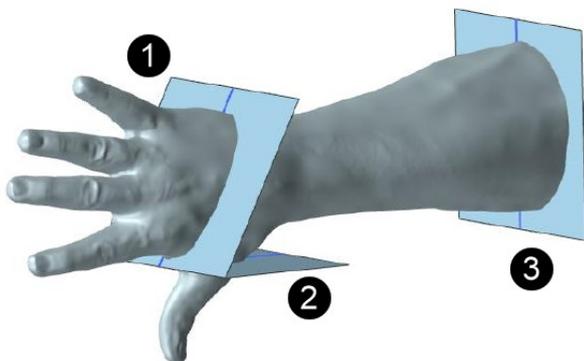


Figure 8: Cut boundaries



Figure 9: Anatomic surfaces and template for cast geometry

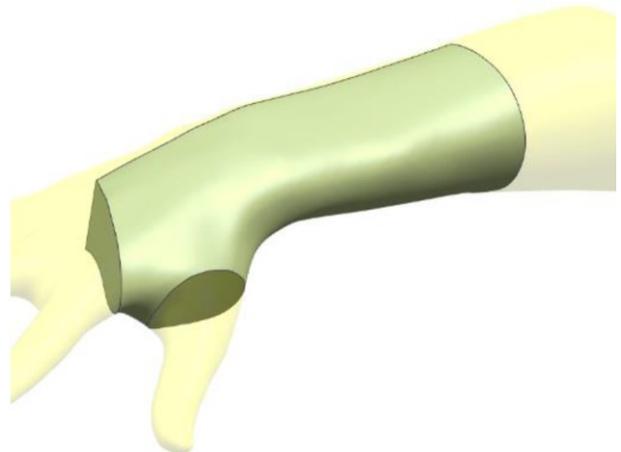


Figure 10: Plaster with all cut-boundaries defined

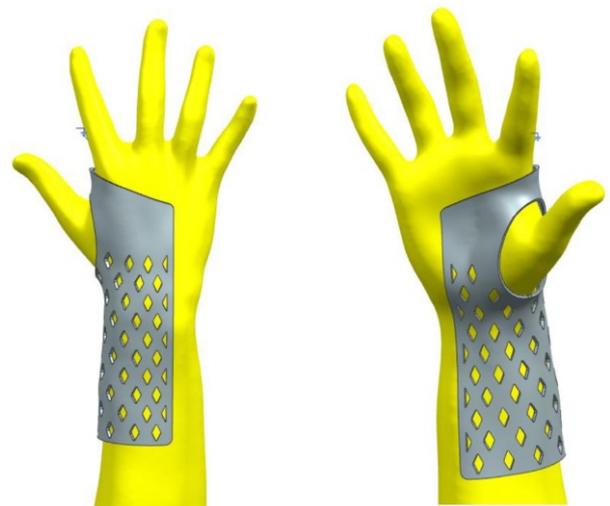


Figure 11: Complete plaster ready for manufacturing

MANUFACTURING AND 3D-PRINTING

Automatization of the design process

Creating 3D-Cast geometry with traditional 3D modelling software (CAD) as described in the procedure above usually requires highly trained CAD operators. However, the procedure is, except for some anatomical deviation between patients, the same for each new plaster. By using a product configurator (Product Template Studio) it is possible to simplify interaction

with scanned data and the 3D CAD-model and thus reduce the need for training. A menu-driven interface is built on top of a 3D-model template. Scanned data for each new patient are imported, and the doctors or a trained operator can do geometric adjustments on the fly in an intuitive interface. Figures 12 to 14 show the menu driven interface processing and the generated 3D-geometry. A total of four tabs have been created to interact with the model.

- 1) **Placement:** Geometry adjustments can be done by moving and adjusting the angle of the cut planes.
- 2) **Cast geometry:** Thickness of the cast can be defined here: usually 2mm. In addition, clearance from the skin can be modified. Usually between 0.2mm and 0.5mm.
- 3) **Hole pattern:** In this menu it is possible to turn the hole pattern on and off and adjust the hole size. To enable 3D-printing of a cast with holes and avoid support during printing, a diamond shaped hole pattern is available, with width adjustments only. The height is automatically adjusted in a 2:1 ratio.
- 4) **Advanced tab:** Contains detailed adjustments for geometry in the thumb area. This is usually not necessary, but enables the user to adjust the angle between thumb hole and the centerline of the cast.

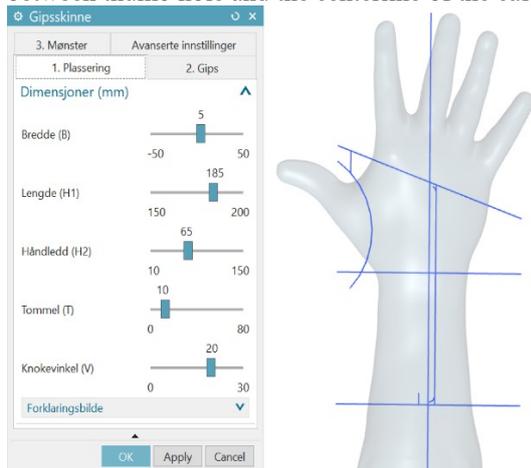


Figure 12: Geometry adjustments for plaster

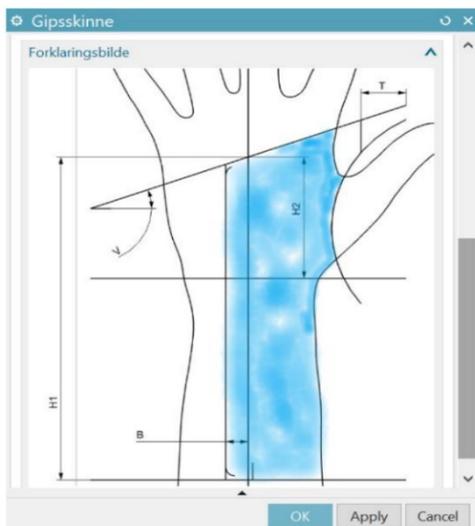


Figure 13: Outline of the brace

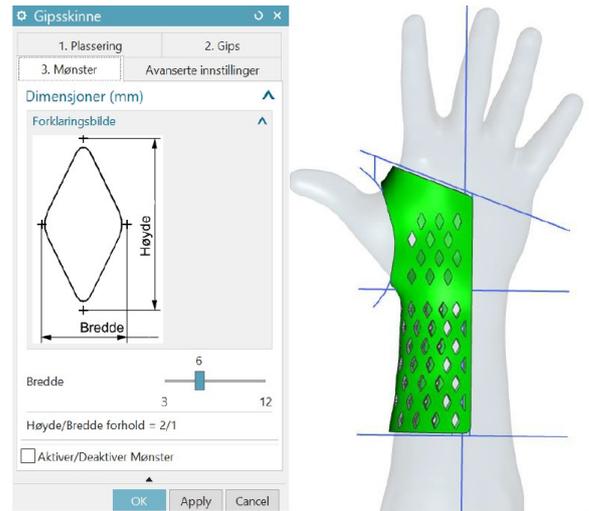


Figure 14: Hole pattern

Further optimization

- Connect the menu interface directly to the scanner to invoke scanning process and import the scan data directly.
- Automate the geometry cleanup post-processing of scanned data.
- Further automate the digital value chain and manufacturing pipeline and develop the interface to the 3D-Printer pool.
- Enable color selection from the menu interface.
- Vision, augmented reality and machine learning to speed up the process and further minimize the need for human interaction.

In an ideal world, the doctor/operator interacts with the machine in several simple steps:

- 1) The patient places their hand in the machine and the operator clicks on a button to start scanning. The scanning process takes less than a second.
- 2) The machine generates a digital model of the hand and iterates to create plaster geometry and suggest a design. This might take some minutes depending on the computing power. The operator approves the generated design and makes modifications if needed.
- 3) The patient selects a color from the library while the machine is generating design suggestions.
- 4) The operator sends the cast to production. Manufacturing data are instantly sent to a pool of 3D printers.

ADAPTATION IN THE CLINIC

The production process and the hardware and software selected are tailored to this special application. However, the production equipment is quite complex, so the main challenge has been the integration of the different parts into a complete system. Many solutions have been examined at each stage of the process with the focus on selecting the best options, regarding both hardware and software, and not least, the interfacing between the

different stages. In this way, we have designed a production system for a product tailored to a particular customer. The implementation of this process in the clinic will have to be done in close co-operation between the engineer and the medical staff. This stage will also involve changing of the traditional routines in the clinic.

Software and interfacing

The software needed in such a complex production unit will consist of many parts. In this case, there are independent software components for the scanning, the preparation of the point cloud of data, the modelling process and the 3D printing. Most of the software covering our needs is available as open source products that may be modified and tailored to our use. The OEM manufacturers of the parts used in the product also deliver the necessary software used with this part. The main challenge is the interface between the different software components regarding parameter settings and the exchange of data. These components must connect into a single unit. Much effort has gone into making this a seamless software product. In addition, an easy and suitable user interface has been developed.

RESULTS

The purpose of this research project is the realization of a radical new way of producing orthopaedic braces, made possible by today's 3D technology. We have already produced such braces using modified 3D products available on the commercial market. Now we have tailor-made a manufacturing process for this product by adapting OEM hardware and software from different vendors to suit our needs. The interfacing between the different software systems and customization of the hardware components has been the most demanding and time-consuming work in obtaining a functional production process.

DISCUSSION

The production process described here is p.t. technologically seen up and running. However, there are stringent demands regarding both the product and the production procedure when treating patients in a clinic. The greatest challenges are connected to production time and adapting to the daily procedures in the clinic. Faster and better 3D printers are available on the market and the prices are rapidly decreasing. On the other hand, there are many benefits of the new process. The clinic can make a product that is customized to each patient. The production is flexible regarding material qualities and design of the brace. The patient will get a brace that can be removed for cleaning and adjusted to his own personal comfort. The production costs, including material and work hours, will be lower than for a traditional plaster cast. The speed of development in 3D technology today with accompanying reduction in costs, the investment in such equipment will soon be very affordable.

CONCLUSION

The aim of this research has been to use state of the art 3D technology to replace traditional procedures in the treatment of hand fractures. The result has shown that it is fully possible to replace the traditional plastering process with 3D modelling and printing of patient braces. Technologically, this new process has already been realized, and processing costs are less than for the traditional plastering process. It will not be long before investment in this kind of equipment is affordable. Processing time is closely connected to costs and the fastest 3D printers are still quite expensive. However, costs on this type of limiting technology tend to decrease very fast, so this will soon be easier to access. We are quite confident that this will be the kind of technology found in the orthopaedic clinics in most hospitals in the near future.

REFERENCES

- Alvestad, V. A. J.; Nedrelid, O. H.; Sjøstad, D. 2019. "Brukertilpasset gips." B.Sc. thesis, NTNU, Norwegian University of Science and Technology, Aalesund.
- Aarsæther, T.; Dale, A. N. 2019. "3D-tilpasset støtteskinne ved håndleddsbrudd." B.Sc. thesis, NTNU, Norwegian University of Science and Technology, Aalesund.
- Kleppe, P. S.; Dalen, A. F.; Rekdalsbakken, W. 2018. "A novel way of efficient adaption of orthopaedic braces using 3D technology." *1st IICPS 1st IEEE International Industrial Conference on Cyberphysical Systems*. IEEE Xplore. www.ieeexplore.ieee.org
- Gya, M. and A. B. Thorsen. 2017. "Spesialtilpasset gips for håndleddsbrudd bed bruk av dagens 3D teknologi." B.Sc. thesis, NTNU, Norwegian University of Science and Technology, Aalesund.
- Summitid. 2014. <http://www.summitid.com/> Amsterdam, NL.
- Scarano, V.; Chiara R.; Erra, U. 2008. "Meshlab: an Open-Source Mesh Processing Tool." Eurographics Italian Chapter Conference.
- Hahnmann, S.; Brunett, G.; Farin, G.; Goldman, R. 2002. "Geometric Modelling" Springer-Verlag Wien GmbH.
- Mork, O. J.; I. E. Hansen; K. Strand; L. A. Giske; and P. S. Kleppe. 2016. "Manufacturing Education – facilitating the Collaborative Learning Environment for Industry and University." *6th CLF- 6th CIRP Conference on Learning Factories*. Elsevier BV. www.sciencedirect.com

AUTHOR BIOGRAPHIES

PAUL STEFFEN KLEPPE is assistant professor at NTNU, Norwegian University of Science and Technology, in Aalesund. He has a master degree in Mechanical Engineering, and a MBA in Technology Management from NTNU.

WEBJØRN REKDALSBAKKEN is assoc. professor and program leader of the engineering program in cybernetics at NTNU, Norwegian University of Science and Technology, in Aalesund.

NAVIGATION SYSTEM FOR LANDING A SWARM OF AUTONOMOUS DRONES ON A MOVABLE SURFACE

Anam Tahir^{1, a}, Jari Böling^{2, b}, Mohammad-Hashem Haghbayan^{1, c}, and Juha Plosila^{1, d}

¹Autonomous Systems Laboratory, Department of Future Technologies

²Laboratory of Process and Systems Engineering

¹University of Turku, ²Åbo Akademi University

^{1,2}Turku, Finland

Email: ^aanam.tahir@utu.fi, ^bjari.boling@abo.fi, ^cmohhag@utu.fi, ^djuha.plosila@utu.fi

KEYWORDS

Unmanned Aerial Vehicles; Distributed Control; Leader-Follower Hierarchy; Soft Landing

ABSTRACT

The development of a navigation system for the landing of a swarm of drones on a movable surface is one of the major challenges in building a fully autonomous platform. Hence, the purpose of this study is to investigate the behaviour of a swarm of ten drones under the mission of soft landing on a movable surface that has a linear speed with the effect of oscillations. This swarm, arranged in a leader-follower hierarchical manner, has distributed control units based on Linear Quadratic Regulator control with integral action technique. Furthermore, to prevent drones from landing arbitrarily, the leader drone takes the feedback of translational coordinates from the movable surface and adjusts its position accordingly. Hence, each follower tracks the leader's trail with offsets, taking collision avoidance into account. The design parameters of controllers are mapped in a way that the simulations demonstrate the feasibility and great potential of the proposed method.

INTRODUCTION

Unmanned Aerial Vehicles (UAVs), or drones, are increasingly getting attention in the aviation and maritime industries with the evolution of drone technology for both recreational and military grounds [1, 2]. These have innovative impacts in the areas of data collection for inspection purposes and are capable of carrying out tasks in a variety of situational operations. They can shape the future with potential benefits in the fields such as security and surveillance, remote sensing, search and rescue, elimination of human error, and autonomous deliveries and shipping [3–6]. For example, in 2017, the European Maritime Safety Agency (EMSA) issued a contract to Martek, valued at €67M for the usage of drones in European waters to provide assistance with border control activities, pollution monitoring, search and rescue tasks, and detection of illegal

trafficking (drugs and people) and fishing [7, 8].

The landing mechanism of UAVs is one of the challenging problems. An extensive survey based on vision-based autonomous landing methods is elaborated in [9]. Based on the setup of the vision sensors, these methods are divided into two main categories, i.e., onboard vision landing systems and on-ground vision systems. In [10], a detailed review on control based landing techniques (such as from basic nonlinear to intelligent, hybrid and robust control) along with GPS and vision-based landing schemes is presented. A wide literature is available related to vision-based autonomous landing of UAVs [11–17]. For example, in [18], a vision-based net recovery landing system is proposed for a fixed-wing UAV that does not require a runway. Likewise in [19], the landing of a quadrotor on a moving platform is addressed. In [20], a real-time vision-based landing algorithm for an autonomous helicopter is implemented. An on-board behaviour-based controller is used that is subdivided into hover, velocity, and sonar sub-behaviours. Hovering control of the helicopter is implemented using proportional control, whereas velocity and sonar controllers are implemented with proportional-integral control design. In [21], a nonlinear proportional-integral type controller is proposed for vertical take-off and landing of a quadcopter. It exploits the vertical optical flow to facilitate hover and land on a movable platform. In [22], to control a quadcopter's vertical take-off and landing on a moving platform, the image-based visual servoing integrated with the adaptive sliding mode controller is validated. However, this approach requires the landing site to be predetermined and therefore, it is not suitable for operations in unknown terrain.

Due to the rising importance and research effort put into autonomous vehicles and robots, there is broad research on vision-based methods integrated with/without control techniques for landing missions. Hence, this paper focuses on the control design for landing a swarm of drones on a movable surface, and the vision-based approaches are of no interest in this work.

Landing safely, i.e., soft landing, is the key to successful exploration of the assigned missions; in electromechanical systems, the mitigation of the connected effects of collision relies on the conversion of kinetic energy into heat or potential energy. An effective landing-system design should minimize the acceleration acting on the payload. In other words, the major challenges in autonomous landing are; (a) accurate placements as much as possible on the landing platform, and (b) the trajectory following in the presence of disturbances and uncertainties. Portions of this work have been reported in the previous work [23]. However, for present paper, there are additional contributions to face these challenges. This paper addresses the problem based on system modelling and testing of a swarm in a leader–follower hierarchical formation, consisting of ten drones, aiming at executing missions of soft landing on an oscillating surface that can be a vessel or any surface having oscillations. The distributed control units of each quadcopter in the swarm are designed using an Linear Quadratic Regulator (LQR) with integral action technique that can handle a multivariable system.

This paper comprises of 6 sections. Section 1, in addition to introducing the topic, also dwells upon the significance of the study as well as the works already carried out in this particular domain. Section 2 discusses the elements of a swarm formation while Section 3 elaborates the composition of swarm formation of UAVs. Section 4 describes the proposed control design for soft landing, whereas Section 5 builds upon the evaluation of the landing missions. Lastly, Section 6 presents the concluding remarks.

COMPONENTS OF A FORMATION

One of the most vital challenges in multi-agent systems is the formation control. It is defined as an organisation of a group of agents to maintain a formation with a certain shape [6]. Three main components are considered to solve any formation control problem, i.e., system architecture, its modelling, and strategies of formation control [24].

System Architecture

The system architecture delivers the infrastructure upon which formation control is implemented such as:

- **Heterogeneity vs. homogeneity:** Heterogeneous teams consist of either different apparatus or software, whereas homogeneous teams comprise of similar modules of hardware or software.
- **Communication structures:** The communication structures in the swarm can be categorised w.r.t. range, topology, and bandwidth.
- **Centralization vs. decentralization:** In the centralized controlling approach, a single controller possesses all the information required to get the desired control objectives, whereas each agent has its own local control mechanism and is completely autonomous in the decision process of decentralized control. Hybrid centralized/decentralized architectures, in turn, use central planners to provide high-level control over autonomous

robots.

System Dynamics

The dynamics of each drone in the swarm is based on the model of a quadcopter, i.e., a drone that has four propellers and ℓ is a length of the fixed pitch to mechanically movable blades, as shown in Fig. 1.

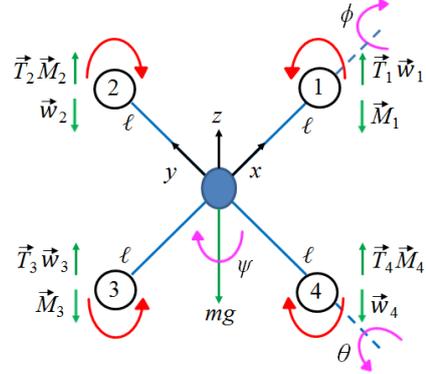


Fig. 1: Kinematics of the quadcopter

The gravity g and the thrust T_i , $i \in \{1, 2, 3, 4\}$, of the propellers are the main forces acting on the quadcopter. In this model, the inertial reference is the earth shown as (x, y, z) that is the origin of the reference frame. The drone is assumed to be a rigid body that has the constant mass symmetrically distributed with respect to the planes (x, y) , (y, z) , and (x, z) . The orientation of a quadcopter reference frame (x, y, z) with respect to an inertial frame $(x, y, z)_0$ can be expressed mathematically in a state variable form [25], where translational and angular accelerations are given by

$$\begin{aligned} \dot{v}_x &= -v_z w_y + v_y w_z - g \sin \theta \\ \dot{v}_y &= -v_x w_z + v_z w_x + g \cos \theta \sin \phi \\ \dot{v}_z &= -v_y w_x + v_x w_y + g \cos \theta \cos \phi - \frac{T}{m} \end{aligned} \quad (1)$$

and

$$\begin{aligned} \dot{w}_x &= \frac{1}{J_x} (-w_y w_z (J_z - J_y) + M_x - \frac{k_w T}{k_{MT}} J_{mp} M_z w_y) \\ \dot{w}_y &= \frac{1}{J_y} (-w_x w_z (J_x - J_z) + M_y - \frac{k_w T}{k_{MT}} J_{mp} M_z w_x) \\ \dot{w}_z &= \frac{M_z}{J_z} \end{aligned} \quad (2)$$

respectively. The thrust produced by each propeller T_i is translated into a total thrust T , and the reactive torque M_i , $i \in \{x, y, z\}$, is affecting the rotations along the corresponding axis. J_i , $i \in \{x, y, z\}$, is known as the moment of inertia along the corresponding axis, and J_{mp} is the moment of inertia of a motor with a propeller. The velocities corresponding to Equations (1) and (2) are

$$\begin{aligned}
\dot{x} &= v_x \cos \psi \cos \theta + v_y (-\sin \psi \cos \phi + \cos \psi \sin \theta \sin \phi) + v_z (\sin \psi \sin \phi + \cos \psi \sin \theta \cos \phi) \\
\dot{y} &= v_x \sin \psi \cos \theta + v_y (\cos \psi \cos \phi + \sin \psi \sin \theta \sin \phi) + v_z (-\cos \psi \sin \phi + \sin \psi \sin \theta \cos \phi) \\
\dot{z} &= v_x \sin \theta - v_y \cos \theta \sin \phi - v_z \cos \theta \cos \phi
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
\dot{\theta} &= w_y \cos \phi - w_z \sin \phi \\
\dot{\phi} &= w_x + w_y \sin \phi \tan \theta + w_z \cos \phi \tan \theta \\
\dot{\psi} &= w_y \frac{\sin \phi}{\cos \theta} + w_z \frac{\cos \phi}{\cos \theta}
\end{aligned} \tag{4}$$

respectively. The Equations (1)–(4) represent the complete nonlinear model of a quadcopter, composed of 12 states, 4 inputs, and 12 outputs. More precisely,

$$\mathbf{x} = [v_x \ v_y \ v_z \ w_x \ w_y \ w_z \ \theta \ \phi \ \psi \ x \ y \ z]^T \tag{5}$$

is the state or system vector,

$$\mathbf{u} = [T \ M_x \ M_y \ M_z]^T \tag{6}$$

is the input or control vector,

$$\mathbf{y} = \mathbf{x} \tag{7}$$

is the output (measured) vector. Furthermore, the performance output

$$\mathbf{y}_p = [x \ y \ z]^T \tag{8}$$

is defined for future use.

Using standard linearization, that is cutting off a Taylor series expansion after the first derivative, the nonlinear dynamic equations can be converted into linear state-space equations. This yields,

$$\begin{aligned}
\dot{\mathbf{x}} &= \left[-g\theta \ g\phi - \frac{T}{m} \frac{M_x}{J_x} \frac{M_y}{J_y} \frac{M_z}{J_z} \ w_y \ w_x \ w_z \ v_x \ v_y \ -v_z \right]^T \\
\mathbf{y} &= \mathbf{x}
\end{aligned} \tag{9}$$

that can further written into the standard state space form

$$\begin{aligned}
\dot{\mathbf{x}} &= A\mathbf{x} + B\mathbf{u} \\
\mathbf{y} &= C\mathbf{x} + D\mathbf{u}
\end{aligned} \tag{10}$$

where A , B , C , and D are known as the state or system matrix, input or control matrix, output (measured) matrix, and feedthrough matrix respectively. Correspondingly, \mathbf{x} , \mathbf{u} , and \mathbf{y} are known as the state or system vector, input or control vector, and output (measured) vector as in Equations (5)–(7). The system parameters

are taken from [25] and illustrated in Table I. The linear model in Equation (10) is used to examine the landing stability and controllability of the system as well as to design an LQR with integral control. The system has twelve eigenvalues at the origin, and all twelve states are controllable.

TABLE I: System parameters

Symbol	Quantity	Value
g	gravitational force	9.81 m/s ²
ℓ	length of the fixed pitch to mechanically movable blades	0.2 m
m	mass of quadcopter	0.8 kg
J_{mp}	moment of inertia of motor with propeller	≈ 0
J_x, J_y	moment of inertia w.r.t. axis x, y	1.8×10^{-3} kgm ²
J_z	moment of inertia w.r.t. axis z	1.5×10^{-3} kgm ²
k_{MT}	ratio of the reactive moment and thrust	0.1 m

Formation Control Schemes

A formation control scheme defines how a group of robots can be controlled to form and to maintain the desired formation. To control the formation of a drone swarm, the recent studies generally classify the different strategies into following main categories.

- Leader–follower [26–28]: The leader seeks for some group objectives, while the followers track the leader’s coordinates with prescribed offsets.
- Virtual structure [29–31]: A virtual moving structure reflects the complete formation as a rigid body such that the control design for a single agent is derived by defining the virtual structure. It then translates the movement of the virtual structure into the desired movement of each agent. Furthermore, as an actual leader is not needed, each virtual vacant pose can be filled by any agent.
- Behaviour-based [32–34]: each agent is assigned to the process of actuation that is defined as several desired behaviours. In each agent, to form the desired shape of the swarm, the overall control is derived by allocating different weights to behaviours.

The formation control schemes can be further categorised into position-, displacement-, distance-, and angle-based in terms of the requirement on the sensing capability and the interaction topology [35]. In position-based control, each agent is able to sense its own position in the formation that is defined by the desired positions of the different agents with respect to a global coordinate system. In contrast to this, in displacement-based control, each agent is assumed to sense its own as well as its neighbouring agents’ position in the formation that is defined by the desired displacements between pairs of agents with respect to

the global coordinate system. Then again, in distance-based control, the formation is defined by the desired inter-agent distances that are actively controlled. Each agent in the formation is expected to sense relative positions of their neighbouring agents with respect to their own local coordinate systems. Likewise, the actively controlled variable is the bearing between neighbours in angle-based control, rather than the distance to each of the neighbours.

COMPOSITION OF SWARM OF UAVS

Consider a hierarchical formation that has four levels using ten quadcopters, as illustrated in Fig. 2. This formation is based on a tightly coupled leader–follower flying mechanism in which the leader is directly communicating with its followers by providing its position references that are passed on to the followers [6, 36].

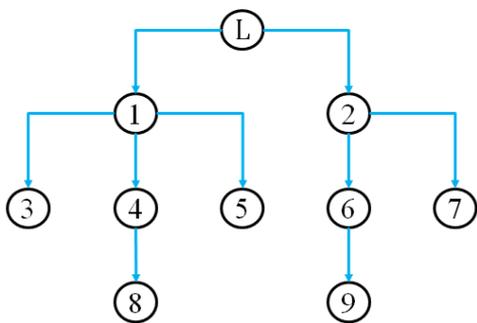


Fig. 2: Organization of the considered swarm of drones

In Fig. 2, the swarm is responsible for tracking the desired trajectories as well as for hovering at desired positions for given time intervals. The straight arrows show the direction in which coordinate variables are shared. The leader’s trajectory is independent and defines the formation’s trajectory. The trajectory of each follower is defined based on the orientation and actions of its respective leader. In terms of movement, each follower is dependent on its respective leader’s movement using a safe distance strategy that is denoted by $d_{\alpha,\beta}$, where $\alpha, \beta \in \{L, 1-9\}$. Each follower is responsible for efficiently tracking the respective leader’s trajectory, maintaining the distance between two respective entities.

To address the research questions, consider a control system that is liable for fine-tuning the movement of each drone in a swarm while maintaining the desired safe travel distance. Each drone is based on the similar system dynamics, which is illustrated in Section 2. In Fig. 3, a simple mechanism of feedback control system is presented in which output is controlled using its measurement as a feedback signal. This feedback signal is compared with a reference signal to generate an error signal which is filtered by a controller to produce the system’s control input.

PROPOSED CONTROL DESIGN FOR SOFT LANDING

In a simple example illustrated in Fig. 4, the swarm of drones aims to land on a vessel (or any type of mov-

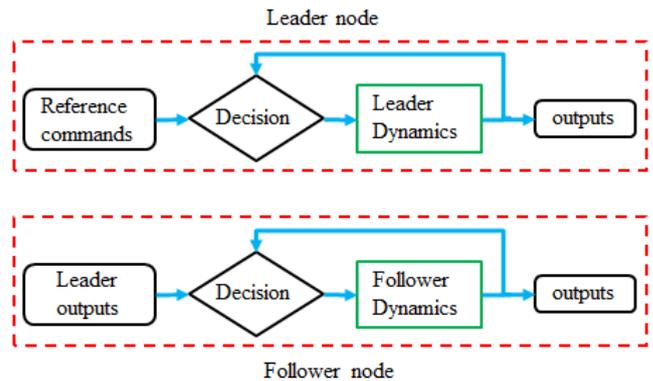


Fig. 3: Transmission topology in swarm formation

able surface) that has continuous speed with oscillations. It is assumed that the data is available through communication link on-board drones and vessel.

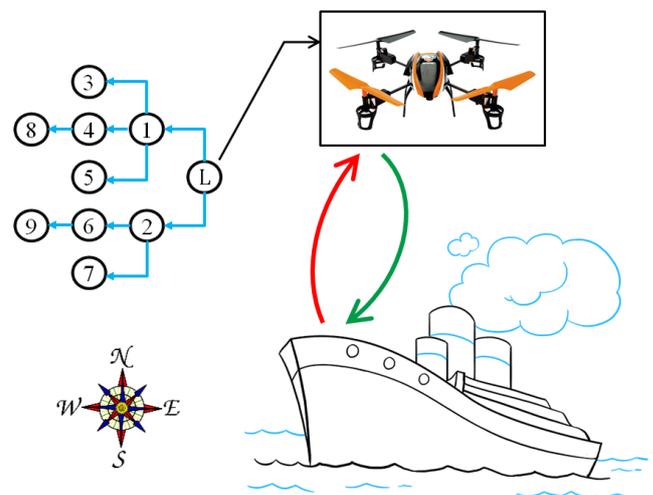


Fig. 4: Arrival of the swarm of drones for landing

In Fig. 4, the reference commands of a moving vessel are continuously sent to the leader drone as a feedback (red arrow line), outlining a tracking phenomenon. The local control unit of the leader then computes the values under its vicinity and generates the force in order to stabilize its landing movement (green arrow line). This process continuous until the desired goal is achieved. Since, the designed formation is based on a leader–follower tightly coupled approach therefore, all the followers track their corresponding leaders, which can minimize the overall computation time of path planning, indicating the fast decision making within the swarm. A filter block (see Appendix) is included in the altitude of the leader drone to slow down its speed on a close arrival by avoiding sudden hit on the surface.

For the controlled movement of each quadcopter in the swarm, the initial step is to construct a balanced drone in the presence of uncertainties and external disturbances with an adaptive computing platform. For this study, a standard LQR with integral action technique has been adapted [36]. Based on the linear model in Equation (10), LQR is a way of finding an optimal

full state feedback controller for each quadcopter. Fig. 5 shows the decision-making process of a drone that is split into two feedback loops, i.e., inner and outer loops. The inner loop is the full state feedback system and the outer loop is responsible for the x , y , and z positions, generating the thrust T and the torques M_i .

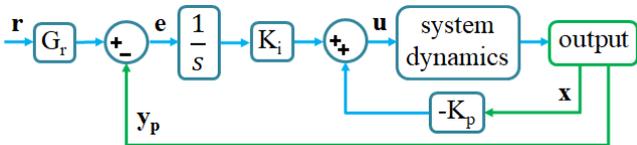


Fig. 5: Block diagram of the control design

The control input \mathbf{u} minimizes the quadratic cost function

$$J(\mathbf{u}) = \int_0^{\infty} (\dot{\mathbf{x}}_a^T Q \dot{\mathbf{x}}_a + \dot{\mathbf{u}}^T R \dot{\mathbf{u}}) dt \quad (11)$$

where Q and R are known as the weight matrices (see Appendix), and \mathbf{u} is given in Equation (6). The Q matrix is a positive semi-definite that defines the weights for the states, whereas the R matrix is a positive definite that indicates the weights of the control inputs. The controller can be tuned by changing the entries in the Q and R matrices to get the desired response. LQR method returns the solution S of the associated Riccati equation

$$(A)^T S + SA - SBR^{-1}(B)^T S + Q = 0 \quad (12)$$

for $S = S^T > 0$. The optimal gain matrix K is derived from S as $K = R^{-1}(B)^T S$. The four control inputs are generated for thrust T , M_x (along x-axis i.e., roll ϕ), M_y (along y-axis i.e., pitch θ), and M_z (along z-axis i.e., yaw ψ) using state feedback law,

$$\mathbf{u} = \frac{K_i}{s} \mathbf{e} - K_p \mathbf{x}, \quad (13)$$

where $\mathbf{e} = r - y_p$, $\mathbf{r} = [x_r \ y_r \ z_r]^T$, \mathbf{y}_p is given in Equation (8), K_i is the integral gain, and K_p is the state feedback gain.

RESULTS

The landing of the swarm of drones on a movable surface (can be defined as a vessel) that has continuous speed with oscillations, which is moving from south-west to north-east, is considered with a smallest margin of error. The simplest model of a movable surface V is defined as a ramp function with a slope of $0.5t$. The oscillations of a movable surface are defined as sine wave with amplitude of 1m, and frequency of 0.1rad/sec. Since the swarm is arranged in a tightly coupled leader–follower hierarchical formation, the reference signal of the leader drone is available to its immediate follower(s). Thus, the followers track the output of the leader with set distance. The initial time

$t = 0$ s while the landing occurs at time $t = 15$ s, are set in the references of the leader drone. The reference positions of the leader drone are $x = y = 2$, and $z = 10 \rightarrow 0$ with step time $t = 15$ s. The initial launching position x of each drone is set to 7m away from its respective neighbouring node(s), and the further data for simulation is shown in Table II. Simulations in Simulink[®] MATLAB are used for the evaluation of the proposed method. In all simulations, the sampling time t_s of 0.01s is used for all the figures.

TABLE II: Initial positions (m) and offsets (m) of drones used in simulation

Drones	Symbol	Initial Position (x, y, z) [*]	Offset (x, y, z) [*]
Leader	L	(0, 0, 10)	–
Follower 1	f1	(7, 0, 10)	(9, 0, 0)
Follower 2	f2	(–7, 0, 10)	(–5, 0, 0)
Follower 3	f3	(14, 0, 10)	(9, 0, 0)
Follower 4	f4	(21, 0, 10)	(16, 0, 0)
Follower 5	f5	(28, 0, 10)	(23, 0, 0)
Follower 6	f6	(–14, 0, 10)	(–5, 0, 0)
Follower 7	f7	(–21, 0, 10)	(–12, 0, 0)
Follower 8	f8	(35, 0, 10)	(16, 0, 0)
Follower 9	f9	(–28, 0, 10)	(–12, 0, 0)

The landing mechanism on a movable surface is shown in Fig. 6(a) and (b). The orientation of the surface is available as a feedback at the leader drone that resulted in accurate landing with negligible errors. Hence, each follower track the reference commands that are defined by its respective leader and the pre-specified formation. To avoid collisions in the swarm, there is a gap of 7m in x positions for all the drones with their neighbouring peer(s). Therefore, it is evident that the landing of all the drones is occurred in a straight line with different x positions. Furthermore, the total kinetic energy, KE , produced by the swarm due to its motion versus the total stored potential energy, PE is described in Fig. 6(c). Total energy possessed and held in the swarm are calculated as $KE = 0.5mv^2$ and $PE = mgh$ respectively. These energies relate how much work is conserved in the process of the swarm movement.

The trajectory errors, $\{e_x, e_y, e_z\}$, between the orientation of the movable surface V and the drones are illustrated in Fig. 7(a), (b), and (c). The trajectory error e_z is sometimes positive in Fig. 7(c) because z position is different depending on y position, and the drones land at different positions and/or time instances.

CONCLUSION

This paper addressed one of the interesting challenges in employment of swarming drones, namely landing softly/safely on a movable surface. More specifically, a setup is considered where a swarm of ten drones in a hierarchical leader–follower formation aims at land-

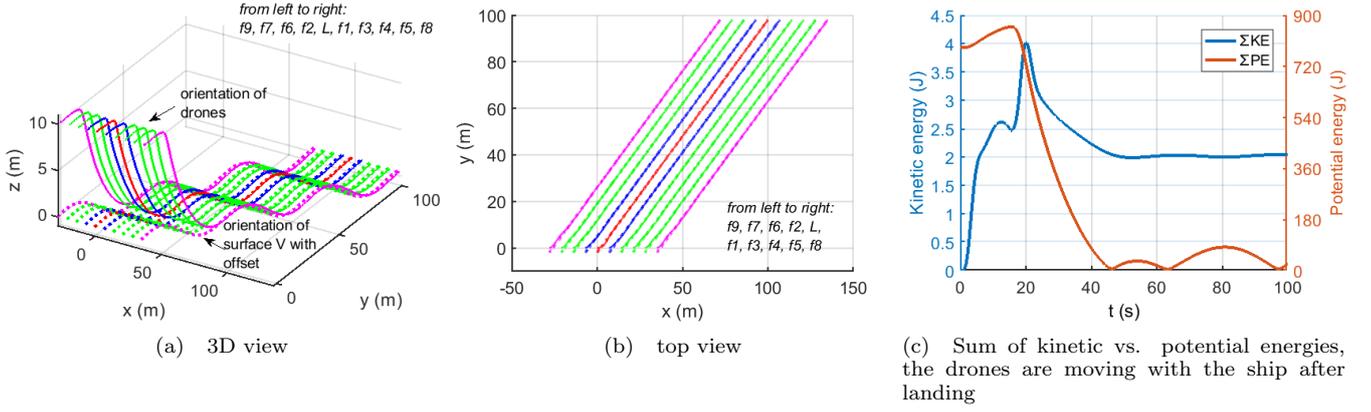


Fig. 6: Landing placements of swarming drones

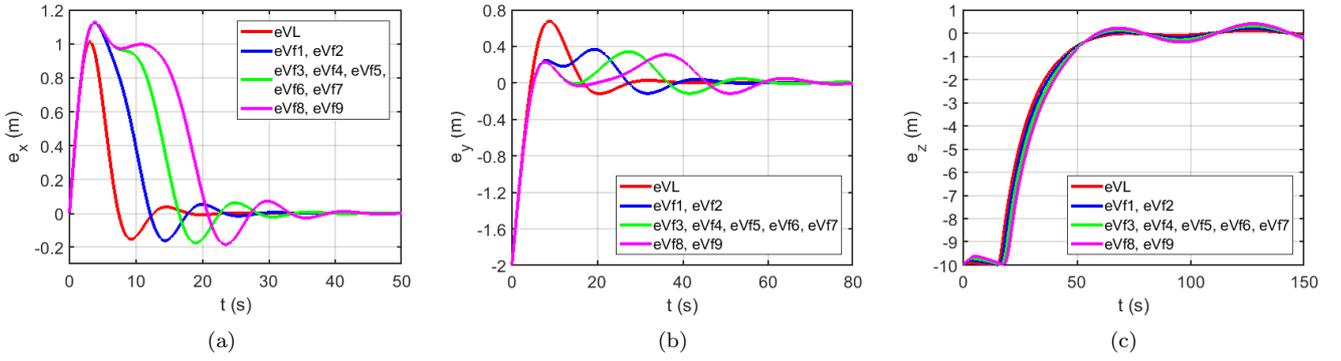


Fig. 7: Trajectory errors between movable surface and corresponding drone

ing on a moving vessel that has a linear speed under the effect of oscillations. In the proposed distributed control system design, each drone in the swarm has a local controller based on an LQR with integral action technique that is an optimal control method providing the smallest possible error to its input. To avoid any collisions among drones in the swarm, a safe travel distance strategy using offsets is employed in the overall system. In the considered scenario, each drone is already at a specific altitude and from there it lands. The swarm is composed of tightly coupled agents, where each drone in the swarm is directly communicating, using shared coordinate variables, with its immediate associate. Therefore, the leader of the swarm is responsible for the execution of the path planning algorithm. It takes the translational measurements of the movable surface as a feedback in order to generate the landing coordinates. It is evident from the simulation results that the proposed system guarantees the convergence of the desired landing missions on the movable surface while minimizing the possibilities for landing errors. The other key advantages of the proposed method are its robustness and scalability. Furthermore, It is understandable from the graphs that the control strategy permits the intuitive execution of an extensive variety of the swarm behaviours.

APPENDIX

$$Q = \text{diag} \left(\begin{array}{c} 0.0885, 4.6064e - 04, 6000, 1080, 1080, \\ 1080, 180, 180, 180, 0.0147, 7.6773e - 05, \\ 1000, 0.4423, 0.0023, 30000 \end{array} \right)$$

$$R = I_4$$

Filter block $G_r = \{G_{x_r}, G_{y_r}, G_{z_r}\}$. For the leader drone, $G_{x_r} = G_{y_r} = 1$, and $G_{z_r} =$ state-space model in which $A = -1/12$, $B = 1/12$, $C = 1$, $D = 0$, and initial conditions = 10. For all the other drones, $G_r = \{G_{x_r}, G_{y_r}, G_{z_r}\} = \{1, 1, 1\}$

ACKNOWLEDGEMENTS

This work has been supported in part by the Academy of Finland, project no. 314048.

REFERENCES

- [1] G. Collins, D. Twining, and J. Wells, "Using vessel-based drones to aid commercial fishing operations," in *OCEANS 2017 - Aberdeen*, 2017, pp. 1–5.
- [2] H. H. Seck. This unmanned rolls royce ship concept could launch drone choppers.
- [3] J. Karpowicz. 4 ways drones are being used in maritime and offshore services.
- [4] J. Hines. The use of drones in shipping and cover implications.
- [5] M. H. Frederiksen and M. P. Knudsen, "Drones for offshore and maritime missions: Opportunities and barriers," in *Denmark: Centre for Integrative Innovation Management*, 2018.

- [6] A. Tahir, J. Böling, M.-H. Haghbayan, H. T. Toivonen, and J. Plosila, "Swarms of unmanned aerial vehicles – a survey," *Journal of Industrial Information Integration*, vol. 16 (100106), 2019.
- [7] Martek marine named on world's biggest ever €67m maritime drone contract - martek aviation. [Online]. Available: <https://www.martekuas.com/martek-marine-awarded-place-remotely-piloted-aircraft-systems-framework-contract-european-maritime-safety-agency/>
- [8] Drones in the deep: new applications for maritime uavs. [Online]. Available: <https://www.ship-technology.com/features/drones-deep-new-applications-maritime-uavs/>
- [9] W. Kong, D. Zhou, D. Zhang, and J. Zhang, "Vision-based autonomous landing system for unmanned aerial vehicle: A survey," in *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, 2014, pp. 1–8.
- [10] A. Gautam, P. B. Sujit, and S. Saripalli, "A survey of autonomous landing techniques for uavs," in *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2014, pp. 1210–1218.
- [11] O. Araar, N. Aouf, and I. Vitanov, "Vision based autonomous landing of multirotor uav on moving platform," *Journal of Intelligent Robotic Systems*, vol. 85, no. 2, p. 369–384, 2017.
- [12] M. Meingast, C. Geyer, and S. Sastry, "Vision based terrain recovery for landing unmanned aerial vehicles," in *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, vol. 2, 2004, pp. 1670–1675.
- [13] S. Huh and D. H. Shim, "A vision-based automatic landing method for fixed-wing uavs," *Journal of Intelligent and Robotic Systems*, vol. 57, p. 217–231, 2010.
- [14] G. Xu, X. Chen, B. Wang, K. Li, J. Wang, and X. Wei, "A search strategy of uav's automatic landing on ship in all weathe," in *2011 International Conference on Electrical and Control Engineering*, 2011, pp. 2857–2860.
- [15] J. Park, Y. Kim, and S. Kim, "Landing site searching and selection algorithm development using vision system and its application to quadrotor," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 2, pp. 488–503, 2015.
- [16] T. K. Venugopalan, T. Taher, and G. Barbastathis, "Autonomous landing of an unmanned aerial vehicle on an autonomous marine vehicle," in *2012 Oceans*, 2012, pp. 1–9.
- [17] T. Templeton, D. H. Shim, C. Geyer, and S. S. Sastry, "Autonomous vision-based landing and terrain mapping using an mpc-controlled unmanned rotorcraft," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 1349–1356.
- [18] H. J. Kim, M. Kim, H. Lim, C. Park, S. Yoon, D. Lee, H. Choi, G. Oh, J. Park, and Y. Kim, "Fully autonomous vision-based net-recovery landing system for a fixed-wing uav," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 4, pp. 1320–1333, 2013.
- [19] P. Serra, R. Cunha, T. Hamel, D. Cabecinhas, and C. Silvestre, "Landing of a quadrotor on a moving target using dynamic image-based visual servo control," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1524–1535, 2016.
- [20] S. Saripalli, J. F. Montgomery, and G. S. Sukhatme, "Vision-based autonomous landing of an unmanned aerial vehicle," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 3, 2002, pp. 2799–2804.
- [21] B. Hérisse, T. Hamel, R. Mahony, and F.-X. Rusotto, "Landing a vtol unmanned aerial vehicle on a moving platform using optical flow," *IEEE Transactions on Robotics*, vol. 28, no. 1, p. 77–89, 2012.
- [22] D. Lee, T. Ryan, and H. J. Kim, "Autonomous landing of a vtol uav on a moving platform using image-based visual servoing," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 971–976.
- [23] A. Tahir, "Autonomous swarming drones — landing a swarm of quadcopters on a vessel." Master's thesis, NOVIA University of Applied Sciences, Turku, Finland, December 2019.
- [24] K. Kanjanawanishkul, "Formation control of mobile robots: Survey," *UBU Engineering Journal*, vol. 4, no. 1, p. 50–64, 2011.
- [25] F. Šolc, "Modelling and control of a quadcopter," *Advances in Military Technology*, vol. 5, no. 2, p. 29–38, 2010.
- [26] P. Wang and F. Hadaegh, "Coordination and control of multiple microspacecraft moving in formation," *Journal of the Astronautical Sciences*, vol. 44, no. 3, p. 315–355, 1996.
- [27] D. Galzi and Y. Shtessel, "Uav formations control using high order sliding modes," in *2006 American Control Conference*, 2006, p. 4249–4254.
- [28] B. Yun, B. Chen, K. Lum, and T. Lee, "Design and implementation of a leader-follower cooperative control system for unmanned helicopters," *Journal of Control Theory and Applications*, vol. 8, no. 1, p. 61–68, 2010.
- [29] M. Lewis and K. Tan, "High precision formation control of mobile robots using virtual structures," *Autonomous Robots*, vol. 4, no. 4, p. 387–403, 1997.
- [30] T. Paul, T. Krogstad, and J. Gravdahl, "Modelling of uav formation flight using 3d potential field," *Simulation Modelling Practice and Theory*, vol. 16, no. 9, p. 1453–1462, 2008.
- [31] Z. Chao, S. Zhou, L. Ming, and W. Zhang, "Uav formation flight based on nonlinear model predictive control," *Mathematical Problems in Engineering*, vol. 2012, p. 1–15, 2012.
- [32] T. Balch and R. C. Arkin, "Behavior-based formation control for multirobot teams," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 6, p. 926–939, 1998.
- [33] J. R. T. Lawton, R. W. Beard, and B. J. Young, "A decentralized approach to formation maneuvers," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 6, p. 933–941, 2003.
- [34] D. Bennet and C. McInnes, "Verifiable control of a swarm of unmanned aerial vehicles," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 223, no. 7, p. 939–953, 2009.
- [35] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, p. 424–440, 2015.
- [36] A. Tahir, J. Böling, M.-H. Haghbayan, and J. Plosila, "Comparison of linear and nonlinear methods for distributed control of a hierarchical formation of uavs," *IEEE Access*, DOI: 10.1109/ACCESS.2020.2988773.

Machine Learning for Big Data

A NOVEL OVERSAMPLING TECHNIQUE TO HANDLE IMBALANCED DATASETS

Ayat Mahmoud
Faculty of Computer Sciences
October University for Modern Sciences
and Arts
Cairo, Egypt
E-mail: eng.ayat@gmail.com

Ayman El-Kilany
Information Systems Department
Faculty of Computers and
Information, Cairo University
Cairo, Egypt
Email: a.elkilany@fci-cu.edu.eg

Farid Ali
Information Technology Department
Faculty of Computers and Information
Beni-suef University, Egypt
E-mail: farid.cs@gmail.com

Sherif Mazen
Information Systems Department
Faculty of Computers and
Information, Cairo University
Cairo, Egypt
Email: s.mazen@fci-cu.edu.eg

KEYWORDS

Imbalance, Oversampling, Classifier.

ABSTRACT

With the amount of data is growing extensively in different domains in the recent years, the data imbalance problem arises frequently. A dataset is called imbalanced when the data of a certain class has significantly more instances than that of other classes of the same dataset. This imbalanced nature of the data negatively affects the performance of a classifier since misclassification of data may cause data analysis results to be inaccurate and hence leads to wrong business decisions. This paper presents a study of the different techniques that are used to handle the imbalanced dataset, and finally proposes a novel oversampling technique to tackle the binary classification of imbalanced dataset problem.

INTRODUCTION

In today's world of internet, there is huge amounts of data generated every day. Therefore, it becomes important to advance the deep understanding of knowledge discovery (KD) and analysis of raw data to support decision-making in businesses. An evolution has been done on classification of data through the learning process. The problem gets more complex when the dataset is imbalanced. A dataset is said to be imbalanced if the class distribution is not uniform (Fernández et al., 2017). In this situation, there are instances from one class are higher in number than the other class. The class having a greater number of samples is named "majority class", and the class having relatively a smaller number of instances is named "minority class".

Researches in the field of machine learning proved that using an uneven distribution of class instances can cause a bias in the performance of the used learning algorithm (Herland et al., 2018). In other words, the classifier

gives high accuracy on the majority class, while giving poor accuracy on the minority class. This happens because traditional training measures such as the overall success are inclined by the larger number of instances from the majority class. In many real-world applications, the minority classes are more important as in cancer diagnosis in the medical field applications. That is why classifying the imbalanced datasets has a growing attention from both academia and industry (Chu et al., 2016).

Most of the traditional methods of machine learning had limitations when applied to imbalanced datasets. They do not work well for imbalanced data classification because they assume equal costs for each class. Therefore, the data classification results may be biased and inaccurate. The reason is that traditional machine learning algorithms aimed at enhancing accuracy by reducing the error without taking into consideration the class distribution or balance (Zhang et al., 2018). Several solutions to the misclassification of imbalanced datasets problem were previously proposed in many researches like oversampling, undersampling and cost-sensitive learning.

Classifiers encounter imbalanced distribution in many real-world applications. For example, the number of legal credit card transactions is significantly greater than that of illegal transactions. Most medical fields have similar conditions, where the number of patients needing special care (e.g. rehabilitation or treatment) is significantly lower than the number of those who don't. In many other fields, such as oil deposits identification in satellite images (Cai et al., 2018), there were also class imbalances.

The common problem of these real-life applications is that every class contains completely different number of instances. Since traditional classifiers were designed to get higher accuracy regardless of the distribution of classes, they gave a less attention to the minority class and therefore caused a noticed bias towards the majority class.

In this paper, we argue that oversampling techniques can yield better results to handle imbalanced dataset problem if the majority data was considered during the oversampling process. The proposed technique tackles the imbalanced dataset problem by using a sample of the majority data to create a better new sample of minority data. The evaluation results show that such oversampling technique outperform the standard oversampling algorithm.

The rest of this paper is organized as follows: Section 2 contains the related work and explains the basic techniques that used to solve the problem of imbalanced dataset. In Section 3, we introduce our proposed technique to handle imbalanced data problem while Sections 4 and 5 show the details of the performance evaluation and conclusion remarks.

RELATED WORK

There are three main approaches to tackle imbalanced data problems. Data-level methods -also called external methods- that work on the data in a way to adapt the number of data instances to more balance the distribution. On the other hand, the algorithm-level methods (also called internal methods) that adapt the traditional algorithms of learning to minimize the bias, increase accuracy, and get benefit of mining data that have skew distributions. Hybrid methods combine both data and algorithm level methods. All approaches are summarized in Figure1 and further explained in the next sections.

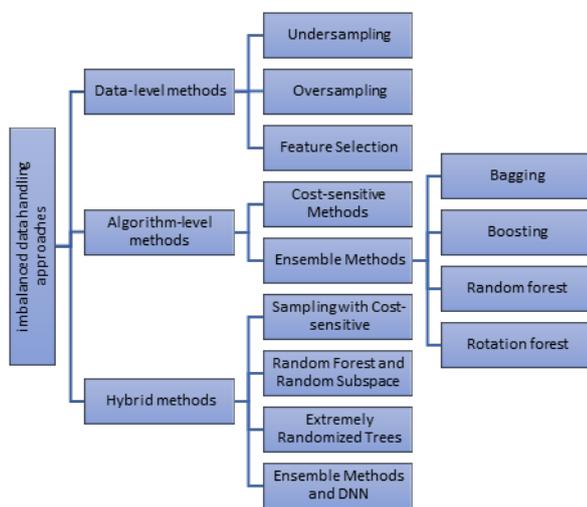


Figure 1: Taxonomy of Imbalanced Data Learning Approaches

Data-level Methods

In data-level methods, the goal is to modify the dataset to make it more suitable to apply a traditional learning algorithm. Three sub-approaches are used to modify datasets, undersampling, feature selection, and oversampling. Undersampling is to remove samples from majority class whereas oversampling is to generate new objects for minority class (Devi & Purkayastha,

2017; Ha & Lee, 2016; Ng et al., 2014). Feature selection means the algorithms that output a subgroup of the input feature set that are more relevant and help a classifier to enhance its performance (Pant & Srivastava, 2015). Traditional approaches use random techniques to select the target samples for sampling (Lin et al., 2017). But this frequently leads to exclusion of important samples or appearance new samples that are meaningless. Consequently, more adapted approaches were proposed to try to maintain structure of classes and generate new data samples that conform to the original distribution. These new adapted algorithms contain also some methods for cleaning the overlapping samples and removing dirty samples that may affect learning process in a negative way (Hu et al., 2015).

In contrast to undersampling, the oversampling adds synthetic samples to the minority class with the aim of balancing the distribution of the classes (Abdi & Hashemi, 2015). The simplest method of oversampling is replication of instances of minority class (Liu et al., 2007). This reduces the class imbalance but on the other hand, it can cause the problem of overfitting. SMOTE is a fundamental oversampling approach that uses data synthesis. SMOTE algorithm implements an oversampling approach to rebalance the training dataset. As an alternative to applying a simple duplication of the minority class examples, the main idea of SMOTE is to generate synthetic instances. This novel data instance is created by interpolation between several minority class instances that are inside a well-defined neighborhood. Therefore, the process is said to be concerned about the feature space instead of on the data space, in other words, the algorithm is based on the values of the features and their relationship, in place of considering the data points as a whole (Sáez et al., 2016).

Sampling is the most widely used approach to deal with the problem of imbalanced datasets. The sampling of data turns the unbalanced distribution into a better class distribution. This is achieved by generating data samples or by deleting them. As stated before, there are two key techniques for sampling namely oversampling and undersampling (Bunkhumpornpat & Sinapiromsaran, 2017). The benefits and drawbacks of each will be discussed here.

There are considerable drawbacks of oversampling that can lead to the issue of overfitting, to that the time required to create the classifier or even to harm the learning process. Undersampling approach makes the balancing by the removing class samples. While the particular space can be defined, data loss can be caused by the reducing data size. The bias in the datasets, which influences on minority groups more than a majority, is also an important issue for sampling. Researchers should understand the scope of the problem being tackled and the appropriate classifier for this situation. When supplemented by sampling methods, several classifiers achieve better performance.

Algorithm-level Methods

They modify traditional learning algorithms to lessen the bias found towards the majority class. To achieve this, a good understanding of the learning algorithm is needed, and a clear analysis of reasons for its failure in learning from imbalanced datasets. The most prevalent division is cost sensitive approaches. In cost sensitive approaches the traditional learning algorithm is adapted to include varying penalty for each of class of samples (Khan et al., 2017). This is done by giving a higher cost to the group of instances that is less represented in the dataset. We increase its significance through the mining process (Cheng et al., 2017).

While the approaches of sampling try to achieve more balanced dataset, considering the representativeness of the class instances in the data, cost-sensitive methods take into account the cost of misclassified samples (Khan et al., 2017). Cost-sensitive learning addresses the problem of imbalanced datasets by using different cost formulae that assign some cost for a particular data sample (Sáez et al., 2016).

There are a lot of methods to ensemble algorithms such as bagging, boosting, random forest and rotation forest. Till now several approaches have been developed and improvements to traditional methods have been designed to solve the issue of imbalanced distributions (Cai et al., 2018). Bagging, is a machine learning approach that is used to improve accuracy while reducing variance in classifying samples. Boosting means that poor classifications can be combined to create a more correct decision. That is to say, boosting means a number of algorithms that use weights to make weak learners more accurate. Unlike bagging that has run separately from each classifier and at last merges the output without any classifier being preferred.

Hybrid Methods

Sampling-based Approaches with Cost-Sensitive Learning

In these methods, a preprocessing is done to the data samples with imbalanced distribution. This is done by using over or undersampling at first, and then using cost-sensitive approach. Some remarkable researches of this area are Akbani et al. and L'opez et al. (Akbani et al., 2004; López et al., 2012).

Random Forest and Random Subspace Methods

Random trees are still growing, mostly because of their flexibility and good performance. The random forest is treated as a simple to tune technique, unlike other techniques (e.g. GBM) which require careful tuning. Random forests utilize a large number of integrated decision trees.

Extremely Randomized Trees

Extremely random trees (Geurts et al., 2006) use randomness in the training stage in order to produce different sets. In addition to the random subgroup of attributes that choose the most distinctive feature,

defining attributes are randomized when extremely random trees are applied.

Ensemble Methods and Deep Neural Networks

In the area of machine learning, DNNs are considered nowadays a major force. In many fields such as speech recognition and object detection and many other fields, DNN has improved dramatically in the last few years (LeCun et al., 2015). DNNs are made up of many layers of un-linear procedures. They allow us to examine complicated patterns, and if used with big data, it can learn high-level concepts.

The research in (Batista et al., 2004) presented a respectable study of sampling methods. Several strategies of over and under sampling and dynamic / hybrid processes have been tested and examined carefully on thirteen datasets. While most of them had improved performance, in all experimental datasets there has been no method overwhelming others. The experiment findings have revealed that random over-sampling yielded great findings relative to more complicated approaches. Additional research argued that the improvements made by undersampling and oversampling approaches have a greater impact on the highly imbalanced distribution datasets (Japkowicz & Stephen, 2002).

PROPOSED TECHNIQUE

The proposed technique main focus is to generate new synthetic minority samples that lessen the difference in number between majority and minority data. Towards this goal, majority data samples are considered while generating the new minority data samples. The detailed steps of the proposed technique is described in Algorithm 1.

ALGORITHM 1: Proposed Technique (MODIFIED SMOTE)

```
1: Function Modified_SMOTE (K,N,A,B,T,R)
   Input: K,N,A,B,T,R where K:#neighbors,
   N:Percentage of required oversampling,
   A:Majority class samples, B:Minority class
   samples, T:# Minority samples, R:#iterations
   Output: Original Data + (N/100) * Minority class
   samples
2: If N<100 then
3:   N=100
4: End if
5: For i=1 to T do
6:   y← B(i)
7:   Get k-nearest neighbors of y from A along
   with their distances
8:   minA← the nearest neighbor of A to y
9:   Get k-nearest neighbors of y from B along
   with their distances
10:  minB← the nearest neighbor of B to y
11:  x← Randomly select one of the nearest
   neighbors of y from B
12:  dist_x_minA← calculate difference in distance
   between x and minA
```

```

13:  dist_x_minB← calculate difference in distance
    between x and minB
14:  While r<R do
15:    For j=1 to T do //loop to generate samples
16:      s(j) ← x +(minA - minB)*α
17:    End for
18:    dist_x_s←calculate distance between s
    and x
19:    If dist_x_s < dist_x_minA and
    dist_x_s > dist_x_minB then
20:      Go to 24
21:    End if
22:    r = r +1
23:  End while
24:  synth ← concatenate s with B
25: End for
26: Write file
27: End

```

The algorithm starts by considering samples in minority class. For each sample, we get the k-nearest neighbors of it from the majority class and also the k-nearest neighbors from minority class. We randomly select one of the minority neighbors and calculate the distance between it and both the nearest majority neighbor and the nearest minority neighbor multiplied by random number. We then generate the new synthetic sample depending on the randomly selected minority neighbor adding to it the difference in distance between the nearest majority neighbor and the nearest minority neighbor. Before writing the new generated sample to the generated samples file, we calculate the difference in distance between it and the randomly selected minority neighbor to make sure that it is already in the area between the closest majority neighbor and the closest minority neighbor.

PERFORMANCE EVALUATION

The objective of the evaluation is to prove the effectiveness of the proposed oversampling technique which considers the majority data samples while generating new minority data samples. Towards this goal the proposed technique is evaluated on different datasets over different classifiers against SMOTE method which is the standard oversampling approach in the literature. The following subsections describes the evaluation details.

Datasets

For our experiment, we used three numerical datasets which can be downloaded from (Alcalá-Fdez et al., 2011). The first dataset is Poker dataset which contains 1485 objects for 2 classes, the first class is 25 (values 1) and the second class is 1460 (value 0). Each object contains 11 attributes. The second dataset is Yeast which contains 514 objects for 2 classes, the first class is 51 (values 1) and the second class is 463 (value 0). Each object contains 9 attributes. The third dataset is Cleveland which contains 173 objects for 2 classes, the first class is 13 (values 1) and the second class is 163

(value 0). Each object contains 9 attributes. For each dataset, cross validation procedure was used to split the data into training and testing sets. The number of cross validation folds was set to 10 folds. Details of the datasets is summarized in Table 1.

Table 1: Number of Samples in each Dataset Before and After Oversampling

	Poker		Yeast		Cleveland	
	Before	After	Before	After	Before	After
Minority	25	650	51	459	13	169
Majority	1460	1460	463	463	160	160

Evaluation Method

Six evaluation matrices were used for performance evaluation, which are, accuracy, sensitivity, specificity, precision, f-score, and error. Sensitivity measures the proportion of actual positives that are correctly identified as such, while Specificity measures the proportion of actual negatives that are correctly identified as such. Sensitivity, therefore, quantifies the avoidance of false negatives and specificity does the same for false positives. Evaluation metrics were obtained after training of three different classifiers on the three datasets in different settings. Classifiers were trained once on the imbalanced datasets without applying oversampling and once trained on the datasets after applying traditional SMOTE algorithm for oversampling and finally trained on the datasets after applying the proposed technique for oversampling. Three different classifiers were used which are K-Nearest Neighbors, Fuzzy K-Nearest Neighbors and Support Vectors Machines classifiers. The KNN classifier takes only one parameters and the best results were when k= 10 While the FKNN classifier takes two parameters k and m, with k=10 and m=0.5.

Evaluation Results

Results of the proposed technique against SMOTE algorithm are summarized in Figures 2, 3, and 4.

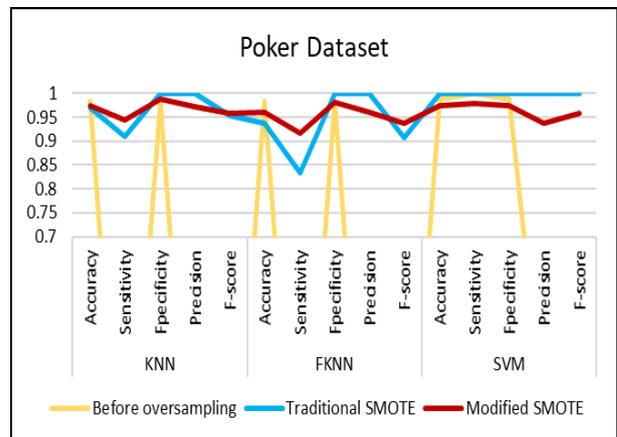


Figure 2: The Evaluation Metrics of Poker Dataset with the Three Classifiers

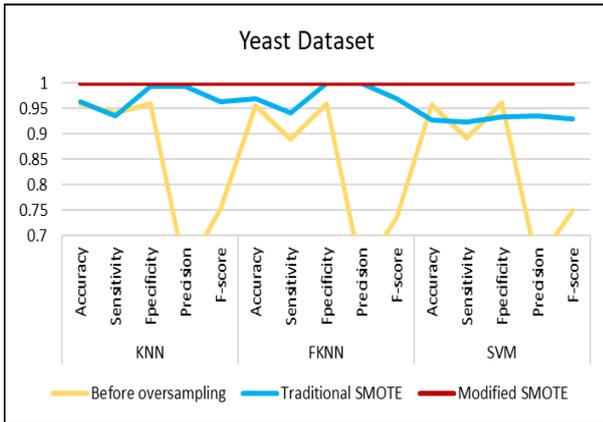


Figure 3: The Evaluation Metrics of Yeast Dataset with the three Classifiers

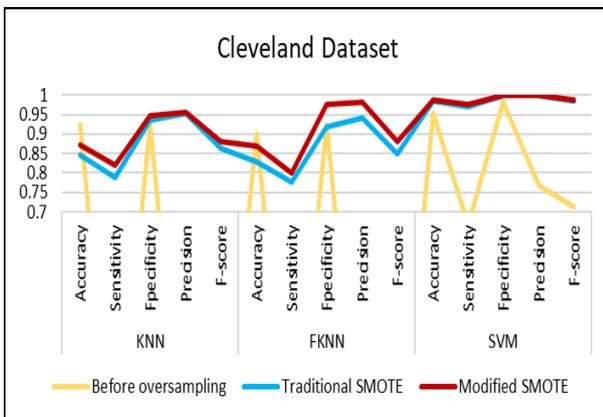


Figure 4: The Evaluation Metrics of Cleveland Dataset with the Three Classifiers

We can see that the proposed technique got almost better performance in all three datasets with different classifiers. We may notice that before oversampling, the classifiers showed high value in the accuracy metric as you see in Figure 2,3 and 4. Generally speaking, the accuracy metric measures the fraction of all instances that are correctly categorized. But here, this metric value is fake if used alone for evaluation (Akosa, 2017). As due to the imbalance of data, the classifiers tend to get all majority samples as correct and all minority samples as incorrect. The other metrics especially precision which is low explains this. The outperformance of the proposed technique against SMOTE algorithm may be related to the difference between how the two methods work. While SMOTE algorithm generates new samples in the space of the minority data space, the Modified SMOTE algorithm generates new samples in the space separating minority and majority data. Those new samples were able to explain the difference between majority and minority points in a better way and consequently lead to better classifications results after using them during training.

CONCLUSION AND FUTURE WORK

The paper explored the nature of the imbalanced data and its current real-life applications. We provided a taxonomy for the solutions found in the literature. Then, we presented a comparative study for the efforts done with the aim of addressing the challenge of the classification of imbalanced data. At last, we introduced our proposed technique for handling the imbalanced data problem along with our experiment which showed a noticeable higher performance results. In future, we aim to further apply it to categorical datasets as we applied it with only numerical ones. Another direction is to apply our approach to multiclass datasets.

REFERENCES

- Abdi, L. Hashemi, S. (2015). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering*, 28(1), 238-251.
- Akbani, R., et al. (2004). *Applying support vector machines to imbalanced datasets*. Paper presented at the European conference on machine learning.
- Akosa, J. (2017). *Predictive accuracy: a misleading performance measure for highly imbalanced data*. Paper presented at the Proceedings of the SAS Global Forum.
- Alcalá-Fdez, J., et al. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.
- Batista, G. E., et al. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.
- Bunkhumpornpat, C. Sinapiromsaran, K. (2017). DBMUTE: density-based majority under-sampling technique. *Knowledge and Information Systems*, 50(3), 827-850.
- Cai, T., et al. (2018). Breast cancer diagnosis using imbalanced learning and ensemble method. *Applied and Computational Mathematics*, 7(3), 146-154.
- Cheng, F., et al. (2017). Large cost-sensitive margin distribution machine for imbalanced data classification. *Neurocomputing*, 224, 45-57.
- Chu, X., et al. (2016). *Data cleaning: Overview and emerging challenges*. Paper presented at the Proceedings of the 2016 International Conference on Management of Data.
- Devi, D. Purkayastha, B. (2017). Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance. *Pattern Recognition Letters*, 93, 3-12.
- Fernández, A., et al. (2017). An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2), 105-120.
- Geurts, P., et al. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3-42.
- Ha, J. Lee, J.-S. (2016). *A new under-sampling method using genetic algorithm for imbalanced data classification*. Paper presented at the Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication.
- Herland, M., et al. (2018). Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(1), 29.

- Hu, Y., et al. (2015). An improved algorithm for imbalanced data and small sample size classification. *Journal of Data Analysis and Information Processing*, 3(03), 27.
- Japkowicz, N. Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5), 429-449.
- Khan, S. H., et al. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8), 3573-3587.
- LeCun, Y., et al. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Lin, W. C., et al. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409, 17-26.
- Liu, A., et al. (2007). *Generative Oversampling for Mining Imbalanced Datasets*. Paper presented at the DMIN.
- López, V., et al. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7), 6585-6608.
- Ng, W. W., et al. (2014). Diversified sensitivity-based undersampling for imbalance classification problems. *IEEE transactions on cybernetics*, 45(11), 2402-2412.
- Pant, H. Srivastava, R. (2015). A survey on feature selection methods for imbalanced datasets. *International Journal of Computer Engineering and Applications*, 9(2).
- Sáez, J. A., et al. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57, 164-178.
- Zhang, C., et al. (2018). A cost-sensitive deep belief network for imbalanced classification. *IEEE transactions on neural networks and learning systems*, 30(1), 109-122.

MODELLING INTERLEAVED ACTIVITIES USING LANGUAGE MODELS

Eoin Rogers, Robert J. Ross, John D. Kelleher
Applied Intelligence Research Centre
Technological University Dublin
Dublin, Ireland

eoin.rogers@tudublin.ie, robert.ross@tudublin.ie, john.d.kelleher@tudublin.ie

KEYWORDS

Activity discovery ; Activity Recognition ; Interleaving ; Neural language modelling ; Behaviour modelling

ABSTRACT

We propose a new approach to activity discovery, based on the neural language modelling of streaming sensor events. Our approach proceeds in multiple stages: we build binary links between activities using probability distributions generated by a neural language model trained on the dataset, and combine the binary links to produce complex activities. We then use the activities as sensor events, allowing us to build complex hierarchies of activities. We put an emphasis on dealing with interleaving, which represents a major challenge for many existing activity discovery systems. The system is tested on a realistic dataset, demonstrating it as a promising solution to the activity discovery problem.

INTRODUCTION

Given the increasing ubiquity of computer systems in our everyday lives, using them to model, monitor and analyse human behaviour becomes increasingly possible and useful. The field of *activity recognition* studies the design and implementation of such systems (Kwapisz et al., 2011; Kim et al., 2009), which can be useful for applications as diverse as elder care and security.

One persistent issue facing activity recognition is the difficulty in finding suitably annotated datasets. Labelling such datasets can be time consuming, and in many cases is quite difficult due to differences in levels of abstraction and ambiguities about the precise start and end times of activities. In order to address this, the field of *activity discovery* (AD) has proposed the unsupervised extraction of plausible human activities from unannotated datasets (Gjoreski and Roggen, 2017; Cook et al., 2013; Rogers et al., 2016).

Real-word activities are often *interleaved*, meaning that they take place at the same time. This results in sensor readings for multiple activities showing up in quick succession on the data stream. Recognising that this is happening, separating the activities from each other, and recognising that sensor events that do not

occur adjacent to each other may be part of the same activity is a significant challenge for existing activity discovery systems.

In this paper, we propose a novel approach to activity discovery that makes use of *neural language models*, developed by the natural language processing (NLP) community, to model the sensor feeds from activity discovery systems and extract useful activities from them. This work builds upon a previous related system presented in (Rogers et al., Forthcoming), but with major changes including the discovery of proper, non-binary activities, and clustering of activities into distinct types. Our approach is also designed to be aware of, and disentangle, interleaved activities. This paper is divided into five remaining sections, outlining the activity discovery problem in more detail, covering prior work in the field, a description of our approach, a description of the experiments we ran to test our approach, and a presentation of our results respectively.

ACTIVITY DISCOVERY

In order to have a clean model description, it is necessary for us to briefly introduce the terminology that we will use later. Formally, an activity discovery system can be modelled as a 5-tuple (Σ, D, A, f, g) , where:

- Σ is a set of *event types*;
- D is an ordered sequence of events, $D = \langle d_1, d_2, \dots, d_L \rangle$ of length L , such that each $d_i \in D$ is drawn from the set Σ . We call this the *dataset*;
- A is a set of *activity types*;
- f is a mapping $f : D \rightarrow X^*$, which takes a sequence of events D as input, and returns a set of (possibly non-contiguous) sub-sequences of D as output; and
- g is a mapping $g : X \rightarrow A$, where $X \subset D^*$, which takes a sub-sequence produced by f as input, and returns an activity type $a \in A$ as output.

This definition can be made clearer with a concrete example. Supposing we have a dataset $D = \langle d_1, d_2, \dots, d_N \rangle$. Each $d_i \in \Sigma$ is a sensor event drawn from Σ , our full set of sensor events. In an environment where sensors have been set up in a home, for instance, Σ could consist of events such as *open front door*, *turn oven on*, *flush toilet* and similar domestic events. An *activity*, then, is simply a sub-sequence of D consisting of events that appear to the activity discovery system

to be semantically related. For instance, we would expect that events such as *turn oven on*, *open kitchen cupboard*, *open refrigerator* might occur in an activity together, since they tend to occur together temporally. It should be noted that D is not a set of sequences as might be the case in a supervised learning setting; D is a single large dataset from which we extract activities.

Multiple similar activities can then be clustered or lifted into one *type*. The activity discovery system might notice that an activity similar to the one mentioned in the previous paragraph seems to occur nightly, and may cluster them all into a single *making dinner* activity type. The concrete sub-sequences of D are referred to as the *instances* of the *making dinner* activity.

Note that we don't generally expect an activity discovery system to operate with human-like semantic knowledge or expectations in the basic case. Thus, it would not be expected to be able to name the new activity type as *making dinner*, only to identify that the instances involved can be sensibly clustered together. A commercial activity discovery system might well be supplemented with real-world knowledge, with the intention of biasing towards the sort of activities we would expect to find in the environment in which it operates. For instance, knowledge that events relating to a fridge or oven indicates activities relating to food preparation such as *making dinner* are taking place. In many ways this would stray over into being a form of activity recognition as well as discovery. For this reason, we stick to a pure form of activity discovery without any real-world knowledge. We do still expect to be able to discover *making dinner* as an activity, just not to be able to give it a label (*making dinner*) that would be semantically meaningful to a human observer.

PRIOR WORK

A number of existing approaches to activity discovery exist in the literature. Cook et al. (2013) provides a good overview of the field. This paper also introduces an activity discovery system that applies a beam search algorithm using an operator called *ExtendSequence* to discover activities in an unlabelled dataset. Like a number of other systems in the field, this algorithm utilises the *minimum description length* (MDL) principle (Rissanen, 1978, 1989), which proposes evaluating machine learning models by measuring the degree to which they *compress* their input dataset. This is an important principle, and one which will turn out to be useful to our own work also.

Activity discovery can also be carried out by relatively simple systems that utilise topic models (Huynh et al., 2008). Here, the *latent Dirichlet allocation* (LDA) topic model (Blei et al., 2003) is used to model the relationship between sensor events and latent variables which are presumed to represent activities. The model is shown to have good performance, even on a complex dataset. More recently, other models based on statistical models have been proposed: for example, Fang et al. (2019) proposes activity discovery by means

of a hierarchical mixture model. Saives et al. (2015) propose using activity discovery to build a model of normal behaviour patterns of a person in order to detect anomalous behaviours that may be of interest to medical professionals.

Related fields also provide an important source of ideas. Grammar induction is a concept from computational linguistics which refers to the derivation of grammar productions for a language given only a dataset. Some forms of grammar induction require labelled input, distinguishing positive and negative examples, but others require only positive examples. In the general case, grammar induction is not a tractable problem, regardless of whether the dataset is labelled or not (Gold, 1967), but tractable approximations have been demonstrated which solve the problem to a degree (Cramer, 2007). This problem is by no means equivalent to activity discovery (and is in fact in many ways harder), but it does involve the induction of structure from a one-dimensional input vector. *Adios* (Solan et al., 2005) induces a grammar by loading a dataset into memory as a graph, with words represented as vertexes and sentences represented as directed edges between these. This representation allows for the identification of *equivalence classes* between words and phrases which share the same input and output edges, which can then be added to the graph as nonterminals. A variant of the *Adios* approach, which supplements the basic grammar induction algorithm with logical predicates to allow for more accurate induction in a limited linguistic domain is presented by (Gaspers et al., 2011).

The *eGrids* grammar induction algorithm (Petasis et al., 2004) bears a resemblance to the beam search-based system mentioned previously from (Cook et al., 2013). It also uses an MDL-based objective function to guide the search. An interesting deep learning-based grammar induction model using convolutional networks to determine *syntactic distance* (the degree to which two neighbouring words or symbols belong to the same POS phrase) is similar to the approach we present in this paper (Shen et al., 2017). Finally, our approach can also be understood as a non-local variant of the tree structure induction algorithm *Sequitur* (Nevill-Manning and Witten, 1997), which groups input symbols together even if they do not appear contiguously.

Alshammari et al. (2017) present a 3D simulation of a house that can be used to automatically generate datasets for use in activity discovery, activity recognition and related fields. This could prove useful for validating activity discovery systems.

APPROACH

We will now outline the approach that we have taken to activity discovery. From the perspective of somebody using our system, we assume that the input is a dataset consisting of a finite series of discrete sensor events $D = \langle d_1, d_2, \dots, d_L \rangle$. Each individual sensor event has an associated *type* from a fixed set of types T . We write the type of d_i as $t(d_i) \in T$. The primary out-

put from the system is a series of discovered *activities* $\langle Act_1, Act_2, \dots, Act_N \rangle$, where each activity is a tuple of the form $(Indexes_{Act_j}, Type_{Act_j})$. $Indexes_{Act_j}$ is a set of indexes into the dataset D , $\{x_1, x_2, \dots, x_{ActSize(j)}\}$, of length $ActSize(j)$, which can be different for each activity output by the system. $Type_{Act_j}$ is a *type* associated with the discovered activity, analogous to the type of a sensor event discussed above. Act_j and Act_k may share the same type, but they may not share the same set of indexes.

The basic internal operation of our model proceeds according to the following three steps:

- We build a probabilistic model by analysing the dataset. Given a subset of the dataset $D_{i:i+n-1}$, which we call the *sliding window*, we use the model to predict the probability distribution over sensor events $P(d_{i+n+l}|D_{i:i+n-1})$ for all $l \in \{0, 1, \dots, m-1\}$. We call the subset of the dataset $D_{i+n:i+n+m-1}$ the *lookahead window*, m the *lookahead length*, and l the *lookahead offset*.
- We use the probabilistic model to construct *links* between sensor events if the model is confident that one event can be predicted from the other. Links are grouped together to create the $Indexes_{Act_j}$ part of the output described above.
- Activities are then *clustered* together based on the *similarity of the sensor types present within them*. The clusters are used as the $Type_{Act_j}$ part of the output described above.

A more detailed outline of these stages now follows.

NLP practitioners will of course recognise that the probabilistic model that we describe is a form of *language model*. As is now common in that field, we use a *neural language model* (Bengio et al., 2003) or NLM. We use a modern recurrent design – specifically the LSTM-based (Hochreiter and Schmidhuber, 1997) approach described in (Zaremba et al., 2014) (which is itself adapted from (Graves, 2013)), which allows for the modelling of long-distance dependencies and robustness to noise. The LSTM consists of an input gate i , a forget gate f , and an output gate o . If h_t^l denotes the output of layer l at timestep t , g denotes the *modulated input* and c_t^l denotes the value of an LSTM cell in layer l at time t , activation of the gates is defined as:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} W \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix} + b \quad (1)$$

We then update the value in the cell:

$$c_t^l = f \odot c_{t-1}^l + i \odot g \quad (2)$$

Where \odot denotes the *Hadamard product* (i.e. elementwise multiplication, as opposed to matrix multiplication). The output h_t^l is then:

$$h_t^l = o \odot \tanh(c_t^l) \quad (3)$$

We train m networks, one for each lookahead offset. As a result, one network is responsible for predicting

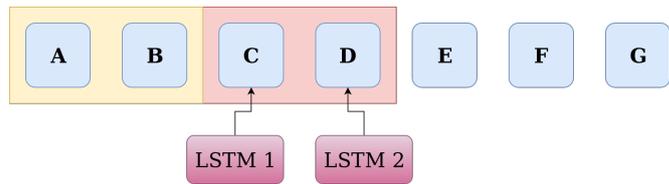


Fig. 1: In this example, events A and B are inside the sliding window, and are fed as input into two LSTMs. Each LSTM has to predict a probability distribution over the corresponding offset within the lookahead window containing C and D.

the first item in the lookahead window, another for predicting the second and so on. This process is illustrated in Fig. 1.

Building links between activities is carried out in two stages: building simple binary links, and grouping the links together to form activities. The binary links are built using the NLM. First we run the networks over the entire dataset. We build a binary link between the final event in the sliding window and the l th event in the lookahead window if the l th LSTM successfully predicts the l th event from the sliding window (for example, between events B and D in Fig. 1 if LSTM 2 predicted event D). Doing this naively would make the system vulnerable to building links between common events. In the NLP community, this approach is usually solved by removing common words (stop words), but this could reduce the quality of the resulting activities discovered.

As a result, we don't work with probabilities directly, but rather with *probability deltas*. Again looking at Fig. 1, this is the average probability that LSTM 2 predicts D when event B is present in the sliding window *minus* the probability when it isn't present. Thus, we only build a link between B and D if the presence of event B makes the LSTM more confident that event D is to follow. The exact calculation we use is presented in equation 4, where P_{i+n}^l denotes the probability vector produced by the l th LSTM of the $i+n$ th item in the dataset (which we denote above as $P(d_{i+n+l}|D_{i:i+n})$).

$$\text{delta}(P_{i+n}^l) = \frac{\sum_{k=i+n}^{i+n+n-1} P_k^l}{k} - \frac{P_i^l + P_{i+n}^l}{2} \quad (4)$$

A link is built between d_{i+n} and d_{i+n+l} *if and only if this probability delta exceeds a certain threshold*. The thresholds are computed at runtime, since different event types need different associated thresholds for the link-building to work correctly. We experimented with a number of automatic discretisation algorithms. One commonly used method to compute a binary threshold is Otsu's method, which is commonly used in the image processing community. This works by converting an image to greyscale, and then producing a histogram over its pixel intensities. The threshold is the point in the histogram where the integral of pixel densities to both the left and right of the threshold are equal. This is mathematically equivalent to computing the k-means clustering with $k = 2$ for the pixel

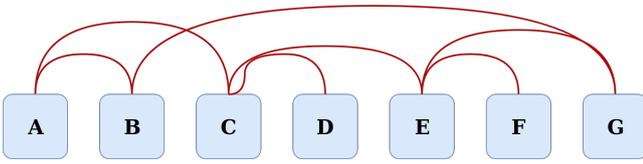


Fig. 2: Links are built between events if an LSTMs probability delta passes a threshold. This threshold is computed dynamically at runtime.

intensities and then taking the average of the two centroids as a threshold. However, this method assumes that the histogram in question has two distinct humps, one for light pixels and another for dark pixels. This is unlikely to be the case for our dataset. In this case, thresholds can be computed automatically by applying Otsu’s method, dividing the image into bright and dark sub-images based on the threshold, running Otsu’s method over both sub-images, and then using the average of these two thresholds as the threshold for the next iteration. We stop iterating when the threshold begins to converge, i.e. when the threshold computed in two iterations falls below a certain epsilon. This is the method we used to automatically compute thresholds per event type, using probability deltas in place of pixel intensity values. The linking process results in a tangle of binary links between events, as illustrated in Fig. 2.

The links are then grouped together to form activities. For example, if a link is built between the sensor events at indexes 1 and 5, and another link is built between 5 and 8, the system will output an activity consisting of 1, 5 and 8.

The final stage of the three introduced above is clustering. We say two activities have the same type as each other if they share at least 50% of the same event types. We have looked at other, more sophisticated types of clustering, but we have found that the type used has surprisingly little effect on the performance of our system.

Building Hierarchies of Activities

So far, we have maintained a sharp distinction between the types of sensor events and the types of activities. Removing this distinction has an obvious advantage: we can replace the discovered activities in the dataset with their activity types, producing a new dataset that the above process can be repeated on. This allows us to produce *activity hierarchies*, where activities can contain previously discovered activities as members, which allows for the modelling of activities that contain other activities. These are abundant in the real world: for example, a *making dinner* activity could contain a smaller activity such as *chopping vegetables*.

Fig. 3 presents a possible output from such a process. If events B, D, F and G are all found to belong to the same activity, they can be removed and replaced with a new event representing the discovered activity. We can then train and run the system again, which allows this



Fig. 3: If events B, D, F and G are all part of the same activity, we can remove these events from the dataset and replace them with a new activity. We can then train and run the system from scratch again, allowing us to build rich hierarchies of activities.

activity to be detected as belonging to other activities. This allows for the building of complex hierarchies of activities, such as the one described in the previous paragraph.

Note that the discovered activity includes events that are not adjacent to each other in the original dataset. For example, event B does not directly neighbour the other events in the activity. This is one of the major strengths of our approach: it does not assume or require that the activities discovered are contiguous. This allows us to deal with one of the most pressing issues in the field of activity discovery, which is that of *interleaving*, where the person or people under observation are carrying out multiple activities in parallel. From the viewpoint of an activity discovery system, interleaved activities usually look like a person switching back and forth between activities, in much the same way that a modern operating system context switches between running processes to allow multitasking. Our approach aims to explicitly address interleaving by disentangling interleaved activities from each other.

One non-obvious aspect of this process is why the new event was placed after activity E. For example, event B was also part of the discovered activity the event is replacing, so would it not make equal sense to place the new event between A and C? We decide where to place the new event based on the number of events it removes from various locations of the original dataset. The event in Fig. 3 removed one event between A and C (event B), one between C and E (event D), and two after E (F and G). Thus, we place the new event after event E.

The process can be repeated as many times as needed to produce a many-layered hierarchy of activities.

EXPERIMENT SETUP

We implemented our system in Python using the TensorFlow (Abadi et al., 2015) machine learning library. We used a four-layer neural language model, with 150 cells per layer for the lowest level of the hierarchy. As we ascended the hierarchy we found that the size increase of the vocabulary due to the addition of discovered activities was straining the network. As a result, we increased the size of the network by 50% for each level: level 1 had 150 LSTM cells, level 2 had 225, level 3 had 337 and so on.

We use the Kyoto 3 dataset from the CASAS smarthome project (Cook and Schmitter-Edgecombe, 2009). This dataset consists of readings from a range of

sensors installed in a small apartment. The dataset was gathered by asking a number of participants to perform *activities of daily living* (ADLs) in a natural manner in the apartment. Most of the sensor readings are either binary (they have a simple on/off state), or can only enter one of a handful of states. This means they can be easily converted to the sequence of events format our system expects by creating event types of the form *SensorName_SensorState*. For example, one of the sensors are referred to as *M17* in the dataset, and can take the state *ON*, so *M17_ON* becomes an event type in the dataset. For the few sensor types that did have continuous values, we used the *Jenks natural breaks algorithm* (Jenks, 1967) to discretise the data. Our system does not take temporal distance into account, so it cannot, for instance, see large gaps between events. This makes the system’s task substantially harder, but it allows us to put our system through much more challenging testing than most activity discovery practitioners settle for.

RESULTS

Evaluating activity discovery systems can be a challenge for a number of reasons. Human annotators may not come to an agreement with each other over the start and end points of activities, which makes working from a gold-standard ground truth quite difficult. For example, when does the activity of *Making_Dinner* start? When a person enters the kitchen? When they turn on the oven? In many cases, a ground truth may not even be available (although that isn’t an issue for the Kyoto dataset). The output from an AD system may be on a different level of abstraction from the ground truth: for example the system may discover an activity that could be called something like *chop_vegetables*, but the ground truth instead has an activity called *make_dinner*, which *chop_vegetables* would be a constituent of. A good overview of evaluation for activity discovery can be found in (Cook and Krishnan, 2015).

Since we do have access to a ground truth in this experiment, it makes sense to use it, although we must keep the above issues in mind. Because of the abstraction issue mentioned before, we argue that both raw accuracy and F-measures are inappropriate for evaluating this system. Instead, we compare each new event type from each level of the hierarchy using the *precision* metric, i.e. the true positives divided by the sum of the true and false positives. Each event type is then matched with the ground truth activity with which it achieves the highest associated precision.

Given a window length n and lookahead length m of 10, and building a hierarchy of 4 levels, the average precision per level is shown in Table I.

We can see that the results improve the higher up through the hierarchy we ascend. This is expected, since the events become more abstract and thus closer to the (very abstract) activities in the ground truth. We also compute the results per discovered activity (the results presented in Table I is the average over these scores.) These are too large to be presented in

Level number	Average precision
Level 1	0.7694124354455739
Level 2	0.8183002393181837
Level 3	0.82831171138164
Level 4	0.8341719705663135

TABLE I: Average precision score achieved for a 4-level hierarchy

Event type	Precision
new_event_45	1.0
new_event_46	1.0
new_event_47	0.75
new_event_48	1.0
new_event_49	0.5
new_event_51	0.5

TABLE II: Extract of the full results, showing the precision of each activity type found

full in this paper, but an extract of the full results are presented as Table II.

We also experimented with different hyperparameter values. In particular, we studied the effect of adjusting both the window and lookahead lengths. Increasing the window length has a moderate negative impact on the observed results, as shown in Table III. This indicates that the extra information provided in the longer sliding window actually ends up confusing the LSTM networks, since they observe conflicting signals as a result of now having multiple activities visible in their input at any one point in time. Most activities, even when highly interleaved, tend to be very “local”, with events that constitute the activity being located quite close together in the dataset.

Perhaps less surprising is the strong negative correlation observed between lookahead length and precision, shown in Table IV. This is also to be expected: our system is very good at linking nearby events, but struggles to confidently link distant events, which is very much to be expected when processing sequential data, but the extent of the negative correlation is worth pointing out.

A particularly interesting and encouraging result is the difference in performance when dealing with interleaved and non-interleaved datasets. Table V shows the average precision for the system when given a *non-interleaved* dataset as input. Compared to the results for the interleaved dataset (Table I) we see the lack of interleaving is actually confusing the system, which

Window length	Average precision
10	0.80
15	0.80
20	0.79
25	0.77

TABLE III: Relationship between window length and precision

Lookahead length	Average precision
10	0.80
15	0.64
20	0.56
25	0.48

TABLE IV: Relationship between lookahead length and precision

Level number	Average precision
Level 1	0.7376020200520275
Level 2	0.7744498540343416
Level 3	0.7964471494200084
Level 4	0.8018082799378542

TABLE V: Average precision score achieved when using a non-interleaved dataset

is the opposite to what is usually observed in activity discovery systems. This demonstrates that we have a strong reason to claim that this system is well suited to dealing with interleaved datasets. It also means that performance could likely be boosted further by an ensemble of this system and traditional activity discovery systems, since these results demonstrate that they arguably have different strengths, and combining models with different strengths is generally a good idea when carrying out ensemble learning.

We have already mentioned minimum description length (MDL) in the prior work section on this paper on page 2. (Cook et al., 2013) suggest using this as the basis for another metric for activity discovery systems, namely that of *compression ratio*. Since our system is constructing a hierarchy of activities by removed the sensor events that are found to belong to the discovered activities, the dataset reduces in size over time. Compression ratio alone can serve as a metric, since having a high compression ratio can be a sign that the system is correctly finding activities present in the dataset. Our system compresses the original Kyoto3 dataset to around 36% of its original size.

(Cook and Krishnan, 2015) proposes that the concept of compression ratio could be converted into a more principled metric by measuring how well compression ratio *generalises*. In traditional machine learning, we may be more concerned with how a system deals with novel input compared to how it deals with input seen in the dataset. This allows us to be sure it is learning a signal present in the dataset, rather than just memorising the contents of the dataset. This is typically measured using techniques like holding out a validation dataset from the main dataset. If a system is generalising well, its performance on the testing dataset should be similar to the performance on the training dataset. Likewise, if an activity discovery system compresses the dataset by a certain amount, it should also compress a testing dataset by the same amount. We tested our system using ten-fold cross-validation. The results, presented in Table VI, show clearly that the system is finding

Cross-validation	Compression ratio
Fold 1	0.3927149342694845
Fold 2	0.32898505905289627
Fold 3	0.3363159811817841
Fold 4	0.34688667906136067
Fold 5	0.35323690998006696
Fold 6	0.38618623218659726
Fold 7	0.315419492673265
Fold 8	0.32600179330230683
Fold 9	0.4082227112380515
Fold 10	0.3370837650161199

TABLE VI: Ten-fold cross validation of the compression ratio produced by the system

activities that generalise well to the test dataset.

CONCLUSION

This paper introduced a deep learning-based activity discovery system. We have described the system, and also presented results illustrating its performance on a realistic dataset, and an analysis of how the results change in response to a change in the system’s meta-parameters. We feel that these results show that the system performs favourably compared to other systems in the field, and could be adapted for use in real-world activity discovery applications.

A number of changes could be made to this system to improve its performance and further test it. We have already mentioned the work of (Alshammari et al., 2017) as a possible means to generate datasets for very in-depth testing. Testing using the Opportunity dataset (Chavarriaga et al., 2013) could also be useful. Changing the design of the network to take into account the temporal information included in the Kyoto dataset, but not utilised by our model, is another possible way to improve the system in the future.

REFERENCES

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- N. Alshammari, T. Alshammari, M. Sedky, J. Champion, and C. Bauer. Openshs: Open smart home simulator. *Sensors*, 17(5):1003, 2017.
- Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet

- allocation. *Journal of machine Learning research*, 3 (Jan):993–1022, 2003.
- R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digu-
marti, G. Tröster, J. d. R. Millán, and D. Roggen. The
opportunity challenge: A benchmark database for on-
body sensor-based activity recognition. *Pattern Recog-
nition Letters*, 34(15):2033–2042, 2013.
- D. J. Cook and N. C. Krishnan. *Activity learning: dis-
covering, recognizing, and predicting human behavior
from sensor data*. John Wiley & Sons, 2015.
- D. J. Cook and M. Schmitter-Edgecombe. Assessing the
quality of activities in a smart environment. *Methods
of information in medicine*, 48(05):480–485, 2009.
- D. J. Cook, N. C. Krishnan, and P. Rashidi. Activity
discovery and activity recognition: A new partnership.
IEEE transactions on cybernetics, 43(3):820–828, 2013.
- B. Cramer. Limitations of current grammar induction
algorithms. In *Proceedings of the 45th annual meeting
of the ACL: student research workshop*, pages 43–48.
Association for Computational Linguistics, 2007.
- L. Fang, J. Ye, and S. Dobson. Discovery and recogni-
tion of emerging human activities using a hierarchical
mixture of directional statistical models. *IEEE Trans-
actions on Knowledge and Data Engineering*, 2019.
- J. Gaspers, P. Cimiano, S. S. Griffiths, and B. Wrede.
An unsupervised algorithm for the induction of con-
structions. In *2011 IEEE International Conference on
Development and Learning (ICDL)*, volume 2, pages
1–6. IEEE, 2011.
- H. Gjoreski and D. Roggen. Unsupervised online activ-
ity discovery using temporal behaviour assumption. In
*Proceedings of the 2017 ACM International Symposium
on Wearable Computers*, pages 42–49, 2017.
- E. Gold. Language identification in the limit. *infor-
mation and control*, 10: 447–474, 1967.[11] s. *Jain*.
*An infinite class of functions identifiable using minimal
programs in all Kolmogorov numberings*. *International
Journal of Foundations of Computer Science*, 6(1):89–94,
1967.
- A. Graves. Generating sequences with recurrent neural
networks. *arXiv preprint arXiv:1308.0850*, 2013.
- S. Hochreiter and J. Schmidhuber. Long short-term
memory. *Neural computation*, 9(8):1735–1780, 1997.
- T. Huynh, M. Fritz, and B. Schiele. Discovery of activ-
ity patterns using topic models. In *UbiComp*, volume 8,
pages 10–19, 2008.
- G. F. Jenks. The data model concept in statistical
mapping. *International yearbook of cartography*, 7:186–
190, 1967.
- E. Kim, S. Helal, and D. Cook. Human activity recogni-
tion and pattern discovery. *IEEE pervasive computing*,
9(1):48–53, 2009.
- J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activ-
ity recognition using cell phone accelerometers. *ACM
SigKDD Explorations Newsletter*, 12(2):74–82, 2011.
- C. G. Nevill-Manning and I. H. Witten. Identifying
hierarchical structure in sequences: A linear-time al-
gorithm. *Journal of Artificial Intelligence Research*, 7:
67–82, 1997.
- G. Petasis, G. Paliouras, V. Karkaletsis, C. Halat-
sis, and C. D. Spyropoulos. e-grids: Computationally
efficient grammatical inference from positive examples.
Grammars, 7:69–110, 2004.
- J. Rissanen. Modeling by shortest data description.
Automatica, 14(5):465–471, 1978.
- J. Rissanen. *Stochastic complexity in statistical inquiry*.
World Scientific, 1989.
- E. Rogers, J. D. Kelleher, and R. J. Ross. Towards
a deep learning-based activity discovery system. In
AICS, pages 184–191, 2016.
- E. Rogers, J. D. Kelleher, and R. J. Ross. In *Pro-
ceedings of the Second IFIP International Conference
on Machine Learning for Networking*. Springer, Forth-
coming.
- J. Saives, C. Pianon, and G. Faraut. Activity discovery
and detection of behavioral deviations of an inhabitant
from binary sensors. *IEEE Transactions on Automa-
tion Science and Engineering*, 12(4):1211–1224, 2015.
- Y. Shen, Z. Lin, C.-W. Huang, and A. Courville. Neu-
ral language modeling by jointly learning syntax and
lexicon. *arXiv preprint arXiv:1711.02013*, 2017.
- Z. Solan, D. Horn, E. Ruppín, and S. Edelman. Un-
supervised learning of natural languages. *Proceedings
of the National Academy of Sciences*, 102(33):11629–
11634, 2005.
- W. Zaremba, I. Sutskever, and O. Vinyals. Recur-
rent neural network regularization. *arXiv preprint
arXiv:1409.2329*, 2014.

Predicting Business Process Bottlenecks in Online Events Streams under Concept Drifts

Yorick Spennath Marwan Hassani
Department of Mathematics and Computer Science
Eindhoven University of Technology, The Netherlands
y.spennath@tue.nl m.hassani@tue.nl

KEYWORDS

Gradual Concept Drift; Recurrent Concept Drift; Concept Drift Detection; Business Process Bottlenecks; Event Streams

ABSTRACT

Process performance analysis is an important sub-task of process mining that aims at optimizing the discovered process models. In this paper we focus on improving process throughput by predicting congestions in the process execution (bottlenecks). We discuss an ongoing work on incorporating gradual and seasonal concept drift in this bottleneck prediction. In the field of process mining, we develop a method of predicting whether and which bottleneck will likely appear based on data known before a case starts. We introduce GRAHOF, a Gradual and Recurrent Adaptive Hoeffding Option Forest approach, which adapts to gradual and seasonal concept drifts when predicting bottlenecks of business processes in an online setting. We evaluate the parameters involved in GRAHOF using a synthetic event stream and a real-world event log.

INTRODUCTION

Concept drift expresses the occurrence of a shift in relations between input and output data over time. The challenge with concept drift is that it is difficult to assess, detect, and adapt to such evolutions. One method of solving concept drift is to retrain models periodically. This method poses two major disadvantages. First, this reduces the quantity of available data, which is a big issue when the number of data points per time unit is low, or when dealing with unbalanced datasets. The combination of these two causes over-fitting of the model on larger classes. Second, retraining a model with new data inherently discards former data, which could remove valuable information on seasonality.

In this paper, we present an ongoing work on predicting bottlenecks in business processes under (gradual and seasonal) concept drift. Making predictions about process execution is one form of predictive process mining. In predictive process mining two types of concept drift exist. First there can be a drift in predictions (how the relations between process execution information and bottlenecks evolve) or a drift in the

process itself (how the process execution model evolves over time), see for example Hassani (2019); Martjushev et al. (2015). In this paper we are concerned with concept drift of the former, assuming the process execution model to be constant.

The advantage of being able to predict bottlenecks, whether in general or under concept drift, is that it allows process managers to make informed decisions about the process execution, as well as early identification of possible delays. Figure 1 shows the example process used in this paper. Mechanics in an installations services company do repair work. The financial administration of the company is then responsible for invoicing the repairs. This process consists of four activities *Check Repairs*, *Collection Information*, *Make Invoice*, and *Send Invoice*. To increase the throughput of the example process (thus minimising the time between the repair and sending the invoice), we aim to predict if and where in the process bottlenecks occur. We make these predictions based on information available after repairs. This is data on the size of the company the repair was at, the distance the mechanic travelled, and the type of repair executed.

In this work, the goal is to make a prediction of the bottleneck *before* the case starts. In particular, we do not incorporate process execution information. This paper contributes as an extension to the previous work in Spennath and Hassani (2019). The extended method, GRAHOF, is capable of handling *online* analysis of event streams (rather than post-mortem event logs). The extension uses incremental learners to adapt to gradual concept drift. The specific model, the Adaptive Hoeffding Option Tree Bifet et al. (2009), has the advantage of being very fast by design, which allows the online analysis of large event streams. The full implementation of GRAHOF, the simulation, and the experiments, including a visualisation, is available under: <https://github.com/yorick-spennath/GRAHOF>.

The remainder of this paper is organised as follows. The following section introduces related work. The third section sets out some necessary preliminaries. The fourth section provides the algorithm used to assess bottlenecks. The fifth Section presents GRAHOF, the streaming algorithm used to analyse event streams while adapting to gradual and recurrent drift. The sixth section describes a simulation of a synthetic

event stream followed by an experimental evaluation of GRAHOF using that dataset and then using a real-world dataset. Finally, the last section concludes the paper and outlines future work.

RELATED WORK

This section presents related work on Stream Process Mining and on Concept Drift Adaption.

Predictive and Stream Process Mining

Predictive process analytics is becoming increasingly popular in literature. Opposed to classical process mining, predictive process mining aims to assist by making predictions about future behaviour. In a study on handling concept drifts in business processes, Maisenbacher and Weidlich (2017) investigates the different solutions to adapt to concept drift in predicting process outcomes. The authors use information on both process executing (i.e. past events in a case), and on involved data (case data, and data that is learned during process execution). In particular, they show the effectiveness of incremental machine learning models in handling different types of concept drift. In Bose et al. (2011, 2014), the authors consider drift detection in processes. Contrary to the drift looked at in this paper, their focus is on drift detection in the process itself, rather than in the relations between process data and for example bottlenecks. They do so by extracting information about process execution from event logs, and defining features that allow them to detect drift in the underlying process. They argue that their technique can be used successfully on data streams as well. Finally, Carmona and Gavalda (2012) also develops a drift detection method on event streams. The idea behind their method is to abstractly estimate the concept, by spanning a subspace containing the frequency of activities in traces of the currently looked at event log. Part of the traces will be part of this abstract space, and as such a measure of how well the space represents the event log can be derived. After stability of this measure, new traces are evaluated, if as such the measure considerably changes, drift is assumed to have taken place.

In recent years, analysing events on-line instead of post-mortem has become increasingly important Hassani et al. (2019). The authors of van Zelst et al. (2019) discuss an online conformance checking technique. While receiving events from an event stream, they verify whether the cases follow the intended business process model to observe deviations the moment they occur. The works in Hassani et al. (2015) and Burattin et al. (2014) discuss a different part of process mining on event streams: discovery. They consider analysing event streams, but refrain from analysing each event. Instead, they keep a maximum to their memory size, and prune old events and results when needed, periodically using all information in memory to discover process models at that time. The authors of the latter show that this allows concept drift detec-

tion as well. While the concept drift detection effect of the former has been discussed in Hassani (2019).

This paper contributes to the existing literature by analysing and predicting bottlenecks in the processes under concept drift. It can help process managers identify potential process delays early, even under seasonal fluctuations.

Concept Drift Detection and Adaptation

Concept drift is the phenomenon where relations between input (features) and output (labels in the case of the experiments in this paper) change over time, because the underlying models and/or distributions change, see for example Gama et al. (2014). The authors note that this applies to a variety of research areas, including process mining. In our use case, the predictions on bottlenecks will become less accurate if no adaptations to concept drift are made. As such, the effectiveness in preventing bottlenecks is reduced.

Most existing literature focuses on *detecting* concept drift. Detection is a first step, adapting to it is the next step. Literature has a wide variety of detection techniques. The authors of Baena-García et al. (2006) discuss a technique based on changes in classification accuracy to detect abrupt and gradual concept drift. They check how the difference in classification error progresses over time. If the difference is far smaller than a previously measured distance in the same concept, they suppose the occurrence of concept drift. A different technique is to consider the importance of features over time. In Blum (1997), the authors present a technique which can adapt to concept drift. They train ‘experts’, each expert takes a specific combination of feature values and recommends the label based on the latest labels for that specific combination. All experts are then weighted, where incorrect experts lose their weight over time, whereas correct experts increase in weight. One way to adapt to concept drift is by simply retraining the model whenever a drift in concept is detected. A disadvantage of this technique occurs if the drift is periodic, i.e. a drift of concept is detected, but the previous concept may still be of use at a later point in time, just not right after the detected drift. In Wang et al. (2003), the authors do so by training a new model for every specified number of data points, creating an ensemble of different models. Each model is then weighted by the accuracy score on recent datasets. Challenges in seasonal concept drifts have been extensively addressed in literature as well. The method in Jr and de Barros (2013) keeps an ensemble of models, and selects the model that is expected to be best adapted to the current effects concept drift. Finally, Ramamurthy and Bhatnagar (2007) also considers an ensemble of classifiers. Batches of data are tested against this ensemble. If no classifier can explain the new batch, the ensemble is weighted based on how each model explains the batch. If this weighted ensemble is again unable to adequately explain the new batch, the batch is considered as a new concept, and a new model is trained.



Fig. 1. An installation services company has four activities that are part of the invoicing process, starting after repair activities complete. We use this process as a running example.

PRELIMINARIES

We define an *event* e as a tuple $(e.cid, e.act, e.time)$, stating that the activity $e.act \in \mathcal{A}$ was completed for the case with id $e.cid$ at time $e.time \in \mathbb{R}_+$. An *event log* \mathcal{L} is a set of such events. A case c is a sequence of events $\langle c(0), c(1), \dots, c(|c| - 1) \rangle$. Cases further have a unique id $c.cid$. The trace or variant of c , is the sequence of activities in c , i.e.

$$variant(c) = \langle c(0).act, c(1).act, \dots, c(|c| - 1).act \rangle$$

The total duration of a case is defined as

$$duration(c) = c(|c| - 1).time - c(0).time$$

Finally, $c.start$ denotes the start of the case, i.e. $c(0).time$. We will write $c \in \mathcal{L}$ if there is at least an event $e \in \mathcal{L}$ for which $e.cid = c.cid$. In other words, we will represent the elements of \mathcal{L} as cases or events. From a case c we can derive the duration of each activity in c from the events in c . In particular, for $i = 1 \dots |c| - 1$, we have that activity $c(i).act$ took $c(i).time - c(i - 1).time$ time. As such we cannot compute the duration of the first activity in a case; this is generally not a problem since this activity only indicates the start of the case.

Cases further have feature data $c.x$. This feature data is any case information that is available at the start of the case, which will be used to predict bottlenecks. An event stream \mathcal{S} is a (possibly infinite) set of events. We assume that events are chronological, i.e. for any event e that we receive from \mathcal{S} and for any event e' we received from \mathcal{S} before e , we have that $e'.time \leq e.time$. We define $\mathcal{S}.next()$ to retrieve the next event in \mathcal{S} . We further assume that, given an event e , the predicate $\mathcal{S}.end(e)$ tells whether an event e was the last event of the case. In our example of Figure 1, we could have $\mathcal{S}.end(e) \Leftrightarrow e.act = Send\ Invoice$. Finally, we assume that the first event we receive for a case contains the feature data of that case. As such we refer to such events as a *start event*. A start event hence contains all required information to start a case (feature data, id, and start time).

BOTTLENECK COMPUTATION

In this Section we present how we decide on the bottlenecks in a case. The way in which we define these is that we point out a *single* activity to be *the* bottleneck. The idea is that this activity was responsible for a considerable delay compared to other cases that started around the same time. Addressing this specific (predicted) bottleneck should be seen as the best opportunity on improving the throughput of a case.

First, we split the event log in \mathcal{L}_{short} and \mathcal{L}_{long} . \mathcal{L}_{short} contains all cases that are completed within a given time duration, set by domain experts. Let this duration be d , i.e. $\mathcal{L}_{short} = \{c \in \mathcal{L} | duration(c) \leq d\}$, and $\mathcal{L}_{long} = \{c \in \mathcal{L} | duration(c) > d\}$. We next compare cases in \mathcal{L}_{short} with cases in \mathcal{L}_{long} . We do so per variant. Let there be N variants in \mathcal{L} , and let V_i be these variants, $i \in 1 \dots N$. We partition \mathcal{L}_{short} into sets $\mathcal{L}_{short}^i = \{c \in \mathcal{L}_{short} | variant(c) = V_i\}$. We similarly split \mathcal{L}_{long} into sets \mathcal{L}_{long}^i .

For each variant V_i , we use \mathcal{L}_{short}^i as a ‘benchmark’ for how long events should take, and then assess the duration of events in \mathcal{L}_{long}^i . We require that bottlenecks are both resolvable and relevant. We formalise ‘Resolvability’ by requiring that for an activity to be a bottleneck, the duration should be longer than the average for that activity in \mathcal{L}_{short}^i . We formalise ‘Relevance’ by requiring that the event takes at least α part of the total duration of the case. $\alpha \in [0, 1]$ is a parameter, also set by domain experts. Setting it too low will remove the requirement, setting it too high make no event meet it. Of all events in a case meeting both requirements, we select the one that deviates the most from the benchmark derived from \mathcal{L}_{short}^i .

Handling the cases per variant has multiple advantages. First, the technique is not limited by more complex process structures, such as loops and choices. Second, the algorithm is robust against cases that do not conform the (expected) process. Finally, cases that deviate from the process (they do not conform) are compared to cases that deviate from the process in the same way. We refer to our previous work Spenrath and Hassani (2019) for a detailed formalisation of the above.

OUR APPROACH: GRAHOF

In this Section we introduce **GRAHOF**. **GRAHOF** is an acronym for Gradual and Recurrent Adaptive Hoeffding Option Forest. Similar to the work presented in Spenrath and Hassani (2019), **GRAHOF** uses an ensemble to adapt to gradual and recurrent concept drift. Different to it, **GRAHOF** requires incremental models, and uses them to adapt to gradual drift. Our approach does not require a specific incremental learner, we implemented it using the Adaptive Hoeffding Option Tree (AHOT) Bifet et al. (2009) as incremental learner. The choice of AHOT is inspired by Maisenbacher and Weidlich (2017), where they use the learner in predicting process outcomes. In this paper we focus on how **GRAHOF** uses the AHOT as incremental learner. We hence refrain from the details of AHOT and we use the default settings of AHOT in Bifet et al. (2010).

In the following, we formalise the execution algorithm, provide the model updating and creation, and

discuss the making of predictions.

Main GRAHOF execution

The full method is presented in Algorithm 1. In the following, the line numbers refer to that Algorithm.

Algorithm 1 GRAHOF on an Event Stream \mathcal{S}

```

1:  $M \leftarrow \emptyset$ 
2:  $B^* \leftarrow \text{Batch}(0, S) // S$  is the batch size
3:  $\mathcal{B} \leftarrow \{B^*\}$ 
4: while True do
5:    $e \leftarrow \mathcal{S}.\text{next}()$ 
6:   if  $e$  is start event then
7:     while  $B^*.t_e \leq e.\text{time}$  do
8:        $B^* \leftarrow \text{Batch}(B^*.t_0 + S, B^*.t_e + S)$ 
9:        $\mathcal{B} \leftarrow \mathcal{B} \cup \{B^*\}$ 
10:    end while
11:     $c \leftarrow$  new case with id  $e.cid$ 
12:    Add  $c$  to  $B^*$ 
13:    Predict  $c.y_{pred}$  using  $c.x, M$  and Equation 1
14:  else
15:    Find batch  $B$  and case  $c$  with  $c \in B \wedge c.cid = e.cid$ 
16:    Add  $e$  to  $B$ 
17:    if  $S.\text{end}(e)$  then
18:       $c.\text{closed} = \text{True}$ 
19:      if  $\forall c.\text{cases} \in B : c.\text{closed} \wedge B \neq B^*$  then
20:         $\text{close}(B)$ 
21:      end if
22:    end if
23:  end if
24: end while

```

We define a *batch* as follows:

Definition 1: A batch B is an extension to an event log, with cases that start in a specific interval. A batch has the following properties:

- A batch is initiated as $\text{Batch}(t_0, t_e)$, with $t_0, t_e \in \mathbb{R}_+$ and $t_e - t_0 = S$. The batch size $S \in \mathbb{R}_+$ is a parameter for GRAHOF .
- A batch contains a set of cases $B.\text{cases}$ that start on or after $B.t_0$ and before $B.t_e$: $\forall c \in B.\text{cases} : c.\text{start} \in [B.t_0, B.t_e)$.

We create the first batch at the start of the event stream. This first batch has interval $[0, S)$ (Line 2). Since we assume the stream to produce events chronologically, we create a new batch whenever a case starts outside of the interval of the current batch. In the remainder, we use B^* to refer to the most recently created batch. (Lines 2, 8).

We receive the events from the event stream one-by-one. For any received event e , we first decide if that event is a start event. If so, we create a new case (Line 11), and add the case to the current batch (or create a new batch if the case start is after the current batch end). We then make the prediction (see the subsection on making predictions) (Line 13). If the event is not a start event, then the case the event belongs is already added to a batch (as per Line 12). We as such add it (Line 16) to the corresponding batch (Line 15). If

e is the final activity of a case, we mark the case as complete (Line 18).

Whenever we complete a case we check if 1) all cases in B are closed, and 2) whether $B \neq B^*$ (Line 19). If both predicates hold, then we neither expect more events for existing cases, nor new cases for B . The batch is therefore complete and we can close it. Closing the batch ($\text{close}(B)$) consists of computing the true labels (See the section on bottleneck computation), updating or existing an AHOT (See the section on updating models), and removing B from \mathcal{B} . The batch B is therefore no longer relevant after it is used for training.

Algorithm 2 Finding the best model in GRAHOF

Ensure: For a dataset \mathbb{ID} we update an existing model or create a new one

```

1: Get a balanced, stratified 80%  $\mathbb{ID}_{\text{train}}$  and 20%  $\mathbb{ID}_{\text{test}}$  from  $\mathbb{ID}$ 
2:  $M' \leftarrow M$ 
3: while  $M' \neq \emptyset$  do
4:    $m^* = \arg \max_{m \in M'} f(\mathbb{ID}_{\text{test}}, m)$ 
5:   if  $f(\mathbb{ID}_{\text{test}}, m^*) \geq \phi f_{m^*}$  then
6:      $f_{m^*} \leftarrow \phi f_{m^*} + (1 - \phi) f(\mathbb{ID}_{\text{test}}, m^*)$ 
7:     Update  $m^*$  with  $\mathbb{ID}_{\text{train}}$ 
8:     Add  $\mu_B$  to  $m^*.\mu_{\text{train}}$ 
9:     return
10:  else
11:     $M' = M' \setminus \{m^*\}$ 
12:  end if
13: end while
14:  $m^* \leftarrow$  new AHOT
15: Update  $m^*$  with  $\mathbb{ID}_{\text{train}}$ 
16:  $f_{m^*} = f(\mathbb{ID}_{\text{test}}, m^*)$ 
17: Add  $\mu_B$  to  $m^*.\mu_{\text{train}}$ 
18:  $M \leftarrow M \cup \{m^*\}$ 

```

Updating Models

In the following, $F_1(\mathbb{ID}, m)$ denotes the F_1 score of a model m on a dataset \mathbb{ID} . Whenever all cases that belong to a batch are complete, we compute the bottleneck labels as per the section on bottleneck computation for the cases in the batch. This creates a dataset \mathbb{ID} . There is a possibility that one of the activities is labelled as a bottleneck more often than others, i.e. that we are dealing with an unbalanced dataset. We therefore apply a k -medoid reduction on \mathbb{ID} , to reduce the size of overrepresented to at most twice the size of the smallest class. We split the resulting balanced dataset into a stratified $\mathbb{ID}_{\text{train}}$ and $\mathbb{ID}_{\text{test}}$. Initially, we have no models (the ensemble, M , is empty), and we create a new AHOT m , which we proceed to train with $\mathbb{ID}_{\text{train}}$. We store $f_m = F_1(\mathbb{ID}_{\text{test}}, m)$, and add m to M . After the first model creation M is no longer empty, so we first verify if there is a model m^* that can be improved with \mathbb{ID} . We therefore first evaluate all existing models starting with the model $m^* \in M$ that has the highest F_1 on $\mathbb{ID}_{\text{test}}$. If $F_1(\mathbb{ID}_{\text{test}}, m^*)$ is at least $\phi \cdot f_{m^*}$, then we update f_{m^*} to $\phi \cdot f_{m^*} + (1 - \phi) \cdot F_1(\mathbb{ID}_{\text{test}}, m)$ and train m^* with $\mathbb{ID}_{\text{train}}$. $\phi \in [0, 1]$ is a parameter for

GRAHOF . If m^* does not meet the requirement, we try the model with the next-best score, and so on. If none of the models meet the requirement, we create a new AHOT as above. The procedure of updating a model is presented separately in Algorithm 2.

Apart from setting or updating f_m , we also update a set $m.periods$. This set tracks all of the batches that have been used to train m . More specifically, $m.periods$ is initialised as \emptyset , and $[B.t_0, B.t_e)$ is added to it for every batch B that has been used to update m . This set is used when computing the weights of the model for a new batch.

Making predictions

The weight a model receives for a batch depends on how many seasonally-similar training periods there are in the model. Formally, if we make a prediction at time t , the weight of model m , will be¹

$$|\{t'|t' \in m.periods \wedge t - t' \equiv_{\rho} 0\}| \quad (1)$$

$\rho \in \mathbb{R}_+$ is a parameter for GRAHOF . In particular, we expect the period of the recurrent drift to be ρ , which would equal year for seasonal drift.

When making a prediction, each model computes its predicted probability for each bottleneck label; the predicted label is then the label with the highest weighted probability. This label can be used during operations by process managers, and is later evaluated against the true label computed by the bottleneck algorithm (Line 13).

EXPERIMENTAL EVALUATION

We simulate an event log based on the running example. We define three features, the repair *Type* (Electric, Water, Gas, Mechanical), the travelled *Distance*, and the client company *Size*. We simulate 250 cases per *Type* per month for 57 years.

The *Dist* and *Size* values are drawn from a discrete uniform distribution from 1 to 100. The values of *Dist*, *Size*, and *Type* determine if and where a bottleneck will occur in a case.

Gradual drift The idea behind the gradual concept drift is as follows: we take about $\frac{1}{5}$ of the combinations of values for **Dist** and **Size**. If a given case has one of these combinations, the case will have no bottleneck, otherwise it will. Formally, let $A(t) \subseteq S_{Dist} \times S_{Size}$, with $S_{Dist} = S_{Size} = 1..100$. A case c with *Dist* and *Size* has a bottleneck if and only if $(Dist, Size) \in A(t)$. $A(t)$ is a time-dependent subset of $S_{Dist} \times S_{Size}$, and is visualised in Figure 2.

Seasonal drift We simulate recurrent concept drift by defining four concepts, each of which occurs during a different quarter. In the first quarter each year, the repair type *Electric* causes a bottleneck in activity *Check Repair*, *Water* in *Collection Information*, *Gas* in *Make Invoice*, and *Mechanical* in *Send Invoice*. We then rotate this mapping each quarter, as depicted in Table I.

We make this switch over a period of two weeks, using the sigmoid function presented in Bifet et al. (2009).

Qt.	E	W	G	M
1	CR	CI	MI	SI
2	SI	CR	CI	MI
3	MI	SI	CR	CI
4	CI	MI	SI	CR

TABLE I: Bottlenecks as determined by topic and quarter of the start of the case. For example, we have that a (*W*)ater repair causes a bottleneck in the activity (*M*)ake (*I*)nvoice in the fourth quarter.

Results on the Simulated Dataset

In this Section we evaluate GRAHOF , as a verification of the method. We do so using the simulation previously described . We analyse the effect of ϕ and S separately. When analysing one parameter, we set the other parameters to $\phi = 0.75$ or $S = 3$ months (see our main algorithm section for details.). The results presented are the averages over 10 simulations.

The effect of ϕ Figure 3 shows that there is a trade-off in ϕ . For lower values, models can be updated with data that the model can not properly explain. In fact, only a single model is used for $\phi \leq 0.2$. For $\phi > 0.75$, F_1 decreases as a result of the increase in the number of models. Because models become harder to update, more models are created. As such, older models (which due to gradual concept drift are no longer updated) have a negative influence on the prediction, while newer models have seen fewer datapoints.

The effect of S Figure 4 shows the results for different values of S . As expected, $S = 3$ has a higher F_1 . For larger batch sizes, models are created and updated on multiple concepts. As such, GRAHOF cannot properly differentiate between the concepts, and as such unable to properly predict the correct bottleneck labels. For $S = 2$, there are two models that are trained on multiple (2) seasonal concepts, whereas other models correctly get a single concept. This slightly decreases the F_1 . $S = 1$ also has a slightly lower F_1 than $S = 3$. This is likely caused by the increase in the number of models.

Results on a Real-World Dataset

We applied the described method in an initial case study on a real-life event log used in the BPI Challenge 2017 Van Dongen, B.F. (Boudewijn) (2017). The original event log contains many different variants and events, we limit ourselves to work-flow events only (removing application and offer events, see the original dataset), and limiting to the most common variant. As a result, we have a total of 6664 cases, each having three potential bottlenecks. We applied GRAHOF for different values of S and ϕ . The results are shown in Figure 5.

The results indicate the potential for GRAHOF to work on real-life event logs. Though there seem to be small

¹ \equiv_{ρ} is the module operator with divisor ρ .

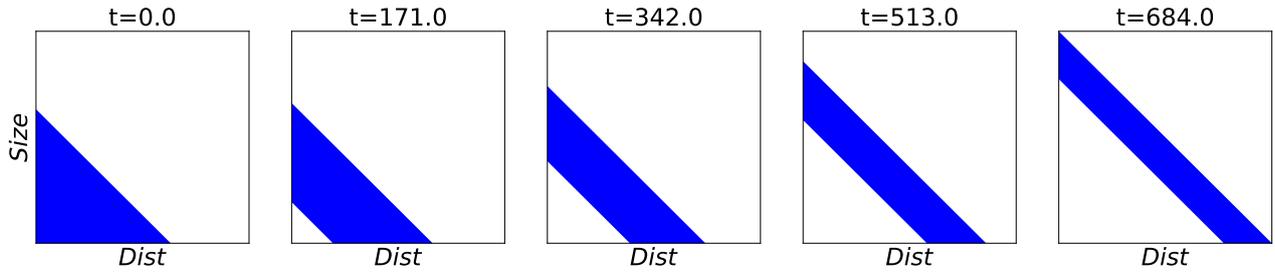


Fig. 2. Visualisation of $A(t)$

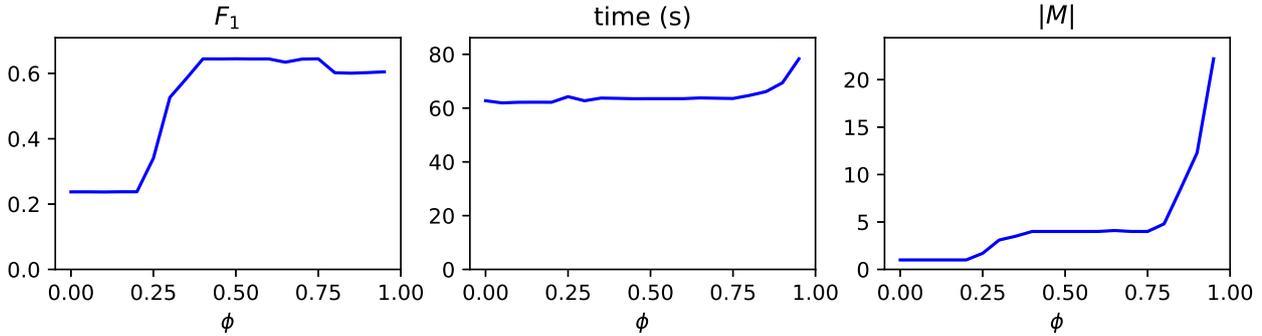


Fig. 3. F_1 , simulation time, and number of created models for different values of ϕ

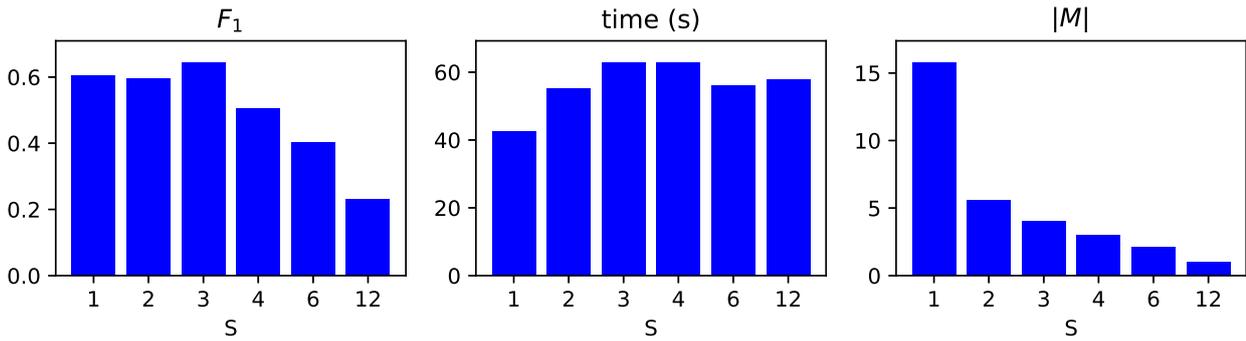


Fig. 4. F_1 , simulation time, and number of created models for different values of S

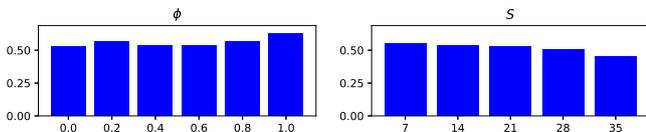


Fig. 5. Results for the case study. The different values of ψ use $S = 14$, the different values of S use $\phi = 0.6$.

differences between the values for both parameters, future work on more extensive event logs should make more conclusions about the stability of the approach with respect to the required parameters.

CONCLUSION AND FUTURE WORK

In this paper, we presented **GRAHOF**, an extension from earlier work. The analysis in previous section show the potential of **GRAHOF** on event stream. We have presented a first evaluation of two hyper-parameters of **GRAHOF** in Sections and . The results present the effect of each parameter (ψ and S) on its

own, future work should elaborate on the dependence between their values.

Several improvements for **GRAHOF** are targeted in the future. The bottleneck algorithm has the robustness that is required to deal with event logs that do not conform with the process model. On one hand, this removes restrictions on the quality of the event log. On the other hand, the algorithm does not allow to account for more complex process patterns such as choices and loops. This paper has specifically targeted seasonal drift with a known recurrence frequency, keeping ρ constant. Including dedicated recurrent drift detectors or finding ways to remove the need for ρ will be the next step in improving **GRAEC**. As is clear from the simulation results, picking the right value for S is important. Such a value can be estimated from an initial event log. An extension to **GRAHOF** is to make S dynamic. Future work on **GRAHOF** could be evaluating whether and how the performance decreases when the number of events or models is limited. The way in which **GRAHOF** is de-

financed cases that span multiple batches to be included in a single batch. A further direction for future work could be to not only add cases to batches they start in, but also to batches they run in. Finally, we would like to test GRAHOF in concrete process mining applications focusing on customer journey optimization and changes of consumer behavior (Terragni and Hassani (2019) and Goossens et al. (2018)).

REFERENCES

- M. Baena-García, J. Campo-Ávila, R. Fidalgo-Merino, A. Bifet, R. Gavaldà, and R. Morales-Bueno. Early Drift Detection Methods. *Intl. Workshop on Knowledge Discovery from Data Streams*, pages 77–86, 2006.
- A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavaldà. New Ensemble Methods for Evolving Data Streams. In *KDD'09*, pages 139–148, 2009.
- A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: massive online analysis. *JMLR*, 11, 2010.
- A. Blum. Empirical Support for Winnow and Weighted-Majority Algorithms: Results on a Calendar Scheduling Domain. *Mach. Learning*, 26(1):5–23, 1997.
- R. P. J. C. Bose, W. M. P. van der Aalst, I. Žliobaitė, and M. Pechenizkiy. Handling Concept Drift in Process Mining. In *AISE*, pages 391–405, 2011.
- R. P. J. C. Bose, W. M. P. van der Aalst, I. Žliobaitė, and M. Pechenizkiy. Dealing With Concept Drifts in Process Mining. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1):154–171, jan 2014. ISSN 2162-237X. doi: 10.1109/TNNLS.2013.2278313.
- A. Burattin, A. Sperduti, and W. M. P. van der Aalst. Control-flow discovery from event streams. In *CEC'14*, pages 2420–2427. IEEE, jul 2014.
- J. Carmona and R. Gavaldà. Online Techniques for Dealing with Concept Drift in Process Mining. In *IDA*, pages 90–102, 2012.
- J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, 2014.
- J. Goossens, T. Demewez, and M. Hassani. Effective steering of customer journey via order-aware recommendation. In *ICDM Workshops*, pages 828–837, 2018.
- M. Hassani. Concept Drift Detection of Event Streams Using an Adaptive Window. In *33rd International ECMS Conference On Modelling And Simulation*, 2019.
- M. Hassani, S. Siccha, F. Richter, and T. Seidl. Efficient process discovery from event streams using sequential pattern mining. In *SSCI 2015*, pages 1366–1373, 2015.
- M. Hassani, S. J. van Zelst, and W. M. P. van der Aalst. On the application of sequential pattern mining primitives to process discovery: Overview, outlook and opportunity identification. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 9(6), 2019.
- P. M. G. Jr and R. S. M. de Barros. RCD: A recurring concept drift framework. *Pattern Recognition Letters*, 34(9):1018–1025, 2013.
- M. Maisenbacher and M. Weidlich. Handling Concept Drift in Predictive Process Monitoring. In *SCC'17*, pages 1–8, jun 2017.
- J. Martijushev, R. P. J. C. Bose, and W. M. P. van der Aalst. Change Point Detection and Dealing with Gradual and Multi-order Dynamics in Process Mining. In *Change*, volume 229 of *Lecture Notes in Business Information Processing*, pages 161–178, Cham, 2015.
- S. Ramamurthy and R. Bhatnagar. Tracking recurrent concept drift in streaming data using ensemble classifiers. In *ICMLA 2007*, pages 404–409, dec 2007.
- Y. Spenrath and M. Hassani. Ensemble-Based Prediction of Business Process Bottlenecks With Recurrent Concept Drifts. *Workshop proceedings of the EDBT/ICDT 2019 Joint Conference*, 2019.
- A. Terragni and M. Hassani. Optimizing customer journey using process mining and sequence-aware recommendation. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, Limassol, Cyprus, April 8-12, 2019*, pages 57–65, 2019.
- Van Dongen, B.F. (Boudewijn). BPI Challenge 2017, 2017.
- S. J. van Zelst, A. Bolt, M. Hassani, B. F. van Dongen, and W. M. P. van der Aalst. Online conformance checking: relating event streams to process models using prefix-alignments. *Int. J. Data Sci. Anal.*, 8(3):269–284, 2019.
- H. Wang, W. Fan, P. S. Yu, and J. Han. Mining Concept-drifting Data Streams Using Ensemble Classifiers. In *KDD '03*, pages 226–235. ACM, 2003.



Yorick Spenrath is a PhD Candidate at the Process Analytics group in the department of computer science at Eindhoven University of Technology (TU/e), The Netherlands. Previous to that, he got his bachelor's and master's degrees in software science at TU/e, graduating on the topic of bottlenecks in stream process mining under concept drift.



Marwan Hassani is assistant professor at the Process Analytics group in the department of computer science at Eindhoven University of Technology (TU/e), The Netherlands. Previous to that, he worked as a postdoc at the Data Management and Data Exploration Group at RWTH Aachen University, Germany. His research areas are: data mining, data science, process mining and big data analytics. Current interests are: stream data mining, real-time process mining, sequential pattern mining, privacy-aware process mining, subspace clustering and evolving graph mining. Marwan is using customer journey optimization as a use case in several running projects. He has co-authored more than 65 scientific publications and is acting as an organizer and a regular reviewer in several international conferences and journals.

Estimating Relationships in Multi-Dimensional Data Sets by Means of Asymmetric Fuzzy Regression

Raphael A. Krauthann, Tobias Kruse, Hinnerk Jannis Müller, Michael Stumpf, and Peter Rausch

Department of Computer Science

Nuremberg Institute of Technology Georg Simon Ohm

Keßlerplatz 12, 90489 Nuremberg, Germany

{krauthannra64754|kruseto64083|muellerhi65413|michael.stumpf|peter.rausch}@th-nuernberg.de

KEYWORDS

Fuzzy Regression Analysis, Modeling, Machine Learning, Data Mining, Predictive Analytics, Outlier Detection, Asymmetric Fuzzy Regression

ABSTRACT

In spite of all progress of AI and Machine Learning, making predictions based on real-world data is still a challenging task. For this purpose, Tanaka's approach of symmetric fuzzy linear regression is explained, and open issues are outlined. These issues occur if the instances of data sets are not symmetrically distributed. For this purpose, new solutions based on enhancements of Tanaka's approach are discussed. A real-world scenario for predicting house prices is used to illustrate the ideas. It is shown that the new asymmetric approach works at least as well as the symmetric version but is superior in certain situations.¹

I. INTRODUCTION

Although a lot of research has been done in the field of AI and Machine Learning, making predictions based on real-world data is still a challenging task. Thus, a new approach is presented to estimate relationships in multi-dimensional data sets by applying an asymmetrical fuzzy regression technique. Regression approaches are prevalent and applied in numerous areas, for instance, in the fields of financial risk measurement (Valaskova et al. 2018), sales revenue prediction in the telecommunications industry (Welc and Esquerdo 2018), or stock price prediction (Patel, Patel, and Darji 2018). Nevertheless, in many cases, a point forecast is not always appropriate (Rausch and Jehle 2013). For instance, if the impact of weather parameters on sales of ice cream is analyzed, no clear causal relationship between independent and dependent variables is apparent, but a fuzzy linear correlation can be observed (Rausch and Jehle 2013). To solve this type of real-world issues, fuzzy versions of regression approaches were developed, see for instance (Tanaka, Uejima, and Asai 1982). Approaches providing lower

and upper boundaries for fuzzy linear relationships are already available. However, this paper shows that these approaches are not appropriate for all types of data sets, for instance, if symmetric lower and upper boundaries are an inadequate representation of a fuzzy relationship. Thus, in the following sections, a new solution to overcome this issue is presented. Section II describes a real estate data set which will be used for illustration purposes. Section III provides a brief survey of available solutions and their limitations. In Section IV, details of Tanaka's popular approaches are provided, and enhancements to handle issues in asymmetrically distributed data sets are introduced as a new approach. To illustrate its features in Section V, it is applied to the data set, and the impact of predictors on sales prices of houses is analyzed. In Section VI, the findings are evaluated. It is worked out in which situations the associated features are beneficial. Note that a comparison to other types of AI techniques is not within the scope of this paper and should be handled in a follow-up study. Finally, plans for future research into further enhancements and other fields of application are presented.

II. DATA SET

The data set used in the following sections to illustrate the new approach represents sales of individual residential property in Ames, Iowa, from 2006 to 2010. It contains 2930 transactions and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) that play a role in the valuation of property values (Cock 2011). This data set is the basis of the Kaggle Competition "House Prices: Advanced Regression Techniques". The contest's goal is to predict the sales price for each house. In order to discuss the features of the new fuzzy method presented in this paper, only a few attributes of the data set are used. The temporal attributes considered are the year of construction, the year of renovation, and the month and year of sale. Furthermore, the data set contains information on lot size, the quality of the material used, and its condition, which are also considered.

III. RELATED WORK

The data set described has been used in several publications, such as (Fan, Cui, and Zhong 2018; Vik-

¹The algorithms in this paper were implemented in Python 3 and are freely accessible at (https://github.com/TheGarkine/fuzz_regression_py).

torovich et al. 2018; Yan 2017). The data set analyses followed a similar process:

First, the data set was preprocessed, removing outliers and handling missing values, as well as removing unwanted features and possibly adding custom features. Subsequently, several base methods and one or more ensemble methods were selected. Fan et al. used Lasso and Ridge linear regression models, which introduce penalty terms. Besides, random forest, support vector regression with linear and Gaussian kernel, and extreme gradient boosting trees as base methods are applied (Fan, Cui, and Zhong 2018). As ensemble strategy, they used a simple weighted linear combination of the best performing models. Viktorovich et al. used Lasso linear regression, ElasticNet, extreme gradient boosting trees and neural networks as base methods. Different combinations were assessed as integrating ensemble methods (Viktorovich et al. 2018). Whereas Yan used gradient boosting trees, random forest, and regularized regression with both Lasso and Ridge penalization as base methods (Yan 2017). On top of that, different ensemble methods were applied, whereby a multi-class ROC random forest was proposed.

After selecting appropriate base and ensemble methods, a hyperparameter optimization took place, where the best performing combination of the selected methods was determined. Since the output of the base models is the input for the ensemble methods, this is a multi-layered process or stacked generalization as described in (Wolpert 1992). Decent results are achievable with this practice. However, in many cases, the results of ensemble methods are difficult to understand and are often treated as opaque “black-boxes” (Cortez and Embrechts 2011). There are approaches and ongoing research summarized under the term Explainable Artificial Intelligence (Arrieta et al. 2020), which try to make “black-box”-models more understandable, for instance, with sensitivity analyses (Cortez and Embrechts 2011), by extracting rules (Tickle et al. 1998) or by visualizing different layers (Yosinski et al. 2015). To avoid the issue mentioned above, the focus in the following sections is on fuzzy linear regression.

While conducting the literature review, it became clear that developments on the research topic of fuzzy regression have not received much attention in the last two decades. Thus, there is not much preliminary work on this research subject. However, a few sources could be identified: D’Urso and Gastaldi presented an asymmetric fuzzy regression approach for crisp data sets with a fuzzy output component using a new metric (D’Urso and Gastaldi 2001). In (Neubauer 2010), much knowledge about fuzzy regression and further developments can be found. The methods discussed can handle fuzzy input, which is not the focus of this paper.

One of the first publications in this field is from Tanaka et al., who also tried to model house prices (Tanaka, Uejima, and Asai 1982). They used available crisp data in five dimensions to estimate prices for Japanese prefabricated houses using an algorithm with a linear optimization problem.

Diamond also investigated fuzzy regression in 1988 (Diamond 1988). He solved the problem of fuzzy-to-fuzzy regression by implementing a new metric and proving its properties. Additionally, his ideas of crisp-to-fuzzy regressions were presented. His regression resulted in symmetric triangular fuzzy numbers.

In 1997, Tanaka and Lee published a fuzzy regression approach to process crisp data sets (Tanaka and Lee 1997). Their work is the basis for the approaches presented in this paper. Therefore, their ideas are outlined in Section IV. Additionally, Lee and Tanaka presented a non-symmetric fuzzy regression idea with crisp input and output, in 1999 (Lee and Tanaka 1999). However, the approach presented has the significant drawback that it weights the fuzzy properties to the linear correlation unequally. This feature is explained in Section VI.

IV. APPROACH

This section presents the general approaches, starting with a definition of fuzzy linear functions, their properties, and other fundamentals. Afterward, Tanaka’s approaches are reviewed. Note that many regression algorithms are often denoted using matrices and vectors (see (Neubauer 2010) or (Tanaka, Uejima, and Asai 1982)), while this paper relies on sums over real values, iterating through input dimensions. This is done to make the general idea of the algorithms and the underlying quadratic optimization problem more understandable. Based on this, a new asymmetric approach is introduced. Tanaka’s so-called optimization without expert knowledge (Tanaka and Lee 1997) is also integrated into this approach.

A. Fundamentals

A.1 Data Set Notation and Definition

The data set shall be defined as D containing p instances. Each instance of D has $n + 1$ dimensions or attributes. For the sake of this paper, x_j means the j -th data point with all its attributes. Therefore, the notation x_{ji} represents the value of the i -th dimension of the j -th instance within D . The algorithms expect numerical, non-negative values (\mathbb{R}_0^+). Concluding, the data set can be defined as

$$D \subset \mathbb{R}_0^{+n+1}, |D| = p \quad (1)$$

A.2 Fuzzy Linear Regression

The goal is to find a fuzzy linear function $f : \mathbb{R}^n \rightarrow \tilde{\mathbb{R}}$. These functions can be generally described as:

$$Y = \sum_{i=0}^n \gamma_i X_i \quad (2)$$

where

$$X_0 = 1 \quad (3)$$

$$X_i \in \mathbb{R}, i = 1, \dots, n + 1 \quad (4)$$

$$\gamma_i \in \tilde{\mathbb{R}}, i = 0, \dots, n \quad (5)$$

In Equation (2), n denotes the number of input dimensions in the data analyzed. The X_i represents the crisp i -th dimension of the data set, while γ_i values

are fuzzy coefficients calculated by the linear regression algorithm for the i -th dimension. For algorithmic purposes, X_0 is initialized with the value of *one* so that the sum includes the γ_0 case. In accordance with the data set notation, X_{n+1} denotes the dimension analyzed that is to be approximated.

In the following sections, the performance of three algorithms inspired by Tanaka's approach (Tanaka and Lee 1997) is compared. Additionally, these solutions are transferred to an asymmetrical algorithm to solve the issues mentioned in Section I. The goal is to compare the approaches in both the performance and relevance of their results. Each algorithm gives a fuzzy linear approximation, including a centerline and lower and upper boundaries for expected values. These boundaries may also be referenced as a tunnel. This concept is typically not implemented by other algorithms, such as neural networks.

Asymmetric approaches could be superior in representing real-world scenarios (described in Subsection IV-D). This work tests this hypothesis. All of the procedures presented provide a linear function with triangular fuzzy numbers, which are either symmetric or asymmetric. Symmetrical triangular fuzzy numbers, which can be used to represent the γ_i parameters are, denoted as $(a; c)_S$, where a is the center of the number and c the symmetric spread. Asymmetrical triangular fuzzy numbers are denoted as $(a; l; u)_{LR}$, where a is the center of the number, l , and u describes the lower and upper boundary distances.

A.3 h -Level

Some fuzzy algorithms have an h -level parameter (Tanaka and Lee 1997). The h -level defines the minimum membership for each data point of the data set with the resulting linear regression. The closer the h -level comes to 1, the wider the tunnel becomes. Thus, the fuzziness of the resulting regression increases.

B. Linear Regression with Symmetric Triangular Fuzzy Numbers, Ignoring Linear Correlations

Using Equation (2) the linear goal function for symmetric triangular fuzzy numbers can generally be described as follows:

$$Y = \sum_{i=0}^n (a_i; c_i)_S X_i \quad (6)$$

For the first version of the algorithm, the linear correlation within the data set is ignored. Instead, the spread of the resulting symmetric triangular fuzzy coefficients is reduced, so that all values are within the h -level of the fuzzy function. Therefore, the following linear programming problem has to be solved.

$$\min_{(a_0, c_0), \dots, (a_n, c_n)} \sum_{i=0}^n \sum_{j=1}^p c_i x_{ji} \quad (7)$$

subject to:

$$\sum_{i=0}^n \left((a_i + (1-h)c_i) x_{ji} \right) \geq x_{j(n+1)}, \quad j=1, \dots, p \quad (8)$$

$$\sum_{i=0}^n \left((a_i - (1-h)c_i) x_{ji} \right) \leq x_{j(n+1)}, \quad j=1, \dots, p \quad (9)$$

$$c_i \geq 0, \quad i=0, \dots, n \quad (10)$$

The constraints defined in (8) and (9) cause the regression to have all values within the spread with a minimum membership of h . Additionally, it is desirable to prevent spreads to become less fuzzy for higher numbers, which is achieved by (10) (Tanaka and Lee 1997).

C. Linear Regression with Symmetric Triangular Fuzzy Numbers, Using Linear Correlations

Ignoring the data set's linear correlation between dimensions can result in errors in the interpretation of the resulting linear regression. Hence, this property should not be ignored. Tanaka provides a solution combining the properties of least-squares regression with the fuzzy effect of Section IV-B (Tanaka and Lee 1997). Again, the goal is to create a function similar to Equation (6).

The resulting quadratic optimization problem can now be written as:

$$\min_{(a_0, c_0), \dots, (a_n, c_n)} k_1 \sum_{j=1}^p \left(x_{j(n+1)} - \sum_{i=0}^n a_i x_{ji} \right)^2 + k_2 \sum_{j=1}^p \left(\sum_{i=0}^n c_i x_{ji} \right)^2 \quad (11)$$

subject to (8), (9) and (10).

Since the same constraints are still valid, those from the previous approach are reused. The factors k_1 and k_2 are user-specified positive real values and enable adjustments of the function. The ratio between k_1 and k_2 influences the weight of the linear property in contrast to the slimness of the resulting tunnel. Therefore, k_1 and k_2 can be seen as the importance of the linear property and the crispness, respectively.

D. Linear Regression with Asymmetric Triangular Fuzzy Numbers, Using Linear Correlations

In the asymmetric case, the definition of the linear asymmetric fuzzy numbers is taken from Subsection IV-A and transformed into the following general function:

$$Y = \sum_{i=0}^n (a_i; l_i; u_i)_{LR} X_i \quad (12)$$

Thus, a modified version of Tanaka's symmetric fuzzy regression algorithm (Tanaka and Lee 1997) is used. Asymmetric approaches have the advantage that they are not so much affected by outliers lying only above or beneath the centerline. In this case, an extreme point only affects the side it occurs on (above or below the actual linear relationship). For the optimization, the

squares of both boundaries are combined. Since the fuzziness has now potentially twice the importance in the optimization, the combination of the two parts is halved. Finally, this result is combined with the least square regression component, and both are weighted with the two crisp factors, k_1 and k_2 .

$$\min_{(a_0, l_0, u_0), \dots, (a_n, l_n, u_n)} k_1 \sum_{j=1}^p \left(x_{j(n+1)} - \sum_{i=0}^n a_i x_{ji} \right)^2 + \frac{k_2}{2} \sum_{j=1}^p \left(\left(\sum_{i=0}^n l_i x_{ji} \right)^2 + \left(\sum_{i=0}^n u_i x_{ji} \right)^2 \right) \quad (13)$$

$$\sum_{i=0}^n \left((a_i + (1-h)u_i) x_{ji} \right) \geq x_{j(n+1)}, \quad j=1, \dots, p \quad (14)$$

$$\sum_{i=0}^n \left((a_i - (1-h)l_i) x_{ji} \right) \leq x_{j(n+1)}, \quad j=1, \dots, p \quad (15)$$

$$u_i \geq 0, l_i \geq 0, \quad i=0, \dots, n \quad (16)$$

As in the symmetric case, the lower boundary needs to be less than or equal to, and the upper boundary greater than or equal to all recorded values. Therefore, Equations (8), (9), and (10) are modified to use the new lower (l) and upper (u) boundaries.

Figure 1 shows the asymmetric approach applied to the test values from Tanaka's work (Tanaka and Lee 1997). The boundaries of the asymmetric fuzzy regression are closer to the values of the data set compared to Tanaka's symmetrical fuzzy regression approach, since each side is only limited to their respective extrema. Note that this is limited to the space of the training set and is not valid for predictions outside of this area.

In addition, Figure 1 shows the impact on regression. $k_1 = 1$ and $k_2 = 10$ are set for the symmetrical and asymmetrical approaches, meaning that the slimness of the tunnel is more important than the least-squares regression of the centerline. Note that the upper and lower boundaries are not drawn for the symmetrical case, since they are very similar to the asymmetrical case with these parameters. The centerline of the asymmetric case (plotted solid) better matches the actual linear relationship. This effect can be measured using the root mean squared error, which is measured between the actual data points and the centerline. The symmetrical case scores 1.88, and the asymmetrical reduces this by approximately 8% to 1.73 for the given data set.

E. Optimization Using Tanaka's Reliable and Suspicious Sets

In the symmetric cases of Subsection IV-C and Section IV-D, filtering untypical values is also possible. For this purpose, Tanaka's definitions of *suspicious* and *reliable* values are used (Tanaka and Lee 1997). First, a crisp least-squares regression is computed, and the resulting coefficients are defined as $\alpha = \alpha_0, \alpha_1, \dots, \alpha_n$. Subsequently, the standard deviation σ based on the resulting regression is calculated. A point of the data set is defined *reliable* when it is within the interval:

$$[\alpha x_j - t\sigma; \alpha x_j + t\sigma], \quad j=1, \dots, p \quad (17)$$

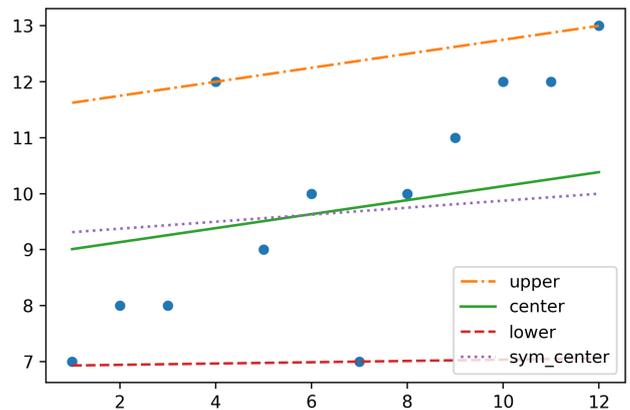


Fig. 1: Visualization of an Asymmetric Triangular Fuzzy Regression and Comparison to the Symmetrical Center Line

t is a factor used to widen or tighten the set of *reliable* values (Re). Values outside of this interval are members of the set of *suspicious* points (Su).

The fuzzy objective function Equation (2) can now be extended by the fuzzy error parameter E , which is of type $(0; e)_S$.

$$Y = \sum_{i=0}^n \gamma_i X_i + E \quad (18)$$

This error E should be minimal. Hence, it is considered within the optimization process as shown in the new goal (19):

$$\min_{(a_0, c_0), \dots, (a_n, c_n), e} k_1 \sum_{j=1}^p \left(x_{j(n+1)} - \sum_{i=0}^n a_i x_{ji} \right)^2 + k_2 \sum_{j=1}^p \left(\sum_{i=0}^n c_i x_{ji} \right)^2 + k_3 e^2 \quad (19)$$

subject to Equation (10) and

$$\sum_{i=0}^n \left((a_i + (1-h)c_i) x_{ji} \right) \geq x_{j(n+1)} \quad \forall x_j \in Re \quad (20)$$

$$\sum_{i=0}^n \left((a_i - (1-h)c_i) x_{ji} \right) \leq x_{j(n+1)} \quad \forall x_j \in Re \quad (21)$$

$$\sum_{i=0}^n \left((a_i + (1-h)c_i) x_{ji} \right) + e \geq x_{j(n+1)} \quad \forall x_j \in Su \quad (22)$$

$$\sum_{i=0}^n \left((a_i - (1-h)c_i) x_{ji} \right) - e \leq x_{j(n+1)} \quad \forall x_j \in Su \quad (23)$$

Note that (22) and (23) are relaxed versions of the previously known constraints since they consider e . These constraints create two *tunnels*. The inner one requires all *reliable* values to be within the boundaries, while all *suspicious* values can be further away within the outer boundaries.

F. Integrating the Optimization Technique into the Asymmetric Approach

Finally, Tanaka's method of reducing the impact of outliers can be applied to the asymmetric adaptation. The general regression function is equal to Equation (18). All coefficients γ_i are LR fuzzy numbers.

Note that the fuzzy error parameter E is now asymmetric as well and can be represented by $(0; e_l; e_u)_{LR}$. This results in the following quadratic optimization problem:

$$\min_{(a_0, l_0, u_0), \dots, (a_n, l_n, u_n), e_l, e_u} k_1 \sum_{j=1}^p \left(x_{j(n+1)} - \sum_{i=0}^n a_i x_{ji} \right)^2 + \frac{k_2}{2} \sum_{j=1}^p \left(\left(\sum_{i=0}^n l_i x_{ji} \right)^2 + \left(\sum_{i=0}^n u_i x_{ji} \right)^2 \right) + k_3 (e_l^2 + e_u^2) \quad (24)$$

subject to Equation (16) and

$$\sum_{i=0}^n \left((a_i + (1-h)u_i) x_{ji} \right) \geq x_{j(n+1)} \quad \forall x_j \in Re \quad (25)$$

$$\sum_{i=0}^n \left((a_i - (1-h)l_i) x_{ji} \right) \leq x_{j(n+1)} \quad \forall x_j \in Re \quad (26)$$

$$\sum_{i=0}^n \left((a_i + (1-h)u_i) x_{ji} \right) + e_u \geq x_{j(n+1)} \quad \forall x_j \in Su \quad (27)$$

$$\sum_{i=0}^n \left((a_i - (1-h)l_i) x_{ji} \right) - e_l \leq x_{j(n+1)} \quad \forall x_j \in Su \quad (28)$$

The calculation of the reliable and suspicious sets Re and Su , according to Tanaka's approach, is described in Section IV-E.

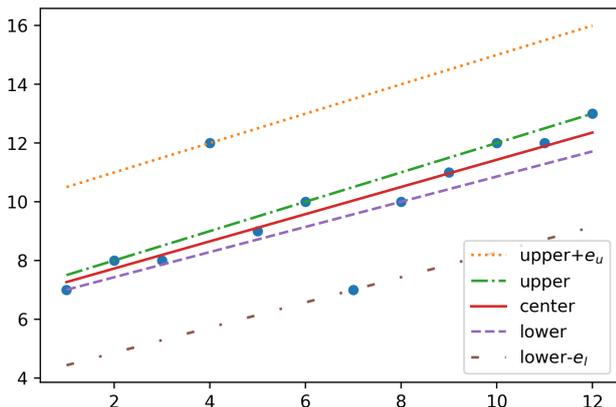


Fig. 2: Asymmetric Triangular Fuzzy Regression with Outlier Detection

In Figure 2, an asymmetric fuzzy regression with outlier elimination is visualized. The algorithm was executed with the parameters $k_1 = 1$, $k_2 = 10$, $k_3 = 1$ and $t = 2$. Obviously, the tunnel becomes significantly thinner to match the data set more closely than before since it is allowed to ignore the outer two data points. The calculated coefficients for the approaches with outlier detection (OD) and without (\overline{OD}) and – if applicable – error values are described in Table I.

TABLE I: Coefficients and Error Values

Alg.	e_l	l_0	l_1	a_0	a_1	u_0	u_1	e_u
OD	2.57	0.23	0.03	6.8	0.46	0.2	0.03	3
\overline{OD}	-	1.97	0.11	8.88	0.12	2.6	3e-6	-

V. EXPERIMENTS

After the implementation of all approaches, their performances on the data set were tested. At the beginning, only seven of the original eighty attributes have

been used. The selection was based on experimenting and calculating their Pearson correlation to the target value.

The best combination of features was found by trying various permutations of features in combination with different values of k_1 , k_2 , k_3 and t . The values of the weights $k_{\{1,2,3\}}$ were in $\{1, 10\}$ while t was selected from the set $\{0.5, 1, 2\}$. Since the model calculations run rather fast, the complete permutations of the seven input features were tested. This approach resulted in 28,585 initial tests.

After the first batch was evaluated, it was assumed that the fuzzy linear model is more suitable when the number of features increases. Therefore, a second batch has been prepared using additional five feature dimensions, increasing the input dimension to $n = 12$. Besides, the set of the $k_{\{1,2,3\}}$ has been increased to $\{1, 10, 100, 1000\}$, and t could be any of $\{0.2, 0.5, 1, 2, 3, 5\}$.

Since this confirmed a hypothesis about the influence of parameterization on the metrics observed, a third batch of experiments was executed, using all of the numerical input data except those features which can be derived from other features. This resulted in an input dimension $n = 34$.

VI. EVALUATION

After testing many configurations of the algorithms, enough data was gathered to evaluate the results. The evaluations were conducted after each batch. Two metrics were used for comparison: the root mean squared error (RMSE) and the root mean squared logarithmic error (RMSLE). The latter is also used as a comparison metric in the Kaggle Challenge for this data set.

In the first batch, the asymmetric approach was significantly superior when compared to the symmetric approaches. In the following batches, when the number of input dimensions was increased, the symmetric approach presented by Tanaka (Tanaka and Lee 1997) performed as well as the extended asymmetric approach. Table II shows the best results that were gathered.

Rows 1 to 4 represent the best results of the first batch, which were achieved by using all seven features. Note that the first experiments also tested all permutations between the seven dimensions selected. Row 1 and 2 include the best (lowest) RMSE, and the next two rows the best RMSLE values.

The best performing symmetrical (Tanaka's) approach was the one with outlier detection (SOD). This has proven to be the best solution for the new asymmetrical version (abbreviated as AOD), too. It should be mentioned that the symmetric and asymmetric models without outlier detection were tested as well, but they never achieved the highest score of any batch.

Rows 5 to 8 show the impact of more extreme values of $k_{\{1,2,3\}}$ and t on the results. Five additional input features extended these tests. The next rows (9 to 12) show how the algorithms perform on all available, non-linear dependent input features.

TABLE II: Experimental Results

Alg.	Feat.	k1	k2	k3	t	RMSE	RMSLE	
1	SOD	7	1	1	10	1	43807	0.730
2	AOD	7	10	1	1	2	43630	0.779
3	SOD	7	1	10	1	1	61499	0.310
4	AOD	7	1	10	10	2	49326	0.269
5	SOD	12	1000	10	1	5	35048	0.733
6	AOD	12	10	1	100	3	37786	0.185
7	SOD	12	1	1	10	2	37296	0.187
8	AOD	12	1	10	1000	2	39618	0.185
9	SOD	34	100	1000	1	1	30683	0.199
10	AOD	34	10	1000	1	1	30689	0.199
11	SOD	34	1	1	100	1	32901	0.166
12	AOD	34	1	1	100	1	33186	0.167
13	L99	34	1	1	-	-	31071	0.216
14	LSR	34	-	-	-	-	31071	0.218

In row 13 (L99), the results of Lee’s algorithm from 1999 (Lee and Tanaka 1999) are used with the parameters optimized for RMSE. The final row 14 shows the RMSE and RMSLE scores of the least-squares regression (LSR). Lee’s approach and the LSR yield almost identical results. This can also be seen when comparing the a -component of the fuzzy coefficients with the crisp coefficients of the LSR. Hence, the difference between them is never higher than 0.02%. Since the boundaries are not squared, the weighting between the least-squares and the tunnel optimization is skewed. Thus, k_1 and k_2 can not be seen as equal weights, and the influence of k_2 depends on the data set. For comparison to Equation (24), Equation (29) describes the fuzzy component of the optimization problem of Lee and Tanaka’s algorithm, which is similar to the approach.

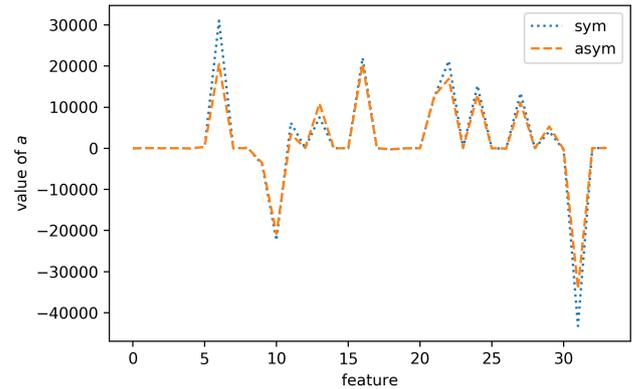
$$L99_{\text{fuzzy}} = k_2 \sum_{j=1}^p \left(\sum_{i=0}^n l_i x_{ji} + \sum_{i=0}^n u_i x_{ji} \right) \quad (29)$$

As already mentioned, the following experiments with a higher number of input dimensions yielded the same results for symmetric and asymmetric approaches. This insight is clearly outlined in Table II (see rows 9 to 12). The final experiments resulted in an almost identical configuration for the SOD and AOD parameters.

Interpretation of the Results

From the experimental results, it can be concluded that the new asymmetrical approach has the advantage of boundaries, which can be derived from a richer solution space due to their independence. If a data set has only outliers above or below the linear tendency, the algorithm is not forced to give both boundaries more room in an equal, symmetric fashion. As the number of input features increased, no noticeable linear correlation between many of them and the target dimension could be found. Since the test set is rather large, this may result in a symmetrical distribution of the data points beneath and above the linear tendency of the sales price. Eventually, this explains the algorithms’ pattern of returning the same results ($c_i = l_i = u_i$) for many dimensions. This results in a_i , which is cal-

culated the same way. The results were measured and are shown in Figure 3.

Fig. 3: Visualization of the Asymmetric and Symmetric Algorithm a_i Results

VII. CONCLUSION AND FUTURE WORK

In this work, a new algorithm has been presented using Tanaka’s (Tanaka and Lee 1997) approach to generate fuzzy linear regressions based on crisp input data and outlier detection. It returns asymmetrical triangular fuzzy numbers instead of symmetrical ones. Furthermore, different fuzzy regression approaches have been tested in a real-world scenario, predicting house prices for a given data set. Based on these experiments, it can be said that the asymmetrical approach should be preferred when calculating fuzzy linear regression with crisp input, since it performs at least as good as the symmetrical idea and outperforms other fuzzy regression approaches which are dependent on the input data. Apart from these promising results, the transparent functionality of the new approach is a remarkable advantage, due to the explainability of its results. At any point during the analysis process, the results of the algorithm is comprehensible.

In the future, it has to be investigated, which parameters affect the performance of the algorithm. The asymmetric fuzzy coefficients have approximately 50% more parameters. This impacts the runtime of the QP solving algorithms. The computational effort depends on the approach chosen. A comparison with other regression algorithms with respect to transparency and accuracy, as well as algorithmic runtime would be conceivable. Furthermore, it would be interesting to apply the approach to different scenarios. Another idea could be to expand fuzzy linear regression with crisp input data to other fuzzy distributions. Currently, triangular fuzzy numbers are used to create the fuzzy linear functions, although in many cases, the data are not distributed linearly. Using Gaussian bell curves could improve the performance of the approaches, which needs to be implemented and further investigated. Additionally, a comparison of the results from this study to the output of other types of AI techniques would be very interesting.

VIII. REFERENCES

- Arrieta, Alejandro Barredo et al. (2020). “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58, pp. 82–115.
- Cock, Dean De (2011). “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project”. In: *Journal of Statistics Education* 19.3.
- Cortez, Paulo and Mark J. Embrechts (2011). “Opening black box Data Mining models using Sensitivity Analysis”. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE.
- Diamond, Phil (1988). “Fuzzy Least Squares”. In: *Information Sciences*, pp. 141–157.
- D’Urso, Pierpaolo and Tommaso Gastaldi (2001). “Linear Fuzzy Regression Analysis with Asymmetric Spreads”. In: *Advances in Classification and Data Analysis*. Springer Berlin Heidelberg, pp. 257–264.
- Fan, Chenchen, Zechen Cui, and Xiaofeng Zhong (2018). “House Prices Prediction with Machine Learning Algorithms”. In: *Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018*. ACM Press.
- Lee, Haekwan and Hideo Tanaka (1999). “Fuzzy Approximations with Non-Symmetric Fuzzy Parameters in Fuzzy Regression Analysis”. In: *Journal of the Operations Research Society of Japan* 42.1, pp. 98–112.
- Neubauer, Dagmar (2010). “Fuzzy-Regression bei Fehlern in den Daten : Modellierung und Analysepotentiale”. PhD thesis. Johann Wolfgang von Goethe University.
- Patel, Janki, Miral Patel, and Mittal Darji (2018). “Stock Price Prediction Using Clustering and Regression: A Review”. In: *Sustainability* 3.1, pp. 1967–1961.
- Rausch, Peter and Birgit Jehle (2013). “Data Supply for Planning and Budgeting Processes under Uncertainty by Means of Regression Analyses”. In: *Business Intelligence and Performance Management: Theory, Systems, and Industrial Application*. Ed. by Peter Rausch, Alaa F. Sheta, and Aladdin Ayesh. Springer U.K., pp. 163–178.
- Tanaka, Hideo and Haekwan Lee (1997). “Fuzzy Linear Regression Combining Central Tendency and Possibilistic Properties”. In: *Proceedings of 6th International Fuzzy Systems Conference*. IEEE.
- Tanaka, Hideo, Satoru Uejima, and Kiyoji Asai (1982). “Linear Regression Analysis with Fuzzy Model”. In: *IEEE Transactions on Systems, Man and Cybernetics* 12, pp. 903–907.
- Tickle, A.B. et al. (1998). “The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks”. In: *IEEE Transactions on Neural Networks* 9.6, pp. 1057–1068.
- Valaskova, Katarina et al. (2018). “Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis”. In: *Sustainability* 10, p. 2144.
- Viktorovich, Parasich Andrey et al. (2018). “Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning”. In: *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*. IEEE.
- Welc, Jacek and Pedro J. Rodriguez Esquerdo (2018). *Applied Regression Analysis for Business*. Springer International Publishing.
- Wolpert, David H. (1992). “Stacked Generalization”. In: *Neural Networks* 5.2, pp. 241–259.
- Yan, Jiaju (2017). “Multi-Class ROC Random Forest for Imbalanced Classification”. PhD thesis.
- Yosinski, Jason et al. (2015). “Understanding Neural Networks Through Deep Visualization”.



RAPHAEL A. KRAUTHANN

Raphael Krauthann studies computer science at Nuremberg Tech since 2015. In his studies, he focuses on big data and modeling.

In his professional life, he assists companies in improving their process by digitizing and automating them. Additionally, he researches in the area of embedded systems and control.



TOBIAS KRUSE

Tobias Kruse is a master’s student in computer science at Nuremberg Tech since 2019. He earned his bachelor’s degree with his thesis about time-traveling debuggers. During his bachelor studies, he worked on the conception and implementation of programming languages, compilers, and interpreters. Currently, he focuses his studies on scalable software architectures, big data systems, and data visualization.



HINNERK JANNIS MÜLLER

Hinnerk Müller studies Information Systems & Management at Nuremberg Tech since 2015. He designed various accounting interfaces for tour operator booking systems during his employment as a working student. His bachelor’s degree focused on visualization and creation of dashboards for tour operators’ data. Upcoming research for his master’s thesis deals with Explainable AI.



MICHAEL STUMPF

Michael Stumpf is a research assistant at Nuremberg Tech. He holds a master’s degree in Information Systems, and his current work and research are focused on business intelligence, business analytics, and fuzzy technologies. Additionally, he is interested in the digital transformation of processes and process automation.



PETER RAUSCH

Peter Rausch is a professor of Information Systems at Nuremberg Tech and has published many papers, articles, and book chapters. He holds a Ph.D. in business administration and has spent several years working in the fields of software development, business process optimization, and consulting. His current research activities are focused on fuzzy technologies, business planning, and process automation.

Open and Collaborative Models and Simulation Methods

FUNDAMENTALS OF DIGITAL TWINS APPLIED TO A PLASTIC TOY BOAT AND A SHIP SCALE MODEL

Ícaro A. Fonseca and Henrique M. Gaspar
Department of Ocean Operations and Civil Engineering
Norwegian University of Science and Technology
Larsgårdsvegen 2, 6009, Ålesund, Norway
E-mail: icaro.a.fonseca@ntnu.no

KEYWORDS

Digital twin, Web-based, Simulation, Visualization, Standardization, Open source.

ABSTRACT

The objective of this paper is to present some fundamentals of digital twins that can be applied to examples ranging in different degrees of complexity. The paper presents a common definition of the digital twin concept to examine what are its main elements and how they interact with each other. Such elements are applied to a simple example with a digital twin of a floating body based on computer vision, developed with open source libraries and a web-based approach.

Moving towards a more complex example, the paper presents a digital twin of a ship scale model in waves. The model is equipped with a dynamic positioning system, allowing remote control of the desired setpoint from the digital twin interface. Finally, as a direction for future work, the paper discusses the early efforts on the creation of a digital twin of the research vessel Gunnerus based on the aggregation of data from various instrumentation devices.

ELEMENTS OF A DIGITAL TWIN

The origins of the digital twin concept can be traced to the aerospace and defense industries, with proponents such as NASA and the US DoD. In a draft roadmap from 2010 outlining planned developments (Shafto et al. 2010, p. 18), NASA defines the concept as:

“an integrated multiphysics, multiscale simulation of a vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its corresponding flying twin.”

By aggregating such information, the digital twin would provide operational support in various ways. The report presents four of its expected use cases: simulate a mission before it is actually executed, mirror the behavior of its physical twin during operation, perform in-situ forensics of a potentially catastrophic fault or damage, and serve as a platform for studying the effects of modifications in mission parameters which were not considered during design phase.

To translate the digital twin concept to practical applications, it becomes relevant to identify what are its main principles and elements in the context of engineering problems. In fact, the digital twin is guided by the underlying principle of using a simulation to reproduce the physical constitution and behavior of a physical asset, with the purpose of supporting its operation.

Attempts to classify the composition of a digital twin data contents usually converge to a typology based on three main groups: asset representation, behavioral models and measured data (Cameron et al. 2018; Cabos and Rostock 2018). This last category can be further broken down into data describing the asset’s state and data describing its surrounding context, whether operational, environmental or other (Erikstad 2017). Figure 1 presents the elements of a digital twin according to that framework.

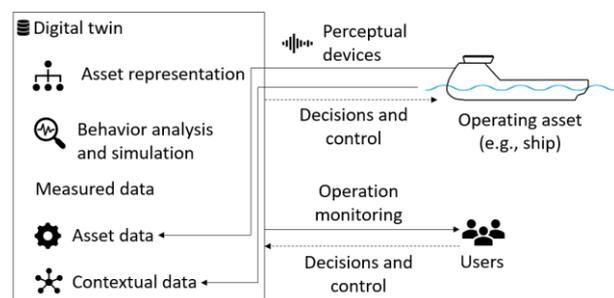


Figure 1: A General Digital Twin Framework

These three main data groups need to be closely aligned in order to interact effectively and fulfill the intended digital twin purpose, so their contents will be to some degree dependent on the domain of interest. In general, the asset representation will commonly include a geometric 3D model of the asset, which can be complemented with relevant metadata about the component description, weight distribution and material models.

The behavior models help the digital twin users to derive operational insight about the asset. Such models offer analyses and simulations that use the measured data to bridge the digital asset representation to the physical reality. The linking of behavior to the

measured data can happen in different manners. A model which feeds directly from sensor log streams can be used for near real-time operation support. On the other hand, historic data can be used to estimate the current asset condition or analyze its performance in previous operations, among others.

The states modelled with measured data can include inspection reports and other information manually entered to the digital twin system. With the advances in technologies for smart devices and internet of things, the novel value of a digital twin will be on extracting insight from sensors and other perceptual devices rather than simply archiving report documents. Some digital twins may also offer the possibility to control the physical asset, though this will not necessarily be a feature of all implementations. For this reason, Figure 1 represents this functionality with dashed lines.

Figure 1 will be used as a template to instantiate the digital twin examples presented in the next sections. The examples employ an open approach as far as possible, allowing collaboration and modification of the source code by interested parties. The digital twin graphic interfaces are developed as web applications. The choice to use a web-based approach is due to reasons which were developed in a previous paper by the authors (Fonseca and Gaspar 2019). In short, web simulations are highly compatible across devices and operating systems due to their reliance on widely adopted open standards such as HTML and JavaScript; they can be shared and accessed across geographically distributed users and their development can make use of various open libraries for multiple purposes, such as performing analyzes and creating visualizations.

A SIMPLE DIGITAL TWIN OF A FLOATING BODY – A PLASTIC TOY BOAT

Digital Twin Setup and Functionality

The digital twin aims to provide a monitoring interface for the motion of a floating object, in this case a toy boat. Figure 2 illustrates the digital twin data flow. The physical setup of the experiment consists in a small aquarium inside which the boat floats. During the data collection (1.), a consumer webcam captures the boat moving in the scene. The webcam streams the captured video to a client (2.). The client executes the digital twin

simulations in real-time on a web browser. In (3.), the client processes the webcam video with a computer vision algorithm to identify and track the boat on a two-dimensional plane with translation (surge, heave) and rotation (pitch). For simplification of the setup, the image gathering and client execution were performed in the same machine, a basic consumer laptop. Alternatively, the camera image could be streamed over a network to share the digital twin with various users.

Once the digital twin can parse the boat movement automatically, it is possible to use the position coordinates to support different functionalities. In (4.), they are monitored with a 3D visualization and a motion plot updated in real-time. This monitoring can be linked to automated reasoning based on the tracked variables, e.g., by making the digital twin automatically emit a warning in case the boat motion crosses a user-specified threshold.

Asset Representation

In a digital twin, the digital asset representation is used to mirror the “life” of a real asset. Given the boat’s “life” in this example is simply its two-dimensional kinematics as a rigid body, a 3D geometry with the same physical proportions as the toy boat suffices as asset representation. The 3D visualization presented in this work are created with Three.js, an open source library that simplifies creation of WebGL scenes and animations (<https://threejs.org/>).

Simulation and Visualization of Asset Behavior

The digital twin should include behavioral models that make use of the data perceived from the physical reality. Thus, in a digital twin the behavioral model is closely related to the perceived data: it should receive a data log and represent it as a meaningful behavior. According to the overall purpose of this example, the digital models for asset behavior should digitally represent the boat motion captured with the camera. To fulfill that purpose, a 3D visualization and a motion plot were created to show the boat motion with the digital twin.

The 3D visualization shows an ocean environment representing the aquarium. It is rendered with water textures and reflection, sky textures and illumination positioned to represent sunlight. It is based on a work

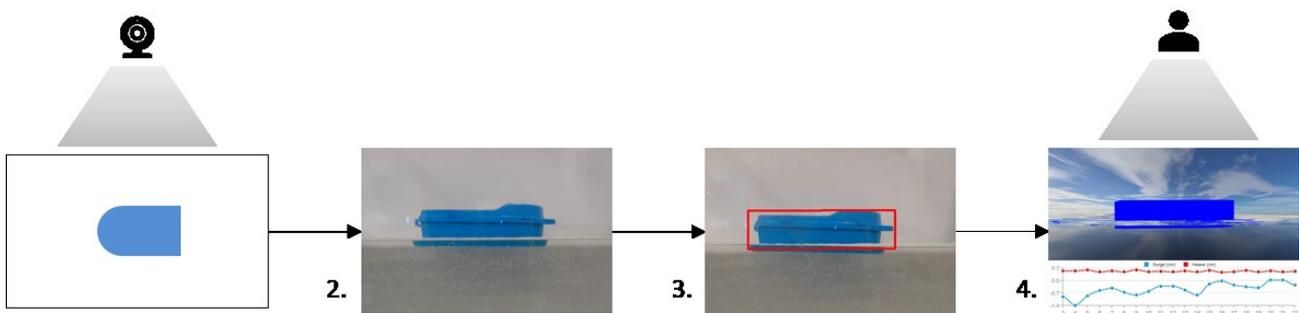


Figure 2: Data Gathering, Transmission, Processing and Usage for the Digital Twin of the Toy Boat

which allowed the user to quickly visualize the motion response of a ship to regular waves (Chaves and Gaspar 2016). In another previous work, it also adapted for usage with the Vessel.js open source library for ship design and simulation (Fonseca et al. 2019). The visualization reused several open source scripts, whose authorship is attributed inside the digital twin repository. Figure 3 shows the 3D digital twin visualization, with the boat representation described in the previous section floating on the water surface. Since the digital twin presented here is not capable of perceiving the water motion inside the aquarium, the water surface in this example is always still.

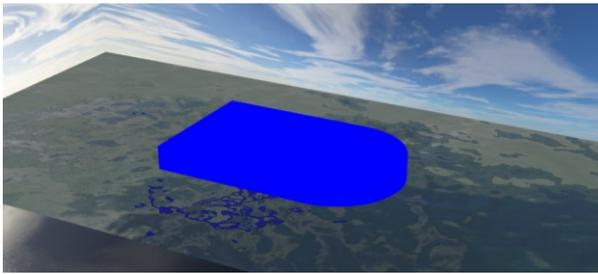


Figure 3: Perspective of the Digital Twin Visualization

As the digital twin receives the coordinates for the three degrees of freedom (DOF) considered in this example, i.e., surge, heave and pitch, the visualizations are updated accordingly. The 2D motion plot displays the current boat coordinates in the time series, while the 3D visualization animates the boat model with new geometry position. The web application executes these operations several times per second, resulting in a responsive simulation showing the motion behavior in near real-time.

Measured Data: Rendering Raw Image into Net Positions

Having the asset representation and the behavior visualizations as the basis for the motion tracking digital twin, let us now investigate the measured data, its role, capture, processing and usage. The digital twin in this example aims to represent the boat motion in real time, so the measured data needs to be handled as a stream log of coordinates (two planar and one rotational) that can be immediately linked to the visualizations. However, in order to obtain that stream, the raw data as captured by the perception device needs to be processed into net data that grasps the physical variables of interest. In this example, this translates to converting the VGA video captured by the web camera to the boat movement in spatial coordinates. This conversion can be performed by calibrating a physical setup for the image recording and then applying a computer vision algorithm to the captured image.

The calibration of the physical setup ensures a consistent and known relation between the physical motion of the boat and the motion of the track box in the video processed from the camera. For that purpose, the

camera was installed in a fixed position by the side of the aquarium in order to capture the boat movements of interest. The mid-section of the aquarium was measured with rulers and a cardboard sheet was placed on that position. The cardboard contained lines of known length that allowed the correspondence between the physical distance on the middle plane of the aquarium and the corresponding pixel distance on the captured video. This calibration method does not account for eventual radial or tangential distortions on the image recorded by the camera. To minimize inaccuracies in the motion tracking, the boat is tied to mooring lines which avoid it from drifting away from the region used as calibration target.

Once the physical setup is arranged to ensure consistent and reliable gathering of raw data, we can proceed to render the raw data into net data, i.e., obtain the movement coordinates from the video source. The OpenCV.js library was used to perform the image processing and motion tracking. OpenCV (Open Source Computer Vision Library) is an open source library for computer vision and machine learning (<https://opencv.org/>). The library is written in C++, but the JavaScript binding offers a subset of the available algorithms, allowing them to be executed directly on the web browser.

More specifically, the example tracks the boat motion with the Camshift algorithm (Bradski 1998). Camshift is based on another algorithm, Meanshift, which identifies objects on video by performing a histogram analysis of image colors and then tracks its motion on the following video frames. Camshift adapts Meanshift by calculating also the size and rotation of the window that best fits the object in the scene. This allows its usage to track also the rotational motion of the boat but tends to make the tracking less stable compared to Meanshift. For the example presented here, Camshift worked reliably given the appropriate object color and scene illumination. The reader can consult the algorithm configuration parameters in the repository linked by the end of the paper.

Results

With all digital twin elements working in conjunction, it is possible to recognize the object on the image, convert its pixel positions to physical coordinates and display those physical coordinates in the visualizations in real-time. For this reason, it is possible to say that the digital twin worked as conceived: it was able to reliably recognize and track the ship model, while the visualizations followed the movement in near real-time for monitoring purposes.

Figure 4 illustrates the digital twin functionality. It shows a screenshot of the web interface with the object tracking box, on the left side, and the corresponding 2D and 3D digital twin visualizations, on the right one. All

the simulation algorithms were executed on a web browser running on a basic consumer laptop.

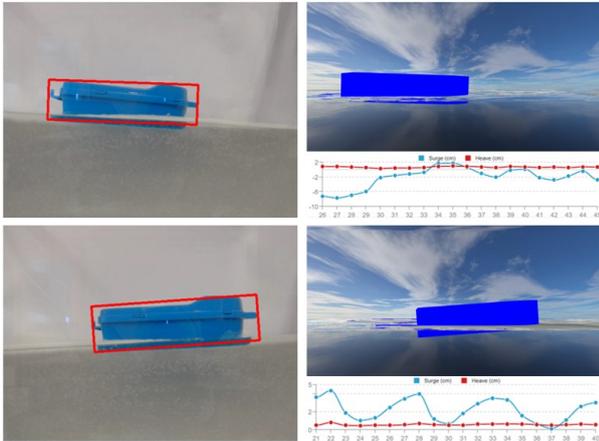


Figure 4: Screenshots of the Digital Twin Interface

Despite this application being a very simple example of a digital twin, it can still be traced to the use case of mirroring the behavior of an asset during operation. It can also be expanded to cover in-situ forensics. For example, the boat positions retrieved as numeric values can be compared with specified thresholds so the operator can be warned in case there is a large response leading to excessive drift or risk of green water.

It is interesting to note that, in this case study, the usage of a single imaging sensor allowed the extraction of three motion modes of a floating object. These results emphasize the importance of purposeful planning and usage of perceptual devices for the creation of digital services, a finding that resonates with other works in the area (Erikstad 2019; Nokkala et al. 2019).

In order to make the jump from mirroring and analyzing current behavior to also predicting and achieving a desired future behavior, the digital twin needs to include data and models that estimate or, ideally, control the asset behavior in prospective situations. The following section gives an example of how this can be achieved.

DIGITAL TWIN OF A SHIP SCALE MODEL

Digital Twin Setup and Functionality

The digital twin in this study case aims to monitor and control a ship scale model navigating in a wave basin. The experiments were performed in the Numerical Offshore Tank at University of São Paulo (TPN-USP). Figure 5 depicts the wave basin, measuring 14 meters on each side and 4.1 meters of depth. It is equipped with flaps that allow generation of regular and irregular waves from a user-specified direction surrounding the scale model. The flaps also work as wave absorbers to minimize interference of wave reflection in the desired wave characteristics.



Figure 5: The Numerical Offshore Tank at University of São Paulo (TPN-USP) (Mello 2012)

Figure 6 shows the scale model used in the experiment, which represents a platform supply vessel. It is actuated with a dynamic positioning (DP) system comprising two azimuth propellers and a bow tunnel thruster.



Figure 6: PSV Scale Model Used in the Experiment (Ilanagui 2019)

The development of the digital twin system followed a bottom-up approach where several subsystems, models and data files already in use on the TPN workflow were aggregated into an overarching architecture. Figure 7 outlines the digital twin framework applied to this example, listing the elements considered for each category. They are detailed in the following sections.

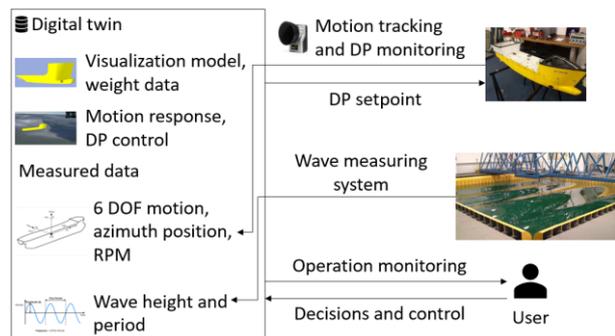


Figure 7: Digital Twin Framework Applied to Experiment with the Scale Model

Asset Representation

The asset representation is again based on a visualization of the ship model, but this time also detailing the installed DP system. It is composed of three different models: hull, azimuth case and propeller.

The hull was obtained by converting the original CAD files to the STL format, which is suitable for 3D visualization and printing. The other two models were already obtained as STL files ready for the intended usage. These three models were replicated and arranged in order to assemble the final asset representation, observed in Figure 8. The assembly considered the movements necessary to represent the DP system operation: the stern systems move on the azimuth plane and the propellers rotate around their central axes.

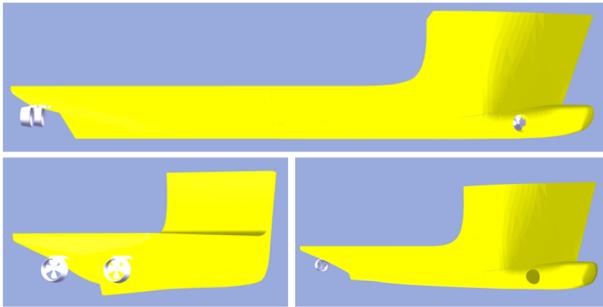


Figure 8: Visualization of the Ship Scale Model with DP System

Besides the geometric data used to create the asset visualization in this example, the manufacturing files of the ship scale model also included Excel tables containing weight data such as weight distribution and center of gravity. This data is not used in the digital twin example presented in this section, but in the future it will be incorporated to the asset representation as metadata for completeness and centralization.

Behavioral Models: Wave Motion and DP Control

The digital twin considers three main groups of relevant behaviors, two directly related to the asset itself and one to its context. The first behavioral model includes the asset motion response to waves. A visualization like the one presented in the previous study case was prepared to show the ship position on its six degrees of freedom. Additionally, a data file with the results of a motion response analysis for several wave conditions, obtained with an external software package, was linked to the digital twin so it could be used for validation and optimization functionalities.

The second group of behavior models and controls the dynamic positioning system. A separate algorithm reads the position on the scale model and controls the propulsion system to attain a desired setpoint (Ianagui 2019). The digital twin communicates with that algorithm both by receiving data with current propulsion parameters and by sending data with desired ship positioning, i.e., the setpoint. The received data allows the digital twin to show the propulsion behavior in the visualization with propeller rotation and azimuth positions. The sent data allows the user to control the ship positioning by sending a setpoint with the three coordinates on the navigation plane: x or surge, y or sway and heading or yaw.

The third and last model accounts for the incident wave. Given the wave height and period, the algorithm calculates the wavelength using the dispersion relation for deep waters. Then, the simulation animates the 3D visualization with a regular wave of corresponding characteristics. By aggregating these three models, the digital twin has the capacity to offer a centralized interface for monitoring and control of the scale model experiment, given that it is provided with the appropriate data measurements.

Measured Data

Each one of the groups of the behavioral models in the previous section is animated by a corresponding set of measured data. The TPN workflow already relied on existing systems for gathering and processing of data for each of the three groups. Such systems were incorporated to the overall digital twin functionality.

The scale model motion is tracked with a stereoscopic system which recognizes five reflective targets fixed to the object, then uses their positions to derive the model motion on six degrees of freedom. The data measuring the behavior of the dynamic positioning system was also easily obtained. The DP control system was configured to stream five parameters to the digital twin: the rotation of the three propellers per minute and the azimuth angle of both stern propellers. The motion and propulsion readings were directly linked to the 3D visualization.

On the other hand, the rendering of water elevation data into wave characteristics was not as straightforward. A probe floating inside the tank, placed near the ship model, was used to measure the wave elevation, but the measuring system had two limitations: first, the wave probe was not capable of sensing the direction of the incident wave, second, the water elevation raw data still needed to be processed into the wave height and period. A few simplifications were adopted to overcome these hindrances. The experiments were performed with regular waves coming from a single direction, and the characteristics of these waves were reconstructed based on the stream of the water elevation log. This reconstruction is performed by identifying the latest wave cycle with one crest and one valley, then calculating the wave height and period so that it can be used on the digital twin simulations.

Results

Figure 9 shows a screenshot of the digital twin visualization mirroring the behavior of the PSV scale model. The visualization can be used to monitor the motion response in 6 DOF, the propulsion system operation and the incident wave.



Figure 9: Screenshot of the Digital Twin Visualization During the Experiment

As in one of the use cases suggested by NASA, the monitoring functionality allows the digital twin to be used as a proxy for the safe and effective operation of the ship model. As the second use case suggests, several times during the experiment the digital twin was used to identify if the ship was able to keep station and test whether the propulsion system was working correctly. For instance, by inspecting the visualization, it is possible to note that the azimuths are locked to the neutral position, a detail that is difficult to observe while the physical scale model is floating during the experiment. In this case, this does not happen by an operational flaw, but by design: the azimuth positions are locked for simplification of an over-actuated problem, so only the tunnel thruster actuates in the direction transversal to the ship.

Another use case mentions the possibility of using the digital twin to predict asset behavior in a future operation. The scale model allows the user to specify a desired setpoint position for the vessel, which will be attained with the corresponding algorithm. This type of control can be allied to optimization algorithms: in one of the functionalities, the user is able to minimize one of the six motion modes of the vessel. Once the user selects the desired mode on a dropdown list, the algorithm searches for the heading that minimizes it according to the stored wave response data and automatically positions the ship with that heading in relation to the incoming wave.

The fourth and final use case mentions the possibility of using the digital twin to study effects of mission parameters that were not considered during the operation. A clear application of that example is the possibility to use the digital twin for validation by comparing the expected motion responses calculated with numerical analyses to the actual empirical response measured during operation. On a real vessel, this type of functionality may help to “close the loop” between design and operation by allowing usage of operational data to guide decisions in future designs.

When these digital twin principles are applied to real operations, it becomes desirable to include techniques to ensure that the sensors are tracking the asset behavior accurately and reliably. In that sense, research on sensor redundancy and diagnostics of erroneous readings may play an important role.

TOWARDS COMPLEX DIGITAL TWINS - ASSEMBLING A DIGITAL TWIN OF A RESEARCH VESSEL

Future research steps will focus on developing a digital twin of NTNU Research Vessel (R/V) Gunnerus, depicted in Figure 10.



Figure 10: R/V Gunnerus, Photo by Fredrik Skoglund

Several vessel systems are already instrumented, and the collected data is currently shared among university members, grouping logs according to their originating sensor. A digital representation of the vessel to be used in the digital twin is already being prepared. Figure 11 shows its preliminary visualization.

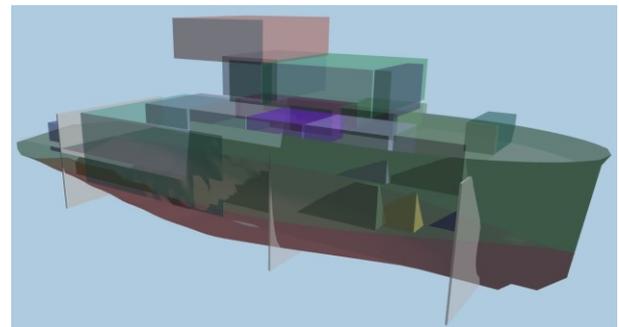


Figure 11: Preliminary Visualization for a Gunnerus Digital Twin

The research will formulate the digital twin requirements based on the research projects carried in the department, which relate to the functionalities requested by the faculty, and on the measured data currently available, which limit the types of behavioral models that can be implemented. This problem formulation will be used to analyze alternatives for the digital twin development, with the resulting models lying on the intersection between the desired and feasible use cases.

Once the stakeholders converge to a digital twin outline, a concept will be implemented with standardized taxonomies and data formats. The concept performance will be evaluated on a practical setup, opening way for identification of improvements and features for further development.

CONCLUSIONS

This paper presented a general framework based on the digital twin definition. The framework can be applied to the modelling of different digital twin examples. Here, it was applied to two study cases following a progression in complexity, the first a toy boat and the second a ship scale model. The asset representation evolves from a simple hull geometry in the first example to a ship model including hull and propeller system in the second. Similarly, the behavioral model is expanded from motion response in 3 DOF to motion response in 6 DOF with monitoring of incident wave and control of the dynamic positioning system. All the behaviors are linked to data measured from the corresponding physical experiments.

Given the novelty of the concept of an integrated digital twin and of its application to the domain of ship operations, we expect the work to contribute to future developments in this area. The study case with R/V Gunnerus, outlined as future work, will provide the opportunity to extend that contribution.

The research makes use open standards to create simulations that can be easily accessed and executed by the users. This is accomplished with web interfaces based on HTML and JavaScript that perform analyses and display visualizations tracking physical behavior in real-time. Future research efforts will focus also on standardization of digital twin data as a method to simplify development and enable reuse of digital models across projects.

SOURCE CODE

The source code for the first example is available on: https://github.com/icarofonseca/dt_cv. The source code for the second example is being prepared for publication and should be available later this year.

ACKNOWLEDGEMENTS

This research is connected to the Ship Design and Operation Lab at NTNU in Ålesund (<http://www.shiplab.ntnu.co/>). The research is partly supported by the EDIS project, in cooperation with Ulstein International AS (Norway) and the Research Council of Norway, and by the INTPART Subsea project in cooperation with the Numeric Offshore Tank at the University of São Paulo (TPN-USP, <http://tpn.usp.br/>) and the Research Council of Norway.

REFERENCES

- Bradski, G. R. 1998. "Real time face and object tracking as a component of a perceptual user interface." In *Proceedings Fourth IEEE Workshop on Applications of Computer Vision*, 214–219.
- Cameron, D. B.; A. Waaler; and T. M. Komulainen. 2018. "Oil and Gas digital twins after twenty years. How can they be made sustainable, maintainable and useful?" In

Proceedings of The 59th Conference on Simulation and Modelling, Oslo, 9–16.

- Chaves, O. S. and H. M. Gaspar. 2016. "A web based real-time 3D simulator for ship design virtual prototype and motion prediction." In *15th Conference on Computer and IT Applications in the Maritime Industries (COMPIT'19)*.
- Cabos, C. and C. Rostock. 2018. "Digital Model or Digital Twin?" In *17th Conference on Computer and IT Applications in the Maritime Industries (COMPIT'18)*.
- Erikstad, S. O. 2017. "Merging Physics, Big Data Analytics and Simulation for the Next-Generation Digital Twins." In *11th Symposium on High-Performance Marine Vehicles*, Zevenwacht.
- Erikstad, S. O. 2019. "Designing Ship Digital Services." In *18th Conference on Computer and IT Applications in the Maritime Industries*, Tullamore, 354–363.
- Fonseca, I. A. and H. M. Gaspar. 2019. "A Prime on Web-Based Simulation." In *33rd International ECMS Conference on Modelling and Simulation*, Caserta.
- Fonseca, I. A.; F. F. de Oliveira; and H. M. Gaspar. 2019. "Virtual Prototyping and Simulation of Multibody Marine Operations Using Web-Based Technologies." In *38th International Conference on Ocean, Offshore and Arctic Engineering*, Glasgow.
- Ianagui, A. S. S. 2019. *Robust system design for consensus control in dynamically positioned vessel fleet*. Ph.D. dissertation, Universidade de São Paulo.
- Mello, P. C. 2012. *Sistema de automação e controle para tanques oceânicos com múltiplos atuadores*. Ph.D. dissertation, Universidade de São Paulo.
- Nokkala, T.; H. Salmela; and J. Toivonen. 2019. "Data Governance in Digital Platforms." In *25th Americas Conference on Information Systems*, Cancún.
- Shafto, M.; M. Conroy; R. Doyle; E. Glaessgen; C. Kemp; J. LeMoigne; and L. Wang. 2010. "Draft modeling, simulation, information technology & processing roadmap." Technology Area 11, National Aeronautics and Space Administration.

AUTHOR BIOGRAPHIES

ÍCARO A. FONSECA is a PhD candidate in engineering at NTNU in Ålesund, developing research on standards for digital twin ships. MSc in Ship Design at the same university with master thesis developed in collaboration with the Marine Research Group at UCL. Engineering degree in Naval Architecture and Marine Engineering at Federal University of Pernambuco (UFPE), Brazil.

HENRIQUE M. GASPAR is an Associate Professor at the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU). The professorship is connected to the Ship Design Chair at the Maritime Knowledge Hub, sponsored by Ulstein Group. Education consists of a PhD degree in Marine Engineering at the NTNU, with research collaboration at UCL (UK) and MIT (USA). Previous professional experience as Senior Consultant at Det Norske Veritas (Norway) and in Oil & Gas in Brazil.

SIMULATION OF THE CONCEPTUAL DESIGN OF OFFSHORE SALT CAVES FOR CO₂ STORAGE

Daniel Prata Vieira
Kazuo Nishimoto
Numerical Offshore Tank
University of São Paulo
São Paulo, SP, Brazil
daniel.prata@tpn.usp.br
knishimo@usp.br

Felipe Ferrari de Oliveira
Henrique Murilo Gaspar
Dept. of Ocean Operations and Civil Eng.
Norwegian University of Science and
Technology
Ålesund, Norway
felipe.oliveira@ntnu.no
henrique.gaspar@ntnu.no

KEYWORDS

Carbon capture and storage, Offshore Salt Cave, Systems Engineering, Web-based Simulation, Virtual Prototype.

ABSTRACT

There is a growing demand for systems to store the excess gas produced by the offshore industry, especially gases rich in CO₂. The geological composition of the Santos Basin in Brazil and the location of the production platforms make it plausible to build caves in the salt layer for this purpose. A system engineering model based on Epoch-Era analyses was employed to analyze the possible solutions to the problem aiming at the conceptual design stage. The system was decomposed and described in means of its several components and subsystems. Some of the most important decisions to be made have been outlined. An estimate of the utility and costs involved was obtained for each epoch. The utility was calculated based on the attributes of each solution and the relevance of each attribute to the stakeholders. An analysis of the solutions tradespace was carried out from the utility and cost estimates to discuss the best alternatives. Also, a three-dimensional web-based simulation model was implemented to provide a more realistic picture of the system's operation.

INTRODUCTION

Oil and gas production in the pre-salt region in the Santos basin is of strategic importance for the development of the Brazilian economy in the coming years. This oil produced in pre-salt has a very high gas-oil-ratio (GOR). While a portion of this gas is treated in the plant and can be used for consumption, another fraction needs to be reinjected into the production well (ANP, 2019) or stored in another location.

Gas storage is a critical problem because a huge volume is required to allocate high production levels. Another problem is that storage sites do not always have robust structures that allow high gas pressurization, further limiting the amount that can be stored

A possible solution to this problem is the construction of underground reservoirs for gas storage and disposal, especially gas highly contaminated with CO₂, considered

polluting and not economically attractive (Costa et al., 2019a).

The works of McCall et al. (2004) and McCall et al. (2005) discuss the potential to use salt layers below the ocean bed for CO₂ storage. The works of Shi et al. (2017) and Londe (2017) present various alternatives and discuss the pros and cons of each one given the technical, economic, environmental, and safety aspects.

A plausible option, specifically for the Brazilian scenario, is the construction of caves in the saline layer located just above the oil-producing fields (Costa et al., 2017).

The innovative project can be achieved by employing modifications in the existing well construction engineering processes and using adaptations in the construction processes of onshore salt caves.

This work presents the modeling of a cave to store the excess of produced gas, using the point of view of systems engineering with a focus on facilitating decision-making in the early stages of the project.

SALT CAVE MODELING

A complete description of the salt-cave system considered in the present work can be found in Costa *et al.* (2019b). A schematic arrangement of the proposed cave for the storage of gas in the salt layer is presented in Figure 1 to illustrate the main components of this system.

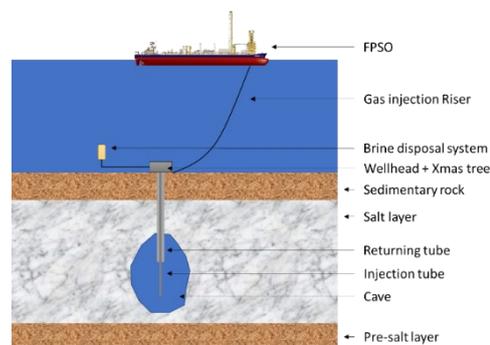


Figure 1: Schematic Arrangement of the Salt-Cave for CO₂ Storage (not in scale; Costa *et al.* 2019b)

The salt cave can be described as a complex system composed of several subsystems that, in turn, have

several components that throughout the project need to be defined and designed. The approach of systems engineering in maritime systems gaining momentum in the last decade (Gaspar *et al.* 2012ab) and seeks to understand how these components work together by hierarchically categorizing the elements.

Exemplifying, the design of a salt cave requires the joint work of various areas of knowledge such as:

- Well construction engineering
- Submarine systems engineering
- Geomechanical study
- Naval systems and operations
- Flow assurance
- Environmental impact
- Logistics

For an integrated analysis of the system, the procedure presented in Figure 2.

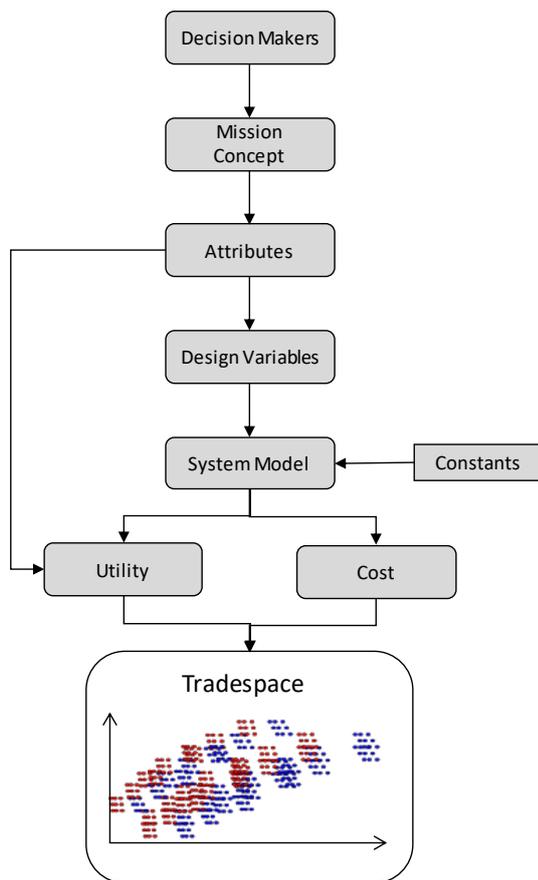


Figure 2: Tradespace Evaluation: Information flow for modeling cost and utility. Adapted from: Ross and Rhodes (2008).

Initially, only the company that has an interest in storing excessive gas will be considered as a stakeholder in the "Decision Makers" block. However, future work may consider multiple stakeholders. Thus, it is possible to state the mission of the concept:

“Build a cave in the salt layer capable of storing excessive gas production, meeting current legal and environmental requirements.”

According to this statement, it is possible to define the system attributes, i.e., the items in which the system will be evaluated. In this way, stakeholders can assess the system's ability to deliver the expected value.

The system attributes were chosen as:

- Flowrate capacity
- Volume capacity
- Environmental Impact
- Safety

Design Variables

Design variables are related to the decisions that define or constrain how the system works and, therefore, its capabilities. Based on the mission statement, it is possible to divide the system life cycle into 4 phases:

1. Well Drilling
2. Construction (Dissolution)
3. Operation
4. Abandonment

The drilling stages comprise the wellhead installation, the well drilling and cementing, Christmas tree installation, and the brine disposal system installation. Their requirements generally define this equipment, so we can identify as essential design variables the nominal values of *diameter* and *pressure* under which the system should operate. An additional

The *driller ship* is the most critical equipment in the drilling stage. The well construction quality and safety are directly related to the capabilities and experience of the company hired to perform the task. These factors have been summarized here as three levels of driller operability, which is linked to the number of days necessary to finish the drilling stage, and consequently, related to the cost of the construction operation.

The dissolution stage comprises two possible solutions for the *dissolution system*: using a dedicated unit (DU) or using the FPSO (Floating Production, Storage, and Offloading) platform available infrastructure.

When using a dedicated unit, a subsea pump will provide water to the cave dissolution. This subsea pump requires an infrastructure consisting of a generator, an umbilical cable to provide high voltage electricity, a water intake pipe, and connection equipment as flowlines, terminals, and jumpers.

The solution using the FPSO structure requires a dry pump located in topside, a suspended riser for water intake, a riser to water injection. For simplicity, at this moment, we can assume that the FPSO power generation module can provide the electricity to the dry pump.

Independent from the dissolution system, after the water pass through the Xmas tree, wellhead, and injection pipe, it will be mixed with dissolved salt. The brine will return to well through the annulus pipe, and it may require a second submerged pump to provide the flowrate until the diffuser located in sea depth with enough current to facilitate the brine dissolution in the seawater.

After defining the equipment set, it is necessary to determine the *dissolution flowrate* and the *dissolution time*, which will result in the cave size.

Once the cave has the desired dimensions, it is necessary to carry out the preparations for the operation phase. In the case of the DU solution, this consists of performing the entire installation of the injection system in the FPSO. In the case of the FPSO solution, it is necessary to disconnect the riser from the water injection system and connect it to the gas production line.

The operating stage consists of all the time when the FPSO will be replacing brine with the gas that will be stored. At this stage, it is necessary to monitor the pressures and structural integrity of the cave and constant care with the discharge of brine into the seabed.

Finally, we have the abandonment phase in which the wellhead is sealed, and all injection equipment is removed. For this phase, it is necessary to ensure the structural stability of the cave and monitoring and safety systems.

Despite the system's various phases, most project decisions are concentrated in phases 1 and 2. In this way, for conceptual design, no additional variable needs to be adopted for phases 3 and 4. Thus, the salt-cave design variables and the parameters used to generate the different solutions are presented in Table 1, which also shows the range and steps of the assumed values.

Table 1: Salt Cave Conceptual Design Variables

Design Variables	Unit	Range	Steps
Selection of Driller	Operability	90% – 99%	3
Nominal Diameter	in	3 – 7	3
Nominal Pressure	ksi	5 – 15	3
Dissolution System	type	DU / FPSO	2
Dissolution Flowrate	m ³ /h	500 – 780	3
Dissolution Time	days	730 – 1095	3

In addition to the design variables, it is necessary to provide the model with the values of some constants. The choice of which parameters to keep constant or variable depends mainly on the designer's experience. However, in the early stages of design, an extensive set of variables can result in long processing time and hinder the process of obtaining insights about the system.

The salt-cave model constants mainly concern in some factors such as cave and FPSO relative location, depth, a vertical dimension of the salt layer in relation to the ocean

bed. In this work, all these things were defined *a priori*. However, choosing the best location to build the cave could be one of the system's variables, for example.

Also, there are some classes of operations and equipment that are always the same, regardless of the solution. For example, all alternatives must pass through the same safety and integrity tests.

System Utility

The system can be assessed by defining a single value that expresses the stakeholder satisfaction. This value is generally defined as the system utility, a value usually taken between 0 and 1, where 0 represents a system that does not deliver value to its user, and 1 represents a system that delivers the highest possible value.

The utility can be determined by evaluating each of the system attributes and using aggregation functions, as presented, for example, in Keeney and Raiffa (1993). This methodology is the so-called Multi-Attribute Tradespace Exploration (MATE).

Each attribute depends on a subset of the Design Variables. For example, the evaluation of Flowrate capacity depends on the Nominal Diameter and the Nominal pressure. Then, for each attribute value X_k will have an associated single-attribute utility function U_k as presented in Figure 3.

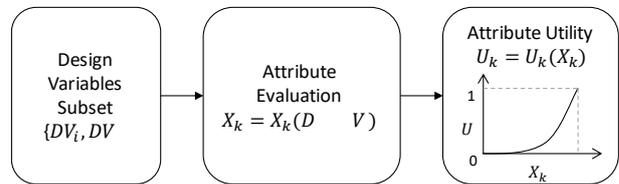


Figure 3: Attribute Utility Evaluation

The multi-attribute utility U is obtained by performing a weighted sum of the attribute's utilities, as shown in Eq. (1):

$$U = \sum_{k=1}^N w_k \cdot U_k(X_k) \quad (1)$$

in which N is the number of evaluated attributes, w_k are the weights defined by stakeholder preferences and by Eq. (2):

$$\sum_{k=1}^N w_k = 1 \quad (2)$$

In present work, it was considered that all four attributes have the same weights $w_k = 0.25$.

Table 2 presents the dependency between the attributes and design variables considered in the present work.

Table 2: Design Variables and Attributes Dependency

Design Variable	Attributes			
	Flowrate capacity	Volume Capacity	Environmental Impact	Safety
Driller				x
Nominal Diameter	x			
Nominal Pressure	x			x
Dissolution System			x	x
Dissolution Flowrate		x	x	x
Dissolution Time		x	x	

System Costs

In addition to calculating the utility, it is necessary to perform an initial estimative of the involved costs for each of the generated solutions. This estimative is just for decision-making purposes, not for budget planning. The most important point is the relative costs of various concepts. The methodology adopted is that presented in Bai and Bai (2018), which consists in to obtain the basic cost of an equipment C_0 and multiply by cost-driving factors f_i . A correction cost C_{corr} can also be applied, as shown in Eq. (3):

$$C_{eq} = C_0 \cdot f_1 \cdot f_2 \cdot f_3 \cdot \dots + C_{corr} \quad (3)$$

The cost-driving factors can be specified, for example, as equipment type, pressure, bore size, or any other characteristic that impacts on equipment price.

Other costs, such as Drill Ship, Platform Supply Vessel (PSV), Offshore Supply Vessel (OSV), Pipe Laying Supply Vessel (PLSV) chartering, or consumable utilization, has been considered according to the average market price and the amount or the time required for each solution.

Besides, the costs can be divided into capital expenditure (CAPEX) and operational expenditure (OPEX) to provide a broader analysis scenario.

For reasons of confidentiality of information, the costs will be presented in a nondimensional way where 1 represents the solution with the highest cost.

The list of materials and equipment used for the cost estimate is shown in Table 3. Some items in this list are just used depending on the solution adopted for dissolution (DU or FPSO).

Table 3: Material and Equipment Considered for Cost Estimation

Drilling phase	Dissolution phase
<i>Casing and tubing</i> <ul style="list-style-type: none"> ▪ External Casing ▪ Intermediary Casing ▪ Internal Casing ▪ Brine return tubing ▪ Seawater injection tubing <i>Consumables</i> <ul style="list-style-type: none"> ▪ Drilling fluid ▪ Cement <i>Drill Ship</i> <ul style="list-style-type: none"> ▪ Drilling, casing, tubing, and integrity tests ▪ Offshore support vessel ▪ Service companies <i>Equipment</i> <ul style="list-style-type: none"> ▪ Wellhead ▪ Christmas Tree 	<i>Lines</i> <ul style="list-style-type: none"> ▪ Umbilical ▪ Riser ▪ Flowline ▪ Contingency riser ▪ Contingency flowline ▪ Brine disposal riser <i>PSV</i> <ul style="list-style-type: none"> ▪ PSV subsea pump installation ▪ PLSV ▪ Mechanical integrity test ▪ Monitoring <i>Equipment</i> <ul style="list-style-type: none"> ▪ Subsea pump (water) ▪ Subsea pump (brine) ▪ Power Module ▪ Diffusers

Tradespace Exploration

The modeling presented in previous sections shown how to determine the stakeholder needs and convert in values. Then, it is possible to perform a tradespace exploration to examine the performance of systems and to verify the relationship between each proposed solution, which helps to better understand the problems related to the key decisions in early design stages.

The results presented in Figure 4 were obtained by applying the utility and costs model to the solutions generated from the combination of the project variables previously shown in Table 1.

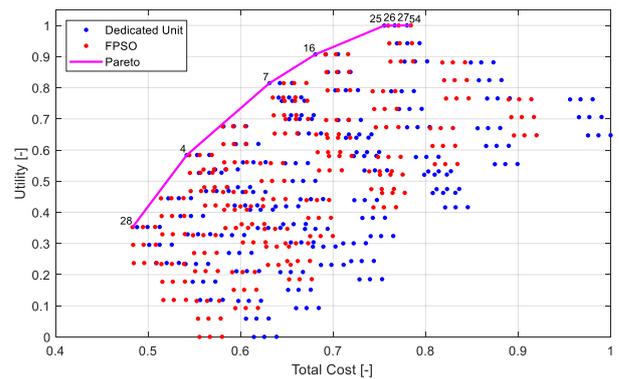


Figure 4: Utility vs. Total Cost for each Generated Solution

In this figure, each point represents a solution; points in blue represent the solutions using a DU, and points in red represent the solutions using the FPSO resources for dissolution. The magenta curve is the Pareto boundary of the simulated cases.

For this simulation, were generated 486 solutions. The number near the Pareto boundary is the ID of the eight

designs contained in the Pareto boundary. These solutions are described in Table 4.

Table 4: Description of Solutions Contained in the Pareto Boundary

ID	Utility [-]	Cost [-]	Driller Oper. [%]	Diam [in]	Press (ksi)	Diss. Sys. [-]	Diss. Flow [m ³ /h]	Diss. Time [days]
4	0.58	0.54	90%	5	5	'DU'	500	730
7	0.81	0.63	90%	7	5	'DU'	500	730
16	0.91	0.68	90%	7	10	'DU'	500	730
25	1.00	0.76	90%	7	15	'DU'	500	730
26	1.00	0.77	95%	7	15	'DU'	500	730
27	1.00	0.78	99%	7	15	'DU'	500	730
28	0.35	0.48	90%	3	5	'FPSO'	500	730
54	1.00	0.78	99%	7	15	'FPSO'	500	730

Essential considerations can be made from the Pareto solutions. First, all solutions presented *Dissolution Flowrate* of 500 m³/h and *Dissolution Time* of 730 days. Both values are the minimum inputs and have a very high impact on system costs. These values will result in caves with less storage capacity, but in the balance of attributes, they are the most efficient.

Six of the eight Pareto solutions consider the *Nominal Diameter* of 7 in. The only change between solutions 4 and 7 is the nominal diameter, which has a small impact on costs (from 0.54 to 0.63) but has a significant influence on utility (from 0.58 to 0.81).

Another interesting comparison is between solutions 27 and 54. The only difference, in this case, is the adopted *Dissolution System*. These solutions deliver practically the same value to the stakeholders (the non-dimensional cost is different just in the third decimal place), which shows that for some parameter combinations, one or another variable may no longer be necessary. In a second step, it is possible to reduce the design variable vector, for example.

WEB-BASED VISUALIZATION

Three-dimensional visualization tools in a virtual environment are essential to the design and development of innovative projects. In possession of these tools, the designers can inform and present the concepts clearly and objectively elucidating many doubts of stakeholders.

The three-dimensional salt cave model was implemented on the Vesseljs library (Gaspar, 2018), which is an open-source JavaScript library to perform the visualization of complex marine engineering systems. Fonseca and Gaspar (2019) present several advantages in using this kind of collaborative platform, as well as offers some examples and functionalities that are already implemented. One of the benefits of using Vesseljs consists of the fact it is a web-based library compatible with the most common web browser, making

unnecessary the necessity for further software installation and increasing its scalability.

As Vesseljs is a platform with continuous development, some new features were implemented to model the salt cave system environment: for instance, the classes for representing the lines and the series of stereolithography (STL) 3D models of the ships and equipment. An auxiliary snippet code was also generated to calculate the position and geometry of the elements throughout the time. In this first implementation, the model has the sole purpose of visualizing the solutions generated by the systems engineering model.

A screenshot of the model is presented in Figure 5, in which it is possible to visualize the equipment involved in the operation, such as submerged pumps, wellhead, and Christmas tree. The arrangement of the well pipes and the dissolution process generating the cave shape was modeled as predicted in Costa et al. (2019b). It is also possible to observe the floating units as well as the umbilical cables and catenary risers.



Figure 5: Screenshot of the Salt Cave Visualization Module Implemented in the Vesseljs Platform. Available in <http://vesseljs.org>

In the control panel is possible to set the flux dissolution rate and time parameters as shown in Table 1, the flow rate of carbon dioxide to the cave, the control button to start, pause or restart the simulation as well as observe the elapsed time. Figure 6 presents a sequence of three different stages of simulation: (a) the beginning of dissolution when the well is drilled, the subsea equipment is installed, and the submerged pump is ready to inject water to dissolve the cave; (b) represents the dissolved cave stage when the desired dimensions were obtained, and the cave is fulfilled with brine; (c) represents the process of substituting the brine by the FPSO gas when the equipment used to the dissolution process was removed, and the gas is filling the cave.

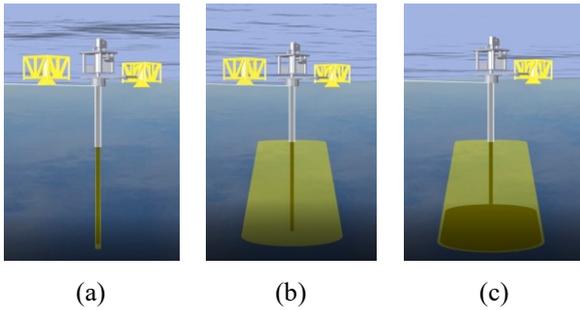


Figure 6: Sequence of Different Stages of Cave Operation: (a) Beginning of Dissolution, (b) Dissolution completed, (c) CO₂ Storage

Although the web-visualization tool itself is not a new implementation, it is the first time it has been applied to the study of complex systems involving concomitantly floating systems, subsea systems, well engineering, and geomechanical systems. The tool application potential can be extended beyond simple visualization, considering a platform for centralization and integration of the areas involved.

From the model parameterization, it is possible to connect to the various databases of hydrodynamic modeling, geotechnics, flow assurance, and obtain a broad model of cave construction and operation.

CONCLUSIONS AND FUTURE WORK

A system engineering model was presented in order to provide insights into decisions making in the early stages of the conceptual design of an offshore salt cave system for CO₂ storage.

The multi-attribute tradespace exploration (MATE) methodology was employed to obtain the utility and costs estimates for a set of possible solutions. Although simplified, the model was able to raise some interesting points and trends about the design of the salt cave system.

The obtained pareto boundary describes two distinct classes of solutions: small caves with a low cost and large caves with a higher cost.

For a more in-depth analysis, future work is intended to complement the model considering the temporal, contextual, and perceptual aspects as proposed by Rhodes and Ross (2010).

Regarding the visualization model, we corroborate in this paper with the call open and collaborative visualization methods made by Gaspar (2018), in his consideration for developing future engineering analysis and simulation in JavaScript (JS). The examples here presented are working in the process, and much of the library structure and methods intend to be improved as the research develops. The main point defended in this example is that technology is not a bottleneck for collaborative design and simulation of offshore systems, neither the speed of the computer processors and memory size, but rather how efficient maritime simulation and design data is able to

be transferred from books and experience to useful reusable models.

Further work could involve not only the demonstration of its components but the mathematical simulation response of the equipment, such as internal pressure in the epoch and pipes. The visualization could have its applicability expanded by incorporating the costs and utility calculations produced in this article.

Furthermore, web-based platforms already had a good response for visualizing data collected in a scale model, as mentioned in Fonseca and Gaspar (2020). In future work, the model could be used as a digital twin of the real operation for monitoring the equipment and for validation of the methodology used.

ACKNOWLEDGMENTS

The authors would like to thank the INTPART “International Partnerships in Excellent Education, Research and Innovation” cooperation program of the Norwegian government that enabled the exchange of knowledge between USP and NTNU. Kazuo Nishimoto gratefully acknowledges the support of the RCGI – Research Centre for Gas Innovation, hosted by the University of São Paulo (USP) and sponsored by FAPESP – São Paulo Research Foundation (2014/50279-4) and Shell Brasil, and the strategic importance of the support given by ANP (Brazil’s National Oil, Natural Gas, and Biofuels Agency) through the R&D levy regulation. Special thanks to Dr. Donna H. Rhodes from SEARI/MIT (Systems Engineering Advancement Research Initiative – Massachusetts Institute of Technology) for discussions and technical support.

REFERENCES

- ANP. 2019. “Oil, natural gas and biofuels statistical yearbook 2019”. URL: <http://www.anp.gov.br/component/content/article/2-uncategorised/5300-oil-natural-gas-and-biofuels-statistical-yearbook-2019>.
- Bai, Y., and Bai, Q. 2018. “Subsea Engineering Handbook.” Gulf Professional Publishing.
- Costa, P.V., Costa, A.M., Szklo, A., Branco, D.C., Freitas, M., and Rosa, L. P. 2017. “UGS in giant offshore salt caverns to substitute the actual Brazilian NG storage in LNG vessels.” *Journal of Natural Gas Science and Engineering*, 46, 451–476.
- Costa, A.M., Costa, P.V., Udebhulu, O.D., Azevedo, R.C., Ebecken, N.F.F., Miranda, A.C.O., de Eston, S.M., de Tomi, G., Meneghini, J.R., Nishimoto, K., Ruggeri, F., Malta, E.B., Fernandes, M.E.R., Brandão, C., and Breda, A. 2019a. “Potential of storing gas with high CO₂ content in salt caverns built in ultra-deep water in Brazil.” *Greenhouse Gases: Science and Technology*, 9 (1), 79–94.
- Costa, A.M, Costa, P.V.M., Miranda, A.C.O., Goulart, M.B.R., Udebhulu, O.D., Ebecken, N.F.F., Azevedo, R.C., de Eston, S.M., de Tomi, G.,

- Mendes, A.B., Meneghini, J.R., Nishimoto, K., Sampaio, C.M., Brandão, C., and Breda, A. 2019b. "Experimental salt cavern in offshore ultra-deep water and well design evaluation for CO₂ abatement." *International Journal of Mining Science and Technology*, 29(5), 641–656.
- Fonseca, Í.A. and Gaspar, H.M. 2020. "A practical approach to digital twin modelling and development." In *Communications of the ECMS*, 34 (1).
- Fonseca, Í.A. and Gaspar, H.M. 2019. "A Prime on Web-Based Simulation." In *Communications of the ECMS*, 33 (1).
- Gaspar, H.M., Rhodes, D.H., Ross, A.M. and Erikstad, S.O. 2012a. Addressing complexity aspects in conceptual ship design - A systems engineering approach. *Journal of Ship Production and Design*, SNAME, 28(4).
- Gaspar, H.M., Rhodes, D.H., Ross, A.M. and Erikstad, S.O. 2012b. Handling temporal complexity in the design of non-transport ships using Epoch Era Analysis. *International Journal of Maritime Engineering*, 154(A3).
- Gaspar, H.M., 2018. "Vessel.js: An Open and Collaborative Ship Design Object-Oriented Library." In *Proceedings of the 13th International Marine Design Conference* (Helsinki, Finland, Jun 10-14).
- Keeney, R.L., and Raiffa, H. 1993. Decisions with multiple objectives: preferences and value trade-offs. Cambridge University Press.
- Londe, L., 2017. "Underground storage of hydrocarbons: Advantages, lessons learnt, and way forward." In *Abu Dhabi International Petroleum Exhibition & Conference*. Society of Petroleum Engineers.
- McCall, M., Davis, J.F., and Kregel, M. 2004. "Offshore salt caverns enable a 'mega' sized LNG receiving terminal." In *Offshore Technology Conference*.
- McCall, M.M., Davis, J.F., and Taylor, C. 2005. "Offshore Salt-Cavern-Based LNG Receiving Terminal." In *International Petroleum Technology Conference*.
- Rhodes, D.H. and Ross, A.M. 2010. "Five aspects of engineering complex systems emerging constructs and methods," In *2010 IEEE International Systems Conference*, 190-195. IEEE.
- Ross, A.M., and Rhodes, D.H. 2008. "Using attribute classes to uncover latent value during conceptual systems design." In *2008 2nd Annual IEEE Systems Conference*, 1–8. IEEE.
- Shi, X., Yang, C., Li, Y., Li, J., Ma, H., Wang, T., Guo, Y., Chen, T., Chen, J., Liu, W., Zhang, N. 2017. "Development prospect of salt cavern gas storage and new research progress of salt cavern leaching in China." In *51st US Rock Mechanics/Geomechanics Symposium*. American Rock Mechanics Association.

AUTHOR BIOGRAPHIES

DANIEL P. VIEIRA is a post-doc researcher at the University of São Paulo, Brazil. Ph.D. and Engineering degree in Naval Architecture and Ocean Engineering at the same university and has experience in the design of ships and offshore structures, experimental and numerical simulation in hydrodynamics. Work as a consultant in several R&D projects in Brazil.

KAZUO NISHIMOTO is a Full Professor at the Department of Naval Architecture and Ocean Engineering of the University of São Paulo, coordinator of Numerical Offshore Tank laboratory specialized in numerical and experimental simulation of ocean systems dynamics, development of new ocean systems, and particle method for continuous media dynamics.

FELIPE FERRARI DE OLIVEIRA is an MSc candidate in Naval Architecture at NTNU, has an engineering bachelor's degree in Naval Architecture and Marine Engineering at the University of São Paulo. Currently employed as a researcher assistant at NTNU for the development of web-based simulations for digital twin ship operation. Industrial experience as ship designer in the shipyard "Bertolini Construção Naval do Amazonas" (BECONAL).

HENRIQUE M. GASPAR is an Associate Professor at the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU). The professorship is connected to the Ship Design Chair at the Maritime Knowledge Hub, sponsored by Ulstein Group. Education consists of a Ph.D. degree in Marine Engineering at the NTNU, with research collaboration at UCL (UK) and MIT (USA). Previous professional experience as Senior Consultant at Det Norske Veritas (Norway) and in Oil & Gas in Brazil.

A MODEL FOR FORECASTING MENTAL FATIGUE IN MARITIME OPERATIONS

Thiago G. Monteiro
Henrique M. Gaspar
Houxian Zhang

Department of Ocean Operations and Civil Engineering
Norwegian University of Science and Technology
6009, Ålesund, Norway
email: thiago.g.monteiro@ntnu.no

Charlotte Skourup
Products and Services R&D
Oil, Gas and Chemicals
ABB As

0666, Oslo, Norway
email: charlotte.skourup@no.abb.com

KEYWORDS

Maritime operations; Mental fatigue prediction and simulation

ABSTRACT

Assessing seafarers' mental fatigue levels helps identifying potential operational risks and the ability to simulate future scenarios can be used during planning and management, to ensure safer operational conditions. In this work, we propose a framework for modelling seafarers' future mental fatigue levels using a combination of both physiological and environmental sensors and model- and data-based techniques. We established building blocks of this framework and presented examples of how it can be applied in different scenarios as soon as enough data is collected to feed the data-based section of the model. Once properly trained, this framework can be used not only to assess human-related operational risks but also to provide the necessary information to ensure that these issues are addressed before potential danger escalates to real accidents.

MENTAL FATIGUE IN MARITIME OPERATION

Safety-critical operations is an increasing concern across all industries dealing with human-machine interaction and systems. Human-related issues are the main cause of accidents in fields such as driving (Williamson et al., 2011), commercial air transport (Suraweera et al., 2013), and maritime operations (Chauvin et al., 2013). Among the most common issues, we can highlight situational awareness and human errors. The main contributing factors leading to these challenges are excessive workload, stress and fatigue, specially mental fatigue (MF).

The maritime industry presents especial fatigue-related challenges connected to the intrinsic nature of maritime operations. This includes long and irregular working hours, long periods away from home, unpredictable environmental factors, and no clear separation between work and leisure. The International Maritime Organization (IMO) defines in its Guidelines on Fatigue (IMO,

2019) several factors influencing fatigue in seafarers. These factors can be categorized as seafarer-specific factors, management factors, ship-specific factors, environmental factors, and operational factors.

Among the seafarer-specific factors we can highlight psychological and physiological characteristics and personal habits. Ship-specific factors cover all aspects related to the design and condition of the ship, such as ship motion and responses, level of automation and reliability, and physical comfort in accommodation and work spaces. Environmental factors include aspects such as noise, vibration, ship motion, ventilation and temperature.

With so many contributing factors, monitoring and controlling MF levels in maritime operations is a complex task. Among the available option to assess MF, several subjective approaches are presented in the literature. Self assessment is the most common subjective approach and can rely on the use of questionnaires, such as Chalder Fatigue Scale (Chalder et al., 1993) and Epworth Sleepiness Scale (Johns, 1991), and sleeping diaries (Wadsworth et al., 2006). Although useful for tracking the user's MF profile, these methods are generally not suitable for real-time applications.

For more a reliable MF assessment, deterministic approaches are recommended. In this case, the use of physiological sensors, such as eye trackers, electrocardiogram (ECG), and electroencephalogram (EEG), is recognized as the best way to reliably assess MF in real-time (Sahayadhas et al., 2012). This is due to the intrinsic relation between changes in physiological signals and variations in MF levels.

After establishing the required understanding about the causes and how to measure MF, a natural follow up question is how can we model the MF development. Answering this question is important since a good MF development model can be used for an early intervention or operational planning. Addressing this task is a challenging issue, due to the effect of external factors in the development of MF and how difficult it is to objectively relate these factors to their effects. In the work, we describe a framework for modelling MF de-

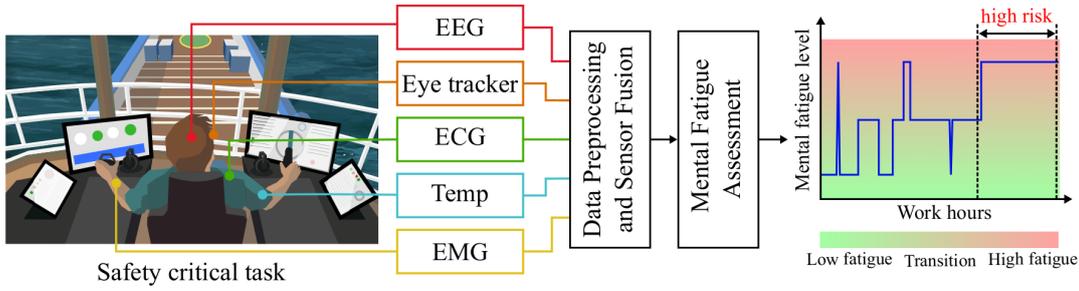


Fig. 1. Framework for Mental Fatigue Assessment

velopment, taking into account not only the seafarer’s physiological condition but also the effects of external factors.

MEASURING AND MODELLING MENTAL FATIGUE SCENARIOS

Being able to monitor the development of seafarers’ MF state in real-time can be beneficial for the safety of demanding maritime operations. A proper MF profile of all operators taking part in a complex operation enables a more precise risk assessment, which in turn can help preventing causalities.

Let’s consider an operational scenario where a pilot is maneuvering a platform supply vessel (PSV) close to an oil rig for cargo unloading. Using a set of physiological sensors, one can monitor this pilot and perform an MF level assessment. This MF profile can be used to perform the operational risk assessment based on how long the operator stayed in a critical MF level. What a critical MF level is and how long the operator needs to be in this state to indicate operational risks need to be defined via experiments. This MF assessment framework is presented in Fig. 1 (Based on (Monteiro et al., 2020)).

Although MF assessment can be a useful tool for reducing the risk of causalities, it presents a limitation regarding how early we can intervene in the operation. This limitation is due to the fact that the assessment system measures only the current state of the operator. Once a dangerous condition is assessed, measures need to be taken in order to mitigate the operational risk. This delay between assessment and action can be sufficient a time window for the assessed risk to turn into a real accident. Thus being able to anticipate the risky period is important to prepare the necessary mitigation measures on time.

Supposing we can assess the operator’s MF state in the current operational scenario (S_0), how can this scenario develop in the next, for example, two hours? If, after the assessment of scenario S_0 , the weather conditions develop to a much rougher sea state, how will the operator’s MF level be affected? If a collision between the PSV and the oil rig happens, how will this stressful situation come into play in the operator’s performance? These different external factors can make the prediction of future scenarios very hard, since the original scenario can lead to several possible future scenarios. This branching is presented in Fig. 2.

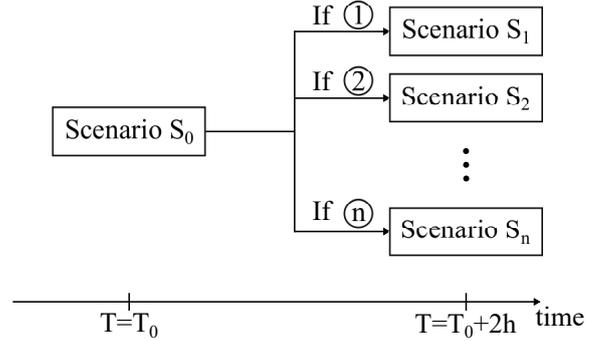


Fig. 2. Branching of Scenarios

The current work aims to extend our previous studies (Monteiro et al., 2019, 2020) by adding a prediction capability to our MF assessment framework. So, the question we want to answer in this paper is: How can we model and simulate other MF scenarios from the current assessed scenario? Can this theoretical model be calibrated and validated with real data gathered from day-to-day operations? In order to be able to answer these questions, we need to address four important points. First, which factors affect the progression of the MF state? Then, how to measure and quantify these factors? Later, how to integrate these measurements and the current MF assessment into an MF prediction algorithm? Finally, how can real data can be used to calibrate these models?

Contributing Factors

As presented in the previous section, there are several factors that affect the development of MF in seafarers. For our analysis we will group these contributing factors in time and distress. Time refers to the natural progression of MF due to physiological and psychological factors, and the prolonged exposition to environmental factors. Distress is related to unexpected and emergency situations. These events can be punctual or have long duration. They elevate MF levels by increasing workload, tension, and stress levels.

Sensors

In this work we propose the use of two different classes of sensors: physiological and environmental sensors. Physiological sensors are the most reliable way to assess MF levels, since the physiological symptoms of MF can be captured as they start to develop. The most

usual physiological sensors used to monitor MF include eye tracker, electrocardiogram (ECG), electroencephalogram (EEG), and body temperature sensors. Ideally, we would like to have as much sensor information available as possible to help in the decision making process. Practically, the use of several sensors attached to a seafarer’s body can hinder the proper execution of complex tasks. So, this trade-off between the amount of data and sensors needs to be taken into consideration when selecting which physiological sensors to use for real-life applications.

Environmental sensors are used to quantify factors that are external to the seafarers. They include gyroscopes, accelerometers, weather sensors, cameras, sound level meters, etc. In opposition to physiological sensors, there is no limit for the number of environmental sensors to apply when monitoring maritime operations. Additionally, most vessels already record data for several of the sensors cited above, so there is little extra setup to be done regarding the environmental sensor. The challenge is how to correlate this kind of sensor data and variations in a seafarer’s MF level.

Mental Fatigue Prediction

There are two main steps when trying to forecast MF scenarios. First, we need to assess the seafarer’s current MF level. Then, we need to predict the expected MF level based on the seafarer’s current MF level and time and distress effects. Fig. 3 presents our proposed framework for forecasting MF scenarios, including the assessment and prediction steps.

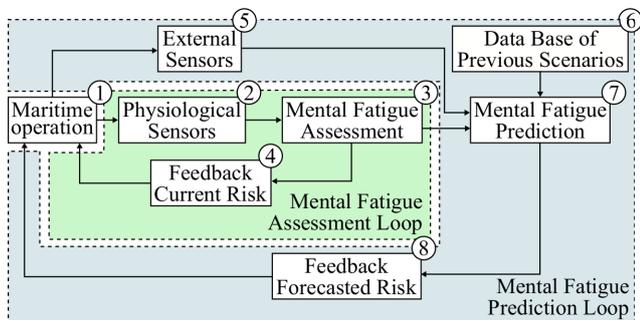


Fig. 3. Framework for Mental Fatigue Prediction

The MF assessment loop is responsible for determining the seafarers’ current MF state at a certain operation (1). The assessment is performed using only data from physiological sensors (2). This data is acquired in real time as time-series. After any required preprocessing, the data can be used as the input to an MF assessment algorithm (3) responsible to determine the current MF level. Here, two different approaches are viable. One possible approach is to use neural networks to classify the input data in different levels in an MF scale. Another alternative is to apply a model-based approach, which defines MF levels in a deterministic way by modeling the MF representation using the sensors data and expert knowledge. With the profiled MF progression, a risk assessment algorithm can be used to inform the seafarer about the current operational risk level (4).

The assessed MF state in a time-stamp t_0 is correlated to all the environmental sensors (5) data collected at that time stamp. The time-series of the MF profile and the time-series of the environmental sensors are stored in a database of previous scenarios (6), which can be used for training the MF prediction algorithm (7). In this case, it is considered a data-driven algorithm. Besides being stored in the database, the MF profile and the environmental sensors data are the inputs for the MF prediction algorithm. Using the combination of recent environmental sensors data and the current MF level of an operator, the trained algorithm can produce a prediction about the expected MF level of said operator in the near future, and a risk assessment algorithm can be used to inform the seafarer about the projected operational risk level (8). The main difference between the algorithms implemented in (3) and (7) is that in (3) we only rely on instantaneous physiological sensor data, while in (7) we consider the instantaneous physiological and external sensors data, while relying on a database of previous scenarios.

Database of Previous Scenarios

For this application, we are handling complex data. The complexity of the data is defined by two main factors: several disparate data sources and data size. Firstly, the environmental sensors can provide data in different domains, which can be hard to fuse in a meaningful way. For example, a camera provides a video feed, while an accelerometer provides accelerations in different directions. In order to efficiently store and handle all external sensor data, having all data in the same domain can be very helpful. Most sensors data are generated as time-series data, which is basically a stream of time-stamp/value pairs. In this case, having time as a common domain is the simplest solution to facilitate the data fusion process. So, in order to ensure that all sensors speak the same language, some preprocessing may be needed for some sensors to extract time-domain features that can be stored in the data base and fused with other sensors data.

With the huge amount of time-series data produced by the environmental sensors, the computational complexity to handle this information is high. In this case, a specialized time-series database can go a long way improving the system’s overall efficiency. In a time-series database, new incoming data is stored in a sequential manner, usually ordered by time-stamp. In this case, new data is inserted in the database instead of old values being updated. This allows for tracking how the data changes with time, making it possible to understand tendencies in the past and predict trends in the future. Traditional databases can be employed to handle time-series data, but usually they lack the tools to handle two important aspects of time-series data: scale and usability.

Regarding the scale factor, the amount of data that needs to be stored when handling time-series can grow very fast. This is specially true when handling several sensors operating at high frequencies. In order to effi-

ciently handle this huge amount of data, the database needs to be optimized to provide bigger ingest rate, faster querying and optimized operations for data compression. This optimization is only possible when the time variable is considered a first priority during the database framework design.

Usually, only storing the time-series data is not enough. Regarding the usability factor, it is important that we are able to perform operations that are characteristic of this kind of data. Some examples of these operations include data retention policies, continuous queries, and flexible time aggregations.

Data-driven Mental Fatigue Prediction

Once enough data is stored in the database for both the MF assessment and external sensors, fully data-driven methods can be effectively applied to perform the MF prediction. The minimum amount of data necessary to perform a good prediction is a relative matter, since it depends on both the complexity of the prediction and its expected accuracy. One natural candidate for this task would be a long-short term memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997). This neural network is capable of learning long-term dependencies in time-series data by using a memory cell to regulate the information flow (Fig. 4). The information flow is controlled by non-linear gating units that include input gates (i_t), output gates (o_t) and forget gates (f_t). A complete formulation of the LSTM algorithm and examples of applications in time-series prediction can be easily found in the literature, for example (Ellefsen et al., 2019). By training the LSTM algorithm in a data stream composed by all the physiological and environmental sensors, it can learn to predict the MF state by applying the penalty functions automatically, without the need to manually tune the penalty functions parameters. The appropriate network structure, including number of layers and neurons, needs to be determined during the training process, according to the desired precision criteria for the prediction task.

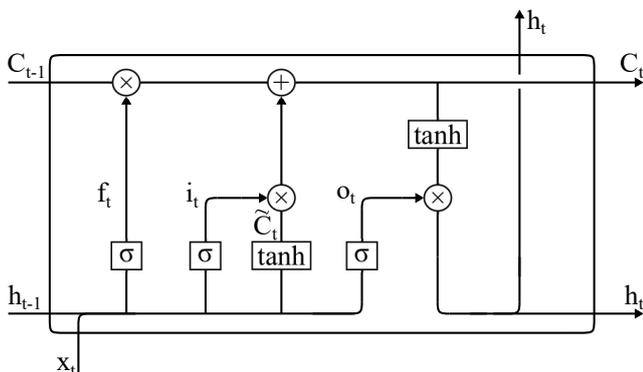


Fig. 4. LSTM Memory Cell

MODELING MENTAL FATIGUE PROGRESSION

As presented in Fig. 3, the prediction of future MF states depends on a database of physiological and en-

vironmental sensors data. In order to establish a foundation for the handling of this data, we can discuss how the integration between physiological and environmental factors can be applied for forecasting MF levels in seafarers. Since we do not have an established database, we need to first model the effects of time and distress over the MF level. This theoretical modeling can be then used to exemplify our proposed approach.

Time Dependant Mental Fatigue Progression

The simplest effect to model is the effect of time. Due to physiological and environmental factors, MF accumulates as time goes by. There is an initial rested state that progresses to a maximum level of MF where staying awake would be almost impossible. Although this situation can sound a little extreme, it is not unlikely to happen, specially during night shifts. Eq. 1 can be used to model this time dependent MF progression

$$MF(t) = \frac{\alpha}{2} + \alpha \frac{\arctan\left(\beta\left(t - \frac{T}{\gamma}\right)\right)}{\pi} \quad (1)$$

where T is the total duration of the prediction, γ indicates the position of the inflection point of the curve, β dictates the inflection angle, π is the non-dimensionalization constant for the arctan function and α scales the function to our desired MF scale.

Fig. 5 presents the proposed model for the time aspect of the MF progression, with the theoretical limits for restedness and tiredness.

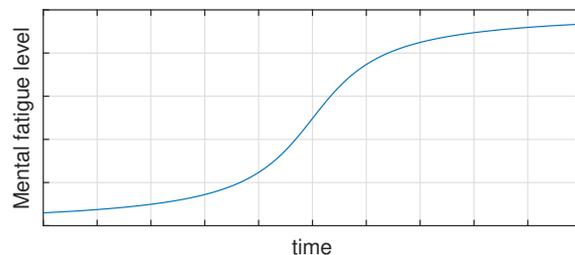


Fig. 5. Time-dependant Mental Fatigue Progression Model

Fig. 5 shows the time-dependant MF progression for $\alpha = 1$, $\beta = 0.2$, and $\gamma = 2$.

Distress Dependant Mental Fatigue Progression

The effects of distress over the time dependent MF progression will be modeled as penalties to the natural MF development. There are several possible causes of distress (IMO, 2019), and for simplicity we will group them into continuous and punctual effects.

Continuous effects represent disturbances that take place continuously, for an extended period of time. Such effects include higher than normal noise levels, excessive vessel motion due to weather conditions, long watch shifts at night, etc. The effects of distress do not push the MF level over the theoretical MF limit. Instead, they accelerate the MF development. For modeling continuous effects, we propose the use of the fol-

lowing Gaussian-like function:

$$P_{const}(t) = a_1 \exp\left(\frac{-(t - \frac{T}{b_1})^2}{2c_1^2}\right) \quad (2)$$

where a_1 is the scale factor for the penalty, T/b_1 dictates the position of center of the normal distribution, and c_1 represents its spread. The penalty is applied in the time dependant MF progression as follows:

$$MF'(t) = MF(t) \cdot (1 + P_{const}(t)) \quad (3)$$

Fig. 6 shows the continuous effect penalty function for $a_1 = 0.15$, $b_1 = 2.5$, and $c_1 = 0.1$.

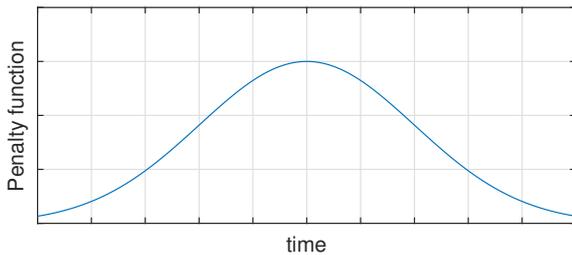


Fig. 6. Penalty Function Model for Continuous Distress

Punctual effects represent disturbances that have short duration but can effect the MF development in the long term. This kind of effect includes, for example, a vessel collision with other vessel or marine structure, accidents with cargo handling, and man overboard scenarios. Punctual effects are more relevant when they have an acute impact on the seafarer's situational awareness and sense of danger, triggering a burst of adrenaline. This phenomenon can, in the short term, increase attention and, in the long term, increase MF progression due to the increase in tension and workload levels.

In order to model punctual effects, we propose the use of a combination of Gaussian-like functions. The negative portion of the equation models the increase in attention after a serious, unexpected event occurs, while the positive portion models the long term increase in the MF progression levels.

$$P_{short}(t) = -a_2 \exp\left(\frac{-(t - \frac{T}{b_2})^2}{2c_2^2}\right) + a_3 \exp\left(\frac{-(t - \frac{T}{b_3})^2}{2c_3^2}\right) \quad (4)$$

where the variables a , b and c are analogous to the ones presented for Eq. 2. The way the punctual penalty is applied to the time dependant MF progression is the same described in Eq 3. Fig. 7 shows the punctual effect penalty function for $a_2 = 0.75$, $b_2 = 2.8$, $c_2 = 0.35$, $a_3 = 0.25$, $b_3 = 2.3$, and $c_3 = 0.15$.

Tuning Models

Previously we described the modeling of the MF development process based on the time-dependant MF progression and penalty functions. So one may ask: how could I tune the parameters that compose the penalty functions for different scenarios? The answer for this

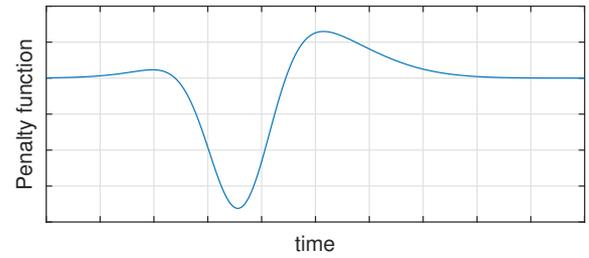


Fig. 7. Penalty Function Model for Punctual Distress

question lays on the time-series complex data stored in the previous scenarios database.

Initially, when no previous scenarios data are available, the prediction capabilities of the proposed framework are poor. Normal operation data is the easiest kind of data to come around. It is produced by the assessment framework when no distress factor is in play. It is essential for implementing a time-dependant MF state forecasting strategy. Using this prediction model as a baseline, the penalty functions for a specific operational conditions (including one or more distress factors) can be approximated using the inverse of the mapping function presented in Eq. 3. This new equation can be written as:

$$P_{const}(t) = \frac{MF'(t)}{MF(t)} - 1 \quad (5)$$

The obtained penalty function can then be approximated to the formulation of either continuous or punctual penalties by defining the appropriate parameters. Defined this way, the parameters can be stored and recovered when a similar distress factor takes place during the MF monitoring of an operator. Once calibrated with real data, the proposed framework can be used as management and planning tools. These applications are exemplified in the next section.

SIMULATING MENTAL FATIGUE SCENARIOS

Management Tool

As a management tool, this framework can be used for assessing, in real-time, the operational risk related to seafarers' MF condition. At any given time during an operation, the MF level assessed from an operator until that time and the data from the environmental sensors can be used to forecast how the MF level is expected to change in the near future. The MF level can fall under good (green), attention (yellow) or dangerous (red) ranges. Fig. 8 shows the composition of the predicted MF state for a seafarer. The assessed MF levels (1) is used to extrapolate the time dependant progression of the MF state (2). Environmental sensor data is used to calculate punctual (3) and continuous (4) penalty functions. Applying all this data in Eq. 3, the system outputs the predicted MF state (5).

With this information, a manager can keep track of the operational risks related to the seafarer's MF state. In a scenario where this risk surpasses some predefined criteria, the manager can act by alerting the seafarer

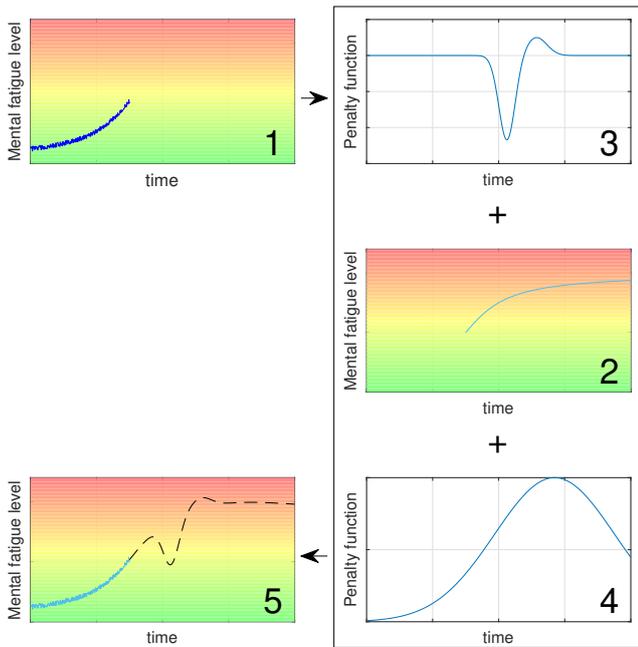


Fig. 8. Mental Fatigue Prediction Composition

about the dangerous condition or even plan a break or a change of operator.

Planning Tool

As a planning tool, this framework can be used for evaluating different possible scenarios during a demanding operation. Based on these possible scenarios, worst case conditions for the seafarers can be identified and the risks involved in the operation can be assessed. Consider for example a crane operation taking place offshore by a construction vessel. The crane operator presents his own time dependant MF progression (S_0), disregarding any external complicating factors. But what if during the operation the weather conditions worsen (S_1), or there is an accident damaging one important component that should be installed (S_2), or an unforeseen delay takes place (S_3) or there is a malfunction in one equipment (S_4). How can these different scenarios effect workload and stress and impact the MF development of the said operator? The utilization of the proposed framework to investigate all these possible scenarios is presented in Fig. 9.

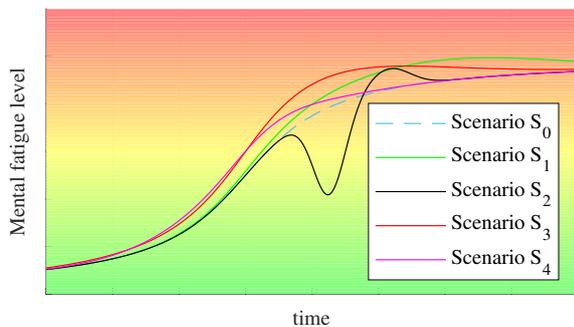


Fig. 9. Comparing Different Possible Scenarios

With the prediction model for the different scenarios at

hand, the operation can be planned to account for the risks related to high MF levels. The operation can be planned for the worst case scenario in order to ensure a higher safety factor. But it also can be planned for less risky scenarios if the probability for the worst case condition is low. Another viable option involves planning for a scenario represented by a weighted average of the possible modeled scenarios, where the weighting factor is the probability of occurrence of each scenario.

TOWARDS AN INTELLIGENT MODEL FOR MENTAL FATIGUE AND HUMAN ERROR PREDICTION

The work here presented was developed as a simple and fundamental approach to put in practice the framework from Fig. 3. Originally, the framework was developed as a way to transform real data from operations into an assessment for MF. The reality is that this real-operations data is not available, and much has been done only from experiments in simulators. The equations here presented can be used as a comparison and benchmark, elegantly in terms of coefficients of well know equations.

We are aware that the presented model is an *educated guess* on how MF seems to behave based on the literature and current research at NTNU. We attempted here to navigate between the two worlds of the literature found in this topic. One extreme, very human and social based, which describes MF qualitatively and is intrinsically connected to the psychological aspects of a human-being performing a safety-critical task in demanding operations at the sea. No wonder IMO uses this side of the spectrum to describe the types and limits of MF, given that no number is able to properly and safely estimate MF in maritime operations. In this way, we understand the problem, but no data is given for decision-making.

On the other extreme, data-driven methods are wide spreading in all fields, promising that a well trained AI will be able to estimate everything with more precision than just a narrow sample of human experience, providing accurate decision-making. The database previously described requires a large amount of complex-data. This intensive data-driven approach demands hundreds of hours of real-data operation that need to be collected, filtered, stored and fed to an advanced AI algorithm. We also observed in recent experiments in simulators (Monteiro et al., 2019; Kari et al., 2019) that physiological factors can add extra complexity to the operations, and demand time and patience to be set up and used. Remote sensor technology, such as Open CV, may be the solution, but they have yet a long path before being commercially applicable in such a narrow niche as maritime operations.

We do believe that having a functional model, able to theoretically predict MF based on time and distress, is a fundamental piece to achieve the framework proposed. In this context, we call for an open and collaborative approach to calibrate, improve and validate our model. The source code is available at an online repository

(<https://github.com/thiagogabrielm/MFM>).

We plan to continue the experiments at NTNU, and currently we are gathering data from our research vessel. This is, however, a small sample yet for properly training AI. In this sense, having a sound mathematical model as the one here described seems a good initial step for other actors to use and adapt our model. If we are right, we will be able to present a library of tangible coefficients, calibrated for each operation. In the future, connecting these coefficients to each MF aspect from IMO, can be the basis for a sound regulation on MF and, therefore, safer operations.

REFERENCES

- Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., & Wallace, E. 1993. "Development of a fatigue scale." *Journal of psychosomatic research*, 37(2), 147–153.
- Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.-P., & Langard, B. 2013. "Human and organisational factors in maritime accidents: Analysis of collisions at sea using the hfacs." *Accident Analysis & Prevention*, 59, 26–37.
- Ellefsen, A. L., Bjørlykhaug, E., Æsøy, V., & Zhang, H. 2019. "An unsupervised reconstruction-based fault detection algorithm for maritime components." *IEEE Access*, 7, 16101–16109.
- Hochreiter, S., & Schmidhuber, J. 1997. "Long short-term memory." *Neural computation*, 9(8), 1735–1780.
- IMO. 2019. *Guidelines on fatigue*. International Maritime Organization London, UK.
- Johns, M. W. 1991. "A new method for measuring daytime sleepiness: the epworth sleepiness scale." *sleep*, 14(6), 540–545.
- Kari, R., Steinert, M., & Gaspar, H. M. 2019. "Eeg application for human-centered experiments in remote ship operations." In *Centric 2019, the twelfth international conference on advances in human oriented and personalized mechanisms, technologies, and services*.
- Monteiro, T. G., Skourup, C., & Zhang, H. 2020. "Optimizing cnn hyperparameters for mental fatigue assessment in demanding maritime operations." *IEEE Access*.
- Monteiro, T. G., Zhang, H., Skourup, C., & Tannuri, E. A. 2019. "Detecting mental fatigue in vessel pilots using deep learning and physiological sensors." In *2019 IEEE 15th international conference on control and automation (icca)* (pp. 1511–1516).
- Sahayadhas, A., Sundaraj, K., & Murugappan, M. 2012. "Detecting driver drowsiness based on sensors: a review." *Sensors*, 12(12), 16937–16953.
- Suraweera, P., Webb, G. I., Evans, I., & Wallace, M. 2013. "Learning crew scheduling constraints from historical schedules." *Transportation research part C: emerging technologies*, 26, 214–232.
- Wadsworth, E. J., Allen, P. H., Wellens, B. T., McNamara, R. L., & Smith, A. P. 2006. "Patterns of fatigue among seafarers during a tour of duty." *American Journal of Industrial Medicine*, 49(10), 836–844.
- Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. 2011. "The link between fatigue and safety." *Accident Analysis & Prevention*, 43(2), 498–515.

THIAGO G. MONTEIRO received the B.Sc. degree in naval architecture and maritime engineering from the University of São Paulo, Brazil, in 2013 and the M.Sc degree in Ship Design from the Norwegian University of Science and Technology (NTNU), Norway, in 2016. He is currently undergoing his Ph.D. research at NTNU in the field of physiological sensors fusion in the maritime domain. His research interests include sensor fusion, machine learning, mental fatigue assessment and human factors in maritime operations.

HENRIQUE M. GASPAR is an Associate Professor at the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU). The professorship is connected to the Ship Design Chair at the Maritime Knowledge Hub, sponsored by Ulstein Group. He has a PhD degree in Marine Engineering at the NTNU, with research collaboration at UCL (UK) and MIT (USA). Previous professional experience as Senior Consultant at Det Norske Veritas (Norway) and in Oil & Gas in Brazil.

HOUXIANG ZHANG received Ph.D. degree in Mechanical and Electronic Engineering in 2003. From 2004, he worked as Postdoctoral Fellow at the Institute of Technical Aspects of Multimodal Systems (TAMS), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Germany. In Feb. 2011, he finished the Habilitation on Informatics at University of Hamburg. Dr. Zhang joined the NTNU (before 2016, Aalesund University College), Norway in April 2011 where he is a Professor on Robotics and Cybernetics. The focus of his research lies on biological robots and modular robotics and on virtual prototyping and maritime mechatronics. In these areas, he has published over 130 journal and conference papers and book chapters as author or co-author.

CHARLOTTE SKOURUP received the M.Sc. degree in Mathematics from the Norwegian University of Science and Technology (NTNU), Norway, in 1994 and the Ph.D. degree in human-machine interaction NTNU, Norway, in 1999. She worked as an associated professor for the Department of engineering cybernetics at NTNU, Norway, from 2004 to 2015. She has been working at ABB AS Oil, Gas and Chemicals, Oslo, Norway, since 2007 and is currently the section manager for products and services R&D.

INTEGRA: AN OPEN TOOL TO SUPPORT GRAPH-BASED CHANGE PATTERN ANALYSES IN SIMULATED FOOTBALL MATCHES

Nicolò Oreste Pinciroli Vago^{1,2}, Yuri Lavinias³, Daniele Rodrigues⁴, Felipe Moura⁵, Sergio Cunha⁶, Claus Aranha³, Ricardo da Silva Torres¹

¹NTNU – Norwegian University of Science and Technology, Ålesund, Norway

²Dipartimento di Elettronica, Informazione e Bioingegneria – Politecnico di Milano, Milan, Italy

³University of Tsukuba, Tsukuba, Japan

⁴College of Computer Engineering, Pontifical Catholic University of Campinas, Brazil

⁵Laboratory of Applied Biomechanics, State University of Londrina, Londrina, Brazil

⁶College of Physical Education, University of Campinas, Campinas, Brazil

KEYWORDS

Simulated football matches, multi-agent systems, temporal graphs, graph visual rhythm.

ABSTRACT

This paper introduces Interactive Graph Analyzer (INTEGRA), a tool to support the comparison of simulated football matches with real ones, through the analysis of dynamic graphs. Our tool supports coordinated views of temporal graphs, benefiting from traditional node-link diagrams and graph visual rhythms, a recently proposed 2D image representation. Our proposal is generic and may be tailored to different applications. We demonstrate the use of this tool in compelling case studies related to the comparison of graph-based measurements obtained from three different kinds of simulated football matches and real ones. In particular, we exploit usage scenarios related to how graph measurements evolve over time.

INTRODUCTION

Multi-Agent (MA) system technologies have been applied successfully in several fields, such as games [Marín-Lora et al., 2020], robotics [Sharma et al., 2016], and medical research [Pathirana et al., 2019]. In the context of sport games, such technologies have been exploited to create realistic events and scenes aiming at improving users' experience. In particular, we highlight applications in players' interaction in football matches [Kitano et al., 1997], [Kurach et al., 2019], [Asada and von Stryk, 2020]

This paper targets the problem of creating more realistic MAs for football matches. The envisioned MA simulation is expected to allow the visualization, validation, and exploration of football player models, leading to a greater understanding of the relationship between the models and the real-world data, and extrapolations of many different scenarios

using the rules derived from the models. The challenging problem is how to aggregate and represent different members of a team and their interactions [Machado et al., 2017] with the purpose of presenting to coaches and researchers a valuable tool to identify key-events and determinant moments of the matches, such as attacking sequences, shots to goal and tackles [Moura et al., 2012]. The resulting simulator is also expected to be used to help coaches and educators. The simulator will assist them with planning and decision making, by giving these professionals the tools to simulate fictional scenarios in the simulation and observe how these scenarios play out. The simulator will also allow them to obtain easily understandable outputs from the model in the form of videos of games between simulated agents, leading to an interactive process of trial, error, and discovery. Qualitatively and quantitatively assessing how different simulated football matches are from real-world ones is, therefore, of paramount importance.

This paper focuses on the comparison of simulated and real-world football matches based on tactical indicators defined in terms of position of players and their interaction (e.g., passes or proximity) on the pitch over time, a subject well studied in the Sport Science community [Duch et al., 2010], [Pena and Touchette, 2012], [Cho et al., 2018], [Mendes et al., 2018], [Oliveira and Clemente, 2018], [Buldú et al., 2018]. A widely used representation to support the identification and analysis of pattern changes associated with objects and their relationships (e.g., interaction of players in a football match) over time relies on the use of *temporal graphs* [Leskovec et al., 2005] and their proper visualization through visual representations [Beck et al., 2016].

In this paper, we propose Interactive Graph Analyzer (INTEGRA), a web-based visualization tool that supports coordinated views of dynamic graphs. This is a generic tool, which can be easily tailored to different applications. INTEGRA provides cus-

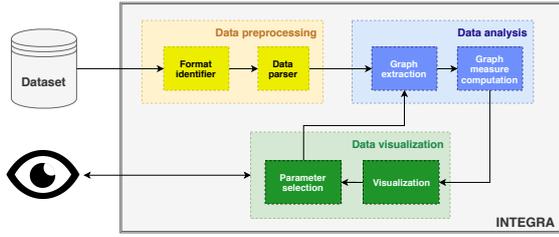


Fig. 1: Schematic architectural overview of INTEGRA.

tomizable interaction mechanisms and multiple visual representations based on Graph Visual Rhythms (GVRs) [Rodrigues et al., 2019] – a bidimensional representation which encodes graph features (e.g., vertex properties encoded through graph measurements) as columns of an image – and node-link diagrams.

Our approach focuses on allowing the user to customize the user interface, to interact with graphs, and to set custom parameters, on the basis of which analysis and visualization are performed. Using multiple views also expands the possibilities of identifying patterns of interest by coordinating different and complementary insights provided by different representations, obtained from both simulated and real-world matches. This paper describes the functionalities and architecture of the tool and illustrates its use in compelling usage scenarios, related to the comparison of data associated with three different kinds of simulated matches and real ones, based on graph measurements.

INTEGRA

INTEGRA architecture (shown in Figure 1) is divided into three modules: data pre-processing, data analysis, and data visualization.

Data pre-processing (in yellow) consists in the identification of the dataset format and in the subsequent parsing. In case of football games, the input data are converted into a matrix, that represents the players’ coordinates over time. Data analysis (in blue), in turn, is in charge of extracting temporal graphs from the parsed dataset and of computing graph measurements. Later, extracted measurements will be used to compose visual structures in the data visualization module. A central feature of the tool consists in giving the user the possibility to write the code to plot custom graphs, using JavaScript. Data visualization (in green) allows the user to visualize the computed data and to customize the analysis and visualization parameters. The user may also interact with the graphs, for instance by selecting a specific range of values, by exporting the graphs as images and by playing animations. INTEGRA also allows creating several workspaces, that

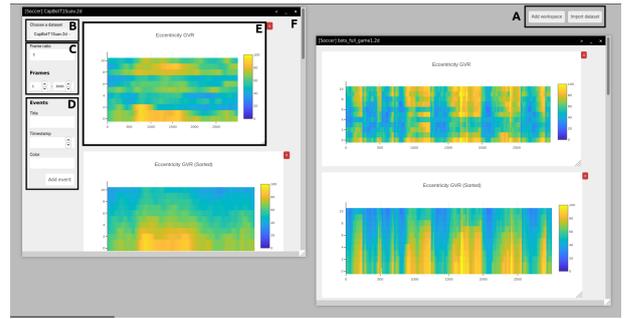


Fig. 2: Screenshot of the INTEGRA’s interface.

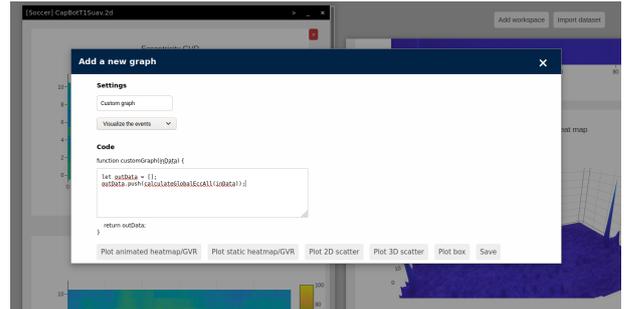


Fig. 3: Screenshot of the INTEGRA’s interface for inserting temporal graphs manually.

are represented as draggable and resizable window-like containers, so that users may compare the graphs of two or more datasets in the same browser window.

GUI Design

Figure 2 shows a screenshot of INTEGRA’s interface. It shows that the GUI is divided into different workspaces. This workspace-based approach allows more sets of data to be studied at the same time, maintaining independent parameters for each workspace.

Components

Figures 2 and 3 show that the interface of the proposed tool is divided into the following sections:

- **Section A** contains the buttons that allow the user to import datasets in supported formats and to add new workspaces.
- **Section B** contains the menu that allows selecting one of the imported datasets for visualization. The modular architecture of INTEGRA allows the implementation of custom modules to parse different datasets, so that data can be properly handled using JavaScript. Currently, parsing is limited to the datasets for which specific parsers are included in the tool.
- **Section C** contains the menus to choose the frame ratio and the frame range. The frame ratio is the ratio between the current framerate and the original data framerate, while the frame range refers to the initial and final frames displayed in the visualization section.

- **Section D** is the *Events* section and it allows to add and delete custom events to the visualization. An event is defined as something that happens in correspondence of a specific timestamp. The events are shown in different ways on the basis of the graph type.
- **Section E** allows the customization of the visualization solutions and shows detailed information about the quantities represented in the graphs, for instance by zooming on a particular area or volume, by performing rotations and in some cases by starting animations.
- **Section F** shows the different visualization alternatives. The user may interact with the proposed graphs using the features available in Section E.
- The modal window *Add a new graph* (Figure 3) allows the user to create a customized graph on the basis of the imported data. In particular, it is possible to define and display the variables as colors, animations or axis variables. In order to create custom graphs, the user has to define the content of a JavaScript function.

Interaction

Initially, the user may create a new workspace or import one or several new files. Once the datasets have been imported, they can be selected from the left menu of a workspace. The result of the analysis will be shown on the graphs displayed next to the left menu.

It is also possible to create a custom graph by clicking on the + button in the graph area, once a specific dataset has been selected.

Implementation Aspects

The tool has been developed using HTML, CSS, and JavaScript in order to guarantee portability. Data analysis is performed using the local computational power, so that it is not necessary to rely on an Internet connection or on remote server availability.

The interface for a new workspace is dynamically generated using JavaScript. Firstly, a new workspace is generated on the basis of a default empty workspace. Then `EventListeners` are added, in order to allow the user to handle time and events.

In order to propose a modular structure, each workspace is described, at a given moment, by a set of status variables, that represent the values of the different parameters.

The proposed approach allows handling different kinds of datasets, since data access is independent of data elaboration and presentation. The tool supports several data formats and it is possible to implement and integrate other parsers in the code.

The tool is openly available at <https://github.com/nicolopinci/INTEGRA> (As of March 2020).

CASE STUDIES

In this section, we present two case studies about the use of INTEGRA for the analysis of simulated matches. In both cases, we use the proposed tool to perform comparative analyses among dataset associated with simulated matches (Google Research Football Environment) and a dataset related to a real football match.

Dataset Details

Simulated Match Dataset

The simulated match data were obtained from the Google Research Football environment, a recently published open-source football simulator [Kurach et al., 2019]. This simulator was originally proposed for the development of Artificial Intelligence Neural Networks. It reproduces a full football match with all of its usual regulations and events, as well as player tiredness, misses, etc. Figure 4 shows a standard image of the running simulator.

The entire simulated match lasts 5 real world minutes, and samples 10 frames from the environment per real world second, for a total of 3000 frames per game. Each frame has information about the position of each player, the position and possession of the ball, and player fatigue. This information is stored as a log file after the match. It is important to note that this simulator does not simulate data during interruptions in the game (fouls, offside, goals, and other referee-related interruptions). In these cases, the players are automatically placed on their positions when the game is re-started, and no information is recorded about the “off-play” period.

The simulator provides a standard, fixed strategy for controlling the agents (players), with three levels of strength (easy, medium, and hard). Additionally, the simulator allows a special controller (usually an AI controller) to take control of the player which is closest to the ball. For this case study, we produced games where all the players are controlled by the standard strategy (Bots Full), games where actions are selected at randomly disregarding any data from the simulator (Random Full) and games where the special controller is controlled by an Artificial Neural Network using the Proximal Policy Optimization algorithm (NN Full) [Schulman et al., 2017].

Figure 5 shows a visual output of the data acquired. The ball is shown as a red *B*, players from the same team are shown with the same letter, *A* and *H*. *X* is the player we control (in this case, a player of team *H*), while *A* is the team we play against.

Real Football Dataset

The considered dataset contains the position of the players for every timestamp. A total of 82,850 timestamps, related to 45 minutes (plus additional time) using a video framerate of 30 Hz, has been considered in the chosen dataset. Positional data were obtained using the DVideo software [Figuerola et al.,



Fig. 4: Screenshot of the football simulator used in this case study.

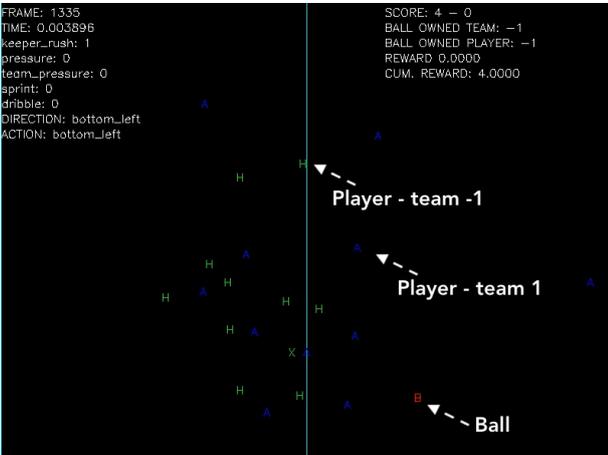


Fig. 5: Visual data from the simulator used by the Artificial Neural Network controller. The ball is shown as a red, the controlled player is shown as the green, team-mates are shown in green. Opponents are shown in blue.

2006b], [Figueroa et al., 2006a] applied to official football matches. In particular, the players’ positions are computed for every timestamp by using a computer-vision-based tracking algorithm.

Tool Usage Overview

Figure 6 represents multiple views in a football match analysis scenario. Figure 6a represents the parameter selection. In particular, this section allows choosing a dataset, the frame ratio and the frame range. In the example, the user has selected a frame ratio of 1 (thus, the framerate is the same as the recording) and chose to visualize temporal graph information related to the period from frame 1 to frame 3000. Figure 6b presents the result of the parameter selection and some of the computed graphs. For each graph, a user can perform additional operations, such as selecting a specific area of interest. An example is shown in Figure 6c, where the GVR associated with

the users’ eccentricity scores is shown.

Graph Visual Rhythm

A Graph Visual Rhythm is a 2D representation that visually encodes temporal graph changes. This is a compact representation intended to highlight patterns of interest related to graph measurements in time.

Let $\mathcal{G} = \langle G_1, G_2, \dots, G_T \rangle$ be a time evolving graph, where $G_t = (V_t, E_t)$ is an instant graph at timestamp $t \in [1, T]$ composed of a set of vertices, V_t , and a set of edges, E_t . A graph visual rhythm image GVR is defined as [Rodrigues et al., 2019]:

$$GVR(t, z) = \mathcal{F}(G_t), \quad (1)$$

where $\mathcal{F}_{G_t} : \mathcal{G} \rightarrow \mathbb{R}^n$ is a function that represents the instant graph $G_t \in \mathcal{G}$ as a point in an n -dimensional space, $t \in [1, T]$ and $z \in [1, n]$. \mathcal{F}_{G_t} can be defined as a graph measurement for each vertex of any other function that encodes relationships among vertices (e.g., degree histogram).

Eccentricity

Our case studies rely on the analysis of different graph patterns over time. In particular, we investigate the use of the eccentricity scores of players. In football matches, it measures the accessibility degree of a level from the other vertices, for a given time frame. Eccentricity, therefore, describes the spread degree of the players on the football field and how central a player is if compared to the other players of the same team.

Given a graph $V(G)$, the eccentricity is a vertex measurement and corresponds to the maximum shortest distance from a vertex i to all others in the graph, as defined by [West, 2000].

$$\epsilon_i = \max(d_{ij}) \quad (2)$$

where d_{ij} is the shortest distance between vertex i and j , where $j \in V(G)$.

On The Comparison Of Different Simulated Matches

Suppose that a user wants to compare the results of three different matches, which have been simulated using the Bots Full algorithm. Figure 7 shows the comparison between the eccentricity values of the three simulated matches based on the same Bots Full strategy, using GVR representations. For each GVR, we compute the eccentricity of players and sort their scores for each timestamp (GVR column). The GVR related to the dataset `bots_full_game1.2d` (top-left GVR) presents a large yellow area, which suggests that the players in that match are more connected to each other when compared to the players in the other two analyzed matches. The reason for this is that the football simulator, as real games, is a stochastic environment [Yue et al., 2008], causing the same action

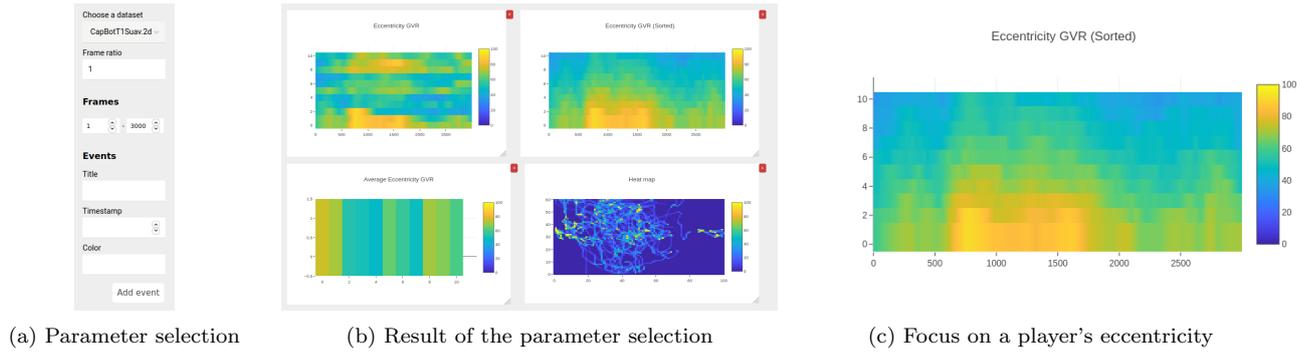


Fig. 6: Multiple views in football analysis scenario.

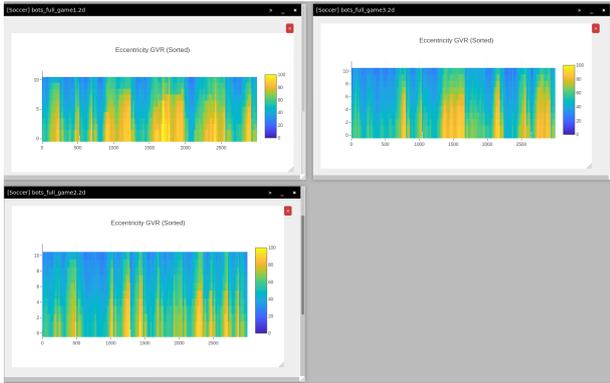


Fig. 7: Comparison between eccentricity GVRs of three Bots Full simulated matches.

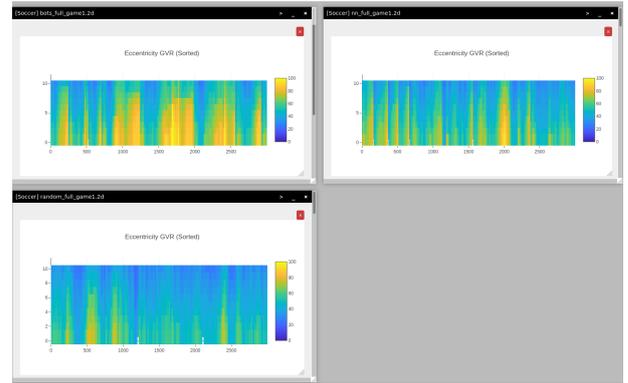


Fig. 8: Comparison between three eccentricity GVRs of three different simulated matches kinds: Bots Full (top-left), NN Full (top-right), and Random Full (bottom-left).

to have different outcomes at different times. Therefore, we can find different behaviors in matches that use the same strategy.

Suppose now that the user is interested in comparing different kinds of simulated matches. Figure 8 presents the comparison between the eccentricity values for a Bots Full simulated match (Bots Full, top-left GVR), a Neural Network simulated match (NN Full, top-right GVR) and a random agent (Random Full, bottom GVR). From those figures, we can observe that eccentricity scores for Bots Full are higher (more yellow areas) when compared to the others. The lack of yellow regions for the Random Full match may indicate that players are more disconnected in this kind of simulation. We may also observe a repetition pattern (yellow columns appear in a periodical frequency) for the NN Full match.

It comes with no surprise that eccentricity scores for Bots Full are high since the main goal of the bots Full player is to provide a realistic game play with reasonable football actions and strategies. On the other hand, the reason why the players in the match of the Random Full show low eccentricity scores might be because the actions of the controlled player are randomly selected, without any information about the environment, leading to a more sparse distribution of the players during the match. Finally, we comment about the repetition pattern of the NN

Full match. This strategy was built with the only focus of scoring as many goals as possible, and the clear repetition pattern suggests that it found a combination of actions that, if repeated, allows the NN Full to score many goals.

On The Comparison Of Simulated And A Real football Match

Figure 9 presents the comparison between the real (bottom-right chart) and three simulated games using Bots Full (the other three plots) in terms of the eccentricity box plots. The main difference consists in the distribution of the eccentricity values: while the players' eccentricity presents several outliers (represented as points outside of the whiskers), the simulated games eccentricity values are almost entirely contained inside the whiskers' limits. The reason why this phenomenon is observed could be that the bots base their position on a common strategy, behaving like a swarm, while the players in real matches do not know or are not completely aware of the intentions of the other players in some of the defensive and offensive actions. In addition, it is important to emphasize that the real match dataset contains information about player positions even when the game is interrupted. Thus, it is possible that at

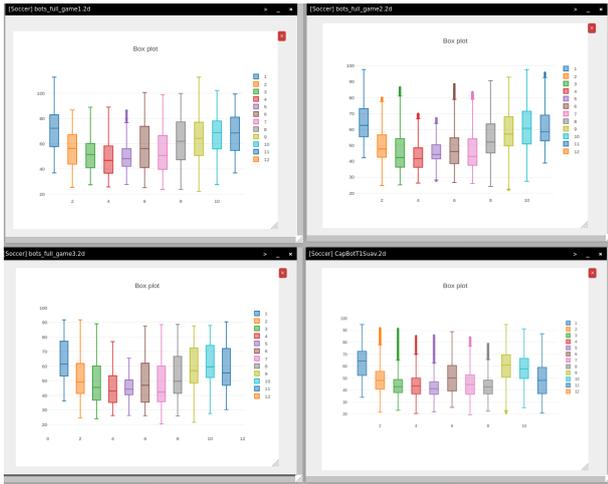


Fig. 9: Comparison between the box plots of completely simulated matches (top-left, top-right, and bottom-left) and of a real match (bottom-right). The real is associated with more outliers in the eccentricity values.

these moments, the values of eccentricity present discrepancies from the rest of the game. It is also possible to notice that the simulated game associated with the top-right plot is the most similar to the real one, with players from the same positions (i.e. goalkeepers and attackers) presenting similar behavior. Thus, this outcome suggests a more realistic simulation.

CONCLUSIONS

This paper introduced INTEGRA, a tool to support the analysis of simulated and real football match data based on visual analytics of temporal graphs. Its main novelty relies on permitting multiple views associated with the same dynamic graph, ranging from node-link diagram to graph visual rhythms. The tool is generic and can be tailored for different applications. It supports the implementation of various graph measures, as well as their visualization in 2D representations using graph visual rhythms.

The use of INTEGRA was illustrated in two scenarios related to the comparisons of graph-based dynamic evolution involving data of different simulated and real football matches. The tool supports the identification of similar and different patterns observed in football dynamics, defined in terms of player positions throughout the match and the behaviour of the players at prominent events in a match. This should allow coaches to have insights about successful and unsuccessful tactical strategies, possibly incorporating lessons learned in future decision-making processes.

Future work includes the evaluation of the tool with possible users. We also plan to investigate its use in different comparison scenarios, involving other tactical indicators proposed in the literature [Moura et al., 2012], [Moura et al., 2016].

ACKNOWLEDGEMENTS

Authors are grateful to CNPq and São Paulo Research Foundation – FAPESP (grants #2014/12236-1, #2015/24494-8, #2018/15178-3, #2016/50250-1, # 2018/19007-9, and #2017/20945-0). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

- [Asada and von Stryk, 2020] Asada, M. and von Stryk, O. (2020). Scientific and technological challenges in robocup. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):null.
- [Beck et al., 2016] Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2016). A taxonomy and survey of dynamic graph visualization. *Computer Graphics Forum*.
- [Buldú et al., 2018] Buldú, J. M., Busquets, J., Martínez, J. H., Herrera-Diestra, J. L., Echevoyen, I., Galeano, J., and Luque, J. (2018). Using network science to analyse football passing networks: dynamics, space, time and the multilayer nature of the game. *Frontiers in psychology*, 9:1900.
- [Cho et al., 2018] Cho, Y., Yoon, J., and Lee, S. (2018). Using social network analysis and gradient boosting to develop a soccer win–lose prediction model. *Engineering Applications of Artificial Intelligence*, 72:228 – 240.
- [Duch et al., 2010] Duch, J., Waitzman, J. S., and Amaral, L. A. N. (2010). Quantifying the performance of individual players in a team activity. *PLoS one*, 5(6):e10937.
- [Figuerola et al., 2006a] Figuerola, P. J., Leite, N. J., and Barros, R. M. (2006a). Background recovering in outdoor image sequences: An example of soccer players segmentation. *Image and Vision Computing*, 24(4):363 – 374.
- [Figuerola et al., 2006b] Figuerola, P. J., Leite, N. J., and Barros, R. M. (2006b). Tracking soccer players aiming their kinematical motion analysis. *Computer Vision and Image Understanding*, 101(2):122 – 135.
- [Kitano et al., 1997] Kitano, H., Tambe, M., Stone, P., Veloso, M., Coradeschi, S., Osawa, E., Matsubara, H., Noda, I., and Asada, M. (1997). The robocup synthetic agent challenge 97. In *Robot Soccer World Cup*, pages 62–73. Springer.
- [Kurach et al., 2019] Kurach, K., Raichuk, A., Stańczyk, et al. (2019). Google research football: A novel reinforcement learning environment. *arXiv preprint arXiv:1907.11180*.
- [Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge discovery in data mining*, pages 177–187. ACM.
- [Machado et al., 2017] Machado, V., Leite, R., Moura, F., Cunha, S., Sadlo, F., and Comba, J. L. (2017). Visual soccer match analysis using spatiotemporal positions of players. *Computers & Graphics*, 68:84 – 95.
- [Marín-Lora et al., 2020] Marín-Lora, C., Chover, M., Sotoca, J. M., and García, L. A. (2020). A game engine to make games as multi-agent systems. *Advances in Engineering Software*, 140:102732.
- [Mendes et al., 2018] Mendes, B., Clemente, F. M., and Maurício, N. (2018). Variance in prominence levels and in patterns of passing sequences in elite and youth soccer players: a network approach. *Journal of human kinetics*, 61(1):141–153.
- [Moura et al., 2012] Moura, F. A., Martins, L. E. B., Anido, R. D. O., Barros, R. M. L. D., and Cunha, S. A. (2012). Quantitative analysis of brazilian football players’ organisation on the pitch. *Sports Biomechanics*, 11(1):85–96. PMID: 22518947.
- [Moura et al., 2016] Moura, F. A., van Emmerik, R. E. A., Santana, J. E., Martins, L. E. B., de Barros, R. M. L., and Cunha, S. A. (2016). Coordination analysis of players’ distribution in football using cross-correlation and vector coding techniques. *Journal of Sports Sciences*, 34(24):2224–2232. PMID: 27079483.

- [Oliveira and Clemente, 2018] Oliveira, P. and Clemente, F. M. (2018). Network properties and performance variables and their relationships with distance covered during elite soccer games. *Journal of Physical Education and Sport*, 18:1045–1049.
- [Pathirana et al., 2019] Pathirana, S., Asirvatham, D., and Johar, M. G. M. (2019). Applicability of multi-agent systems for electroencephalographic data classification. *Procedia Computer Science*, 152:36 – 43. International Conference on Pervasive Computing Advances and Applications-PerCAA 2019.
- [Pena and Touchette, 2012] Pena, J. L. and Touchette, H. (2012). A network theory analysis of football strategies. *arXiv preprint arXiv:1206.6904*.
- [Rodrigues et al., 2019] Rodrigues, D. C. U. M., Moura, F. A., Cunha, S. A., and da Silva Torres, R. (2019). Graph visual rhythms in temporal network analyses. *Graphical Models*, 103.
- [Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- [Sharma et al., 2016] Sharma, K. R., Honc, D., Dusek, F., and Kumar, T. G. (2016). Frontier based multi robot area exploration using prioritized routing. In Claus, T., Herrmann, F., Manitz, M., and Rose, O., editors, *30th European Conference on Modelling and Simulation, ECMS 2016, Regensburg, Germany, May 31 - June 3, 2016, Proceedings*, pages 25–30. European Council for Modeling and Simulation.
- [West, 2000] West, D. B. (2000). *Introduction to Graph Theory*. Prentice Hall, 2 edition.
- [Yue et al., 2008] Yue, Z., Broich, H., Seifriz, F., and Mester, J. (2008). Mathematical analysis of a soccer game. part i: Individual and collective behaviors. *Studies in Applied Mathematics*, 121(3):223–243.

NICOLÒ ORESTE PINCIROLI VAGO is a MSc Simulation and Visualization student at NTNU and Computer Science and Engineering student at Politecnico di Milano, where he obtained a BSc in the same field. His main research interests are genetic algorithms and Artificial Neural Networks applied to computer vision.

YURI LAVINAS is a PhD student at the University of Tsukuba. He graduated in Computer Science (2016) in the University of Brasilia, Brazil, and got his Master Degree at the University of Tsukuba (2020). His research focus on Multi Objective Evolutionary Computation, neuro-evolution and Artificial Life.

DANIELE RODRIGUES obtained her PhD in Computer Science from the University of Campinas (Unicamp) in 2017, Masters degree in Software Engineering from University of Campinas in 2004, and Bachelor Degree in Computer Engineering from Pontifical Catholic University of Campinas in 2001. She is currently a professor at College of Computer Engineering at Pontifical Catholic University of Campinas (PUC-Campinas). Her research focus on complex networks, machine learning, data science and sports science.

FELIPE MOURA is Professor at the Sports Sciences Department at State University of Londrina

(Brazil). He graduated in Physical Education (2003) and got his Masters degree (2006) at São Paulo State University. He got his PhD in Physical Education at the University of Campinas (2011). Nowadays, he is the leader of the Laboratory of Applied Biomechanics in which the main research interests are biomechanics applied to sport and exercise, and signal processing of biological data.

SERGIO CUNHA has a degree in Physical Education, a master's degree in and a Ph'd in Sports Sciences from the State University of Campinas. He was a Visiting Professor at the "University of Calgary" for eight months, working on a research project together with Prof. Walter Herzog. He was also at the "University of Groningen" (Netherlands) as a Visiting Professor during 2019. He is currently an Associate Professor III at the School of Physical Education at Unicamp. He has experience in Physical Education and Sports, with an emphasis on Biomechanical Analysis Methods, working mainly on the following topics: biomechanics, football, futsal and mathematical models for sports.

CLAUS ARANHA obtained his PhD from the University of Tokyo in 2010, and is currently an assistant professor at the University of Tsukuba, and a member of the Center for Artificial Intelligence (CAIR). His research focus on the study and application of Evolutionary Computation, including optimization, procedural generation, and artificial life simulations.

RICARDO DA SILVA TORRES is Professor in Visual Computing at the Norwegian University of Science and Technology (NTNU). 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 Senior Associate Editor of the IEEE Signal Processing Letters and Associate Editor of the Pattern Recognition Letters.

ENABLING PYTHON DRIVEN CO-SIMULATION MODELS WITH PYTHONFMU

Hatledal, Lars Ivar*

Zhang, Houxiang

Department of Ocean Operations and Civil Engineering
Norwegian University of Science and Technology
Postbox 1517, 6025 Aalesund, Norway

Collonval, Frédéric

Modeling & Simulation
Safran Tech

CS80112 Chateaufort
78772 Magny Les Hameaux, France

KEYWORDS

Co-simulation; Modelling; FMI; FMU; Python

ABSTRACT

This paper introduces PythonFMU, an easy to use framework for exporting Python 3.x code as co-simulation compatible models compliant with version 2.0 of the Functional Mock-up Interface (FMI). The framework consists of a set of helper classes and a command line utility for transforming compliant python source into ready to use cross-platform FMUs. PythonFMU seamlessly takes care of a number of low-level FMI functions such as getting and setting variable values, and state handling, including serialization and deserialization. Furthermore it provides pre-built binaries for Windows and Linux 64-bits, generates the required *modelDescription.xml* containing meta-data about the model and packages all related files into a Functional Mock-up Unit (FMU) - ready to be imported into any FMI compatible simulation tool. The framework can be effortlessly installed using de-facto standard Python package managers pip and conda. While PythonFMU is more geared towards ease of use and enabling Python driven co-simulation models, it is shown to have adequate performance compared to much more low-level alternatives targeting other programming languages.

INTRODUCTION

The Functional Mock-up Interface (FMI) [Blochwitz et al., 2012] is a tool independent standard managed by the Modelica Association that supports both Model Exchange (ME) and Co-Simulation (CS) of dynamic models. A key goal of FMI is to improve the exchange of simulation models between suppliers and original equipment manufacturers (OEMs). The current major version of the standard is 2.0, which was released in 2014. A minor revision, 2.0.1, was released in 2019.

An FMU is a model that implements the FMI standard and is distributed as a zip-file with the extension *.fmu*. This archive contains:

- An XML-file that contains meta-data about the model, named *modelDescription.xml*.

*Corresponding author. E-mail: laht@ntnu.no

- C-code implementing a set of functions defined by the FMI standard.
- Other optional resources required by the model implementation.

The FMI standard consists of two main parts, both of which a single FMU may support:

- FMI for ME: Models are exported without solvers and are described by differential, algebraic, and discrete equations with time-, state-, and step-events.
- FMI for CS: Models are exported with a solver, and data is exchanged between subsystems at discrete communication points. In the time between two communication points, the subsystems are solved independently from each other.

The work presented in this paper, however, is only concerned about the co-simulation part of the standard.

Many tools support importing co-simulation FMUs, however, fewer tools supports exporting such FMUs. Many of whom are commercial and or domain specific. Furthermore, FMUs generated with these tools may not support the optional parts of the standard such as state handling, which are required by some more advanced co-simulation algorithms in order to achieve better numerical accuracy and stability during simulations [Broman et al., 2013, Cremona et al., 2016, Tavella et al., 2016].

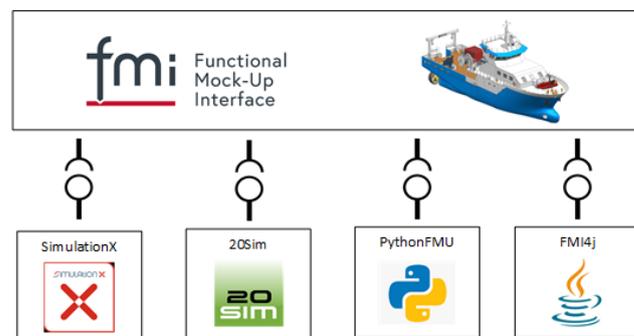


Fig. 1: Possible use of PythonFMU in realizing complex cyber-physical systems using FMI based co-simulation.

Python [Van Rossum et al., 2007] is one of the most popular programming language today [O'Grady, 2020]. The major reasons for that are the ease of learning the language, the huge spectra of libraries covering fields

such as video game, machine learning, web server or scientific computing and the recent explosion of data science in which Python plays a central role. Of particular importance for scientific computing is the creation of the *Numpy*[Oliphant, 2007] library that bridge the gap between efficient code in C or Fortran languages and the ease of a scripting language. That library is now at the heart of all major scientific Python libraries from Pandas for data analysis to Scipy for classical algebra operators or Scikit-learn for machine learning analysis.

This paper introduces PythonFMU, an easy to use Python based framework that allows plain Python code to be exported as FMI compatible co-simulation FMUs. Figure. 1 shows how PythonFMU could potentially be used to implement complex cyber-physical systems that are aggregates of models from different simulation domains.

The paper is organized as follows. Firstly some related works are given. After which PythonFMU is introduced. Then some benchmark results are presented. Finally some concluding remarks and notes on future works are provided.

RELATED WORK

A number of open-source software frameworks for exporting FMUs from source-code have been developed in the recent years. While many more tools are capable of exporting FMUs like 20Sim, OpenModelica, MATLAB and SimulationX, this paper is more focused on frameworks that allows the generation of FMUs from plain source-code. Each of these are described in more detail below.

CPPFMU [SINTEF Ocean, 2017] is a set of interfaces and helper functions for writing FMI-compliant model/slave code in C++ using high-level features such as exceptions and automatic memory management, rather than having to implement the low-level C functions specified by FMI. However, while CPPFMU makes implementing and compiling the shared library required by an FMU, it does not handle generating the *modelDescription.xml* nor packaging of the FMU. CPPFMU was developed as part of the R&D project Virtual Prototyping of Maritime Systems and Operations (ViProMa) Hassani et al. [2016], and is currently maintained by SINTEF Ocean.

FMUSDK [QTronic, 2017] is a free, BSD licensed, software development kit (SDK) provided by QTronic to demonstrate basic use of FMUs for ME and for CS as defined by FMI version 1.0 and 2.0. The software is written in C++, but models are to be implemented in C. The first version of FMU SDK was released already in 2010, with the latest version, *2.0.6*, being released in 2018.

Like CPPFMU, FMUSDK does not auto-generate the *modelDescription.xml*. The main difference between these tools is that CPPFMU provides a more high level and structured API in C++, whereas FMUSDK requires source-code to be written in quite low-level C. On the other hand, FMUSDK supports

ME and provides utilities for packaging the model as an FMU, whereas CPPFMU only provides helper functions to aid in development.

JavaFMI [Galtier et al., 2017] is a set of components for working with the FMI standard using Java, developed by SIANI institute (Las Palmas University) and funded by the European Institute for Energy Research (EIFER). It support both import and export of FMUs. For export, it support FMI 2.0 for Co-simulation. Generated FMUs runs both on Linux and Windows. JavaFMI has been actively maintained since its inception in 2013 and is licensed under the LGPLv3.

FMI4j [Hatledal et al., 2018] is a MIT licensed software package for dealing with Functional Mock-up Units (FMUs) on the JVM. It support both import and export of FMUs. For export, it support FMI 2.0 for Co-simulation. FMI4j is written in Kotlin, which is 100% interoperable with Java. On the native side, FMI4j makes use of CPPFMU to implement the FMI functions. FMUs generated using FMI4j can run on both Linux and Windows.

While both JavaFMI and FMI4j allows FMUs to be created using the Java language, they differ quite a bit in their implementation and usage. JavaFMI uses message-passing to bridge Java and the underlying C functions defined by FMI, while FMI4j relies on the Java Native Interface (JNI) for this. Consequentially, FMI4j generates much faster executables. Another key difference is how users define their model. JavaFMI is imperative, e.g meta-data is defined using API functions. FMI4j on the other hand is declarative, with meta-data defined using annotations.

Evidently, some open-source software for generating FMUs from source code already exists. See Table. I for a summary. However, only the ones targeting the JVM can be said to be easy to use as these manages everything related to the creation of an FMU. Still, the JVM may not be a natural choice for many for implementing models and the barrier for using these tools are high for non Java developers. CPPFMU and FMUSDK both eases the process within the realm of C/C++, but still requires significant know-how in order to produce a ready to use FMU. Furthermore, these tools only covers a small subset of available programming languages.

TABLE I: Open-source framework for exporting source-code as FMUs.

Tool	Target language	Target platform	FMI version
JavaFMI	JVM	Win, Linux	2.0
FMI4j	JVM	Win, Linux	2.0
CPPFMU	C++	Win ^a , Linux ^a	1.0 & 2.0
FMUSDK	C	Win ^a , Linux ^a , OSX ^a	1.0 & 2.0

^a Binaries are only built for the current platform.

PYTHONFMU

PythonFMU is a MIT licensed framework that enables the packaging of Python 3.x code as co-simulation FMUs, currently maintained in collaboration between NTNU and Safran Tech. The library required by users to implement their own FMI co-simulation slaves as well as the Command Line Interface (CLI) required to build the actual FMU is easily retrieved using either the *pip* or *conda* package managers. Unlike some FMU exporters, FMUs built with PythonFMU runs out of the box on both Windows and Linux 64-bit systems. PythonFMU has been implemented using the limited Python API, which makes it compatible with any Python 3.x version. However, PythonFMU does not bundle a Python distribution, which means that a compatible Python distribution must be present on the target system for the FMU to work. The same is true for any imported 3rd party libraries. Consequentially, if the slave makes use of e.g. the *numpy* package for scientific computing, this library must already be present on the target system. To remedy this, PythonFMU allows users to specify any dependency it should have on 3rd party libraries. This information is bundled with the FMU as a standard *requirements.txt* for use with one of Python’s package managers. Thus allowing end-users to easily figure out what kind of libraries that must be installed for it to run on a particular machine.

Listing 1: Writing FMI 2.0 compatible slaves in Python using PythonFMU.

```
from pythonfmu import *

class PythonSlave(Fmi2Slave):

    author = "John Doe"
    description = "A simple description"

    def __init__(self, **kwargs):
        super().__init__(**kwargs)

        self.realOut = 0.1
        self.register_variable(Real("realOut",
                                    causality=Fmi2Causality.output))

    def do_step(current_time, step_size):
        return True
```

Listing. 1 shows the minimal required code to write FMI 2.0 compatible co-simulation models in Python using PythonFMU. Additional FMI functions like e.g. *setupExperiment*, *enterInitializationMode*, *exitInitializationMode* and *terminate* have default no-op implementations and may be overridden on demand. PythonFMU automatically handles getting and setting variables, logging, resetting, state handling, serialization and deserialization as well as generating the required *modelDescription.xml*. The fact that PythonFMU handles state handling makes it possible to use with advanced co-simulation master algorithms that depends on rollback capabilities, like variable step algorithms. This is important in order to achieve numerically stable and accurate simulation results. List-

ing. 2 shows how to build an FMU from Python source that implements the PythonFMU API using the accompanying CLI. Additional options may be specified, such as documentation and associated project files. The FMU built by PythonFMU contains pre-built binaries for Windows and Linux 64-bit. This lowers the threshold for using it tremendously compared to many exporting tools as a C++ compiler does not have to be installed and the user does not have to figure out how to cross-compile.

Like FMI4j, PythonFMU makes use of CPPFMU for implementing the C functions required by the FMI standard. This shows a clear utility for CPPFMU as an enabler for higher-level applications to support the export of FMI compatible co-simulation models.

Listing 2: Building an FMU from Python source using the PythonFMU CLI.

```
pythonfmu-builder -f PythonSlave.py
```

RESULTS

In the following some performance metrics for PythonFMU is given.

Table. II show the performance of PythonFMU compared to other similar tools. The FMUs used all implements the same model. The model does no computation during stepping, but defines a single real, integer, boolean and string variable. These variables are read by the importing tool after each iteration. 100.000 iterations were run. That makes for a total of 400.000 calls through the FMI API. The benchmark was performed on a computer running Windows 10 fitted with an Intel i7-8700k processor.

TABLE II: Time required to step a simple FMU with one integer, real, string and boolean variable 100.000 times. All variables are read after each step.

Tool	Version	Time[s]
FMUSDK	2.0.6	4.6
CPPFMU	-	4.6
FMI4j	0.30.0	6.1
JavaFMI	2.6.0	40
PythonFMU	0.6.0	7.9/7.3 ^a

^a Using lambdas for getters, as demonstrated in Listing. 3.

From the results we can see that FMUSDK and CPPFMU are equally fast, and as expected, faster than both FMI4j and PythonFMU. This is natural as both of these uses CPPFMU internally and have an additional overhead from having to cross the native bridge using JNI and the Python C API respectively. JavaFMI is by far the slowest contender, due to it’s choice of using message passing over direct API calls through JNI. Note that PythonFMU provides two results. By supplying a lambda function to the optional *getter* and *setter* parameters of PythonFMUs *ScalarVariable* as

demonstrated in Listing. 3, users may increase performance of variable read/write. When not specifying lambda functions for the getter and setter, PythonFMU defaults to using the built in Python functions *getattr* and *setattr* respectively. Furthermore, the use of lambdas allows non Python fields to be used as variables.

Listing 3: Supplying a lambda for increased flexibility and performance at the cost of a slight increase in verbosity.

```
self.register_variable(Real("realOut",
    causality=Fmi2Causality.output
    getter=lambda: self.real))
```

Note that the results presented here does not necessarily translate to more complex models with more code evaluation, as the presented benchmark it is mainly interested in measuring the performance of raw FMI calls. As Python is an interpreted language it is naturally slower to run than e.g. C/C++.

CONCLUSIONS

This paper introduces PythonFMU, an easy to use framework for exporting Python code as FMI 2.0 for co-simulation compatible models. The framework is easy to install and requires very little boilerplate code, allowing users to focus on the problem at hand. This coupled with the fact that Python is an easy to use scripting language with a strong standard library and a rich set of 3rd party libraries makes it ideal for fast prototyping. Furthermore, the position Python has as a language for scientific computing should make PythonFMU a natural choice for data scientists that want to take advantage of or contribute to co-simulation technology. In fact, PythonFMU was specifically developed to allow data scientist with little or no background from co-simulation or software engineering at NTNU to contribute with models related to the development of digital twins, as Python and it's strong ecosystem of libraries allows easy integration with e.g. models that is connected to web services or utilizes neural networks. While the focus of PythonFMU is ease of use and being an enabler for Python driven co-simulation models, the performance is shown to be quite adequate compared to more low-level implementations.

Future works includes adding more features, bug-fixes and improving documentation. The source is available at <https://github.com/NTNU-IHB/PythonFMU>, and users are encourage to contribute.

ACKNOWLEDGMENT

The research presented in this paper is supported by the Norwegian Research Council, SFI Offshore Mechatronics, project number 237896.

REFERENCES

T. Blochwitz, M. Otter, J. Akesson, M. Arnold, C. Clauss, H. Elmqvist, M. Friedrich, A. Junghanns, J. Mauss, D. Neumerkel, et al. Functional mockup interface 2.0: The standard for tool independent exchange of simulation models. In *Proceedings of the 9th*

International MODELICA Conference; September 3-5; 2012; Munich; Germany, number 076, pages 173–184. Linköping University Electronic Press, 2012.

D. Broman, C. Brooks, L. Greenberg, E. A. Lee, M. Masin, S. Tripakis, and M. Wetter. Determinate composition of fmus for co-simulation. In *2013 Proceedings of the International Conference on Embedded Software (EMSOFT)*, pages 1–12. IEEE, 2013.

F. Cremona, M. Lohstroh, D. Broman, M. Di Natale, E. A. Lee, and S. Tripakis. Step revision in hybrid co-simulation with fmi. In *2016 ACM/IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE)*, pages 173–183. IEEE, 2016.

V. Galtier, M. Ianotto, M. Caujolle, R. Corniglion, J.-P. Tavella, J. É. Gómez, J. J. H. Cabrera, V. Reinbold, and E. Kremers. Building parallel fmus (or martyrshka co-simulations). In *Proceedings of the 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017*, number 132, pages 663–671. Linköping University Electronic Press, 2017.

V. Hassani, M. Rindarøy, L. T. Kyllingstad, J. B. Nielsen, S. S. Sadjina, S. Skjong, D. Fathi, T. Johnsen, V. Æsøy, and E. Pedersen. Virtual prototyping of maritime systems and operations. In *ASME 2016 35th International Conference on Ocean, Offshore and Arctic Engineering*. American Society of Mechanical Engineers Digital Collection, 2016.

L. I. Hatledal, H. Zhang, A. Styve, and G. Hovland. Fmi4j: A software package for working with functional mock-up units on the java virtual machine. In *The 59th Conference on Simulation and Modelling (SIMS 59)*. Linköping University Electronic Press, Linköpings universitet, 2018.

S. O’Grady. The redmonk programming language rankings: January 2020, 2020. URL <https://redmonk.com/sograd/2020/02/28/language-rankings-1-20/>. (Date accessed 08-March-2020).

T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

QTronic. Fmusdk, 2017. URL <http://www.qtronic.de/de/fmusdk.html>. (Date accessed 06-March-2020).

SINTEF Ocean. Cppfmu, 2017. URL <https://github.com/viproma/cppfmu>. (Date accessed 06-March-2020).

J.-P. Tavella, M. Caujolle, S. Vialle, C. Dad, C. Tan, G. Plessis, M. Schumann, A. Cuccuru, and S. Revol. Toward an accurate and fast hybrid multi-simulation with the fmi-cs standard. In *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–5. IEEE, 2016.

G. Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.

AUTHOR BIOGRAPHIES

LARS IVAR HATLEDAL received the B.Sc. degree in automation and the M.Sc. degree in simulation and visualization from the Norwegian University of Science

and Technology (NTNU), Ålesund, Norway, in 2013 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Department of Ocean Operations and Civil Engineering. Email: laht@ntnu.no

DR. FRÉDÉRIC COLLONVAL is the lead developer of a collaborative multi-systems and multi-physics simulation tool, of the Collaborative System Design team at Safran R&D center. He obtained his PhD in numerical simulation and modeling in gas turbine combustion chamber at TU Munich. He then joined Safran Group to work on airplane engine performance modeling and associated simulation tools. In 2018, he co-founded a new kind of collaborative simulation tool to target multi-physics simulation in pre-design phase for rapid design evaluation at Safran. He is interested in enhancing physical simulation tools by leveraging the features of innovative open-source projects.

PROF. HOUXIANG ZHANG received the Ph.D. degree in mechanical and electronic engineering, in 2003, and the Habilitation degree in informatics from the University of Hamburg, in February 2011. Since 2004, he has been with the Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Institute of Technical Aspects of Multimodal Systems (TAMS), University of Hamburg, Germany. He joined the NTNU, Norway, in April 2011, where he is currently a Professor of robotics and cybernetics. His research interests lie on two areas: one is on biological robots and modular robotics and the other is on virtual prototyping and maritime mechatronics.

Finite - Discrete - Element Simulation

LIGHTWEIGHT INDUSTRIAL TRAILER BY USING COMPOSITE MATERIAL - A NEW CONCEPT DESIGN

Federico Ceresoli

Andrea Buffoli

Department of Mechanical and Industrial Engineering, University of Brescia, via Branze, 38 Brescia 25123, Italy

f.ceresoli002@unibs.it

a.buffoli005@unibs.it

KEYWORDS

composite structures design, carbon fiber, trailer, weight reduction, lightweight vehicle

ABSTRACT

This paper presents a new concept of industrial vehicle trailer frame, suitable for some categories of vehicles, in order to reduce their weight. This allows an increase of the payload and a reduction of the fuel consumption and the pollution emissions. Usually, the structural frame of an industrial vehicle trailer has a ladder shape, in this work, the torsional stiffness of a typical ladder frame, evaluated by FEM analysis, will be the reference for the design of the new frame made of different materials like steel and composite material. For those cases, we have studied a new concept of structure, made by a single central hollow beam. The use of a composite material for the central beam allows a weight reduction of the 70% compared to the ladder steel frame and an increase of modularity. Moreover, even if the construction of the composite frame is more expensive than the classical frame, it is possible to repay the composite frame during the first 2/3 years thanks to the reduction of fuel consumption due to the weight reduction.

1. INTRODUCTION

One of the most important parts of an industrial vehicle trailer is the structural frame Fig.1, usually made of steel with a ladder shape. Generally, the trailer of a vehicle is designed to carry different kind of work equipment, suitable for the type of transport carried out by the customer. For this reason, a structural frame must be capable to interface with a wide variety of additional structures and withstand a wide variety of load histories. The most difficult task entrusted to the frame is to distribute the load between the axles, because this implies high bending moments. In fact, in most cases the frame is designed taking into consideration mainly the bending stiffness and the bending strength; the torsional stiffness is a consequence. The equipment that is added on the main frame is suitable for the type of load that must be transported. The resulting truck assumes different body types configuration according to this additional equipment: box truck, dump truck, tank truck, flatbed truck, concrete mixers, etc. Each equipment has its own bending and torsional resistance that collaborates with the ladder frame ones. In most cases, like in the box truck, the resistance and stiffness of the main frame is essential to support the loads.

Nevertheless, in some cases, like tank trucks or refuse compactors, the added structure could be enough strong and rigid to assure the needed bending and torsional stiffness and resistance. Starting from a case of study, the purpose of this paper is to make a feasibility study of a new kind of structural frame shape dedicated to the last case of trucks. For those cases in which it is not requested a demanding bending stiffness of the structural frame, it is proposed a single beam frame, realized with an elliptical hollow pipe section. Considering the current trends in the vehicle design, the main scope of this new concept of frame is the weight reduction. In fact, the main challenges for the automotive industry today are: the reduction of the pollution emission for the public health, the reduction of CO₂ emission (which is responsible of the greenhouse effect) for the environmental protection and the fuel economy improvement (Santos et al. 2017). The mass of a vehicle influences directly the power requirements for the longitudinal dynamic of the vehicle itself, like acceleration or rolling resistance; so a weight reduction causes an abatement of the power requirement and is considered a key element in the design strategies. For this reason, all the manufacturers go in the direction of lighter vehicles even through material substitution. Some studies have shown an increase in the fuel efficiency from 5% to 8% as a consequence of a 10% reduction of the vehicle weight (Brooke and Evans 2009). Every reduction of 100kg leads to an avoided emission of 12.5 g/km of CO₂. It is important to underline that a weight saving in the structural frame offers two opportunities: one is a downsizing of the other components, like the engine and the brakes, which amplifies the weight saving itself; the other one is increasing the payload of each vehicle (Solazzi et al. 2019), which reduces the number of vehicles required to transport a specific load. In both cases, there are advantages for the fuel economy improvement and the environmental protection (Solazzi 2009; Solazzi 2012). An important role in the strategy for the weight reduction is the material substitution (Baskin et al. 2002; Collotta and Solazzi 2017; Laxman and Mohan 2007) for the components and the greater opportunities for this operation exist in the body and chassis components, which comprise the 60% of the total weight. So even if the low carbon steel has always been the most used material of a vehicle structure, nowadays it is common the use of high strength steel (Solazzi 2010); moreover, other materials (Mallick 2010) like aluminum alloys (5000-series and 6000-series) (Solazzi 2010),

magnesium alloys and polymer matrix composites like CFRE (carbon fiber reinforced epoxy) and GFRE (glass fiber reinforced epoxy) are considered and used (Solazzi 2009; Gay and Suong 2003).



Fig. 1 Vehicle trailer

2. STATE OF THE ART

The structural frame of an industrial vehicle trailer is the “skeleton” of the vehicle and carries all the components keeping them together. It is usually made of steel. The main purpose of the structural frame is to be the interface between the transported load and the engine on one side and the suspensions on the other side. For this reason, the frame supports the static weight and the horizontal and vertical dynamic forces during the motion withstanding to all the internal actions due to these forces. The most important actions that the structure has to support are the bending and torsional moments. The leading bending moments are generated by the vertical forces, which are the sum of the static weight and of the inertial forces coming from the vertical accelerations. These forces are applied on the distributed and concentrated masses on the trailer. The reactions to those forces are applied from the ground to the axles and from the axles to the structural frame. The resulting distribution of forces determines the bending actions on the trailer supported by the structural frame. Currently, almost all the structural frames of the industrial trailers adopt the so called “ladder frame” geometry. This technical solution is composed by two main parallel beams, with an open (“I” or “C”) or closed cross section. These beams or rails lay in a horizontal plane and are aligned to the longitudinal direction of the vehicle.

This technical solution has some advantages:

- good capacity to withstand the bending actions,
- simple realization,
- reduced cost of the equipment,
- easy cable and pipe routing.

There are also some disadvantages:

- low torsional stiffness,
- high number of welds or bolts,
- high mass.

3. LOAD CONDITIONS AND CRITERIA ADOPTED TO DEVELOP THE NEW TRAILER

The load conditions considered in this work are two: bending and torsion. The bending is, in general, due to all the vertical forces acting on the trailer coming from

the weight and from the vertical dynamic of the vehicle. In our study only the static weight forces due to the payload and the trailer itself have been considered. The wheels are ideally fixed, creating a hinge constraint on each axle so that the frame is subjected to bending.

The torsion is, in general, due to the inertial horizontal forces which comes from the vehicle lateral dynamic (for example in curve) and by different vertical forces acting on the left wheel and right wheel of each axle; these forces determine a rotation around the roll axis. The entity of these forces is not important in this work, because the attention has been focused on the torsional stiffness, not on the resistance to a particular value of rolling actions. Therefore, the torsional stiffness has been evaluated in a simplified way imposing that the front axle is rigidly fixed, while the other one has its left extremity fixed and the right vertically loaded with 10000N, so the frame is mainly subjected to torsion.

The criteria adopted to develop the new concept frame is that it must have similar properties to a real average steel ladder frame available on the European market pushing the weight reduction to the limit (Solazzi and Scalmana 2016; Solazzi 2019; Solazzi and Buffoli 2019). Therefore, the first step of this work was to apply the load conditions, exposed before, on an existing trailer to establish a benchmark of bending stiffness, stress level and torsional stiffness; the second step was to design the new frame with similar properties but much lighter.

4. CURRENT STEEL LADDER FRAME TECHNOLOGY

The average size trailer chosen as reference is one of the most widespread in the European market and its frame has the following characteristics:

- the length is 9230mm,
- the distance between the front end and the anterior axle is 1370mm,
- the distance between the axles is 5180mm,
- the maximum carriage length is 7840mm,
- the longitudinal beams have a “C”-shaped cross-section with a height of 255mm, a width of 70mm and a thickness of 7mm,
- the external width of the ladder is 850mm,
- there are 8 crossbeams of two different geometries,
- the rear bumper has a length of 2300mm (it won't be loaded, but it will be used to identify the torsional displacement),
- the material is steel SR235JR EN10025-2 ($E=210\text{GPa}$, $\sigma_r=360\text{MPa}$ and $\sigma_y=235\text{MPa}$),
- the total mass is 870kg (evaluated by the solid model).

A simplified (de-featured) CAD model of the structure was prepared with the software Solidworks®; it is represented in Fig.2, while the finite element analysis was performed with Autodesk Simulation®.

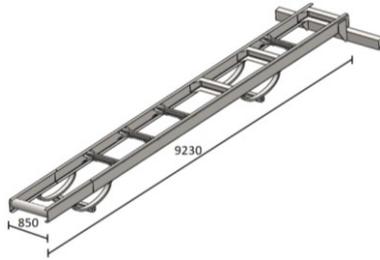


Fig.2 Model of the steel ladder frame

First Load Condition - Bending

The first load condition (bending condition) is composed by:

- the weight of the engine plus the cabin of 20000N uniformly distributed on the anterior part of the chassis,
- the maximum payload of 74000N uniformly distributed on the rear part of the chassis,
- the axles realize an isostatic constraint of hinge and simple-support.

Considering only a half of the system (because of its symmetry) the scheme is represented in Fig.3

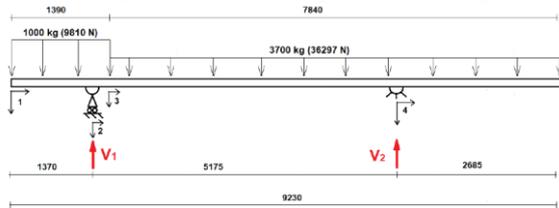


Fig. 3 Load scheme

In the FEM analysis (Vanderplaats and Miura 1986; Jonathan 2017) the constraints have been represented in a more realistic way, modeling each axle with a hollow square beam, measuring 100x100x10mm, positioned on the geometrical axis position of each axles and connected to the longitudinal beams by means of two semi-elliptical leaf springs, whose measures are length 1060mm, height 250mm and width 40mm, the stiffness of the leaf spring is determined on the real frame. The material used is again the steel SR235JR EN 10025-2. The constraints have been obtained fixing the lower faces of the hollow square beams. The finite element analysis model is represented in Fig.4 and is composed by 93783 brick element and 43249 nodes.

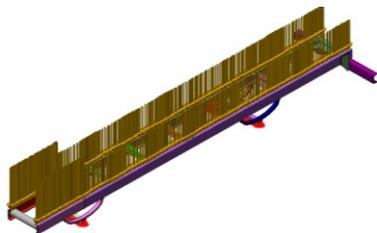


Fig. 4 Bending FEM model

The maximum stress evaluated by Von Mises criteria is 81.5MPa and maximum displacement is 4.5mm as can be seen in Fig.5, so that the safety factor is $\eta_s = 235\text{MPa}/81.5\text{MPa} = 2.88$

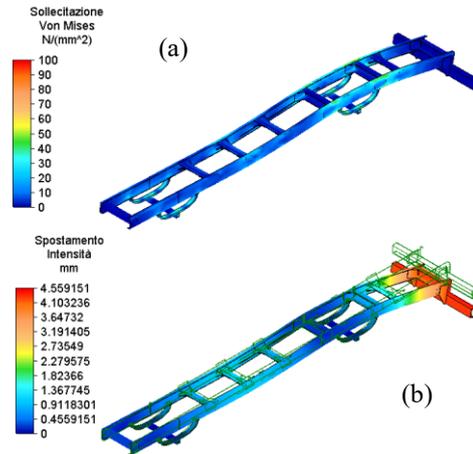


Fig. 5 Von Mises stress (a) and displacement (b) of the ladder frame in the bending load case

Second Load Condition – Torsion

As said before, the torsional stiffness has been evaluated in a simplified way imposing that the front axle is rigidly fixed, while the other has its left extremity fixed and the right vertically loaded with 10000N as can be seen in Fig.6.

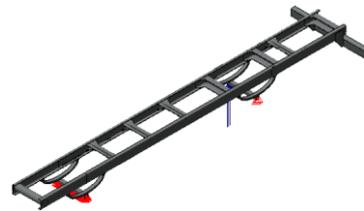


Fig. 6 Torsional FEM model

The displacement value used to evaluate the torsional stiffness is the maximum displacement obtained by the undeformed rear bumper, which is 43.2mm as can be seen in Fig.7.

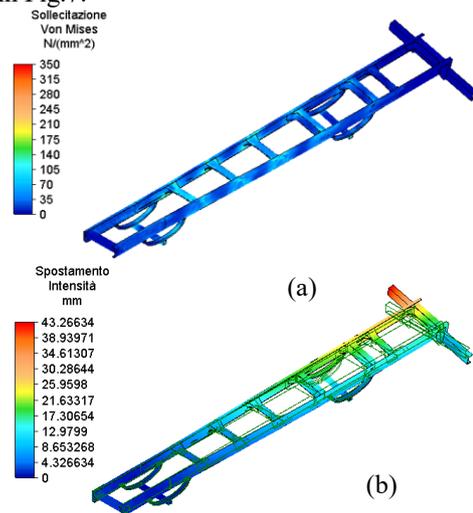


Fig. 7 Von Mises stress (a) and displacement (b) of the ladder frame in the torsional load case

5. NEW CONCEPT FRAME

The new frame proposed Fig.8 in this work is constituted by a single central longitudinal beam (Babamohammadi et al. 2019) onto which are clamped nine mounting brackets in the required position. As can

be calculated in the following paragraphs the total mass of the new concept frame is 267kg.

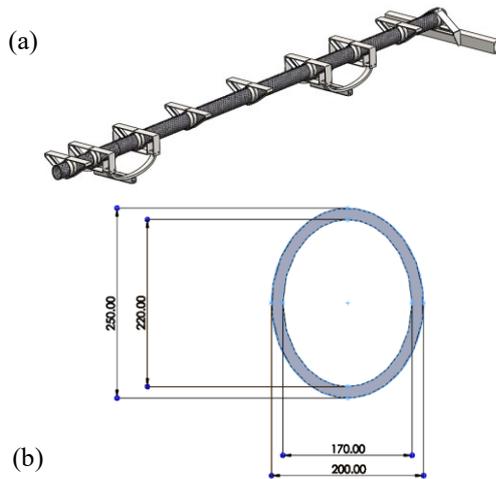


Fig. 8 Geometry of new concept frame (a), elliptical section of the frame (b)

Properties of the beam material

The material is chosen as a compromise between performance and costs (Tiwari et al. 2016). It is made by commercial pre-preg laminas of an intermediate modulus carbon fiber with an epoxy matrix, in which the fiber percentage of each layer is 60% in volume (Agarwal et al. 2006; Mallick 2008; Njuguna 2016).

Table 1 Mechanical characteristics of the materials composing the used lamina

Material	Density ρ [kg/m ³]	Young modulus E [MPa]	Shear modulus G [MPa]	Poisson ratio ν
Epoxidic resin	1200	4500	1600	0.4
Carbon fiber	1750	230000	50000	0.3

The fibers in each lamina are unidirectional and are the long ones; this make the composite oriented, able to optimize the resistance in the main stress direction; the use of long fibers, instead of short ones, is also justified because the technological process allows to realize the beam using specific molds. The short fibers also have a lower fatigue resistance (Garmstedt and Berglund 2003; Talreja and Singh 2012). The principal stress direction is on a helix at 45°, so the right orientation of the fibers is mainly +45° and -45°. This orientation, can be obtained with the filament winding technology (Balasubramanian 2017; Kamal 2017).

The stacking sequence of the layers composing the laminate has to comply the following rules:

- the orientation angle of the single layer has to be chosen according to the principal load directions;
- each lamina orientation (0°, 45°-45°, 90°) should be present in the laminate at least with a percentage of 10%;
- provide a symmetrical distribution of the layers, in order to avoid shear-extension and extension-bending couplings which can induce dangerous curvatures in the laminate;

- consider the protection of the primary layers through their collocation in the internal parts of the laminate;
- guarantee a gradual thickness change where necessary.

With these data is possible to compute the mechanical characteristics of the lamina (Solazzi et al. 2018) reported in Table 2.

Table 2 Resulting mechanical characteristics of the entire lamina

Density ρ [kg/m ³]	E_1 (to fiber) [MPa]	E_2 (\perp to fiber) [MPa]	Shear modulus G_{12} [MPa]	Major poisson ratio ν_{12}	Minor poisson ratio ν_{21}
1530	139800	10929	3817	34	0.026

Some other mechanical characteristics of interest are:

- density $\rho=1530$ kg/m³,
- total mass 140kg.

Central beam

The central beam has a hollow elliptical cross-section Fig. 8b to assure the torsional stiffness to the frame (Solazzi et al. 2019), its dimensions are: height 250mm, width 200mm, thickness 15mm and length 9230mm.

Brackets

The mounting brackets Fig.9 present two different geometries because they perform two different functions: the first kind of bracket, called load bracket, is designed to support only the caisson; the second kind of bracket, called suspension bracket, supports the caisson and the leaf spring of the axle. The brackets have all a width of 853.4mm, like the original frame, an axial length of 150mm and a height of 340mm; they also are divided in two parts symmetrical about the longitudinal vertical plane and have a central “elliptical hole” so that they are bolted one half to the other clamping on the central elliptical beam with M16 8.8 bolts; the elliptical shape of the interface makes the joint irrotational as necessary.

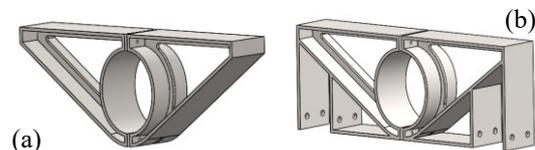


Fig. 9 Load bracket (a) and the suspension bracket (b)

Properties of the bracket material

The brackets are made in an aluminum alloy 7075 T6 (Michael 2016; Michael 2003; Davis 1993), whose commercial name is “Ergal”.

The mechanical characteristics are:

- Young modulus $E=70\ 000$ MPa
- Ultimate tensile strength $\sigma_r=570$ MPa
- Yield strength $R_{p0.2}=500$ MPa

The mass of the load bracket is 12.9kg, the mass of the suspension bracket is 18.9kg, while the mass of the rear

bumper bracket with the bumper is about 40kg. All the total bracket weight is 167kg. The FEM analysis has been performed in the same way followed for the ladder frame, so also in the FEM model have been added the suspensions modeled as a leaf spring with an axle made by a square hollow beam which has been constrained as previously.

First Load Condition - Bending

The first load condition (bending condition) is composed by the same weight of the engine plus the cabin of 20000N, and the maximum payload of 74000N on the rear part of the chassis, but in this case they are not uniformly distributed on the beam, but are distributed onto each bracket considering the area of influence of each one as showed in Fig 10. The axles realize the same isostatic constraint of hinge and simple-support as before.

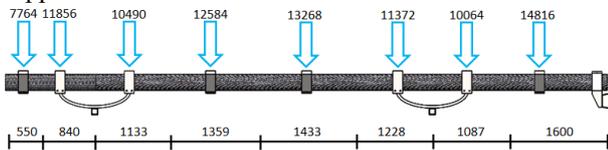


Fig. 10 Scheme of loads on each bracket

The FEM resulting model is shown in Fig.11 it is composed by 162481 brick elements and 86993 nodes.

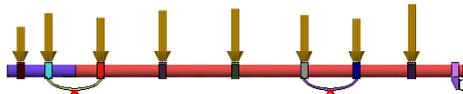


Fig. 11 FEM bending model of the new frame

The maximum stress evaluated by Von Mises criteria is 92.3MPa and the maximum displacement is 6.6mm as can be seen in Fig.12

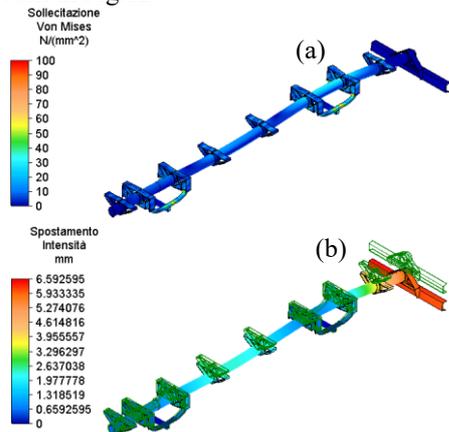


Fig. 12 Von Mises stress (a) and displacement (b) of the new frame in the bending load case

The safety factor has been evaluated with the Tsai-Hill criterion, which is in general expressed by the relation:

$$\left(\frac{\sigma_1}{X}\right)^2 + \left(\frac{\sigma_2}{Y}\right)^2 - \frac{\sigma_1 \cdot \sigma_2}{X^2} + \left(\frac{\tau_{12}}{S}\right)^2 < 1$$

In this case, the stress tensor has only one component σ_1 which is quite similar to the Von Mises stress calculated by the software, so only the directional resistance is need for the verification

$$\left(\frac{\sigma_1}{X}\right)^2 < 1 \quad \left(\frac{92.3MPa}{1270MPa}\right)^2 < 1$$

The safety factor results:

$$\eta_t = \frac{X}{\sigma_1} = \frac{1270MPa}{92.3MPa} = 13.7$$

Much greater than before.

For this new frame it is important to check the resistance of the mounting brackets, so a simplified analysis has been carried on a half of bracket, considering the planar symmetry. The bracket has been constrained on the symmetry plane and the load applied on the top is the half of the maximum foreseen Fig.13.

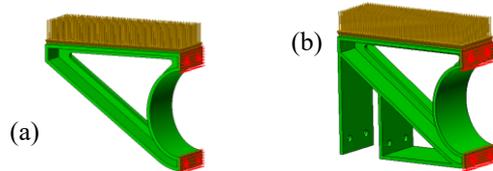


Fig. 13 FEM models of the load bracket (a) and the suspension bracket (b)

The resulting Von Mises stresses are 16MPa for the suspension bracket and 21MPa for the load bracket Fig.14.

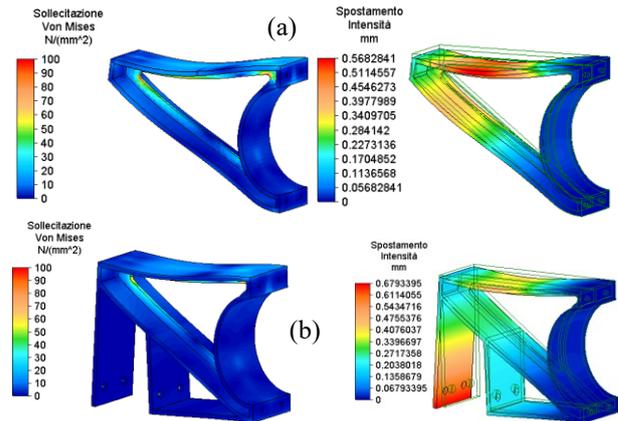


Fig. 14 Stress and displacement for load bracket (a), stress and displacement for the suspension bracket (b)

Second Load Condition – Torsion

As for the model before, the torsional stiffness has been evaluated in a simplified way imposing that the front axle is rigidly fixed, while the other has its left extremity fixed and the right vertically loaded with 10000N as can be seen in Fig.15.

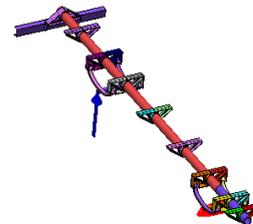


Fig. 15 Torsional FEM model

The displacement value used to evaluate the torsional stiffness is the maximum displacement obtained by the undeformed rear bumper, which is 62mm as can be seen in Fig.16.

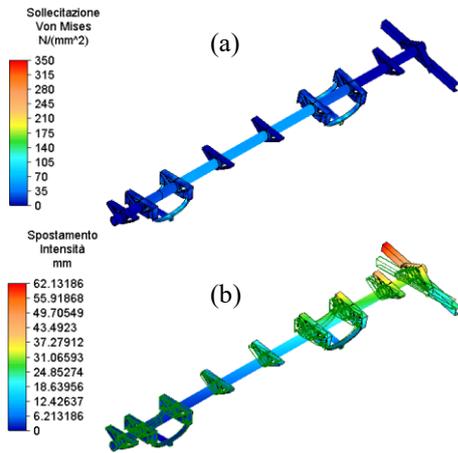


Fig. 16 Von Mises stress (a) and displacement (b) of the new frame in the torsional load case

Economical and industrial evaluations

A comparison between the cost of the various solutions is necessary. For the ladder steel frame proposed, it is possible to make a first estimation, by the rule of a thumb, multiplying its mass for a coefficient of cost, which depends on the price per kilo of the material itself and on the complexity of the technological process (in general cutting, bending, welding). In this case, considering that the price of the steel per kg is less than 1.8€ and that the productive process is conventional and well known, we could estimate a coefficient of cost of 4€/kg; so the result is about 2.200€. In the case of the new concept frame we need to subdivide it in the aluminum made parts and carbon fiber parts. For the aluminum brackets, we could consider a coefficient of cost of 8€/kg, because the current price of the aluminum alloy 7075 T6 is 5,5€/kg and because the welding of this material is more complex; so the result is a cost about 1.250€. Instead, the evaluation of the cost of the carbon fiber frame is quite difficult. In this case, it is not so accurate an estimation by the rule of a thumb, just on the basis of weight of the structure, because the production process is completely different. In fact, it is necessary to build a mold and it is necessary an autoclave; moreover, in this application, the dimensions are really big, this implies that there are only a few producers who have the necessary equipment. In any case, to have a term of comparison, we could try an estimation on the basis of the cost of the pre-preg laminas of this intermediate modulus carbon fiber with an epoxy matrix and on the basis of the cost of the structures in composite material that have already been produced in the past. We can evaluate a cost of about 16.000€ for the new concept frame. Even if our economical consideration is approximate, it is clear that a trailer frame realized in carbon fiber composite has a cost which is of one order of magnitude greater than the cost of a classic steel solution.

6. RESULTS

The Table 3 resumes the results of the analyses:

Table 3 Comparison between the solutions

Frame	Steel ladder	Composite	Difference
Mass [kg]	870	267	-603 (-70%)
Bending disp. [mm]	4.6	6.6	+2 (+43%)
Torsional disp.[mm]	43.2	62.1	+19 (+44%)
Estimated cost [€]	~2200	~16000	+13800 (+627%)

The new concept frame allows a great weight reduction of the 70% but has a small increase (from 4.6 to 6.6mm similar engineering values) in the displacements indicating minor bending and torsional stiffness (Raugei et al. 2015). For the cases indicated in the introduction like tank trucks or refuse compactors, where the added structure is enough strong and rigid to assure the needed bending and torsional stiffness and resistance, the performance reduction of the new frame is acceptable Fig.17(a). Currently the cost of the composite frame is high but thanks to the weight reduction of the entire truck, from 4500kg (traditional steel solution) to 3897kg (composite material solution) equal to 13%, we can obtain a reduction of the fuel consumption of about 11%. In this way, by assuming to travel 120000km per year per driver, as indicated by the Italian Ministry of Transport, it is possible to repay the investment for the adoption of a composite frame over the first three years of use of the vehicle Fig.17(b). Another important aspect (Lewis et al. 2019; Luk et al. 2018) is also given by the reduction of CO₂ emission equal to 75.4g/km thanks to the weight and fuel consumption reduction. Furthermore, the adoption of the solution presented in this paper has the great advantage of being highly modular thanks to the use of brackets and the geometry of the tubular frame. Finally, thanks to the weight reduction of the vehicle and its modularity, it is also possible to adopt different propulsion system instead to the typical combustion engine (Kim et al. 2016).

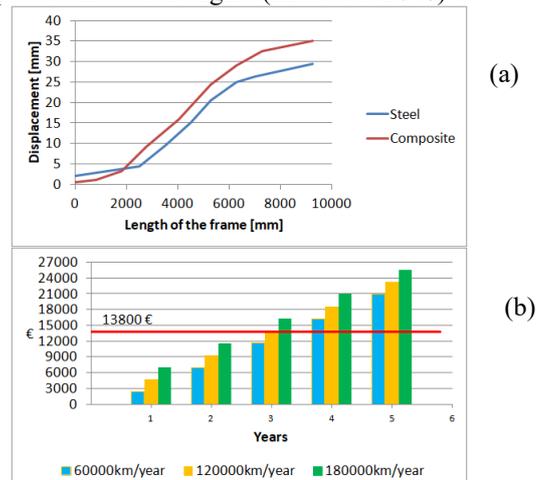


Fig. 17 Trend of the torsional displacement (a), cost savings with reduction of fuel thanks to the composite frame (b)

7. CONCLUSIONS

The purpose of this work is to reduce the weight of the structural frame of a trailer. The proposed geometry of the frame decreases the bending and the torsional

stiffness, which is acceptable only in those cases in which the applied caisson is capable to withstand the actions due to the load, but reduces the weight of the trailer's frame of the 70%. This weight reduction of the frame leads to an increase of the transportable load or to a downsizing of the other components, like the engine and the brakes, which amplifies the weight saving itself. In any case, it leads to a reduction of the fuel consumption and to the emission of CO₂. The great results obtained can also be applied to other fundamental components and can also be used to vehicle that present frames with different dimension and load capacity.

REFERENCES

- Santos J, Gouveia RM, Silva FJG. 2017. Designing a new Sustainable Approach to the Change for Lightweight Materials in Structural Components Used in Truck Industry. *J Clean Prod*; 164:115–23.
- Brooke L, Evans H. Lighten up! 2009. *Society of Automotive Engineers (SAE)*;117:16-22.
- Solazzi L, Ceresoli F, Cima M. 2019. Structural analysis on lightweight excavator arms. *Proceedings - European Council for Modelling and Simulation. ECMS 33(1)*: 351-357.
- Solazzi L. 2009. Design of a new complete industrial trailer using different materials. XIX International conference on "Material Handling, Constructions and Logistics", Belgrade, Serbia, 15-16 October.
- Solazzi L. 2012. Applied research for weight reduction of an industrial trailer. *FME Transactions*; 40: 57-62.
- Baskin D, Dinda S, Moore T. 2002. A Simple Approach to Selecting Automotive Body-in-White Primary-Structural Materials. SAE Technical Paper.
- Collotta, M, Solazzi L. 2017. New design concept of tank made of plastic material for firefighting vehicle. *International Journal of Automotive and Mechanical Engineering*;14: 4603-4615.
- Laxman S, Mohan R. 2007 Structural optimization: achieving a robust and light-weight design of automotive components. SAE Technical Paper.
- Solazzi L. 2010. Design a scrap loader using different alloy steels. II^o International Conference on Super High Strength Steel, Peschiera del Garda (Vr), Italy, 17-20 October.
- Mallick PK. 2010. Materials design and manufacturing for lightweight vehicles. Woodhead Publishing.
- Solazzi L. 2010. Design of an aluminum boom and arm for an excavator. *Journal of Terramechanics*; 47: 201-207.
- Solazzi L. 2009. Influence the design of composite material on the mechanical behavior of the cantilever – type spring. Ninth International conference on Experimental techniques and design in composite materials, Vicenza, Italy, 30 September, 1-2 October.
- Gay D, Suong VH, Tsai SW. 2003. Composite materials- design and applications. CRC press.
- Solazzi L, Scalmana R. 2016. New design concept for a lifting platform made of composite material, *Journal of Applied Composite Materials*; 4: 615-626.
- Solazzi L. 2019. Feasibility study of hydraulic cylinder subject to high pressure made of aluminum alloy and composite material. *Composite Structures*; 209: 739-746.
- Solazzi L, Buffoli A. 2019. Telescopic Hydraulic Cylinder Made of Composite Material. *Applied Composite Materials*; 26(4): 1189-1206.
- Vanderplaats GN, Miura H. 1986. Trends in structural optimization: some considerations in using standard finite element software. SAE Technical Paper 860801.
- Jonathan W. 2017. Finite Element Methods A Practical Guide. 1st Edition. Springer International Publishing.
- Babamohammadi S, Fantuzzi N, Lonardi G. 2019. Mechanical assessment of hollow-circular FRP beams. *Composite Structures*; 227.
- Tiwari A, Alenezi MR, Jun SC. 2016. Advanced Composite Materials. ISBN: 978-1-119-24253-6 Scrivener Publishing LLC.
- Agarwal BD, Broutman LJ, Chandrashekhara K. 2006. Analysis and performance of fiber composites, third edition. John Wiley and sons, Inc.
- Mallick PK. 2008. Fiber Reinforced Composites. Boca Raton, CRC Press.
- Njuguna J. 2016. Lightweight Composite Structures in Transport Design, Manufacturing, Analysis and Performance. Woodhead Publishing Series in composites Science and Engineering: Number 67 ISBN: 978-1-78242-325-6 (print) ISBN: 978-1-78242-343-0 (online).
- Garmstedt EK, Berglund LA. 2003. Fatigue in thermoplastic composites. eds: Bryan Harris, Woodhead Publishing Ltd, Cambridge.
- Talreja R, Singh CV. 2012. Damage and Failure of Composite Materials. ISBN 978-0-521-81942-8 Cambridge University Press.
- Balasubramanian M. 2017. Composite Materials and Processing. 1st Edition. CRC Press.
- Kamal KK. 2017. Composite Materials Processing, Applications, Characterizations. 1st Edition. Springer International Publishing.
- Solazzi L, Assi A, Ceresoli F. 2018. New Design Concept for an Excavator Arms by Using Composite Material. *Applied Composite Material*; 25(3): 601-617.
- Solazzi L, Assi A, Ceresoli F. 2019. Excavator arms: Numerical, experimental and new concept design. *Composite Structures*; 217: 60-74.
- Michael AF. 2016. Materials and Sustainable Development. Butterworth-Heinemann imprint of Elsevier Ltd, Oxford.
- Michael AF. 2003 Materials Selection in Mechanical Design Third Edition. Butterworth-Heinemann. ISBN:0-7506-6168-2.
- Davis JR. 1993. Aluminum and Aluminum Alloys. ASM Specialty Handbook, ASM International.
- Raugei M, Morrey D, Hutchinson A, Winfield PA. 2015. coherent life cycle assessment of a range of lightweighting strategies for compact vehicles. *Journal of Cleaner Production*; 108: 1168-1176.
- Lewis GM, Buchanan CA, Jhaveri KD, Sullivan LJ, Kelly JC, Das S, Taub AI, Keoleian GA. 2019. Green Principles for Vehicle Lightweighting. *Environ. Sci. Technol.*; 53: 4063-4077.
- Luk JM, Kim HC, De Kleine RD, Wallington TJ, MacLean HL. 2018. Greenhouse gas emission benefits of vehicle lightweighting: Monte Carlo probabilistic analysis of the multi material lightweight vehicle glider. *Transportation Research Part D* 62; 1-10.
- Kim HC, Wallington TJ. 2016. Life Cycle Assessment of Vehicle Lightweighting: A Physics-Based Model To Estimated Use-Phase Fuel Consumption of Electrified Vehicles. *Environ. Sci. Technol.*; 50: 11226-11233.

FAILURE ANALYSIS OF A CUSTOM-MADE ACETABULAR CAGE WITH FINITE ELEMENT METHOD

Martin O. Dóczy
Péter T. Zwierczyk

Department of Machine and Product Design
Budapest University of Technology and
Economics
Műegyetem rkp. 3., Budapest 1111, Hungary
doczi.martin@gt3.bme.hu
z.peter@gt3.bme.hu

Róbert Szódy

Péterfy Hospital and Manninger Jenő
National Institute of Traumatology
Fiumei street 17., Budapest 1081, Hungary
robert.szody@gmail.com

KEYWORDS

Acetabular bone defect, Acetabular cage, Patient-specific, Finite element analysis

ABSTRACT

Research significance: In this paper, a custom made acetabular cage was studied in order to calculate the mechanical stresses of the implant. The goal is to have a validated finite element model, which can provide qualitatively accurate results.

Methodology: The geometry models of the acetabular cage, the pelvis and the fixation are based on the patient's computer tomography data. The geometry models were made during a reverse engineering and surface modeling process.

The hemipelvis was modeled, according to the literature research. The boundary conditions, loading model, material properties were from the literature research as well.

Results: Analyzing the von Mises stresses of the acetabular cage and its screws, the most loaded areas could be detected. Viewing the first and the third principal stresses, the tensioned and the compression areas could be separated. With the patient's control CT data, the results could be validated, analyzing the deformation of the cage, and the fracture of an implant.

Discussion: The FE simulations of the acetabular cage provide results, which are consistent with observed implant failure.

INTRODUCTION

The total hip replacement is an effective solution to treat osteoarthritis, reducing hip pain and the patients can have a better daily living. During the procedure, a socket is inserted in the acetabular part of the pelvis, and a metal stem into the hollow part of the femur. A metal or ceramic ball is placed on the stem, and usually a polymer liner inside the socket.

However, this solution usually requires a revision surgery after 10-20 years, where the reparation of damaged prosthesis' elements are done.

In some cases, one has to deal with large acetabular bone defects (Figure 1).

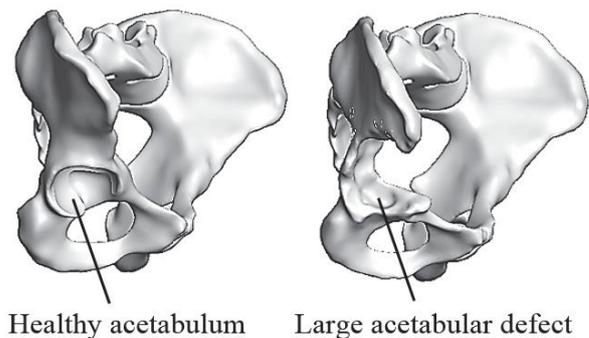


Figure 1: Different clinical states of the pelvis

The treatment of a large acetabular bone defect is a challenging clinical task. There is no consensus about a general treatment, and it is a hard problem because these are always very individual cases. (Ahmad and Schwazkopf 2015; Paprosky et al. 1994)

In this study, the treatment was a cold worked, custom made sheet metal acetabular cage by Róbert Szódy and his colleagues. (Szódy et al. 2017)

The patient could walk again, but after one year from the surgery, small plastic deformation of the cage and a screw's fracture could be discovered on the patient's CT examination.

The task is to make a finite element model which can provide qualitatively accurate results for the stress concentration areas. This can help later to detect the zones where geometry improvement is required.

DATA AND METHOD

Data

There were available three Computer Tomography (CT) data of the patient. The first represented the status before the surgery (pre-operative CT), the second was after the surgery (post-operative CT), and the third was the control after one year (control CT).

Each CT data had a different role for the simulation. The geometry model of the pelvis based on the first CT. For making the geometry model of the acetabular cage and the positions of the screws, the second CT was used.

With the control CT, it was possible to validate the simulation results.

The CT data is a three-dimensional scalar field, where each voxel has a so-called Hounsfield Unit value, which is related to the small volume-parts X-ray attenuation. If the attenuation is great, then the Hounsfield value is higher as well. Hence it is possible to represent the different density regions throughout the CT data.

Geometry Models

The CT data were imported to Slicer 3D for the geometric reconstruction. With this software, one can volume render the relevant areas from the CT and export it in “.stl” file format. After that procedure, the CAD model can be generated with a reverse engineering workflow.

Firstly, the CAD model of the pelvis was made. For this, the pre-operative CT was used.

With a global threshold-based segmentation, the pelvis bone was highlighted. For the accurate geometry model, a manual selection of the relevant parts was required. Then the “.stl” file could be saved.

In MeshMixer, the small defects of the “.stl” file and the holes were eliminated. With a smoothing procedure, it became a manifold hemipelvis model.

For the surface reconstruction, a SolidWorks 2018 CAD system with Scanto3D module was used.

After importing the “.stl” file, an automated surface reconstruction method was chosen, with manually adjusted feature lines. Then, a closed surface model was generated, which could be transformed into a solid body. The pelvis areas which were close to the acetabular defect were separated from the intact parts of the pelvis, because there a different material models were used. The pelvis model can be seen in Figure 2.

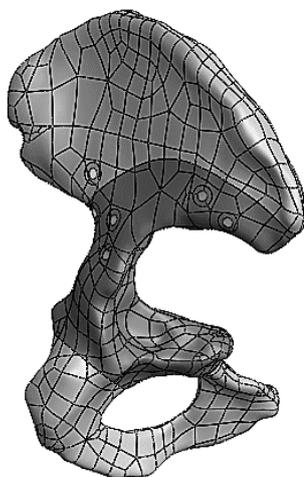


Figure 2: The CAD model of the pelvis

The acetabular cage is a surface model. The post-operative CT data was used to make the CAD model for the cage and its screws. These metal parts were highlighted with a global threshold-based segmentation method. In this case, it was not necessary to use manual segmentation. Then the “.stl” file was exported. The mid surface of the acetabular cage was modeled with reverse engineering and surface modeling tools, using SolidWorks 2018.

The segmented .stl file and the CAD models of the acetabular cage and its screws can be seen in Figure 3.

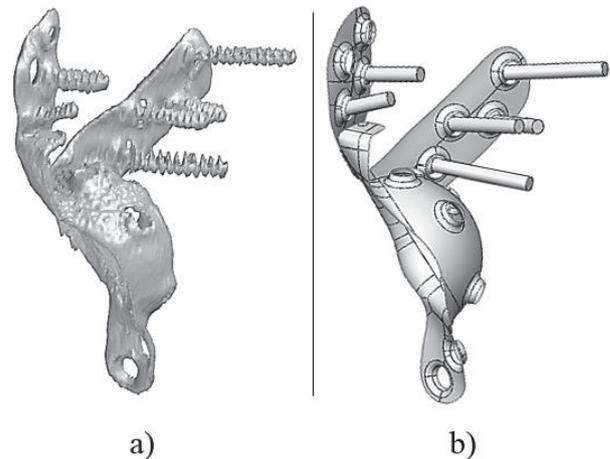


Figure 3: The .stl file of the acetabular cage and its screws (a) and the CAD model (b)

The screws have simplified geometry. Instead of modeling the threads, cylindrical geometries were made. The diameter of the cylinder was the nominal size of the thread (4.5 mm). The head of the screw was made according to the relevant standard (ISO 5831:1991). This was simplified spherical geometry, without the hexagonal drive hole.

For the assembly, the post-operative CT data was used, which helped to position the center of rotation, and the positions of the screws.

The liner and the balls were made as a revolved bodies. It can be seen on Figure 4 the parts and their names in an exploded views.

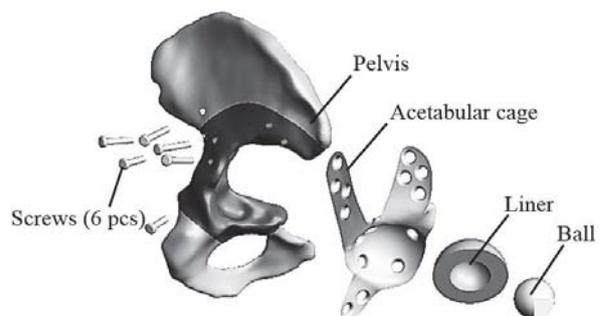


Figure 4: The parts of the fixation in an exploded view

Finite Element Model

For the preprocessing of the finite element model, HyperMesh 2017.2 was used.

For the meshing of the solid models, 10-node tetrahedral elements were used. The finite element mesh of the pelvis has an average 2 mm edge length.

The intact areas of the pelvis are covered by shell elements, representing the cortical bone layer. The thickness of the shell elements was 1 mm. These were linear triangular elements. (Plessers and Mau 2015)

Previously mentioned, the acetabular cage was a surface model. Hence, shell elements were used there as well. The thickness of the shell elements was 1.5 mm, which was the real thickness of the sheet metal. Four node linear quads were used for the meshing of the acetabular cage.

The Optistruct solver does not support the quadratic shell element for large displacement analysis, that was the reason for using linear shell elements.

Manual mesh refinement was only used on the acetabular cage because this part is on the focus of the investigation. The data of the mesh can be seen in Table 1.

Table 1: The data of the FE mesh

Number of nodes	244331
Number of elements	176428
Number of 10 node tetra elements	145562
Number of 4 node quads	14504
Number of 3 node trias	16362
Maximum Aspect ratio (solid elements)	5.28
Maximum Aspect ratio (shell elements)	4.46

The material properties were homogenous, linear elastic and isotropic.

It was previously mentioned that the pelvis had different material models. The healthy pelvis is a so-called irregular bone. On its surface, there is a thin layer of compact bone, called cortical bone, and inside this sandwich structure, there is the trabecular bone.

Due to the acetabular component migration, it can be seen on the pre-operative CT data, that near by the bone defect it cannot be separated easily the cortical and the trabecular bone. For this case, a homogenous part is modeled with averaged Young's modulus. (Anderson et al 2004, Ravera et al. 2015)

The acetabular cage is a cold worked sheet metal part. There was no annealing after the forming. Due to the strain hardening effect, the yield strength is increased in the formed regions. It is a challenging task to calculate the residual stresses and the modified yield strengths for further simulations, thus, linear elastic material properties were used.

The summary of the material properties can be seen in Table 2.

Table 2: Material properties

	Young's modulus [MPa]	Poisson's ratio [-]
Steel (AISI 316L)	192000	0.3
XLPE	1000	0.4
Cortical bone	17000	0.3
Trabecular bone	100	0.3
Homogenous bone	7000	0.3

The material of the ball and the screws were steel, homogenous, linear elastic and isotropic properties were used for the analysis with the steel's Young's modulus and Poisson's ratio.

There were bonded contact between the liner and the acetabular cage because these have a glued connection. The screws with the pelvis had bonded contact as well. The threaded connection was simplified with this modeling procedure.

The other metal – bone interfaces had frictional contact with a 0.3 frictional coefficient, as well as the metal-metal interfaces, but there 0.23 frictional coefficient was used. (Chih-Wei Chang et al. 2014)

There is small sliding friction between the ball and the liner, for numerical stability, a small (0.02) frictional coefficient was used.

According to the literature research, there are only a few publications where the muscular forces were modeled. This study is focusing on the acetabular cage, so the loading model and the boundary conditions were a bit simplified but consistent with the literature. It can be seen in Figure 5.

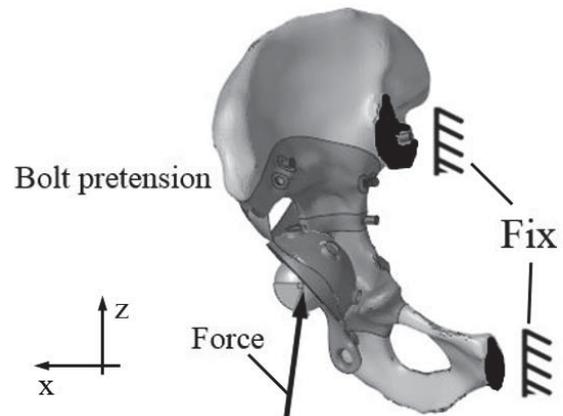


Figure 5: Loading model and boundary conditions

The main load was the maximum of the gait cycle because this is the most common load for this kind of implant. The components of this load were from the literature research. The coordinate system of the pelvis was the same what Bergmann et al. used. (Bergmann et al., 2001) The peak magnitude of the load is 233% Bodyweight. The patient's weight was approximately 75 kg. The components of the main load can be seen in Table 3.

Table 3: Components of the main load

Force component	Value
X	-213.8 N
Y	-194.5 N
Z	1690.8 N

There was a fix boundary condition at the sacroiliac joint and the pubic symphysis. (Plessers and Mau 2016) For all simulations, the large displacement analysis was selected.

Bolt pretension was applied on the screws before the main load, for closing the contacts and modeling the post-operative situation. The magnitude of the bolt pretension was 50N on each screw.

RESULTS

Finite Element Results

The displacement field was as expected, in the direction of the load. The z-direction displacement field after the main load, can be seen in Figure 6.

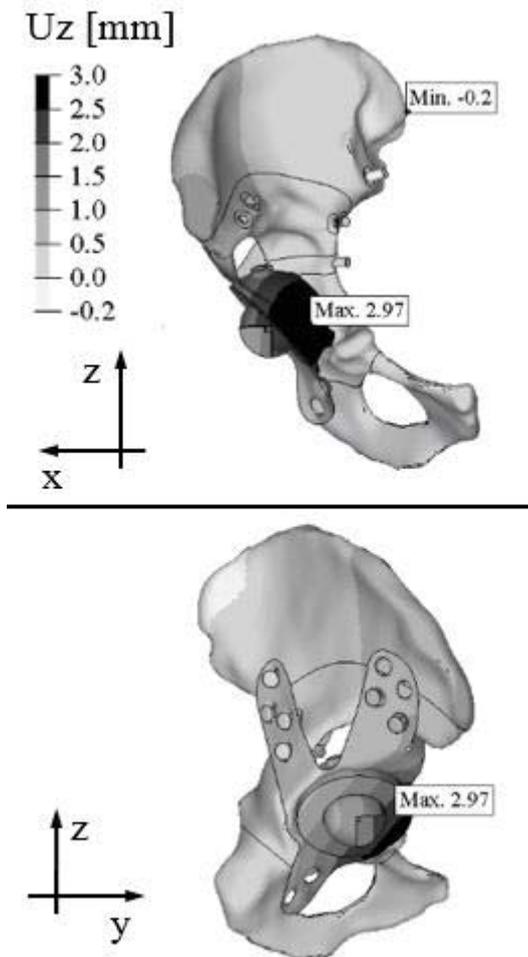


Figure 6: z-direction displacement field of the fixation

The von Mises stresses of the acetabular cage can be seen in Figure 7.

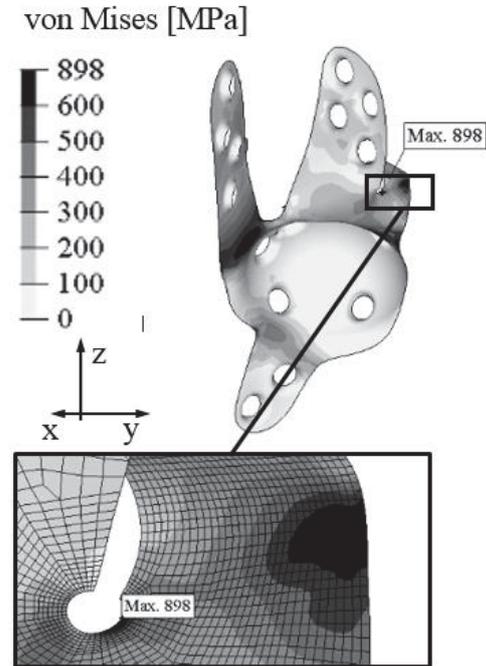


Figure 7: The von Mises stress on the acetabular cage

For investigating the tensioned and compressed zones, the first and third principal stresses were represented. In Figure 8 can be seen the different areas.

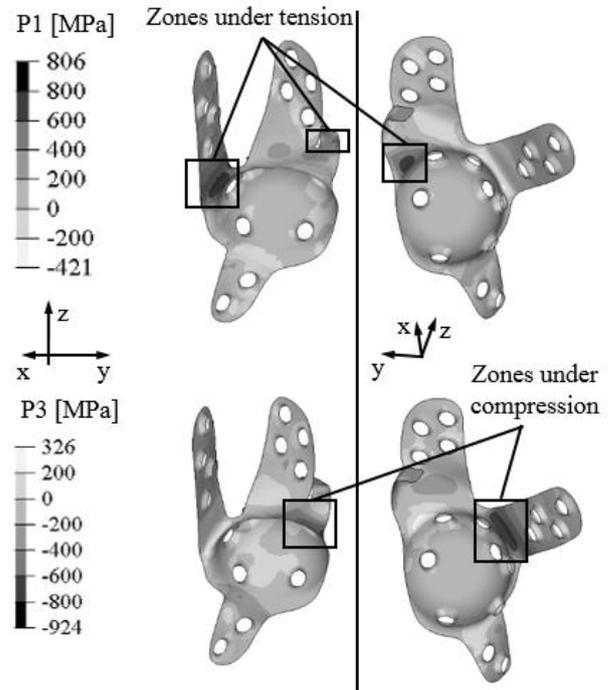


Figure 8: The first and the third principal stresses on the acetabular cage

Investigating the von Mises stresses on the screws, it can be seen which had the largest von Mises stress. It is way below the material's Yield strengths, but it is important that the screws had simplified cylindrical geometry, which did not have so special stress concentrating areas as a threaded geometry, and they had a larger section as well. The loading of the screws was bending, this can be seen after viewing the first and the third principal stresses (Figure 9).

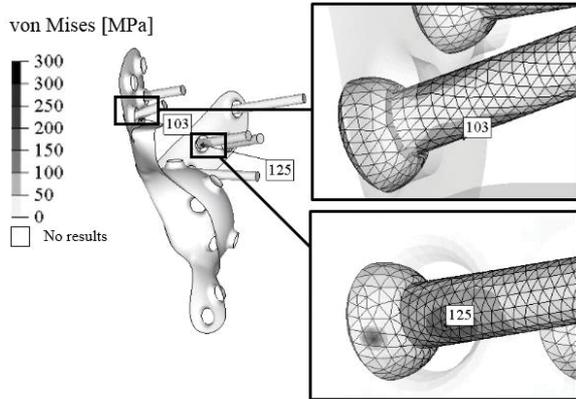


Figure 9: The most loaded screws and its von Mises stresses (cage's visibility transparent)

Validation

After registering the pre-operative and the post-operative stl meshes, there can be seen the main deformation areas. (Figure 10). These are the same, where the large stresses had developed.

The results are qualitatively accurate because the model can predict where are the tensioned or compressed zones. The sheet metal acetabular cage's main loading is bending. Therefore, it can be seen where the sheet metal deform inward or outward (dent or protrusion and deflections). On the side where protrusion occurred, there was the tensioned part, where dent occurred, there was the compressed area.

The lighter gray stl mesh is the deformed cage (from control CT), and the darker one is the undeformed (from post-operative CT).

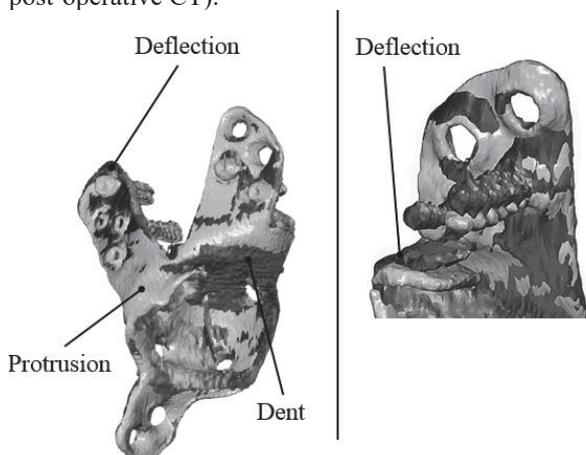


Figure 10: The deformations of the acetabular cage

Another good option for validation is that the fractured screw is the same, which has a large von Mises stress. It can be seen in Figure 11.

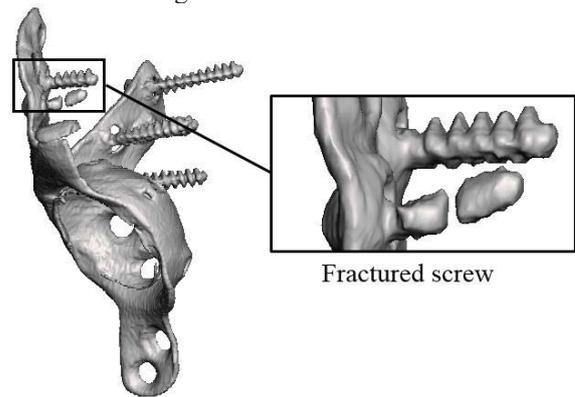


Figure 11: Fractured screw

DISCUSSION

The FE simulations of the acetabular cage provide results which are consistent with observed implant failure. The deflections, dents, and protrusions occurred in the same areas, where were the maximum von Mises stresses. Analyzing the first and the third principal stresses, it could be seen that the results are qualitatively quite accurate because the tensioned and compressed side could be separated, and the characteristic of the deformations could represent the similarity.

There were made geometry simplifications to modeling the pelvis, the screws and the cage. No one of these could provide large effect for the results because the geometries were quite accurate. Because of the larger diameters on the threads, the mechanical stresses were obviously inaccurate, which could provide qualitative results. With the simplification of the cage and the screws, the computational cost was reduced, but with these models, the local stresses, for example near the contacts between the bolt heads and the cage could not be investigated, only the stresses far from these zones.

The more accurate modeling of the material properties of the bone was not a trivial task, but the authors think that this had not a large impact for the stresses of the cage.

Next, the authors want to make more analyses for other cages, and checking the sensitivity of the results for different parameters, for example different material properties, bolt pretension etc.

ACKNOWLEDGMENT

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI), and by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

REFERENCES

- Ahmad, A. and Schwarzkopf, R. 2015. "Clinical evaluation and surgical options in acetabular reconstruction: A literature review." *Journal of Orthopaedics* 12 (2): S238-S243
- Anderson, A. et al. 2005. "Subject-Specific Finite Element Model of the Pelvis: Development, Validation and Sensitivity Studies." *Journal of Biomechanical Engineering* 27 (3): 364-373
- Bergmann, G. et al. 2001. "Hip contact forces and gait patterns from routine activities." *Journal of Biomechanics* 34 (7): 859-891
- Chih-Wei Chang et al. 2014. "Role of the compression screw in the dynamic hip-screw system: A finite-element study." *Medical Engineering & Physics* 37 (112): 1174-1179
- ISO 5831:1991 "Implants for surgery - Metal bone screws with hexagonal drive connection, spherical under-surface of head, asymmetrical thread -Dimensions"
- Paprosky, W., Perona, P. and Lawrence, J. 1994. "Acetabular defect classification and surgical reconstruction in revision arthroplasty: A 6-year follow-up evaluation." *The Journal of Arthroplasty* 9 (1): 33-44
- Plessers, K. and Mau, H. 2016. "Stress Analysis of a Burch-Schneider Cage in an Acetabular Bone Defect: A Case Study." *Reconstructive review*. 6 (1): 37-42
- Ravera, E. et al. 2015. "Combined finite element and musculoskeletal models for analysis of pelvis throughout the gait cycle." *Conference: 1st Pan-American Congress on Computational Mechanics and XI Argentine Congress on Computational Mechanics*
- Szódy, R. et al. 2017. (in hungarian) "Csípőprotézis revízióikor alkalmazott „custom made” vápakosár tervezése és készítése, három esetben alkalmazott eljárás." In *7. Hungarian Conference of Biomechanics* (Szeged, HU, okt 6-7) *Biomechanica Hungarica* 10(2): 20

AUTHOR BIOGRAPHIES

MARTIN O. DÓCZI is a Ph.D. student at the Budapest University of Technology and Economics Department of Machine and Product Design, where he studied mechanical engineering and obtained his degree in 2019. His research area is numerical biomechanics and implant development. His e-mail address is: doczi.martin@gt3.bme.hu and his Web-page can be found at <http://www.gt3.bme.hu>.

RÓBERT SZÓDY is an orthopaedic and traumatology physician. He got his degree at the Semmelweis University in 1995. He made a traumatology professional examination in 2000 and an orthopaedics professional examination in 2005. He works as a surgeon at Péterfy Hospital and Manninger Jenő National Institute of Traumatology. His e-mail address is: robert.szody@gmail.com.

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 member of the finite element modelling (FEM) research group. His e-mail address is: z.peter@gt3.bme.hu and his web-page can be found at: <http://gt3.bme.hu>

A new variable for characterising irregular element geometries in experiments and DEM simulations

Katalin Bagi
Department of Structural
Mechanics
Faculty of Civil Engineering

Budapest University of
Technology and Economics
H-1111, Budapest, Hungary
E-mail: kbagi@mail.bme.hu

Ákos Orosz
Department of Machine and
Production Design
Faculty of Mechanical
Engineering
Budapest University of
Technology and Economics
H-1111, Budapest, Hungary
E-mail: orosz.akos@gt3.bme.hu

KEYWORDS

Fabric tensor, Railway ballast, Crushed rock, Discrete element method, Polyhedron

ABSTRACT

Discrete element method (DEM) has proved to be an excellent tool for modelling bulk materials. Contrarily to the early stages of these simulations when mainly circular and spherical elements were used, extensive research is going on regarding the application of complex element shapes, e.g. polyhedra. Robust and objective geometry characterisation methods are needed to quantify the shape of virtual and real particles in order to assess the effect of particle shape on the global mechanical behaviour.

This paper proposes a weighted fabric tensor that is able to characterise the shape of individual particles in such a way that the preferred potential load-bearing direction(s) are pointed out. The three eigenvalues of this tensor express whether the stone block or grain is compact, flaky, rod-like, or is an intermediate shape in between the three basic shapes. The proposed approach has computational advantages in quantifying the results of imaging processes of stones and grains deterministically without any subjectivity. It has the advance over the traditional bounding box approaches that it is directly based on the orientations of the surface normal vectors, i.e. those directions along which an assembly of stones or grains can best transmit the internal forces.

INTRODUCTION

Particle assemblies (e.g. sand, crushed stone, gravel) and collections of stone blocks are applied as load-bearing structures in several fields of engineering. The shapes of the grains that are used in an assembly significantly affect the overall mechanical behaviour, and hence there are different shape characteristics for the classification of stone geometries that serve as a basis for the design of such structures. Masonry structures are, for example, made up of individual stone blocks whose shape, especially in case of dry joints, have a significant impact on the load-bearing capacity of the whole structure.

Another typical field where the grain shape has high importance is railway ballast stone aggregates (Fischer et al., 2015). The increasing demand for understanding the role of the shape of individual components on the mechanical properties of the whole structure created the need to define quantitative parameters for describing particles shape. Thorough overviews of the different suggestions are given in (Szabó, 2013) and (Guo et al., 2019).

Discrete element method (DEM) simulations (Bagi, 2007; Cundall and Strack, 1979) are frequently applied to study the effect of particle shape. There are two ways to mimic the interlocking effect of irregular geometries in assemblies in DEM models: (i) use simple (i.e. circular or spherical) elements along with a complex contact model (e.g. rolling resistance model) (Wensrich and Katterfeld, 2012), or (ii) apply more complex element shapes. Typically, bonded spheres (i.e. *clumps* or *clusters*), ellipsoids, poly-ellipsoids or convex polyhedra are used. In the case of the second approach, again, robust and subjective shape characterising methods are needed to prove the similarity of grain shapes between reality and virtual models. The aim of the present paper is to characterise not a complete assembly, but the individual shape of each stone in the assembly.

In practice, usually, four main types of shapes are distinguished: (1) “compact” (e.g. a sphere or a cube), (2) “flaky” or “disc-like” (e.g. a disc, or a flat prism), (3) “elongated” or “rod-like” (e.g. a thin column or a needle) and (4) flat-elongated, or blade-like, as done in Zingg’s fundamental publication (Zingg, 1935). The classification is performed according to the relations between the three characteristic sizes of the analysed stone (termed in different ways in the literature, e.g. length, breadth (width) and thickness in (Wadell, 1932)). There are several studies and suggestions in the literature about how these sizes should be defined, but there has been no general agreement on “the” most suitable characterisation.

However, most of these characterisations are based on the “bounding box approach” (Domokos et al., 2015) whose main idea is to include the analysed grain in “the smallest” brick-shaped enclosing domain (“smallest”

according to, e.g. its volume or surface). There is no general agreement today in the literature on exactly how to define this enclosing domain in experiments and simulations, and this introduces ambiguities into the definition of “length”, “breadth” and “thickness” of the analysed particle. In any case, the dimensions $a \geq b \geq c$ are received as the side lengths of the domain, and then based on them, several alternative characteristics can be defined in order to classify the shape of the analysed grain.

Though the bounding box approach has been applied in the literature for a wide variety of purposes, from the point of view of load-bearing capacity of assemblies, it has the disadvantage that the suggested characteristics are based on a surrounding domain, and not directly on those faces of the stone on which it can form contacts with its neighbours for force transmission. In addition, modern imaging techniques (e.g. laser scanning (Asahina and Taylor, 2011) or computer tomography (Juhász and Fischer, 2019)) provide a set of surface triangles (nodal coordinates) as their output, and the bounding box is not entirely straightforward to determine from these data. So, the present study aims to propose a completely deterministic and computationally simple alternative to the bounding box approach in such a way that the orientations of the faces of the stone would serve as the basis of the characterisation and classification.

FABRIC TENSORS

The proposed alternative is based on defining a weighted fabric tensor. Fabric tensors (Satake, 1982, 1983) have been applied for many decades on a wide variety of fields whenever the orientational distribution of a set of unit vectors had to be described. In most applications the contact normal vectors in the assembly gave the basic vector set on which the tensor was built; as an example, a recent application for the characterisation of stress-induced anisotropy can be found in (Shi and Guo, 2018). In the three-dimensional Cartesian (x_1, x_2, x_3) coordinate system, the general definition of the second-order fabric tensor of a set of M vectors of unit length $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(M)}$ is the following (Here \circ denotes dyadic or tensor product.):

$$\boldsymbol{\varphi} = \frac{1}{M} \sum_{k=1}^M (\mathbf{v}^{(k)} \circ \mathbf{v}^{(k)}) \quad (1)$$

This is a symmetric tensor hence its eigenvalues are real numbers and its eigenvectors are perpendicular to each other; in addition, the eigenvalues are nonnegative and sum up to 1. The eigenvector belonging to the largest / smallest eigenvalue expresses the most / least preferred orientation of the $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(M)}$ vector set, and the strength of the bias is expressed by the magnitude of the corresponding largest / smallest eigenvalue.

The terms in the summation may receive some characteristic weight (e.g. for $\mathbf{v}^{(k)}$ being the contact normal in an assembly, the magnitude of compression in the k -th contact can be applied for weighting, which

means that more significant emphasis is given to those contacts in the system which carry larger compression than others), and even on the same set of unit vectors, different weighted fabric tensors can be produced, this way carrying different physical meanings.

DEFINITION OF THE NEW VARIABLE; THE SURFACE ORIENTATION TENSOR

Consider a grain or stone represented by a polyhedron with faces 1, 2, ... k , ..., and denote the corresponding outwards unit normals $\mathbf{n}^{(1)}, \mathbf{n}^{(2)}, \dots, \mathbf{n}^{(k)}, \dots$. They belong to the faces whose areas are $A^{(1)}, A^{(2)}, \dots, A^{(k)}, \dots$ respectively. The surface orientation tensor is defined as follows:

$$\mathbf{f} = \frac{1}{\sum_k A^{(k)}} \sum_{(k)} (A^{(k)} \mathbf{n}^{(k)} \circ \mathbf{n}^{(k)}) \quad (2)$$

where the summations go along the faces of the grain with running index (k) .

This tensor has the following essential properties:

1. The tensor is symmetric. Consequently, its eigenvalues are real, and the eigenvectors are perpendicular to each other.
2. Its trace is equal to 1, and its three eigenvalues are nonnegative. Hence the three eigenvalues are between 0 and 1 so that they sum up to 1.
3. Any of the vectors $\mathbf{n}^{(k)}$ can be turned to its opposite without any effect on the resulting surface orientation tensor: because of the dyadic product of $\mathbf{n}^{(k)}$ with itself in the definition, the calculated \mathbf{f} will not change. Hence, in the computations, the surface normals do not have to point outwards.

Denote the three eigenvalues as $f_1 \geq f_2 \geq f_3$. The eigenvector belonging to f_1 is that orientation about which the stone mostly prefers to have contacts with its neighbourhood. For example, for a thin, flat, disc-shaped polyhedral plate, the eigenvector belonging to f_1 is the orientation of the normal vector of the two large faces (remember Property 3). For a regular polyhedron of isotropic shape (e.g. a cube), the three eigenvalues are equal. For a very long and thin column, the smallest eigenvalue, f_3 , is small, close to zero (belonging to the two distant, small closing faces of the column), and the two large eigenvalues are close to each other approaching 0.5.

Based on the three eigenvalues, the following geometrical quantities can be defined and then applied for the characterisation of individual stone shapes:

$$\text{Compactness: } C := \frac{f_3}{f_1} \quad (3)$$

$$\text{Flakiness: } F := \frac{f_1 - f_2}{f_1} \quad (4)$$

$$\text{Rodness: } R := \frac{f_2 - f_3}{f_1} \quad (5)$$

Note that this method of characterisation can very easily be used in post-processing the output of 3D surface imaging techniques, which produce a point cloud that represents the surface of the grain with a specific density and precision, and from this, build a triangular mesh with suitable software. The surface orientation tensor and its eigenvalues are easy to determine from this triangular mesh and then, based on the eigenvalues, the three characteristics compactness, flakiness and rodness can simply be calculated.

EXAMPLES

As an introductory example, consider a brick with its edges oriented according to the (x_1, x_2, x_3) coordinate axes, so that the sizes of this brick be $a \geq b \geq c$, respectively (Figure 2).

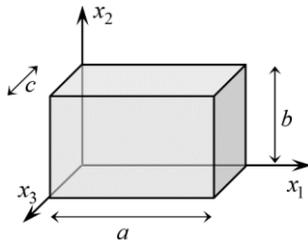


Figure 2: The Introductory Example: a Simple Brick

The areas and outward unit normals of the six faces are:

$$A_{right} = b \cdot c; \quad \mathbf{n}_{right} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad (6)$$

$$A_{left} = b \cdot c; \quad \mathbf{n}_{left} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}; \quad (7)$$

$$A_{top} = a \cdot c; \quad \mathbf{n}_{top} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \quad (8)$$

$$A_{bottom} = a \cdot c; \quad \mathbf{n}_{bottom} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}; \quad (9)$$

$$A_{front} = a \cdot b; \quad \mathbf{n}_{front} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}; \quad (10)$$

$$A_{left} = a \cdot b; \quad \mathbf{n}_{back} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}; \quad (11)$$

From these, the surface orientation tensor is:

$$\mathbf{f} = \frac{1}{a \cdot b + a \cdot c + b \cdot c} \begin{bmatrix} b \cdot c & 0 & 0 \\ 0 & a \cdot c & 0 \\ 0 & 0 & b \cdot c \end{bmatrix} \quad (12)$$

which is a diagonal matrix so the eigenvalues can immediately be seen in the main diagonal. From them, the compactness is $C = f_3 / f_1 = c / a$; the flakiness is $F = (f_1 - f_2) / f_1 = 1 - (c / b)$; and the rodness

$R = (f_2 - f_3) / f_1 = (a - b) \cdot c / a \cdot b = c / b - c / a$. A Python code was developed to analyse general geometries, which is able to process triangulated surface files in .stl format as its input and can compute the shape orientation tensor from this file. A few regular geometries were processed, which were created in a Computer-Aided Design (CAD) system to verify the code and the method. Every geometry was tested both with edges aligned parallel to the axes of the coordinate system and in a randomly chosen skew orientation. The geometries and their corresponding C, F, R values can be seen in Table 2. The verification was successful: even for the sphere being approximated with a finite number of triangular faces, the results provided by the code were in excellent agreement with the expectations, and the change of orientation had only a negligible effect.

Table 2: Regular Geometries in Verification Tests and their Shape Parameters

Geometry ([mm])	Orient.	C	F	R
Cube (10x10x10)	Aligned	1.000	0.000	0.000
	Random	1.000	0.000	0.000
Sphere ($\emptyset 10$)	Aligned	0.9995	0.0000	0.0005
	Random	0.9995	0.0000	0.0005
Rod (50x1x1)	Aligned	0.0200	0.0000	0.9800
	Random	0.0200	0.0000	0.9800
Plate (50x50x1)	Aligned	0.0200	0.9800	0.0000
	Random	0.0200	0.9800	0.0000

For performing a more practical test of the proposed method, three representative grains were selected, one for each type of shape (Figure 3), out of 44 andesite crushed rock particles (source: KőKa andesite quarry, Komló, Hungary). DAVID structured light scanner was applied for the imaging of the surface. The triangulation and the mesh optimisation was executed with the built-in post processor belonging to the scanner.

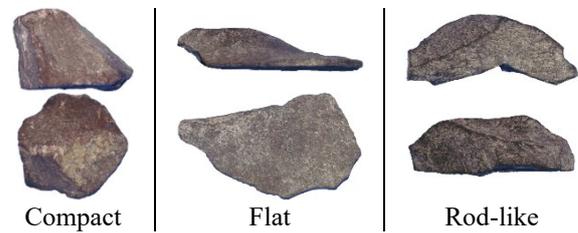


Figure 3: The Three Analysed Stones (Upper Images: Front View, Lower Images: Top View)

These grains were analysed manually as well. The three characteristic dimensions $a > b > c$ of the grains were measured with a calliper, and then their compactness, flakiness and rodness values were determined as:

$$C := \frac{c}{a}; \quad F := \frac{b - c}{a}; \quad R := \frac{a - b}{a} \quad (13)$$

These definitions are very similar to the widely used parameters of Sneed and Folk (Sneed and Folk, 1958), with the difference that they sum up to 1, similarly to those proposed parameters in Section 2 that are based on the fabric tensor. This way, these manually determined C , F and R parameters are directly comparable to those received from the surface orientation tensor, unlike the parameters by Sneed and Folk.

The authors note that while doing the calliper measurements, the subjective nature of measuring the necessary dimensions of the grains was experienced, and hence each dimension of each stone was measured five times, then averages were calculated for each dimension and each stone separately. The length “ a ” was understood in these measurements as the largest size that could be measured for the considered stone. Table 3 and Figure 4 present the results. The shape of a stone can be visualized as the location of a point on a triangular map, using the coordinates C , F and E in a similar manner as was done with other characteristics, e.g. by Sneed and Folk (Sneed and Folk, 1958).

Table 3: Characterisation of the Studied Grains with the Proposed Tensorial Method, and the Manual Method

Grain type	Method	C	F	R
Compact	Tensorial	0.9999	0.0000	0.0000
	Manual	0.6024	0.3383	0.0593
	Difference	0.3975	-0.3383	-0.0593
Flat	Tensorial	0.0710	0.7688	0.1602
	Manual	0.1945	0.4330	0.3725
	Difference	-0.1235	0.3358	-0.2123
Rod	Tensorial	0.2683	0.1032	0.6286
	Manual	0.3344	0.0942	0.5714
	Difference	-0.0661	0.009	0.0572

For all the three stones, the general conclusion can be drawn that the characterisation based on the surface orientation tensor was more sensitive. The results were more extreme in the sense that the three points belonging to the tensorial method were consequently closer to the corresponding vertices of the triangular map than those belonging to the manual calliper method.

This difference is particularly salient in case of the flat stone (middle one in Figure 3) which has large surfaces approximately with the same orientation, while in all other orientations the surfaces are small. The tensorial method gives a significant emphasis to the orientations of the large faces.

Regarding the compact stone (left one in Figure 3), the manual method was rather ambiguous to use: it was challenging to decide what the longest edge of the bounding box is. In addition, the significant difference between the outcome of the tensorial and the manual method can be understood if thinking of a regular cube, a perfectly compact shape according to the tensorial method. In the manual process, according to the practice,

it's main diagonal serves as its longest dimension, size “ a ”. Consequently, the regular cube would be characterised by the manual method as being rod-like to some extent, instead of perfectly compact. The fabric tensor approach, on the other hand, finds that none of the faces is larger than the others. Hence there is no bias in the orientational distribution of the faces, and the regular cube turns out to be perfectly compact. The left stone in Figure 3 is similar.

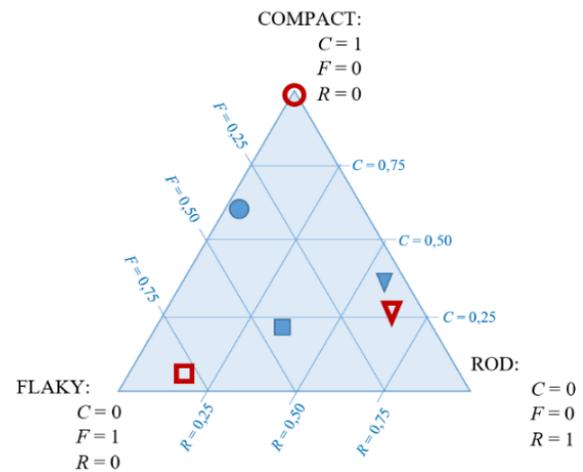


Figure 4: Shape Parameters of the Three Analysed Stones: Characterisation Based on the Surface Orientation Tensor: Hollow Circle \circ , Square \square and Triangle ∇ for the Compact, Flat and Rod Grains; Characterisation Based on the Overall Grain Shape according to Sneed and Folk (1958): Solid Circle \bullet , Square \blacksquare and Triangle \blacktriangledown for the Compact, Flat and Rod Grains Respectively

The elongated stone (right in Figure 3) has tiny faces around the two endpoints. These faces are around perpendicular to the longest axis, while the largest part of the surface is parallel with the longest axis. No wonder that the fabric tensor pointed out the rodness shape. The manual method also found this, and as shown in the triangular map, the two results were close to each other. An important difference between the two methods is their sensitivity to the existence of sharp, small corners. Such corners strongly influence the manual measurements because of the use of the calliper. However, these sharp corners are vulnerable from point of view of mechanics: they typically break for relatively small loads in comparison to the crushing load of the stone. Hence, these domains have low mechanical importance. An advantageous feature of the tensorial method is that since they have only a tiny contribution to the total surface, the existence of sharp corners hardly modify the surface orientation tensor, which is advantageous when the aim is to describe how the analysed grain may get into mechanical interaction with its neighbourhood.

GENERALIZATION TO SMOOTH SURFACES

The definition (Equation 2) can easily be modified to describe smooth surfaces in such a way that instead of a summation over the faces, an integration is done over the entire S surface of the stone:

$$\mathbf{f} = \frac{1}{\oint_{(S)} dS} \oint_{(S)} \mathbf{n} \circ \mathbf{n} dS \quad (14)$$

After determining the eigenvalues of this tensor the same C , F and R characteristics (compactness, flakiness and rodness) can be calculated for the characterisation of the stone as for the polyhedral stones.

Equation (14) is offered basically for theoretical purposes, since recent imaging techniques result in a discretized surface description for which (2) can directly be applied.

SUMMARY OF THE RESULTS

The paper defined the surface orientation tensor as the area-weighted fabric tensor built on the outwards normal vectors of the faces of the analysed polyhedral stone. A generalized version of this tensor was given for stones with a smooth surface. Based on the three eigenvalues of the tensor, the definition of three shape characteristics (the compactness, the flakiness and the rodness) were proposed to serve as the basis of shape classification. The proposed approach has the following advantages to the bounding box methods:

1. The characterisation is completely deterministic. Human subjectivity is excluded. Randomness can be due to the finite resolution of the imaging procedure only.
2. The characterisation is based on the possible orientations of the contacts of the stone with its neighbourhood, in such a way that larger emphasis is given to the larger faces on which contacts can occur with larger probability.
3. Sharp, small corners that may strongly affect the manual calliper analysis receive only minimal weight in the quantitative results.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI), and by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

REFERENCES

- Asahina, D. and M.A. Taylor. 2011. "Geometry of irregular particles: Direct surface measurements by 3-D laser scanner". *Powder Technol.* 213, 70–78.
- Bagi, K. 2007. "The discrete element method". Lecture Notes, Department of Structural Mechanics, Budapest University

- of Technology and Economics, Budapest. (electronic version in English updated continuously), in Hungarian, ISBN: 978-963-4209-29-4
- Cundall, P.A. and O.D.L. Strack. 1979. "A discrete numerical model for granular assemblies". *Géotechnique* 29, 47–65.
- Domokos, G.; F. Kun; A.Á. Sipos and T. Szabó. 2015. "Universality of fragment shapes". *Sci. Rep.* 5, 9147.
- Fischer, S.; B. Eller; Z. Kada and A. Németh. 2015. "Railway construction". Universitas-Győr Nonprofit Kft., Győr, HU. ISBN: 978-615-5298-69-1
- Guo, Y.; V. Markine; X. Zhang; W. Qiang and G. Jing. 2019. "Image analysis for morphology, rheology and degradation study of railway ballast: A review". *Transp. Geotech.* 18, 173–211.
- Juhasz, E. and S. Fischer. 2019. "Specific evaluation methodology of railway ballast particles' degradation". *Sci. Transp. Prog. Bull. Dnipropetrovsk Natl. Univ. Railw. Transp.* 81(3), 96–109.
- Satake, M. 1983. "Fundamental Quantities in the Graph Approach to Granular Materials", In *Studies in Applied Mechanics*, Satake, M. and J. Jenkins. (Eds.), Elsevier, pp. 9–19.
- Satake, M. 1982. "Fabric tensor in granular materials". In *Deformation and Failure of Granular Materials*, Luger, V.A. (Ed.), Presented at the International Union of Theoretical and Applied Mechanics Symposium, A. A. Balkema, Rotterdam, Ne, Delft, Netherlands, pp. 63–68.
- Shi, J. and P. Guo. 2018. "Induced fabric anisotropy of granular materials in biaxial tests along imposed strain paths". *Soils Found.* 58, 249–263.
- Sneed, E.D. and R.L. Folk. 1958. "Pebbles in the Lower Colorado River, Texas a Study in Particle Morphogenesis". *J. Geol.* 66, 114–150.
- Szabó, T. 2013. "A mechanics-based pebble shape classification system and the numerical simulation of the collective shape evolution of pebbles", PhD dissertation. Budapest University of Technology and Economics, Budapest, Hungary, 100.
- Wadell, H. 1932. "Volume, Shape, and Roundness of Rock Particles". *J. Geol.* 40, 443–451.
- Wensrich, C.M. and A. Katterfeld. 2012. "Rolling friction as a technique for modelling particle shape in DEM". *Powder Technol.* 217, 409–417.
- Zingg, T. 1935. "Beitrag zur Schotteranalyse". Schweizerische Mineralogische und Petrologische Mitteilungen 15, PhD dissertation. ETH Zürich, Zürich, Switzerland, 40-133.

AUTHOR BIOGRAPHIES

KATALIN BAGI is a professor at the Department of Structural Mechanics of Budapest University of Technology and Economics (kbagi@mail.bme.hu). She has three fields of expertise: (i) discrete element modelling; (ii) granular micromechanics; and (iii) statics of masonry vaults. More info can be found at researchgate.net/profile/Katalin_Bagi.



ÁKOS OROSZ is a PhD student at the Budapest University of Technology and Economics, Hungary where he received his MSc degree. His research topic is the DEM modelling of crushed stones His e-mail address is: orosz.akos@gt3.bme.hu and his web-page can be found at <https://gt3.bme.hu/>.

ANALYSIS OF THE STRESS STATE OF A RAILWAY SLEEPER USING COUPLED FEM-DEM SIMULATION

Ákos Orosz, Péter T. Zwierczyk
Department of Machine and Product Design
Budapest University of Technology and Economics
Műegyetem rkp. 3., H-1111, Budapest, Hungary
E-mail: orosz.akos@gt3.bme.hu

KEYWORDS

FEM-DEM, coupled analysis, railway ballast, railway sleeper, Yade, ANSYS

ABSTRACT

This paper deals with a coupled finite-discrete element simulation of a railway sleeper. The analysis aimed to show the trends of the stress state of a concrete sleeper during a conventional loading in the case of a crushed stone ballast bed. A typical discrete element analysis helps the engineers to analyze the behavior of the bulk material. The finite element analysis assists in examining the continuum materials. The railway tracks are complex systems, using both of the methods need to understand the interaction between the connected elements. The applied one-way coupled analysis highlighted the trends of the peak stresses on a concrete sleeper caused by the individual stones.

INTRODUCTION

Nowadays, the railway is one of the most popular ways of carrying goods and passengers. Besides the different component systems of the vehicles, the various parts of the railway track, along with their interaction with each other and the surrounding environment, are also in the focus of the researchers, as have been pointed out by Eller and Fischer (2019). The purpose of this paper is to analyze the stress state of sleepers, which are the interface between the rails and the crushed rock ballast as well as to find adequate protection layer.

Numerical simulations are frequently used to obtain data about processes that are hard or cost-demanding to perform experimentally. The finite element method (also known as FEM, Zienkiewicz 1971) gives a tool to the engineers to examine the vehicle and track systems (e.g. Németh et al. 2020) if they are treated as a continuum, while the discrete element method (DEM, Cundall and Strack 1979) helps to understand the role of individual grains on the mechanical behavior of the track system in the virtual space.

For modeling continuum materials, FEM is generally applied. The major fields of application are mechanical (static, dynamic, buckling, fatigue), thermodynamic, and electrodynamic processes. The FEM models consist of a finite number of elements that connect each other with common nodes for which have common

degrees-of-freedom (Zienkiewicz 1971). By contrast, however, the DEM models also consist of finite number of elements. These elements (particles) independent from each other, the DoFs are different on the neighboring elements (Bagi 2007). That is why the DEM is the better way to describe bulk materials while FEM for modeling continua.

Several different approaches exist in the literature for modeling the interaction of sleepers and ballast. If the role of individual grains is not important, FEM can be used alone, as Shahraki et al. (2015) and Paixão et al. (2016) did. The motion of grains can be captured by the application of DEM. The simplest element type for representing the stones is the circle or sphere (Irazábal et al. 2017). The effect of shape can be studied by using more complex element shapes, e.g. glued spheres with rigid (*clump*) or deformable/breakable (*cluster*) connection (Kono and Matsushima 2012; Gao et al. 2015; Khatibi et al. 2017; Laryea et al. 2014; Zhang et al. 2017; Jing et al. 2019; Juhasz et al. 2019). Polyhedra (Ferrellec et al. 2017; Huang and Tutumluer 2011) are even more accurate at the cost of high computational demand and more complicated contact detection.

There are cases when coupled discrete-continuum methods are needed for valuable results (Shao et al. 2017; Song et al. 2019; Shi et al. 2020b, 2020a). Using a coupled analysis, the stress peaks on the sleepers, caused by the stones in the ballast, can be examined, which can help to understand the crack initiation process on a concrete sleeper. This coupled method also allows the simulation of the effect of different under sleeper pads (between the sleeper and the ballast).

SIMULATION ENVIRONMENT

A typical structure of a trackbed, detailing the different layers, can be seen in Figure 1. One sleeper and the connecting ballast environment have been used during the simulation. The sleeper's geometry was created according to an LM-GEO type prestressed reinforced concrete sleeper's dimensions (Lábatlani Betonipari Zrt. 2020). The load on the sleeper was 225 kN static axle load that is the allowable maximum of the examined sleeper (Lábatlani Betonipari Zrt. 2020). The prestressed state of the sleeper was neglected during the simulation. The geometry of the sleeper can be seen in Figure 2.

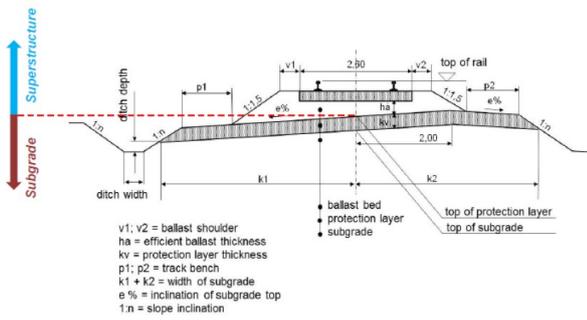


Figure 1: Structure of Trackbed (Fischer et al. 2015)

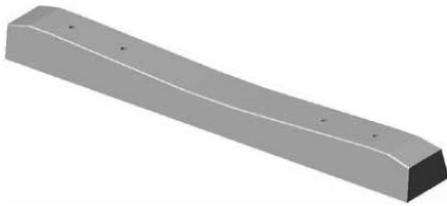


Figure 2: LM-GEO Concrete Sleeper (Lábatlani Betonipari Zrt., 2020)

DEM Model

A 600 mm wide section of the ballast was involved in the model. The boundaries of the model domain were made of *facet* elements (triangle elements with zero thickness). The stones were modeled by polyhedral elements, using Yade (Šmilauer et al. 2016) and the built-in volumetric contact model (Eliáš 2014). Firstly, loose packing was generated with the aid of the Voronoi method (Asahina and Bolander 2011). The size of the elements was set to be between 32 and 50 mm. The assembly consisted 90% compact (size ratio: 2:2:1) and 10% flat (size ratio: 2:1:0.5) elements. The elements fell under the influence of gravity to obtain a dense packing than the shape of the ballast was obtained by deleting sparse elements, including the space for the sleeper. The sleeper was also modeled by facets. It moved downwards with a constant speed until reaching the desired summed load on the sleeper when the acting force on each facet was exported. Falling particles were deleted sequentially. The geometry, after the finishing of the loading process, can be seen in Figure 3.

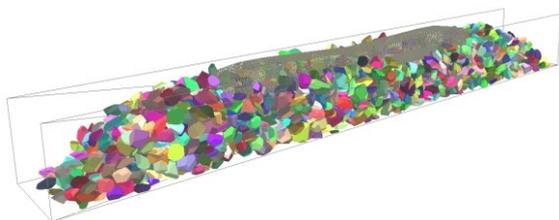


Figure 3: Geometry of the DEM Model

The resultant geometries consisted of 2000-2500 elements, which is lower than in a real ballast domain of the same size. However, it was enough for the authors'

purposes: to obtain qualitative results and to test the methodology. Approximately 20-30 grains came into interaction with the sleeper under it, as can be seen in Figure 4 b). The purple (dark gray in greyscale) lines represent polyhedron-polyhedron contacts, and the green (light gray in greyscale) lines show the feasible polyhedron-facet interactions.

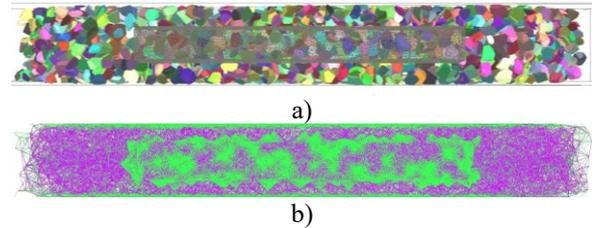


Figure 4: Geometry (a) and Interactions (b) of the DEM Model from Above

The applied model parameters were taken from literature (Eliáš 2014) and can be seen in Table 1. In that paper, the model was calibrated based on oedometric test, which applies similar type of load as in the current study.

Table 1: Micromechanical Parameters of the DEM Model (ρ : density, k_n : volumetric normal stiffness, k_s : shear stiffness, φ : sliding friction angle between elements)

	Stones	Sleeper and walls	Unit
ρ	2600	7800	kg/m ³
k_n	$2 \cdot 10^{13}$	$2 \cdot 10^{14}$	N/m ³
k_s	$2 \cdot 10^8$	$2 \cdot 10^9$	N/m
φ	0.6	0.4	rad

FEM Model

The finite element analysis created in ANSYS Workbench 2020 R1 environment (ANSYS, Inc. 2020). During the simulation 10-nodes, quadratic tetrahedron elements were applied, which mesh arrangement was the same, used during the DEM simulation. The structure of the sleeper's FEM mesh can be seen in Figure 5.

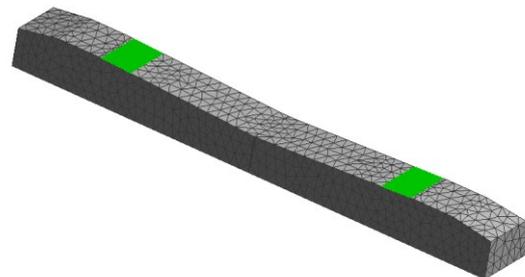


Figure 5: Structure of the Finite Element Mesh Used During the FE Analysis. The Place of the Fix Boundary Conditions Highlighted on the Geometry

Fix boundary conditions were applied on the surfaces where the rails are connecting to the sleeper, neglecting the under sleeper pads (Figure 5). This approach was acceptable because of the acting force came from the DEM simulation on the bottom side of the sleeper. During the simplified simulation, uniform material properties were used. The initial elasticity modulus of the concrete was 37 800 MPa, and the Poisson ratio was 0.2 through the analysis (MSZ 15022-1:1986).

The connection between FEM and DEM

A previously published method (Orosz et al. 2018) was used to establish a one-way connection between Yade and ANSYS. The arising forces on the center of mass of the triangular facets were saved in the appropriate time (when the summarized force on the sleeper exceeded the desired load in more than 100 consecutive timesteps), and were saved in a .csv file with a special format, containing their magnitude (according to components) and coordinates of application. To reduce the effect of numerical errors, the simulation continued for 100 additional timesteps after reaching the maximum force criterion, and the value of force components were averaged over this time, as it was also done in (Vajda et al. 2019). This file was imported into ANSYS after the creation of the proper geometry. The forces were interpolated onto the FE nodes with the so-called “mapping” technique, which is implemented into ANSYS.

RESULTS

The result of the interaction between the andesite ballast and the concrete railway sleeper can be seen in Figure 6. The figure shows that, because of the non-uniform distribution and shape of the trackbed stones, the surface pressure on the bottom of the sleeper is also non-uniform. However, the value of the average pressure is very small, which matches with the requirements of the sleeper-ballast connection (Sysyn et. al. 2019; Figure 7), but there are some peaks which indicate that the trackbed stones overload the rigid concrete in some small surfaces because of its sharp shape. A highlighted overloaded region can be seen in Figure 8. These peaks can help of the crack initiation process on the surface of the railway sleeper, which results damage in long term period.

In reality, contrarily to the simulation, the corners of the crushed stones break off or even the entire grain can split into multiple pieces (Selig and Waters 1994). That phenomenon reduces the magnitude of peak stresses and increases the number and area of sleeper-stone contacts. Therefore, the results of the simulation can be improved by applying a proper stone breakage model.

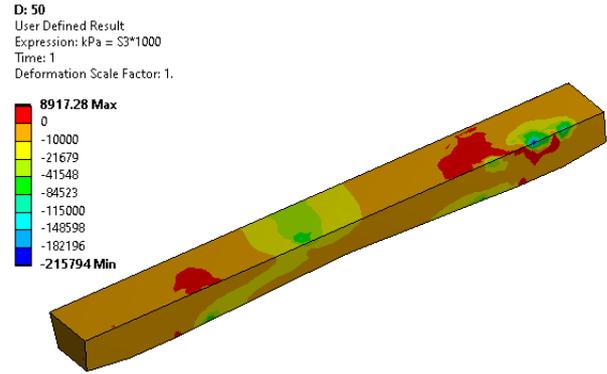


Figure 6: Pressure distribution on the bottom of the examined sleeper (Deformation scale 1:1; unit: kPa)

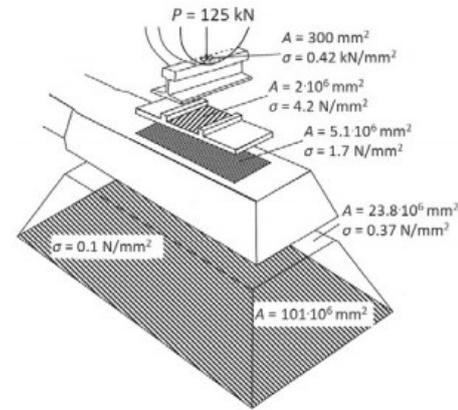


Figure 7: Schematic explanation of the stress distribution of a complete wheel-rail-track connection (Sysyn et. al. 2019, modified after Führer 1978)

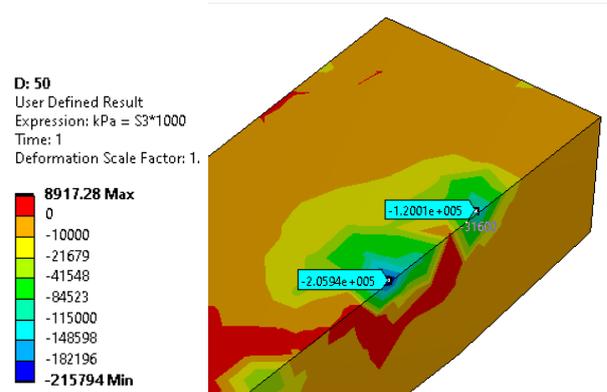


Figure 8: Non-uniform pressure peaks on the bottom of the concrete sleepers (Deformation scale 1:1; unit: kPa)

SUMMARY AND CONCLUSIONS

The examination of the stress field of the railway sleeper due to its interaction with the ballast grains requires a connection between continuum and discrete simulation methods. A simple method was published previously to connect the open-source Yade DEM framework with the commercial ANSYS FEM software without 3rd party extensions. This one-way connection

method has created the possibility to take a closer look at a concrete sleeper and ballast connection. The results highlighted that the pressure distribution is non-uniform on the bottom of the sleeper because of the distinct crushed stones. These pressure peaks might initiate cracks on the surface of the sleeper, or can cause the deterioration of the ballast grains. The elaborated model also provides further opportunity to examine the effect of the different under sleeper pads, which reduce noise emission and produce a more uniform pressure distribution on the sleeper. The described results show trends of the interaction and focused on the modeling technique. Further developments and validation tests are needed to be able to obtain reliable, quantitative results. Both in FEM and DEM simulations, the constitutive models and their parameter values need calibration, as well as the effect of mesh resolution, have to be studied in sensitivity tests. Furthermore, the size and shape distribution of the elements, the value of applied timestep, the force criterion, and the length of force averaging has to be carefully validated in the DEM model and a stone breakage has to be included. Nevertheless, the study proved the applicability of the developed method.

ACKNOWLEDGEMENT

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI), and by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

REFERENCES

- ANSYS, Inc. 2020. “ANSYS 2020R1 Program Help”, Canonsburg, PA, USA
- Asahina, D. and J.E. Bolander. 2011. “Voronoi-based discretizations for fracture analysis of particulate materials”. *Powder Technology* 213., 92–99. Bagi, K. 2007. *The discrete element method*. Lecture Notes, Department of Structural Mechanics, Budapest University of Technology and Economics, Budapest. (electronic version in English updated continuously), in Hungarian, ISBN: 978-963-4209-29-4
- Cundall, P.A. and O.D.L. Strack. 1979. “A discrete numerical model for granular assemblies”. *Géotechnique* 29, 47–65.
- Eliáš, J. 2014. “Simulation of railway ballast using crushable polyhedral particles”. *Powder Technol.* 264, 458–465.
- Eller, B. and Sz. Fischer. 2019. “Review of the modern ballasted railway tracks’ substructure and further investigations”. *Sci. Transp. Prog. Bull. Dnipropetrovsk Natl. Univ. Railw. Transp.* 84., 72-85.
- Ferrellec, J.-F.; R. Perales; V.-H. Nhu; M. Wone; and G. Saussine. 2017. “Analysis of compaction of railway ballast by different maintenance methods using DEM”. *EPJ Web Conf.* 140, 15032.
- Fischer, S.; B. Eller; Z. Kada and A. Németh. 2015. *Railway construction*. Universitas-Győr Nonprofit Kft., Győr, HU. ISBN: 978-615-5298-69-1
- Führer, G. 1978. *Oberbauberechnung*. VEB, Verlag für Verkehrswesen, Berlin. 151 pp. (in German).
- Gao, L.; Q. Luo; Y. Xu; G. Jing; and H. Jiang. 2015. “Discrete element method of improved performance of railway ballast bed using elastic sleeper”. *J. Cent. South Univ.* 22, 3223–3231.
- Huang, H. and E. Tutumluer. 2011. “Discrete Element Modeling for fouled railroad ballast”. *Constr. Build. Mater.* 25, 3306–3312.
- Irazábal, J.; F. Salazar and E. Oñate. 2017. “Numerical modelling of granular materials with spherical discrete particles and the bounded rolling friction model. Application to railway ballast”. *Comput. Geotech.* 85, 220–229.
- Jing, G.; P. Aela; H. Fu and H. Yin. 2019. “Numerical and experimental analysis of single tie push tests on different shapes of concrete sleepers in ballasted tracks”. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* 233, 666–677.
- Juhász, E., R.M. Movahedi; I. Fekete and Sz. Fischer. 2019. “Discrete element modelling of particle degradation of railway ballast material with PFC3D software”. *Sci. Transp. Prog. Bull. Dnipropetrovsk Natl. Univ. Railw. Transp.* 84, 103-116.
- Khatibi, F.; M. Esmaceli and S. Mohammadzadeh. 2017. “DEM analysis of railway track lateral resistance”. *Soils Found.* 57, 587–602.
- Kono, A. and T. Matsushima. 2012. “3D-DEM simulation for shaking table test of ballasted test track”. In: *Advances in Transportation Geotechnics II - Proceedings of the 2nd International Conference on Transportation Geotechnics, ICTG 2012.*, (Hokkaido, JP, Sep. 10-12). CRC Press, pp. 716–721.
- Lábatlani Betonipari Zrt. *Concrete sleepers and other concrete elements for railway construction – Product catalog* – date of download: 02/02/2020.
- Laryea, S.; M. Safari Baghsorkhi; J.-F. Ferrellec; G.R. McDowell; C. Chen. 2014. “Comparison of performance of concrete and steel sleepers using experimental and discrete element methods”. *Transp. Geotech.* 1, 225–240.
- MSZ 15022-1:1986, *Építmények teherhordó szerkezeteinek erőtani tervezése. 1. rész: Vasbeton szerkezetek* - standard, Magyar Szabványügyi Testület, Budapest
- Németh, A.; Z. Major and S. Fischer. 2020. “FEM Modelling Possibilities of Glued Insulated Rail Joints for CWR Tracks”. *Acta Tech. Jaurinensis.* 13, 42–84.
- Orosz, A.; K. Tamas; J.P. Radics and P.T. Zwierczyk. 2018. “Coupling finite and discrete element methods using an open source and a commercial software”, In: *ECMS 2018 Proceedings* (Wilhelmshaven, GER. May 22-25.) ECMS pp. 399–404.
- Paixão, A.; J.N. Varandas; E. Fortunato and R. Calçada. 2016. “Non-Linear Behaviour of Geomaterials in Railway Tracks under Different Loading Conditions”. *Procedia Eng.* 143, 1128–1135.
- Selig, E.T. and J.M Waters. 1994. *Track geotechnology and substructure management*. Thomas Telford Publishing. London, UK. ISBN: 978-0-7277-4982-6
- Shahraki, M.; C. Warnakulasooriya and K.J. Witt. 2015. “Numerical study of transition zone between ballasted and ballastless railway track”. *Transp. Geotech.* 3, 58–67.
- Shao, S.; Y. Yan and S. Ji. 2017. “Combined Discrete-Finite Element Modeling of Ballasted Railway Track Under Cyclic Loading”. *Int. J. Comput. Methods* 14, 1750047.
- Shi, C., C. Zhao; X. Zhang and A. Andersson. 2020a. “Analysis on dynamic performance of different track transition forms using the discrete element/finite difference hybrid method”. *Comput. Struct.* 230, 106187.

- Shi, C.; C. Zhao, X. Zhang and Y. Guo. 2020b. "Coupled discrete-continuum approach for railway ballast track and subgrade macro-meso analysis". *Int. J. Pavement Eng.* 1–16.
- Šmilauer, V.; E. Catalano; B. Chareyre; D. Sergej; J. Duriez; A. Gladky; J. Kozicki; C. Modenese; L. Scholtès; L. Sibille; J. Stránský and K. Thoeni. 2016. *Yade Documentation, 2nd ed.* The Yade Project.
- Song, W.; B. Huang; X. Shu; J. Stránský and H. Wu. 2019. "Interaction between Railroad Ballast and Sleeper: A DEM-FEM Approach". *Int. J. Geomech.* 19, 04019030.
- Sysyn M.; V. Kovalchuk; O. Nabochenko; Y. Kovalchuk and O. Voznyak. 2019. "Experimental study of railway trackbed pressure distribution under dynamic loading". *The Baltic Journal of Road and Bridge Engineering*, 14(4), 504-520.
- Vajda, M.Zs.; Z. Olah and A. Orosz. 2019. "Evaluating The Stress Field On Sweep During Tillage Process Applying Coupled Finite-Discrete Element Method". In: ECMS 2019 Proceedings (Caserta, IT. June 11-14.) ECMS 358–363.
- Zhang, X.; C. Zhao and W. Zhai. 2017. "Dynamic Behavior Analysis of High-Speed Railway Ballast under Moving Vehicle Loads Using Discrete Element Method". *Int. J. Geomech.* 17, 04016157.
- Zienkiewicz, O.C. 1971. *The Finite Element Method in engineering Science*. McGraw Hill, New York

AUTHOR BIOGRAPHIES



ÁKOS OROSZ is a Ph.D. student at the Budapest University of Technology and Economics, Hungary, where he received his MSc degree. His research topic is the DEM modeling of crushed stones. He is also a member of a research group in the field of discrete element modeling. His e-mail address is: orosz.akos@gt3.bme.hu and his web-page can be found at <http://gt3.bme.hu>.



PÉTER T. ZWIERYCZYK 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 modeling (FEM) research group. His e-mail address is: z.peter@gt3.bme.hu and his web-page can be found at <http://gt3.bme.hu>.

A VCCT APPROACH OF CRACK PROPAGATION IN RAILWAY WHEELS

Tamás Máté

Péter T. Zwierczyk

Department of Machine and Product Design
Budapest University of Technology and Economics
Műegyetem rkp. 3., H-1111, Budapest, Hungary
E-mail: mate.tamas@gt3.bme.hu

KEYWORDS

Railway wheel, RCF - Rolling Contact Fatigue, Crack propagation, Thermal cracks, FEM - Finite Element Method, Ansys, VCCT – Virtual Crack Closure Technique

ABSTRACT

In this paper, Rolling Contact Fatigue (RCF) crack propagation in the case of a railway wheel was studied in the presence of significant thermal loading. The complexity of the phenomena and the several assumptions and boundaries of the existing crack propagation modeling methods induce particular difficulties in the creation of this specific kind of contact problem. The primary purpose of the investigation was to reveal the relevancy of the Virtual Crack Closure Technique (VCCT) to see the further implementation opportunities and capabilities of the technique in solving railway RCF based problems.

INTRODUCTION

During the operation of railway vehicles, several reasons can cause wheel and rail failures, which can have a significant effect on passenger comfort and, in a worse case, on the safe operation. Consequently, continuous monitoring is required to be performed to reveal the potential source of failures. Many principles of maintenance rules are in daily use, but none of them are proved scientifically. These, rather practically defined norms, can result in inaccurate timing and also in the unnecessary scale of the maintenance, which significantly increases the costs. Furthermore, latency also can occur in the required maintenance, which can result in more severe damages that influence the operation and cause delays so indirectly affect the costs.

This investigation is a part of my research which goal is to develop a finite element crack propagation model that is able to model specific failure forms in well-defined circumstances to provide more accurate information in order to specify maintenance instructions.

In this investigation, the Virtual Crack Closure Technique (VCCT) (Krueger et al., 2013; Pironi et al., 2015) is scoped and studied to reveal the relevancy in modeling such a complex phenomenon as the contact problem between the rail and the wheel.

The study supposed an intensive braking situation when the stick-slip phenomenon could occur, which causes undesirable high thermal loads on the wheel tread. In the case of those vehicles that are equipped with the Wheel Slide Protection system (WSP), more severe temperatures can arise in some specific hot-spots. This thermal load makes to expand the wheel surface and the thin inner volume under the surface as well (Zwierczyk and Váradi, 2014). The heat expansion and the following rapid cooling make destructive residual stresses in the material, which raises the complexity of the stress situation and the understanding of the crack propagation.

VIRTUAL CRACK CLOSURE TECHNIQUE

The VCCT is based on the assumption that the energy needed to separate a surface is the same as the energy needed to close the same surface. It was initially developed to calculate the energy release rate of a cracked body. Since then, it is widely used in case of investigating interfacial crack-growth or delamination. This method uses interface elements to simulate the fracture by separating the interface elements along a predefined path according to one or more user-specified fracture criteria, for example, the critical energy release rate. (“VCCT-Based Crack-Growth - ANSYS,” Ansys - Help)

Advantages:

- Several fracture criteria are available, including a user-defined option.
- Multiple cracks can be defined in an analysis.
- The crack can be located in the material or along with the interface of the two materials.

Assumptions:

- Crack growth occurs along a predefined crack path.
- The path is defined via interface elements.

- The analysis is quasi-static and does not account for transient effects.
- The material is linear elastic and can be isotropic, orthotropic, or anisotropic.
- Heat loading cannot be defined.

THE FE MODEL

To implement the contact problem with the VCCT method, we had to create a modeling procedure to reveal the difficulties of the investigation process. Since the technique is not able to deal with the direct heat loads in the model, an indirect way was used to include the effect of the braking process. In the first step of the procedure, a coupled transient thermal - stress analysis was performed to calculate the deformation, which is caused by the occurring heat stresses. In the second step, this additional deformation is used as the indirect heat load input of the submodel that was joined with the VCCT crack propagation method. In order to meet the fundamental drawbacks and difficulties of the method and to keep the need for the calculation capacity low at the beginning, we started the investigation with 2D models. These models need further assumptions, and a problem like RCF cannot be appropriately investigated in this way, but on the other hand, 2D models can reveal various problems which, if we are aware of the set-up time of the 3D models can be decreased drastically. The structure of the investigation process can be seen in (Figure 1). In this article, only the 3D analyses are detailed with the result of the VCCT submodel. The simulation was performed in Ansys 18.2 Workbench.

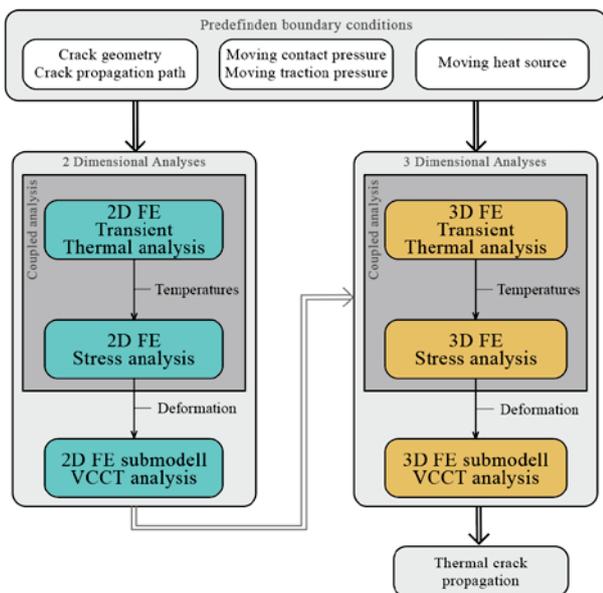


Figure 1 Schematic structure of the modeling procedure

Crack and Path as Boundary Condition

To examine crack propagation, the initial geometry of the crack has to be placed in the model. Furthermore, to perform a VCCT study, the crack propagation path is also needed to be defined.

Thermal cracks initiate from the surface and are oriented perpendicularly on the wheel thread (Figure 2). As it is an RCF phenomenon, not only the thermal loading, but the rolling contact pressure and traction pressure are also needed to develop the cracks. From the experiment of Handa et al., (Handa et al., 2010), it can be clearly seen that thermal cracks only develop within the rail-wheel contact width (Figure 2).

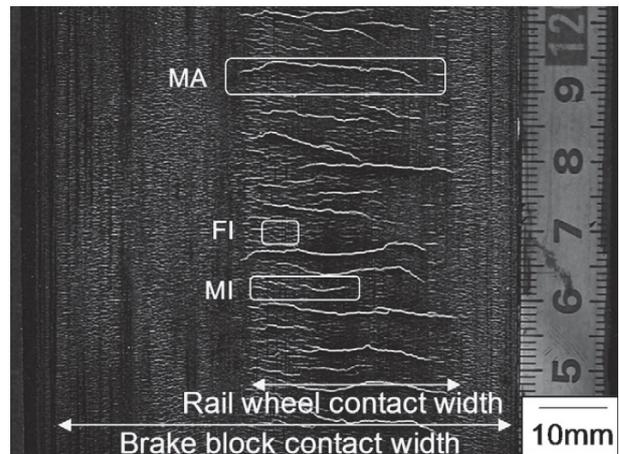


Figure 2 The distribution of thermal cracks on the wheel tread (Handa et al., 2010)

Thermal cracks propagate from the surface of the wheel nearly radial direction, then around in 3-5 mm depth (near the maximum shear stress zone) deviate (and maybe branch) and continue to propagate in the (almost) circumferential direction (Handa et al., 2010). When the cracks meet under the surface, the phenomenon of pitting can occur. Figure 3 presents different life states of thermal cracks that can be seen in the section view.

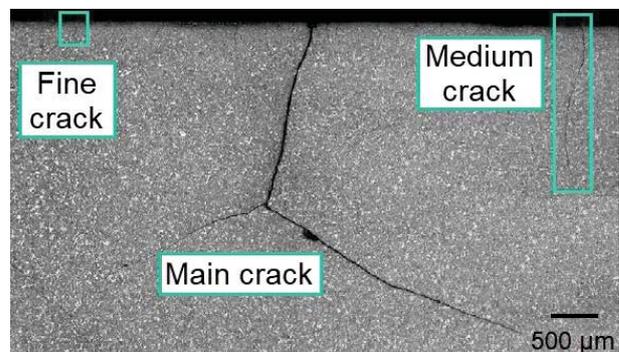


Figure 3 Different life states of thermal cracks from a longitudinal cross-section view (Handa et al., 2010)

As the boundary condition for the simulation a semi-elliptical crack was used, with the semi-axes of 0.5mm and 2 mm, placed in the middle of the rail-wheel contact width in a perpendicular position. The crack path is assumed to start in a radial direction and then deviating into the circumferential direction in the depth of 4 mm. In this study, the first radial direction of the path was investigated.

Operational Factors

In the investigation, the wheel of a passenger car was studied during the braking process. The operational factors of the examined car are the following.

Table 1 Operational factors of the train

Train speed	v	120 km/h
Wheel slip	s	15%
Wheel sliding speed	Δv	18 km/h
Traction coefficient	μ	0.15
Vehicle weight	m	51 t

In a railroad wheel, the contact stress distribution is very complex and depends heavily on the friction forces between the two contacting surfaces and on the applied tangential forces.

In the study, the contact loading conditions of a 4-axle passenger carriage car (510 kN) were used, which were previously calculated and validated by Zwierczyk (Zwierczyk, 2015). The vertical load per wheel is $F = 63,750$ N. The contact pressure distributions - $p(s_t; s_p)$ on the wheel tread can be seen in Figure 7. The maximum pressure value is 1011 MPa. The size of the semi-axes of the ellipsoidal contact patch is 10 mm and 12.6 mm.

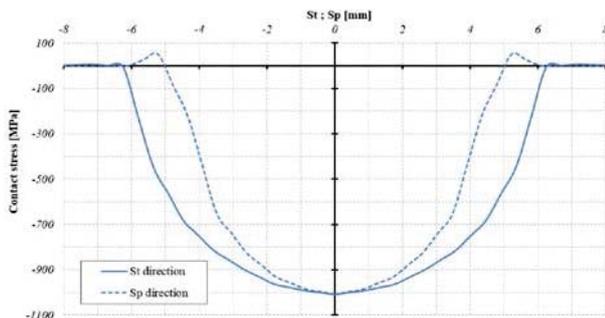


Figure 4 Contact stress distribution on the wheel tread – St - tangential; Sp - perpendicular (Zwierczyk, 2015)

The Geometry of the Examined Wheel and the Predefined Initial Crack

In the investigation, a 920 mm railway wheel was used with simplifications as the flange of the wheel rim, and the conical shape of the tread was neglected. The investigation focuses only on the near vicinity of the

crack, so a smaller 20° piece (Figure 5) of the wheel was examined during the simulations.

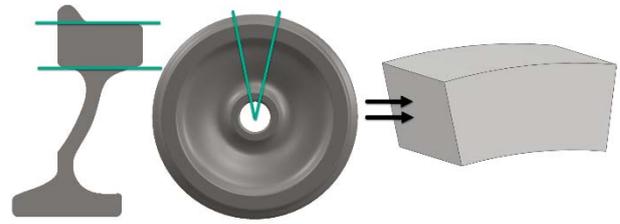


Figure 5 The major simplifications of the wheel

The simplified 20° piece was sliced up to more sections in order to ensure the proper sectors for the meshing (Figure 6) and to apply the loads and boundary conditions.

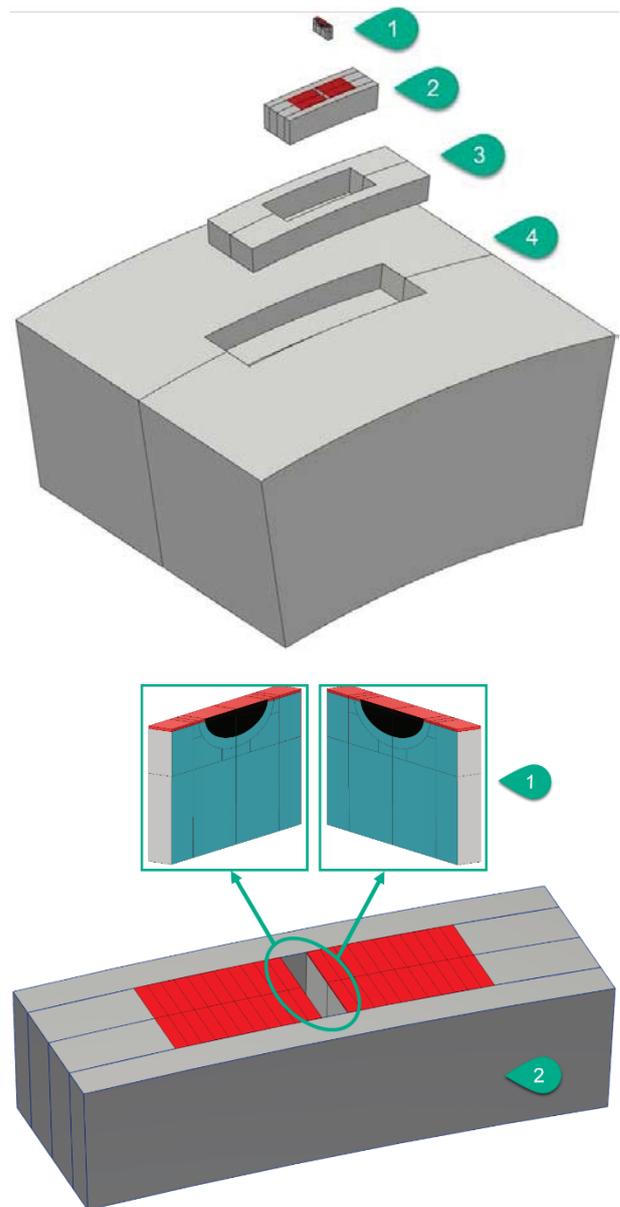


Figure 6 The sectioned wheel geometry is representing the four zones. The loading zone (red)

and the pre-defined semi-elliptical thermal crack (black) in the Crack zone

Four sections were defined according to the distance from the crack:

- 1 - *Crack Zone*: where the crack (black) and the path (turquoise) is placed.
- 2 - *Pressure Zone (red)*: where the mechanical and thermal loading is applied.
- 3 - *Further Vicinity*: transition to the further parts of the wheel.
- 4 - *Wheel Zone*: further areas of the wheel.

Loads and Boundary Conditions

The loading condition consists of three main elements:

- rolling contact pressure from the weight of the vehicle,
- tangential traction pressure caused by the sliding,
- thermal load also caused by the wheel sliding.

The Hertzian contact between the rail and wheel assumes the elliptical contact area, but to make the implementation more straightforward, the shape of the contact patch was considered to be rectangular in the model. The area of the simplified patch matches the size of the elliptical one (Figure 7) (Zwierczyk, 2015). To be able to discretize the loading on the patch, it was divided into 12 sections by 1 mm (Figure 8). A further simplification in the model is that the rectangular patch only assumes the tangential pressure distribution, so the evaluation is based on the tangential distribution of the loadings.

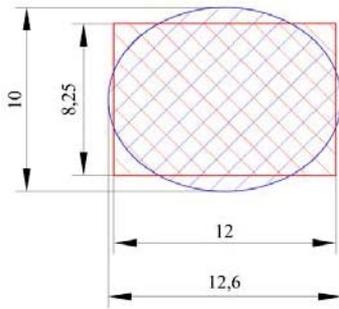


Figure 7 The simplification of the contact patch (Zwierczyk, 2015)

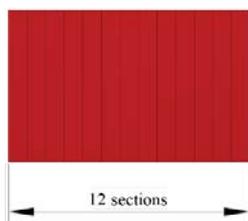


Figure 8 Sectional contact patch

The defined loads in the study were the followings:

- the contact pressure (Figure 9 / red) by Zwierczyk (Zwierczyk, 2015) in discretized form,
- the traction pressure calculated from the contact pressure (Figure 9 / blue),

$$p_t(s_t; s_p) = \mu \cdot p(s_t; s_p) \quad (1)$$

- the discretized heat flux (Figure 9 / orange), which was based on the friction that is caused by the speed difference between the rail and the wheel during the braking condition,

$$q = \frac{Q}{A} = \frac{\mu \cdot \Delta v \cdot F}{A} \quad (2)$$

- and the cooling effect of the air as thermal conduction all around on the surface of the wheel.

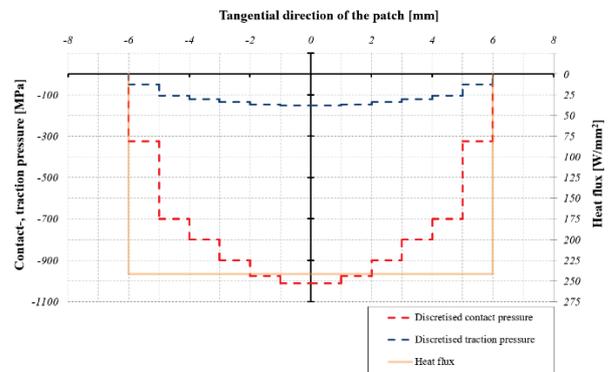


Figure 9 The discretized loading on the contact patch and their response pressure distribution

All the loads, except the cooling effect, which was constant during the study, were defined on the moving patch. The patch starts from exactly before the crack (Figure 10 / green), and in the last step, it just passes. The contact patch moves 1 mm in every load cycle. The time steps were set according to the relative speed between the rail and the wheel.

To complete the model, a fixed constraint was placed on the lower edge of the geometry (Figure 10).

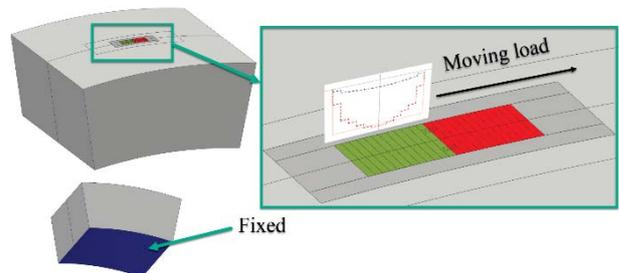


Figure 10 The applied moving loads and the fixed constraint on the bottom

Material properties

During the analysis, linear elastic material model was used. To model the wheel, the SSW-3QS wheel steel was supposed. The mechanical and thermal material properties are listed in Table 2. Also, the temperature-dependent thermal expansion coefficient was taken into account.

Table 2 Material properties of SSW-3QS steel

Poisson ratio	0.3
Young's modulus	206,000 MPa
Density	7,850 kg/m ³
Thermal conductivity	54 W/mK
Specific heat	465 J/kgK

Contacts in the FE model

All the segments of the geometry (Figure 6) were connected with bonded contacts, with Multi-Point Constraint (MPC) contact formulation theory, except the semi-elliptical thermal crack. Between the crack faces, frictionless contact was defined with Augmented-Lagrange contact formulation theory and increased stiffness to avoid the significant penetration of the surfaces into each other.

The mesh of the FE model

The mesh size changes in every sector of the geometry. In order to focus on the calculation capacity, the further areas from the crack meshed with less density, but in the crack vicinity, finer and finer mesh was defined. To match the requirements of the VCCT crack propagation method, linear hexagonal 8-nodes elements were used to mesh the geometry.

The Result of the Transient Thermal Study

The surface reached its' highest temperature in that specific load steps and locations where the whole length of the load passed. The temperature rise was only significant in close surface range in 0.1-0.2 mm deepness. The highest temperature zone was moving together with the load, but it was late by 6 segments compared to the magnitude of the mechanical load. The peak temperature was ~380 °C. After the load passed, the heat was dissipating to the air and to the deeper parts of the wheel. The temperature distribution can be seen in Figure 11.

The VCCT submodel

In the sub-simulation, the aim was to combine the results of the previous coupled study with the VCCT crack propagation method. Since the method cannot deal with direct heat loads, we imported the deformation of the coupled thermal stress study and used it as a heat representation input for the model. The sub-geometry contained the Pressure zone and the Crack zone (Figure 6), which were derived from the

master model. The investigation was restricted to one load cycle because of calculation capacity restrictions. Therefore, only one load cycle was examined.

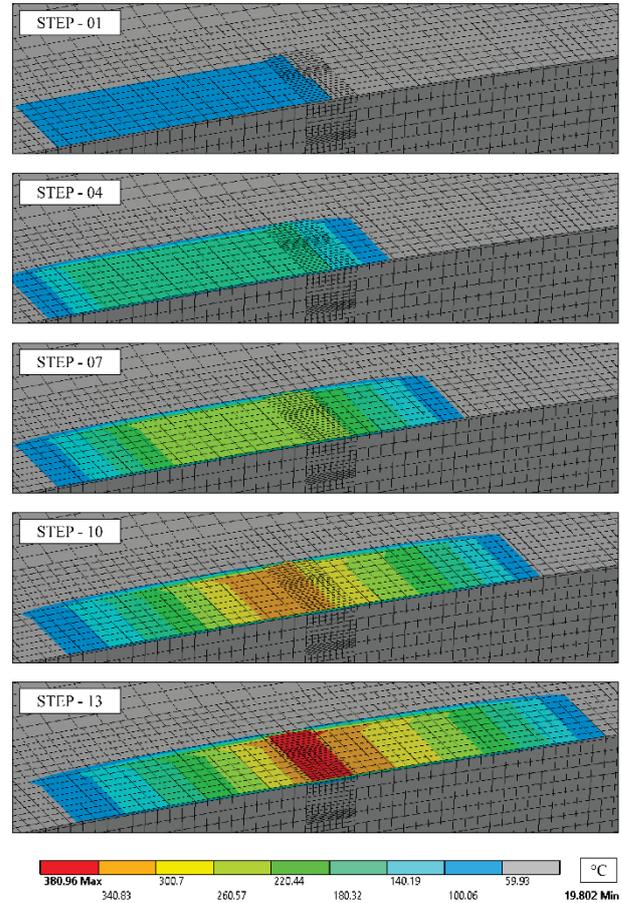


Figure 11 Temperature distribution during the passing of the load

To force the crack to propagate in already one cycle, a different Energy Release Rate had to be defined to move the propagation forward and show the behavior of the crack. The value of the Energy Release Rate was set to 100 J/m² in all three dimensions.

To set up the VCCT study, the crack path had to be defined. Below the crack, a 4 mm deep surface was set with interface elements that can separate in case of enough energy accumulation. Furthermore, a frictionless contact was also defined between the surfaces to avoid penetration after the moving of the crack front (Figure 12).

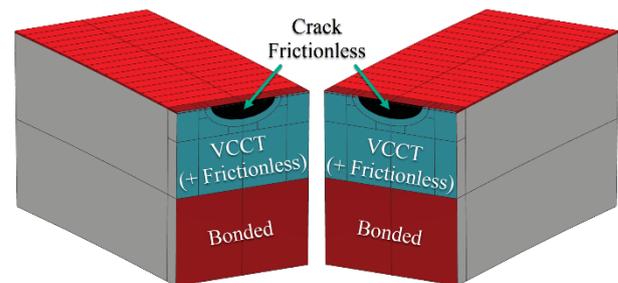


Figure 12 The delamination zone as the crack path and the bonded zone in the plane of the crack

RESULTS - CONCLUSION

The stress results are highlighted during the load pass in every time step during the load movement because the crack also propagates during the first time steps as well. The results are displayed by half geometry, split on the symmetry plane to get insight into the under-surface state of deformation and stress. The stress distribution can be seen in Figure 13.

It can clearly be seen that the crack started to propagate when the magnitude of the load approached the crack-zone, and when it was right there, the propagation stopped. When the load moved forwards and left the Crack zone, the propagation started again. From the results, the crack propagation length ratio between the two phases could be estimated. The first propagation phase from STEP-02 to STEP-05 and the second phase from STEP-09 to STEP-13 in crack growth length is proportional to each other as 1:2. The first phase is mainly driven by Mode II. shear propagation, while the second phase is rather a mixed-mode (Mode I. – Mode II.) propagation. Furthermore, the second phase propagation is longer in time because of the crack-closing effect of the braking traction pressure.

The crack does not propagate in the axial direction only in the radial. This fact is a failure in the model because the crack meant to spread in the width of the contact pressure.

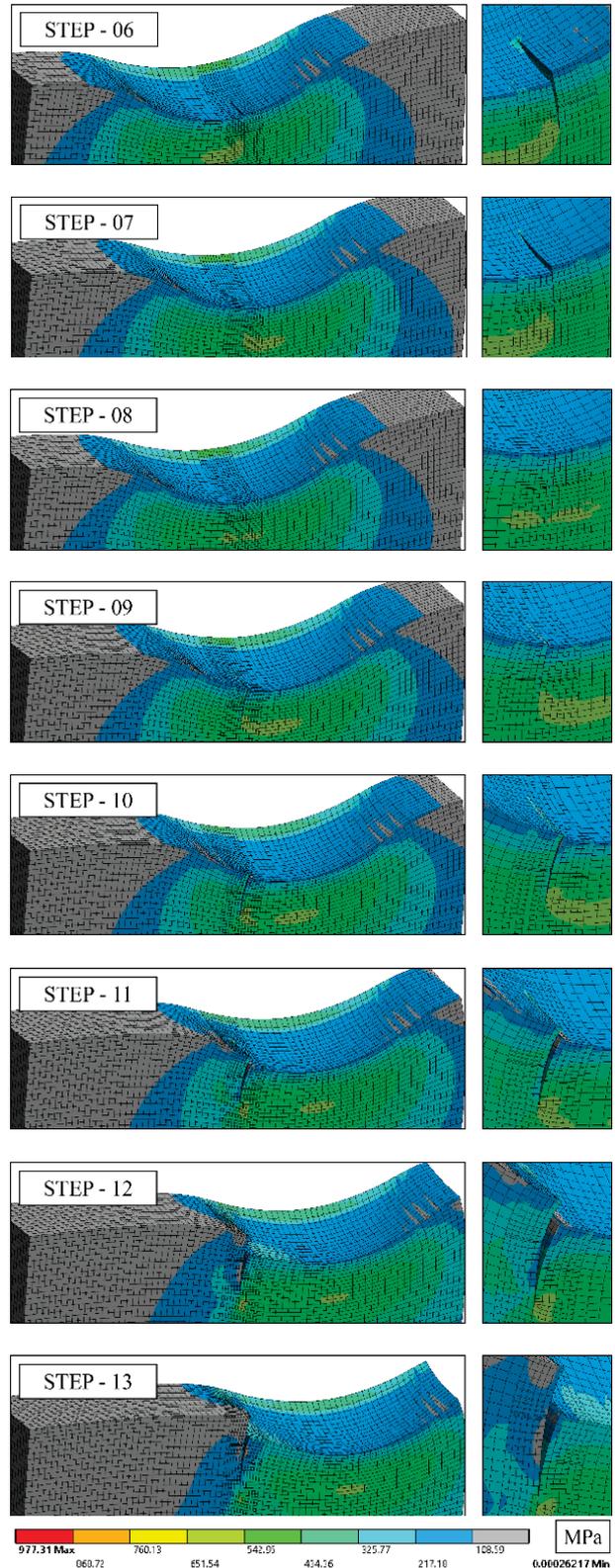
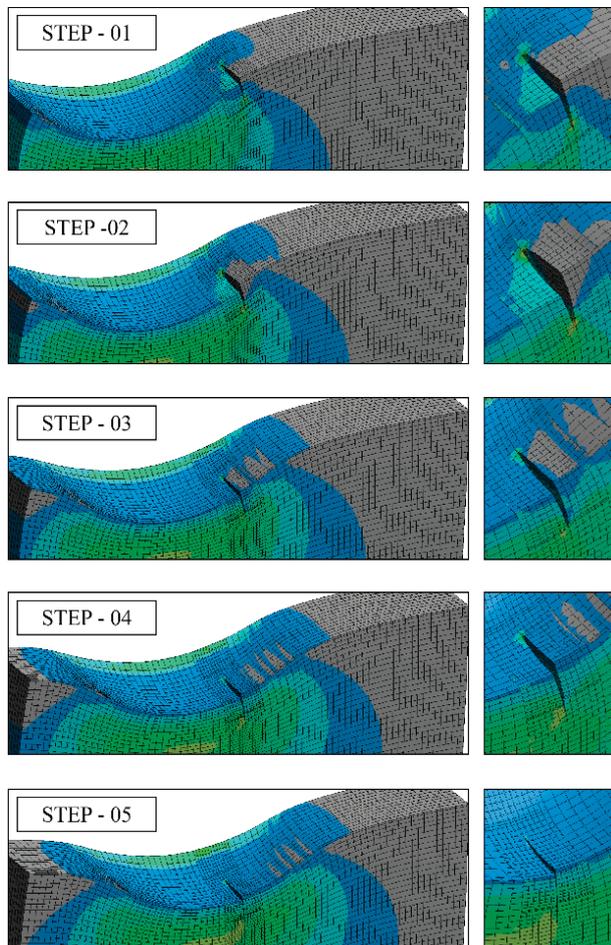


Figure 13 The von Mises stress during the passing of the load (scale: 125)

SUMMARY

The developed FE procedure is successfully implemented to the 3D environment to examine thermal crack propagation with VCCT crack growth simulation method.

The 3D model provides accurate deformation response for the contact pressure and has an appropriate answer in Hertzian stress distribution. Furthermore, the predefined energy release rate – to make the growth process even more spectacular – which helps to open the crack during one cycle of load passing showed better propagation results in the quasi-static study. Therefore, conclusions could be drawn from the connection between the two phases of the propagation.

For further developments of the model, the followings could be applied:

- Longer evaluation path to separate the different loading zones more clearly, to get a more stable temperature distribution on the surface, and to investigate the effect of the moving load when it is in the further vicinity of the crack.
- Examining more than one cycle of load pass with a modified energy-release rate to lengthen the crack propagation for the interval of the extended cycles. (If there is opportunity to use a vast amount of calculation capacity the typical energy-release rate could be used with more thousands of load cycle in the study).
- Examining a slightly bigger sub-environment of the wheel and using finer mesh in the crack vicinity.
- Application of frictional contact between the crack faces.
- Assumption of discretized pressure in the axial direction as well.
- Evaluation of the circumferential part of the thermal crack.
- Using non-linear material models.

This model can be optimized and improved from many aspects, but it has to be kept in mind that in Ansys WB environment, there is no option to perform fatigue analysis with the VCCT crack growth simulation method. Consequently, to approach the problem from fatigue point-of-view, the study has to be programmed in Ansys APDL environment.

ACKNOWLEDGEMENT

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI), and by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

REFERENCES

- Handa, K., Kimura, Y., Mishima, Y., 2010. Surface cracks initiation on carbon steel railway wheels under concurrent load of continuous rolling contact and cyclic frictional heat. *Wear* 268, 50–58. <https://doi.org/10.1016/j.wear.2009.06.029>
- Krueger, R., Shivakumar, K.N., Raju, I.S., 2013. Fracture Mechanics Analyses for Interface Crack Problems - A Review, in: Structures, Structural Dynamics, and Materials and Co-located Conferences. American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2013-1476>
- Pirondi, A., Giuliese, G., Moroni, F., Bernasconi, A., Jamil, A., 2015. Simulating the mixed-mode fatigue delamination/debonding in adhesively-bonded composite joints, in: Vassilopoulos, A.P. (Ed.), . pp. 369–400. <https://doi.org/10.1016/B978-0-85709-806-1.00013-6>
- VCCT-Based Crack-Growth - ANSYS - Help. https://ansyshelp.ansys.com/account/secured?returnurl=/Views/Secured/corp/v193/ans_frac/Hlp_G_STR_VCCT.html
- Zwierczyk, P.T., 2015. Thermal and stress analysis of a railway wheel-rail rolling-sliding contact (PhD Thesis). Budapest University of Technologies and Economics, Budapest.
- Zwierczyk, P.T., Váradi, K., 2014. Thermal Stress Analysis of a Railway Wheel in Sliding-Rolling Motion. *J. Tribol.* 136, 031401-031401–8. <https://doi.org/10.1115/1.4027544>

AUTHOR BIOGRAPHIES



TAMÁS MÁTÉ is a PhD student at Budapest University of Technologies 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 research field is engaged to crack propagation in railway wheels and rails. His e-mail address is: mate.tamas@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 member of the finite element modelling (FEM) research group. His e-mail address is: z.peter@gt3.bme.hu.

Simulation and Optimization

GLOBAL STABILITY OF FRACTIONAL POSITIVE NONLINEAR FEEDBACK SYSTEMS WITH INTERVAL STATE MATRICES

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

KEYWORDS

Global stability, fractional, positive, feedback, nonlinear system, interval state matrix.

ABSTRACT

The global stability of fractional positive continuous-time nonlinear systems with positive linear parts, positive feedbacks and interval state matrices is investigated. New sufficient conditions for the global stability of the classes of fractional positive nonlinear systems are established. The new stability conditions are demonstrated on simple examples of fractional positive nonlinear systems with interval state matrices.

1. INTRODUCTION

In positive systems inputs, state variables and outputs take only nonnegative values for any nonnegative inputs and nonnegative initial conditions (Kaczorek 2002, 2019a, Berman et.al. 1994). Examples of positive systems are industrial processes involving chemical reactors, heat exchangers and distillation columns, storage systems, compartmental systems, water and atmospheric pollutions models. A variety of models having positive behavior can be found in engineering, management science, economics, social sciences, biology and medicine, etc. An overview of state of the art in positive systems theory is given in the monographs (Berman et.al. 1994, Farina, et. al. 2000, Kaczorek 2002, 2011b, Kaczorek, et. al. 2015).

Descriptor positive systems have been analyzed in (Borawski 2017, Kaczorek 2012). Linear positive electrical circuits with state feedbacks have been addressed in (Borawski 2017, Kaczorek et.al. 2012). The superstabilization of positive linear electrical circuits by state feedbacks have been analyzed in (Kaczorek 2017) and the stability of nonlinear systems in (Kaczorek, et. al. 2017). The global stability of nonlinear systems with negative feedbacks and positive not necessary asymptotically stable linear parts has been investigated in (Kaczorek 2015a, 2019b). The global stability of positive standard and fractional nonlinear feedback systems has been analyzed in (Kaczorek 2020).

In this paper the global stability of fractional nonlinear feedback systems with positive linear parts with interval state matrices will be addressed.

The paper is organized as follows. In section 2 the basic definitions and theorems concerning the fractional positive linear systems are recalled. The stability of fractional positive systems with interval state matrices is addressed in section 3. New sufficient conditions for the global positive nonlinear systems with interval state matrices are established in section 4. Concluding remarks are given in section 5.

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

2. FRACTIONAL POSITIVE LINEAR SYSTEMS

Consider the fractional continuous-time linear system

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

$$y(t) = Cx(t) + Du(t), \quad (1b)$$

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

$$\frac{d^\alpha x(t)}{dt^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^t \dot{x}(\tau) (t-\tau)^{\alpha-1} d\tau, \quad \dot{x}(\tau) = \frac{dx(\tau)}{d\tau} \quad (1c)$$

is the Caputo fractional derivative and

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \quad \text{Re}(z) > 0 \quad (1d)$$

is the gamma function (Kaczorek 2011b, Kaczorek, et. al. 2015).

Definition 1. (Kaczorek 2011b, Kaczorek, et. al. 2015) The fractional system (1) is called (internally) positive if $x(t) \in \mathfrak{R}_+^n$ and $y(t) \in \mathfrak{R}_+^p$, $t \geq 0$ for any initial conditions $x(0) \in \mathfrak{R}_+^n$ and all inputs $u(t) \in \mathfrak{R}_+^m$, $t \geq 0$.

A real matrix $A = [a_{ij}] \in \mathfrak{R}^{n \times n}$ is called Metzler matrix if its off-diagonal entries are nonnegative, i.e. $a_{ij} \geq 0$ for $i \neq j$. The set of $n \times n$ Metzler matrices will be denoted by M_n .

Theorem 1. (Kaczorek 2011b, Kaczorek, et. al. 2015) The fractional system (1) is positive if and only if

$$A \in M_n, B \in \mathfrak{R}_+^{n \times m}, C \in \mathfrak{R}_+^{p \times n}, D \in \mathfrak{R}_+^{p \times m}. \quad (2)$$

Definition 2. (Kaczorek 2011b, Kaczorek, et. al. 2015) The positive fractional system (1) (for $u(t) = 0$) is called asymptotically stable (the matrix A is Hurwitz) if

$$\lim_{t \rightarrow \infty} x(t) = 0 \text{ for any } x(0) \in \mathfrak{R}_+^n. \quad (3)$$

Theorem 2. (Kaczorek 2011b, Kaczorek, et. al. 2015) The positive system (1) is asymptotically stable if and only if one of the equivalent conditions is satisfied:

1) All coefficient of the characteristic polynomial

$$\det[I_n s - A] = s^n + a_{n-1} s^{n-1} + \dots + a_1 s + a_0 \quad (4)$$

are positive, i.e. $a_k > 0$ for $k = 0, 1, \dots, n-1$.

2) All principal minors \bar{M}_i , $i = 1, \dots, n$ of the matrix $-A$ are positive, i.e.

$$\begin{aligned} \bar{M}_1 &= |-a_{11}| > 0, \bar{M}_2 = \begin{vmatrix} -a_{11} & -a_{12} \\ -a_{21} & -a_{22} \end{vmatrix} > 0, \\ \dots, \bar{M}_n &= \det[-A] > 0. \end{aligned} \quad (5)$$

3) There exists strictly positive vector $\lambda^T = [\lambda_1 \ \dots \ \lambda_n]^T$, $\lambda_k > 0$, $k = 1, \dots, n$ such that

$$A\lambda < 0 \text{ or } A^T\lambda < 0. \quad (6)$$

3. STABILITY OF FRACTIONAL INTERVAL POSITIVE LINEAR SYSTEMS

Consider the fractional interval positive linear continuous-time system

$$\frac{d^\alpha x}{dt^\alpha} = Ax, \quad 0 < \alpha < 1, \quad (7)$$

where $x = x(t) \in \mathfrak{R}^n$ is the state vector and the matrix $A \in M_n$ is defined by

$$A_1 \leq A \leq A_2 \text{ or equivalently } A \in [A_1, A_2]. \quad (8)$$

Definition 3. The interval positive system (7) is called asymptotically stable if the system is asymptotically

stable for all matrices $A \in M_n$ satisfying the condition (8).

By condition (6) of Theorem 2 the positive system (7) is asymptotically stable if there exists strictly positive vector $\lambda > 0$ such that the condition (6) is satisfied.

For two fractional positive linear systems

$$\frac{d^\alpha x_1}{dt^\alpha} = A_1 x_1, \quad A_1 \in M_n \quad (9)$$

and

$$\frac{d^\alpha x_2}{dt^\alpha} = A_2 x_2, \quad A_2 \in M_n \quad (10)$$

there exists a strictly positive vector $\lambda \in \mathfrak{R}_+^n$ such that

$$A_1 \lambda < 0 \text{ and } A_2 \lambda < 0 \quad (11)$$

if and only if the systems (9), (10) are asymptotically stable.

Theorem 3. If the matrices A_1 and A_2 of fractional positive systems (9), (10) are asymptotically stable then their convex linear combination

$$A = (1-k)A_1 + kA_2 \text{ for } 0 \leq k \leq 1 \quad (12)$$

is also asymptotically stable.

Proof. By condition (6) of Theorem 2 if the fractional positive linear systems (9), (10) are asymptotically stable then there exists strictly positive vector $\lambda \in \mathfrak{R}_+^n$ such that

$$A_1 \lambda < 0 \text{ and } A_2 \lambda < 0. \quad (13a)$$

Using (6) and (13) we obtain

$$A\lambda = [(1-k)A_1 + kA_2]\lambda = (1-k)A_1\lambda + kA_2\lambda < 0 \quad (13b)$$

for $0 \leq k \leq 1$. Therefore, if the positive linear systems (9), (10) are asymptotically stable then their convex linear combination (12) is also asymptotically stable. \square

Theorem 4. The interval positive systems (7) are asymptotically stable if and only if the positive linear systems (9), (10) are asymptotically stable.

Proof. By condition (6) of Theorem 2 if the matrices $A_1 \in M_n$, $A_2 \in M_n$ are asymptotically stable then there exists a strictly positive vector $\lambda \in \mathfrak{R}_+^n$ such that (6) holds. The convex linear combination (12) satisfies the condition $A\lambda < 0$ if and only if (13) holds. Therefore, the interval system (8) is asymptotically stable if and only if the positive linear system is asymptotically stable. \square

Example 1. Consider the fractional interval positive linear continuous-time system (8) with the matrices

$$A_1 = \begin{bmatrix} -2 & 1 \\ 2 & -3 \end{bmatrix}, A_2 = \begin{bmatrix} -3 & 2 \\ 4 & -4 \end{bmatrix}. \quad (14)$$

Using the condition (6) of Theorem 2 we choose for A_1 $\lambda_1 = [1 \ 1]^T$ and we obtain

$$A_1 \lambda_1 = \begin{bmatrix} -2 & 1 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} < 0, \quad (15a)$$

and for A_2 , $\lambda_2 = [0.8 \ 1]^T$

$$A_2 \lambda_2 = \begin{bmatrix} -3 & 2 \\ 4 & -4 \end{bmatrix} \begin{bmatrix} 0.8 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.4 \\ -0.8 \end{bmatrix} < 0. \quad (15b)$$

Therefore, the matrices (14) are Hurwitz. Note that

$$A_1 \lambda_2 = \begin{bmatrix} -2 & 1 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} 0.8 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.6 \\ -1.4 \end{bmatrix} < 0. \quad (16)$$

Therefore, for both matrices (14) we may choose $\lambda = \lambda_1 = \lambda_2 = [0.8 \ 1]^T$ and by Theorem 4 the fractional interval positive system (7) with (14) is asymptotically stable.

4. GLOBAL STABILITY OF FRACTIONAL NONLINEAR FEEDBACK SYSTEMS WITH POSITIVE LINEAR PARTS

In this section sufficient conditions for the global stability of fractional nonlinear systems with interval state matrices of asymptotically stable positive linear parts and positive state feedbacks with gain h will be proposed.

Consider the fractional nonlinear system shown in Fig. 1 which consists of the positive linear part with interval asymptotically stable state matrix, the nonlinear element with characteristic $u = f(e)$ and positive feedback with gain h .

The linear part is described by the equations

$$\begin{aligned} \frac{d^\alpha x}{dt^\alpha} &= Ax + Bu, \\ y &= Cx, \end{aligned} \quad (17)$$

where $x = x(t) \in \mathfrak{R}_+^n$, $u = u(t) \in \mathfrak{R}_+$, $y = y(t) \in \mathfrak{R}_+$ is the state vector, input and output and $A \in M_n$, $B \in \mathfrak{R}_+^{n \times 1}$, $C \in \mathfrak{R}_+^{1 \times n}$.

It is assumed that the positive linear part is asymptotically stable (the matrix $A \in M_n$ is Hurwitz) for A belonging to the interval (8).

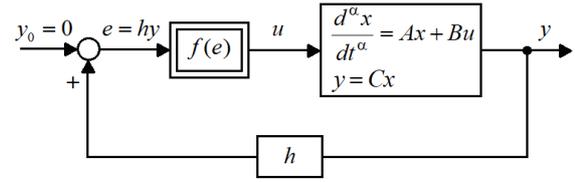


Figure 1: The nonlinear feedback system

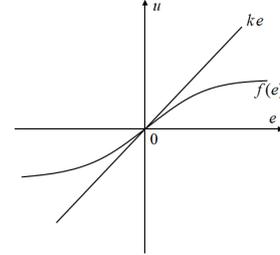


Figure 2: Characteristic of the nonlinear element

The characteristic of the nonlinear element is shown in Fig. 2 and it satisfies the condition

$$0 \leq \frac{f(e)}{e} \leq k < \infty. \quad (18)$$

Definition 4. The nonlinear positive system with interval state matrix $A \in [A_1, A_2] \in M_n$ is called globally stable if it is asymptotically stable for all nonnegative initial conditions $x(0) \in \mathfrak{R}_+$.

The following theorem gives sufficient conditions for the global stability of the positive nonlinear feedback systems.

Theorem 5. The nonlinear system consisting of the positive asymptotically stable linear part with interval state matrix $A \in [A_1, A_2]$, the nonlinear element satisfying the condition (18) and positive feedback with gain h is globally stable if

$$\begin{aligned} &(1-q)A_1 + qA_2 + khBC \\ &= \begin{cases} A_1 + khBC \in M_n & \text{for } q = 0 \\ A_2 + khBC \in M_n & \text{for } q = 1 \end{cases} \end{aligned} \quad (19)$$

is the Hurwitz Metzler matrix.

Proof. The proof will be accomplished by the use of the Lyapunov method (Lyapunov 1963, Leipholz 1970). As the Lyapunov function $V(x)$ we choose

$$V(x) = \lambda^T x \geq 0 \text{ for } x \in \mathfrak{R}_+^n, \quad (20)$$

where λ is strictly positive vector, i.e. $\lambda_k > 0$, $k = 1, \dots, n$.

Using (20) and (17) we obtain

$$\begin{aligned} \frac{d^\alpha}{dt^\alpha} V(x) &= \lambda^T \frac{d^\alpha x}{dt^\alpha} = \lambda^T (Ax + Bu) \\ &= \lambda^T (Ax + Bf(e)) \leq \lambda^T (A + khBC)x \end{aligned} \quad (21)$$

since $u = f(e) \leq ke = khCx$ and $A = (1-q)A_1 + qA_2$ for $q \in [0, 1]$.

From (20) it follows that $\frac{d^\alpha}{dt^\alpha} V(t) < 0$ if the condition (19) is satisfied and the nonlinear system is globally stable. \square

To find the maximal value of k_1 for which the nonlinear system is globally stable the following procedure can be used.

Procedure 1

Step 1. Find the value of k_1 for which the matrix

$$A_1 + k_1 hBC \in M_n \quad (22)$$

is asymptotically stable.

Step 2. Find the value of k_2 for which the matrix

$$A_2 + k_2 hBC \in M_n \quad (23)$$

is asymptotically stable.

Step 3. Find the desired value of k as

$$k = \min(k_1, k_2). \quad (24)$$

Example 2. Consider the nonlinear feedback system with the positive linear part with the interval matrix (14) and

$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, C = [0.4 \quad 0.8] \quad (28)$$

and the nonlinear element satisfying the condition (18) and $h = 0.5$. Find the maximal value of the coefficient for k for which the nonlinear system is globally stable.

Using Procedure 1 and (14), (28) we obtain

Step 1. Using (14), (28) and (22) we obtain

$$\begin{aligned} A_1 + k_1 hBC &= \begin{bmatrix} -2 & 1 \\ 2 & -3 \end{bmatrix} + 0.5k_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} [0.4 \quad 0.8] \\ &= \begin{bmatrix} -2 + 0.2k_1 & 1 + 0.4k_1 \\ 2 + 0.2k_1 & -3 + 0.4k_1 \end{bmatrix} \end{aligned} \quad (29)$$

and the maximal value of k_1 for which the matrix (29) is Hurwitz is $k_1 < \frac{5}{3}$ since for this value the coefficients of the polynomial

$$\begin{aligned} \det[I_2 s - A_1 - k_1 hBC] \\ = s^2 + (5 - 0.6k_1)s + 4 - 2.4k_1 \end{aligned} \quad (30)$$

are positive.

Step 2. Using (20), (28) and (23) we obtain

$$\begin{aligned} A_2 + k_2 hBC &= \begin{bmatrix} -3 & 2 \\ 4 & -4 \end{bmatrix} + 0.5k_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} [0.4 \quad 0.8] \\ &= \begin{bmatrix} -2 + 0.2k_2 & 1 + 0.4k_2 \\ 2 + 0.2k_2 & -3 + 0.4k_2 \end{bmatrix} \end{aligned} \quad (31)$$

and the maximal value of k_2 for which the matrix (31) is Hurwitz is $k_2 < 1$ since for this value the coefficients of the polynomial

$$\begin{aligned} \det[I_2 s - A_2 - k_2 hBC] \\ = s^2 + (7 - 0.6k_2)s + 4 - 4k_2 \end{aligned} \quad (32)$$

are positive.

Step 3. Using (24) and the results obtained in Steps 1 and 2 we obtain

$$k = \min(k_1, k_2) = \min\left(\frac{5}{3}, 1\right) = 1. \quad (33)$$

Therefore, the nonlinear system is globally stable for k less than given by (33).

5. CONCLUDING REMARKS

The global stability of positive continuous-time nonlinear feedback systems with interval state matrices has been investigated. New sufficient conditions for the global stability of this class of positive nonlinear systems are established. A procedure for computation of the value of the coefficient satisfying the condition (18) has been proposed. The new stability conditions are demonstrated on simple examples of positive nonlinear systems with interval state matrices. The considerations can be extended to fractional discrete-time nonlinear positive systems with all interval matrices of the linear parts.

ACKNOWLEDGMENT

This work was supported by National Science Centre in Poland under work No. 2017/27/B/ST7/02443.

REFERENCES

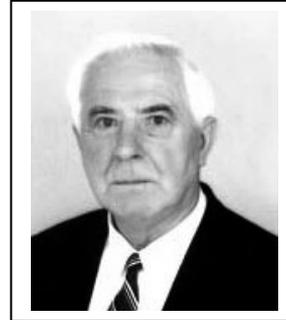
Berman, A. and R.J. Plemmons. 1994. *Nonnegative Matrices in the Mathematical Sciences*, SIAM.
 Borawski K. 2017. Modification of the stability and positivity of standard and descriptor linear electrical circuits by state feedbacks, *Electrical Review*, vol. 93, no. 11, 176-180.
 Busłowicz, M. and Kaczorek T. 2009. „Simple conditions for practical stability of positive fractional discrete-time linear

- systems”, *Int. J. Appl. Math. Comput. Sci.*, vol. 19, no. 2, 263-169.
- Farina, L. and Rinaldi S. 2000. *Positive Linear Systems; Theory and Applications*, J. Wiley, New York.
- Kaczorek T. 2019a. “Absolute stability of a class of fractional positive nonlinear systems”, *Int. J. Appl. Math. Comput. Sci.*, vol.29, no.1, 93-98.
- Kaczorek T. 2015a. “Analysis of positivity and stability of discrete-time and continuous-time nonlinear systems”, *Computational Problems of Electrical Engineering*, vol. 5, no. 1, 11-16.
- Kaczorek T. 2016. “Analysis of positivity and stability of fractional discrete-time nonlinear systems”, *Bull. Pol. Acad. Sci. Techn.*, vol. 64, no. 3, 491-494.
- Kaczorek T. 2019b. “Global stability of nonlinear feedback systems with positive linear parts”, *Intern. J. of Nonlinear Sciences and Numerical Simulation*.
- Kaczorek T. 2002. *Positive 1D and 2D Systems*, Springer-Verlag, London.
- Kaczorek T. 2010. “Positive linear systems with different fractional orders”, *Bull. Pol. Acad. Sci. Techn.*, vol. 58, no. 3, 453-458.
- Kaczorek T. 2011a. “Positive linear systems consisting of n subsystems with different fractional orders”, *IEEE Trans. on Circuits and Systems*, vol. 58, no. 7, 1203-1210.
- Kaczorek T. 2012. “Positive fractional continuous-time linear systems with singular pencils”, *Bull. Pol. Acad. Sci. Techn.*, vol. 60, no. 1, 9-12.
- Kaczorek T. 2011b. *Selected Problems of Fractional Systems Theory*, Springer, Berlin.
- Kaczorek T. 2015b. “Stability of fractional positive nonlinear systems”, *Archives of Control Sciences*, vol. 25, no. 4, 491-496.
- Kaczorek T. 2020. “Global stability of positive standard and fractional nonlinear feedback systems”, *Bull. Pol. Acad. Sci. Techn.*, vol.30, in Press.
- Kaczorek T. 2017. “Superstabilization of positive linear electrical circuit by state-feedbacks”, *Bull. Pol. Acad. Sci. Techn.*, vol. 65, no. 5, 703-708.
- Kaczorek T. and Borawski K. 2017. „Stability of Positive Nonlinear Systems”, *22nd Intern. Conf. Methods and Models in Automation and Robotics*, Międzyzdroje, Poland.
- Kaczorek T. and Rogowski K. 2015. *Fractional Linear Systems and Electrical Circuits*, Springer, Cham.
- Kharitonov V.L., 1978. “Asymptotic stability of an equilibrium position of a family of systems of differential equations”, *Differentsialnye uravnenia* vol.14, 2086-288.
- Kudrewicz J. 1964. “Ustoicivost nelinejnyh sistem z obratnoj swjazju”, *Avtomatika i Telemekhanika*, vol. 25, no. 8.
- Lyapunov A.M. 1963. “Obscaja zadaca ob ustoicivosti dvizenija”, *Gostechizdat*, Moskwa.
- Leipholtz H. 1970. *Stability Theory*, New York Academic Press.
- Mitkowski W. 2008. “Dynamical properties of Metzler systems”, *Bull. Pol. Acad. Sci. Techn.*, vol. 56, no.4, 309-312.
- Ruszewski A. 2019. “Stability conditions for fractional discrete-time state-space systems with delays”, *24th Intern. Conf. Methods and Models in Automation and Robotics*, Międzyzdroje, Poland.
- Sajewski L. 2017a. “Decentralized stabilization of descriptor fractional positive continuous-time linear systems with delays”, *22nd Intern. Conf. Methods and Models in*

Automation and Robotics, Międzyzdroje, Poland, 482-487.

- Sajewski L. 2017b. “Stabilization of positive descriptor fractional discrete-time linear systems with two different fractional orders by decentralized controller”, *Bull. Pol. Acad. Sci. Techn.*, vol. 65, no.5, 709-714.

AUTHOR BIOGRAPHIES



TADEUSZ KACZOREK

received his MSc, PhD and DSc degrees in electrical engineering from the Warsaw University of Technology in 1956, 1962 and 1964, respectively. In the years 1968–1969 he was the dean of the Electrical Engineering Faculty, and in the period of 1970–1973 he was a deputy rector of the Warsaw University of Technology. In 1971 he became a professor and in 1974 a full professor at the same university. Since 2003 he has been a professor at the Białystok University of Technology. In 1986 he was elected a corresponding member and in 1996 a full member of the Polish Academy of Sciences. In the years 1988–1991 he was the director of the Research Center of the Polish Academy of Sciences in Rome. In 2004 he was elected an honorary member of the Hungarian Academy of Sciences. He has been granted honorary doctorates by 13 universities. His research interests cover systems theory, especially singular multidimensional systems, positive multidimensional systems, singular positive 1D and 2D systems, as well as positive fractional 1D and 2D systems. He initiated research in the field of singular 2D, positive 2D and positive fractional linear systems. He has published 28 books (8 in English) and over 1100 scientific papers. He has also supervised 69 PhD theses. He is the editor-in-chief of the Bulletin of the Polish Academy of Sciences: Technical Sciences and a member of editorial boards of ten international journals.

SCENARIO-BASED SIMULTANEOUS INVESTMENT, FINANCING AND OPERATIONAL PLANNING

Mike Steglich
University of Applied Sciences Wildau
Hochschulring 1, 15745 Wildau, Germany
Email: mike.steglich@th-wildau.de

KEYWORDS

Simultaneous investment, financing and operational planning, decisions under uncertainty, scenario based linear integer programming

ABSTRACT

Investment, financial as well as operational decisions influence each other and may lead to suboptimal decisions if decided separately. Nevertheless, they are often dealt with separately, both in literature and in practice, or they are considered under rather unrealistic assumptions. This paper describes a new approach for simultaneous investment, financial and operating planning. It is an approach under uncertainty considering different scenarios with given probabilities taking into account individual risk preferences on the basis of suitable criteria like the von Morgenstern-Neumann expected utility. It contains the choice of different investment alternatives and their disinvestment, different financing alternatives as well as the determination of the sales and production quantities on the basis of a mixed-integer linear programme with the aim of maximising the net present value (NPV) of periodical project dividends.

INTRODUCTION

Investment and financial decisions as well as the operational strategies belong to the strategic planning processes. Each of these sub-plans influences the other depending decisions. Separating these planning parts may lead to sub-optimal decisions. The problem includes a set of investment alternatives which can be realised multiple times. There should also be an opportunity to disinvest these investments within their lifetime. These investments build the environment for the operations which deal with the sales and production quantities as well the stock of finished goods. The investments and the operations have to be financed using different financial sources. The results of such long-term oriented decisions are not certain. Therefore, the problem has to be solved under the considerations of a set of scenarios with given probabilities including an individual risk-preference of the decision maker.

This paper proposes a new approach for simultaneous investment, financial and operating planning. A linear, mixed-integer model will be introduced which maximises the NPV of the periodical cash-flow surpluses available as dividend payouts for the shareholders subject to several investment, financial and operations constraints.

The decision is based on scenarios with given probabilities as well as the risk-preference of a decision maker and is determined in a three-stage process. In the first step, the mixed-integer linear model is solved for all scenarios. These

scenario-optimal solutions can lead to completely different results if another scenario occurs. In this context, the horizon of all periods to be planned is to be divided into a fixed and a flexible horizon. The frozen horizon contains the first periods of the entire planning horizon. The decisions made for one particular scenario for the frozen horizon cannot be changed if another scenario occurs. Therefore, the mixed-integer linear model is solved again for all other scenarios on the basis of the solution of the fixed horizon of the given scenario to find the solutions for the following flexible horizon.

The scenario-optimal solutions for the frozen horizon can be understood as alternatives and the corresponding NPVs of the project dividends for all scenarios as random variables in a discrete decision model under uncertainty with given probabilities. The last step of the suggested approach is to solve this problem considering specific risk preferences by applying suitable criteria like the von Morgenstern-Neumann expected utility (Neumann and Morgenstern, 1953).

This paper starts with a literature review which is the basis for the proposed new approach described in the following section. The paper ends with conclusions.

LITERATURE REVIEW

Several aspects of simultaneous investment, financial and operating planning have been discussed for more than 50 years.

Approaches of simultaneous *investment and financing planning* want to select several investment and financing alternatives simultaneously often in order to maximise the NPV, the terminal wealth or the dividend payouts of all chosen investment and financing alternatives (Blohm et al., 2012, p. 269f). One of the first models was presented by Dean (1969) in which the investment and financing alternatives are selected by comparing the internal rates of returns of the investment alternatives and the interest rates of the financing alternatives (Götze et al., 2015, p. 216). Beside this rather simple model, there are a couple of proposals using linear programming techniques. A good overview of such approaches at that time are given in Bernhard (1969). E.g., Hax (1964) and Weingartner (1963) introduced similar dynamic models with more realistic assumptions than Dean (1969) in which the investments and financing alternatives are chosen simultaneously using a linear programming model in order to maximise the terminal wealth. These models are extended by Park (2008) by repetitive multi-phase decisions sequences. Albach (1962) developed a linear programming model which maximises the sum of the net present values of the selected investment and financial alternatives. Bernhard (1969) introduced a model with assump-

tions and constraints similar to Hax (1964) and Weingartner (1963) with an objective function that maximises the terminal wealth and the stream of dividend payments. Byrne et al. (1967) introduced a chance-constraint approach to capital budgeting including liquidity constraints. Majumdar and Chattopadhyay (1999) proposed an application orientated model for an integrated planning of capacity expansions and financial planning to maximise the net worth for electricity generation. They consider different financial sources (e.g. debt, cash flow from projects, equity) but a choice between different external financial sources is not designated.

One of the first models for integrated *investment and operational planning* was proposed by (Förstner and Henn, 1957, p. 119ff). They extended a production programme model by flexible constraints influenced by investments in order to maximise the terminal wealth. Another approach was introduced by Jacob (1964) with more realistic assumptions regarding the production system. The capacities of several types of machines measured in time units are flexible due to the opportunity to invest and also to disinvest in these machines depending on financial constraints. The objective function maximises the total profit including the sales revenues, the variable and fixed costs, the capital expenditures and the proceeds of liquidation (Jacob, 1964, p. 34ff). The author extended the objective function by interest earnings of the gross profits of the previous period in a second version of his model (Jacob, 1964, p.62f). More recently, Bradley and Arntzen (1999) developed a model for simultaneous capacity and inventory investment planning as well as production schedule in order to maximise the return on assets. Rajagopalan and Swaminathan (2001) proposed an mixed integer nonlinear production planning model with capacity expansion and inventory management which minimises the discounted costs of capacity expansion, inventories and production. Another nonlinear mixed integer programming model was introduced by Hsu and Li (2009) for high-tech manufacturers to determine the optimal supply chain network design including economy of scale effects. The model minimises the total average product costs per unit. In a wider range, models which deal with strategic design decisions in supply chain in combination with operational planning are to be mentioned. E.g., an integrated multi-objective supply chain model was introduced by Sabri and Beamon (2000) which considers the design of a supply chain system simultaneously with the selling and production decisions.

There are also simultaneous *operational and financial planning* models. For instance, Charnes et al. (1959) proposed an approach dealing with warehouse operations of goods to be bought and sold subject to financial constraints. Another recently published example is Guillén et al. (2006). They expanded a supply chain model maximising the profit of all products by cash balance constraints including the opportunity to borrow external capital. Different types of external financial sources are not considered.

All of these approaches consider different aspects of simultaneous investment, financial and operating planning, but none integrates the sub-plans into one approach. It can also be criticised that most of these models assume unrealistically only one relevant scenario and therefore certain

results. Additionally, some of the approaches use an exogenous weighted average cost of capital (WACC) and do not consider the impact of the different chosen financial sources on the WACC (Blumentrath, 1969, p. 271ff).

Despite applicable sub-aspects of these approaches, only models for simultaneous *investment, financial and operating planning* are suitable for the described problem. Blumentrath (1969) was one of the first authors who developed simultaneous investment, financial and operating planning models which maximise the terminal wealth for single- and multi-stage production systems. The models include the opportunity of investment and disinvestment of different investment alternatives and also different financial sources (Blumentrath, 1969, p. 334ff and 412ff). The model by Grundmann (1973) extends the second version of the model by Jacob (1964) by multiple financial sources and corresponding constraints as well as tax aspects (Grundmann, 1973, p. 59-66). Other models introduced by Jääskeläinen (1966) and Haberstock (1971) use rather simple assumptions for the operations (only one product, one type of machine and one raw material) and therefore they are less interesting for further discussion. All of these models consider only one scenario and assume therefore unrealistically all results as certain.

The most relevant model was introduced by Hahn and Kuhn (2012) which extends an own earlier approach (Hahn and Kuhn, 2011). It is a robust optimisation approach which combines investment, operations, and financial planning simultaneously in supply chains. The objective function takes into account both value-based performance and risk aspects by maximising the expected value of the economic value added (EVA) minus the downside risk of EVA. It is to be noted that this objective function does not allow risk-seeking preferences. Only risk preferences between risk-neutral and risk-averse can be chosen by selecting a risk preference parameter (Hahn and Kuhn, 2012, p. 562). Another aspect to be criticised is the handling of external financial sources. The model invokes different internal and external financial sources. Regarding the external financial sources, only the amount of a long-term debt can be chosen, but different types of external financial sources are not considered (Hahn and Kuhn, 2012, p. 565). Additionally, the variable structure and amount of the used financial sources should have an impact on the WACC used for the EVA in the objective function, which was not taken into account in the model.

CONCEPT OF THE SCENARIO-BASED SIMULTANEOUS INVESTMENT, FINANCING AND OPERATIONAL PLANNING APPROACH

Conceptual overview

This section is intended to describe a new approach for simultaneous investment, financial and operational planning.

There is a set of investment alternatives which are not only single machines but rather investment packages which can contain for instance machines, new facilities, licences, investment in markets, etc. (Blumentrath, 1969, p. 341). The investment decisions have to be made at the beginning of each period whereby each investment package can be re-

alised multiple times with the opportunity to disinvest it during its lifetime.

The investments and the operations have to be financed. This can be done by using a set of financial alternatives. These long-term orientated financial resources can be equity, loans, venture capital, etc., for which the periodical cash-flows, containing the raising and repayment of the capital as well as the capital costs, are known at the beginning of the first planning period. The decision about the utilised financial alternatives has to be made once at the beginning of the first planning period. There is also the opportunity of short-term loans which can be borrowed at the beginning of each of the periods and have to be paid back including the interests in the following period. Other financial sources are the sales revenues and the proceeds of liquidation of disinvested investment packages.

The operational decisions deal with the sales and production quantities in a rolling planning approach based on the resources built by realised investment alternatives and financed by the mentioned financial sources.

All these decisions have to be made simultaneously under the perspective of the shareholders. Therefore, a mixed-integer linear model is introduced which maximises the NPV of the periodical cash-flow surpluses available as project dividends for the shareholders subject to several investment, financial and operations constraints.

Since such long-term oriented decisions are usually problems under uncertainty, the new approach covers also a set of scenarios with given probabilities. The final decision is based on the scenarios and the individual risk-preference of a decision maker. It is determined in a three-stage process.

In the first step, the mentioned mixed-integer linear model is solved for all scenarios. However, such an optimal decision can lead to different results in the other scenarios, whereby the horizon of all periods to be planned is to be divided into two horizons. There is a frozen horizon at the beginning of the entire planning horizon in which the decision made for one scenario cannot be changed if an other scenario happens. That means that the values of the variables found for a scenario-specific optimal solution are fixed for the frozen horizon in the other scenarios for which in the second step of the proposed approach the mixed-integer linear model is solved again. These are the scenario-based solutions of the flexible horizon which follows the fixed horizon.

The different solutions of the scenarios lead to specific opportunities or risks. If it is for example assumed that the demands of customers are uncertain then the decisions made for a low-demand scenario could lead to a loss of sales revenues (and of the NPV for the cash-flow surpluses) in a high-demand scenario due to the restricted capacities determined for the low-demand scenario. If in the low-demand scenario the maximum customer demands are the restrictive factor then higher sales quantities in a high-demand scenario might occur. Additionally, the capacities determined for a high-demand scenario could be underutilised in a low-demand scenario.

The scenario-optimal solutions for the frozen horizon can be understood as alternatives and the corresponding NPVs of the project dividends for all scenarios as random vari-

ables in a discrete decision model under uncertainty with given probabilities. The last step of the suggested approach is to solve this problem considering specific risk preferences by applying suitable criteria like the von Morgenstern-Neumann expected utility.

Optimisation model

This subsection describes the mixed-integer linear model for simultaneous investment, financial and operational planning using the following indices, sets, parameters, and variables.

Indices and sets

$i \in O$	investment alternatives
$j \in F$	financing alternatives
$n \in P$	products
$(n, i) \in PO$	valid combinations of products and investment alternatives
$t \in Y$	periods $Y = \{0, 1, \dots, T\}$
$u \in LY_i$	life time of investment i
$k \in R$	investment-independent production factors
$m \in RM_i$	production factors depending on investment i

Parameters

is	interest rate for the shareholders
ps_{nt}	unit selling price for product n in period t
ch_{nt}	unit holding cost for product n in period t
c_{nit}	variable unit cost for product-investment combination (n, i) in period t
cs_{nit}	set-up cost for product-investment combination (n, i) in period t
co_{iu}	operating costs of one realisation of investment alternative i at the age of u periods
cre_{mit}	unit costs for the extension of investment-dependent production factor m of investment i in period t
re_{mit}^u	upper bound for the extension of investment-dependent production factor m of investment i in period t
cp_{xi}	capital expenditures for one realisation of investment i
bv_{iu}	book or market value of one realisation of investment alternative i at the age of u periods
yo_i^u	maximum investments in investment alternative i in a period
ydo_i^u	maximum disinvestments of investment alternative i in a period
xo_i^u	maximum number of investment packages of investment alternative i
xs_{nt}^l, xs_{nt}^u	lower and upper bound for the selling quantity of product n in period t
q_{nt}^l, q_{nt}^u	minimum and maximum inventory of product n at the end period t
cff_{jt}	cash-flow of one utilisation of financing alternative j in period t
yf_j^l, yf_j^u	lower and upper bound of utilisations of financing alternative j
ifs_t	interest rate for the short-term loan in period t
$cf_s_t^u$	upper bound of short-term loan in period t
ak_{nt}	use or consumption of investment-independent production factor k per unit of product n in period t

b_{kt}	upper bound of investment-independent production factor k in period t
ai_{mni}	use or consumption of investment-dependent production factor m for investment alternative i per unit of product n in period t
bi_{miu}	upper bound of investment-dependent production factor m for investment alternative i at the age of u periods
fc_t	cash-relevant fixed costs in period t

Variables

cs_t	cash-flow surplus in period t
yo_{it}	number of investments in investment alternative i in period t
ydo_{iut}	number of disinvestments of investment alternative i at the age of u periods in period t
yf_j	number of utilisations of financing alternative j
cfs_t	short term loan borrowed in period t
cf_t	cash-flow in period t
xs_{nt}	sales quantity of product n in period t
x_{nt}	production quantity of product n in period t
xo_{iut}	number of available packages of investment alternative i at the age of u periods in period t
cap_{mit}	available amount of investment-dependent production factor m for investment alternative i in period t
xi_{nit}	production quantity of product-investment combination (n, i) in period t
yi_{nit}	lot realisation variable of product-investment combination (n, i) in period t
q_{nt}	inventory of product n at the end of period t
qa_{nt}	average inventory of product n in period t
re_{mit}	resource extension of investment-dependent production factor m of investment i in period t

In this model, it is assumed that the investments and the operations are fully financed by using the sales revenues, the financial alternatives, the proceeds of liquidation of disinvestments and if necessary also by short-term loans. If there is a positive cash-flow surplus $cs_t; t \in Y$ after all cash consumptions then it can be used to satisfy the shareholder's wish of a risk-adequate return. The objective of the model is therefore to maximise the NPV of the non-negative periodical cash-flow surpluses. These cash-flow surpluses can be interpreted as dividend payouts of the entire investment, financial and operational programme.

$$z = \sum_{t \in Y} cs_t \cdot (1 + is)^{-t} \rightarrow \max \quad (1)$$

The values of the periodical cash-flow surpluses result from the following cash-flow balance constraint (expressions (2-9) which contain all relevant cash generators and consumers. This constraint has to be formulated for all periods and starts in expression (2) with the sum of the sales revenues of the goods sold of all products and the total holding costs for the average stock of finished goods of the products. The total cash-relevant variable costs of the quantities produced and also the corresponding total set-up costs of all valid combinations of products and investments $(n, i) \in PO$ have to subtracted as well as the cash-relevant fixed costs fc_t as shown

in (3). The first term in expression (4) describes the cash-relevant operating costs of the existing investment objects in the different ages. These investment objects provide resources that can be extended. The corresponding extension costs represent the second part in (4). The capital expenditures of the investments alternatives and the proceeds of liquidation of disinvested investment packages are described in (5). The cash-flow impacts of the utilisation of the long-term investment alternatives are shown in (6) and for the short-term loans in (7). The last part of the cash-flow balance constraint (8) contains the non-negative cash-flows of this and the previous year as well as the cash-flow surplus. That means that the previous cash-flow is used in the current year if it is necessary. Only if all cash requirements are satisfied then the cash-flow surplus is positive and available for the shareholders.

$$\sum_{n \in P} ps_{nt} \cdot xs_{nt} - \sum_{n \in P} ch_{nt} \cdot qa_{nt} \quad (2)$$

$$- \sum_{(n,i) \in PO} c_{nit} \cdot xi_{nit} - \sum_{(n,i) \in PO} cs_{nit} \cdot yi_{nit} - fc_t \quad (3)$$

$$- \sum_{i \in O} \sum_{\substack{u \in LY_i \\ u \leq t}} co_{iu} \cdot xo_{iut} - \sum_{i \in O} \sum_{m \in RM_i} cre_{mit} \cdot re_{mit} \quad (4)$$

$$- \sum_{i \in O} cpx_i \cdot yo_{it} + \sum_{i \in O} \sum_{\substack{u \in LY_i \\ 0 < u \leq t}} bvi_u \cdot ydo_{iut} \quad (5)$$

$$+ \sum_{j \in F} cff_{jt} \cdot yf_j \quad (6)$$

$$+ cfs_t - cfs_{t-1} \cdot (1 + ifs_{t-1}) \quad (7)$$

$$+ cf_{t-1} - cs_t = cf_t \quad (8)$$

$$; t \in Y, cf_{-1} = 0, cfs_{-1} = 0, cfs_T = 0 \quad (9)$$

This constraint can be extended easily by cash-flow impacts of changes of the working capital if necessary. But this paper is focused on the main cash-flow impacts as shown above.

The following constraint (10) describes the use or consumption of all investment-independent production factors $k \in R$ by the quantities to be produced for all products $i \in P$ and the corresponding upper bounds in all periods.

$$\sum_{n \in P} a_{knt} \cdot x_{nt} \leq b_{kt} \quad ; t \in Y, k \in R \quad (10)$$

The relationship between the production quantities of the products $i \in P$ and the quantities produced using the several investment alternatives $(n, i) \in PO$ are given in expression (11). These quantities are required for the calculation of the total variable production costs in (3).

$$\sum_{\{i | (n,i) \in PO\}} xi_{nit} = x_{nt} \quad ; n \in P, t \in Y \quad (11)$$

The balance between the products produced, the sales quantities and the stock of finished goods is defined in (12) (Billington et al., 1983). The sales quantities are the basis for the sales revenues in expression (2). The average stock of finished goods equals the average of the beginning and ending inventories of finished goods as in (13). These

average stocks are needed to calculate the holding costs in (2).

$$x_{nt} - q_{nt-1} - q_{nt} = xs_{nt} \quad ; n \in P, t \in Y \quad (12)$$

$$0.5 \cdot q_{nt-1} + 0.5 \cdot q_{nt} = qa_{nt} \quad ; n \in P, t \in Y, q_{n0} = 0, q_{nT} = 0 \quad (13)$$

The relationship between the product variables and the lot variables in (14) for all valid product-investment combinations $(n, i) \in PO$ is required to calculate the set-up costs in the cash-balance constraint (3).

$$xi_{nit} \leq M \cdot yi_{nit} \quad ; (n, i) \in PO, t \in Y \quad (14)$$

The constraints (15) and (16) are intended to determine the number of investment packages at the different ages for all investment alternatives and periods. As shown in (15), the number of investment packages at the age of zero depends on the number of investments in such packages in a given period. The amount of investment objects at the ages $u \in LY_i$ bases on the amount of these investment packages in the year before and the disinvestments as of (16). Similar formulations can be found in Hahn and Kuhn (2012).

$$xo_{i0t} = yo_{it} \quad i \in O, t \in Y \quad (15)$$

$$xo_{iut} = xo_{iu-1t-1} - ydo_{iut} \quad i \in O, t \in Y, t > 0, u \in LY_i, 0 < u \leq t \quad (16)$$

There is a set of investment-dependent production factors RM_i for which the available amounts per period are to be determined. The upper bounds of these production factors depend on the age of the corresponding investment objects. Therefore, the amounts of the available investment objects in the different ages xo_{iut} have to be multiplied by the age-dependent upper bounds bi_{miu} of the investment-dependent production factors to determine in sum the available amount cap_{mit} of an investment-dependent production factor $m \in RM_i$ for an investment alternative $i \in O$ in a period $t \in Y$ as shown in (17).

$$cap_{mit} = \sum_{\substack{u \in LY_i \\ u \leq t}} bi_{miu} \cdot xo_{iut} \quad ; m \in RM_i, i \in O, t \in Y \quad (17)$$

The investment-dependent production factors are used or consumed by the investment-dependent production quantities xi_{nit} as shown in the constraints (18). The available amounts cap_{mit} cannot be exceeded by this consumption or usages, but if necessary extended in a certain interval as of (32) which lead to additional extension costs shown in (4).

$$\sum_{\{n|(n,i) \in PO\}} ai_{mni} \cdot xi_{nit} - re_{mit} \leq cap_{mit} \quad ; t \in Y, i \in O, m \in RM_i \quad (18)$$

The ranges of all variables are defined as follows.

$$cs_t \geq 0 \quad ; t \in Y \quad (19)$$

$$xs_{nt} = \{xs_{it}^l, xs_{it}^l + 1, \dots, xs_{it}^u\} \quad ; n \in P, t \in Y \quad (20)$$

$$x_{nt} = \{0, 1, \dots\} \quad ; n \in P, t \in Y \quad (21)$$

$$xi_{nit} = \{0, 1, \dots\} \quad ; (n, i) \in PO, t \in Y \quad (22)$$

$$yi_{nit} \in \{0, 1\} \quad ; (n, i) \in PO, t \in Y \quad (23)$$

$$cf_t \geq 0 \quad ; t \in Y \quad (24)$$

$$0 \leq cfs_t \leq cfs_t^u \quad ; t \in Y \quad (25)$$

$$q_{nt} = \{q_{nt}^l, q_{nt}^l + 1, \dots, q_{nt}^u\} \quad ; n \in P, t \in Y \quad (26)$$

$$qa_{nt} \geq 0 \quad ; n \in P, t \in Y \quad (27)$$

$$xo_{iut} \in \{0, 1, \dots, xo_{it}^u\} \quad ; i \in O \quad (28)$$

$$yo_{it} \in \{0, 1, \dots, yo_{it}^u\} \quad ; i \in O \quad (29)$$

$$ydo_{it} \in \{0, 1, \dots, ydo_{it}^u\} \quad ; i \in O \quad (30)$$

$$yf_j \in \{yf_j^l, yf_j^l + 1, \dots, yf_j^u\} \quad ; j \in F \quad (31)$$

$$0 \leq re_{mit} \leq re_{mit}^u \quad ; m \in RM_i, i \in O, t \in Y \quad (32)$$

Solving the discrete decision problem under uncertainty

The proposed simultaneous investment, financing and operational decision model leads to a discrete problem under uncertainty for which a set of scenarios $s \in S$ with given probabilities $p_s : \sum_{s \in S} p_s = 1$ exists. All parameters of the model described in the previous section can be uncertain. The optimal investment, financing and operating decision is based on the scenarios and the risk-preference of a decision maker and to be determined in a three-stage process. To carry out the three working steps, the horizon of all periods to be planned $Y = \{0, 1, \dots, T\}$ is to be divided in a fixed planning horizon $\{0, \dots, \tau\}$ and a flexible planning horizon $\{\tau + 1, \dots, T\}$.

Step 1

As already described, all decisions have to be made at the beginning of the periods, with the financing decision being made only once at the beginning of the planning horizon. It is therefore not possible to predict which of the scenarios will occur before the decisions are made. In the first step, the optimisation model described in the previous section is therefore solved for all of the scenarios. The solutions of all variables found for a scenario $w \in S$ are divided into the set of the solutions of all variables for the frozen horizon A_w and the set of the solutions of all variables for the flexible horizon Ω_w which determine in total the objective function value $Z(A_w, \Omega_w)$ of this scenario.

Step 2

However, such an optimal solution for a scenario $w \in S$ can lead to completely different results in the other scenarios $s \in S \setminus \{w\}$. The decision made for one particular scenario cannot be changed if another scenario occurs. That means that the set of the optimal values A_w of all variables found for scenario w in the first working step are also fixed for the frozen horizon in the other scenarios.

$$A_s = A_w \quad ; w \in S, s \in S \setminus \{w\} \quad (33)$$

In the second working step, the optimisation model described in the previous section has to be solved again for all of these scenario combinations $w \in S, s \in S \setminus \{w\}$, whereby the values of the variables of the frozen horizon $A_s = A_w$ are fixed as shown in (33). The values of the variables of the flexible horizon Ω_s have to be determined optimally in order to maximise the objective function (1) resulting in the optimal objective function value $Z(A_w, \Omega_s)$.

Step 3

The outcome of the two previous working steps leads to a discrete decision model under uncertainty with given probabilities. The optimal solutions of the variables A_w of the

frozen horizon of the scenarios $w \in S$ can be understood as alternatives with a distribution of the resulting objective function values in all scenarios $Z(A_w, \Omega_s)$ $s \in S$. The last step of the suggested approach is to solve this problem considering specific risk preferences by applying suitable criteria.

The simplest approach to solve this decision problem is the expected value principle which is applicable for *risk-neutral* decision makers (Klein, 2009, p. 6-12ff). For all of the alternatives $w \in S$ the expected value

$$\mu_w = \sum_{s \in S} p_s \cdot Z(A_w, \Omega_s) \quad ; w \in S \quad (34)$$

is to be calculated and the alternative with the maximum expected value has to be chosen (Drury, 2018, p. 288ff). If additionally a *risk-averse* or *risk-seeking* preference (Klein, 2009, p. 6-12ff) has to be invoked into this decision then a measure of the opportunities and threats has to be determined. This can be done by the standard deviation of the stochastic results of the alternatives (Drury, 2018, p. 288ff).

$$\sigma_w = \sqrt{2 \sum_{s \in S} p_s \cdot (Z(A_w, \Omega_s) - \mu_w)^2} \quad ; w \in S \quad (35)$$

The utility of an alternative w results from the addition of the expected value with the standard deviation weighted by the parameter alpha (Drury, 2018, p. 288ff).

$$U(\mu_w, \sigma_w) = \mu_w + \alpha \cdot \sigma_w, \quad 0 \leq \alpha \leq 1 \quad ; w \in S \quad (36)$$

The parameter α determines the risk-preference of the decision-maker. If α is negative the decision maker evaluates the standard deviation as a threat for results less than the expected value. In contrast to this *risk-averse* attitude, a *risk-seeking* decision-maker uses a positive α because the standard deviation promises in his or her perception higher results than the expected value. An $\alpha = 0$ leads to the expected value principle used for *risk-neutral* decision-makers (Mulvey et al., 1995, p. 126f). The optimal alternative is the alternative with the maximum value of these utilities.

Another criterion to include specific risk-preferences into this decision is the von Morgenstern-Neumann expected utility (vNM) (Neumann and Morgenstern, 1953). For each alternative and all scenarios, the utility of the results $u(Z(A_w, \Omega_s))$ are calculated. The expected utility per alternative is the sum over all scenarios of these utilities multiplied by the probabilities (Winston, 2004, p. 744).

$$vNM_w = \sum_{s \in S} p_s \cdot u(Z(A_w, \Omega_s)) \quad ; w \in S \quad (37)$$

The specific risk-preferences depend on the utility functions. A concave utility function implies a *risk-averse* behaviour because it is assumed that such a decision maker has a decreasing marginal utility (Winston, 2004, p. 750f). A linear function is used for a *risk-neutral* preference and a convex utility function is interpreted as the utility of a *risk-seeking* decision maker (Klein, 2009, p. 6-12ff). As with the other criteria, the alternative with the maximum von Neumann-Morgenstern expected utility is to be chosen as the optimal alternative.

CONCLUSIONS AND OUTLOOK

Investment and financial decisions as well as the operational plans influence each other and have to be solved simultaneously to avoid suboptimal decisions. Unfortunately, these depending problems are often solved (partially) separately or considered under rather unrealistic assumptions.

This paper describes a new approach for simultaneous investment, financial and operating planning. A mixed-integer linear model is introduced which maximises the NPV of the periodical project dividend payouts subject to several investment, financial and operational constraints. This model is used in a three-stage process. In the first step, the mixed-integer linear model is solved for all scenarios. These scenario-optimal solutions can lead to completely different results if another scenario occurs, for which the optimisation model is solved again under considerations of the results and consequences of the original solution. To do this, the horizon of all periods is to be divided into a frozen and a flexible horizon. The decisions made for a frozen horizon at the beginning of the entire planning horizon cannot be changed if another scenario occurs. Therefore, the mixed-integer linear model is solved again for all other scenarios on the basis of the solution of the fixed horizon to find the solutions for the following flexible horizon. This is the second step of the proposed approach. The set of the alternatives with their solutions for the different scenarios can be understood as a discrete decision model under uncertainty with given probabilities. The last step of the suggested approach is to solve this problem considering specific risk preferences by applying suitable criteria.

This approach tries to fully address the problem described and to avoid the problems of other approaches. It solves investment, financial and operating decisions simultaneously based on a reasonable production-mix model including the opportunity to expand the available amount of the related production factors. It offers several internal and external financing sources including the multiple long-term orientated financial alternatives. Since the objective function is only related to the shareholders, the interest rate for the NPV of the project dividends depends only on their request of a risk-adequate return. This interest rate is not affected by the chosen financial sources in contrast to some of the other published approaches which use an exogenous given WACC without considering the impact of the chosen financial sources. The new approach involves several scenarios, whereby every decision criteria applicable for such problems can be used and therefore every kind of risk-preference can be considered.

A problem of the proposed approach is the effort involved in formulating and solving all of the required scenario combinations which should not be carried out by hand. An opportunity are mathematical programming languages like AMPL (Fourer et al., 2003) or (py)CMPL (Steglich and Schleiff, 2018) which enable a user to formulate mathematical models, to manage the parameters and to obtain the solutions of the models. This can be done iteratively in that a model is solved and its solution is used to specify the parameters of a depending model. Another problem is restricted hardware resources to solve a series of the proposed optimisation model for realistic problem sizes in a reasonable

time. To avoid such problems, a distributed or grid optimisation approach can be applied with which large models can be solved remotely on a single optimisation server or in a grid of optimisation servers installed on high performance systems (Steglich, 2016).

REFERENCES

- Albach, H., 1962. *Investition und Liquidität: die Planung des optimalen Investitionsbudgets*. Gabler, Wiesbaden.
- Bernhard, R.H., 1969. Mathematical programming models for capital budgeting - a survey, generalization, and critique. *Journal of Financial and Quantitative Analysis* 4, 2, 111–158.
- Billington, P.J., McClain, J.O., Thomas, L.J., 1983. Mathematical programming approaches to capacity-constrained mrp systems. *Management Science* 29, 10, 1126–1141.
- Blohm, H., Lüder, K., Schaefer, C., 2012. *Investition, Schwachstellenanalyse des Investitionsbereichs und Investitionsrechnung* (2 edn.). Vahlen, München.
- Blumentrath, U., 1969. *Investitions- und Finanzplanung mit dem Ziel der Endwertmaximierung*. Schriften zur theoretischen und angewandten Betriebswirtschaftslehre 7. Gabler, Wiesbaden.
- Bradley, J.R., Arntzen, B.C., 1999. The simultaneous planning of production, capacity, and inventory in seasonal demand environments. *Operations Research* 47, 6, 795–806.
- Byrne, R., Charnes, A., Cooper, W.W., Kortanek, K., 1967. A chance-constrained approach to capital budgeting with portfolio type payback and liquidity constraints and horizon posture controls. *Journal of Financial and Quantitative Analysis* 2, 4, 339–364.
- Charnes, A., Cooper, W.W., Miller, M.H., 1959. Application of linear programming to financial budgeting and the costing of funds. *The Journal of Business* 32, 1, 20–46.
- Dean, J., 1969. *Mathematical programming and the analysis of capital budgeting problems* (8 edn.). Columbia University Press, New York.
- Drury, C., 2018. *Management and Cost Accounting* (10 edn.). Cengage Learning Emea, Andover, Hampshire, NJ.
- Förstner, K., Henn, R., 1957. *Dynamische Produktionstheorie und lineare Programmierung*. Hain, Meisenheim/Glam.
- Fourer, R., GayBrian, D.M., Kernighan, W., 2003. *AMPL: A Modeling Language for Mathematical Programming* (2 edn.). Thompson, Pacific Grove, CA.
- Götze, U., Northcott, D., Schuster, P., 2015. *Investment Appraisal: Methods and Models* (2 edn.). Springer, Berlin Heidelberg.
- Grundmann, H.R., 1973. *Optimale Investitions- und Finanzplanung unter Berücksichtigung der Steuern*. Dissertation, Universität Hamburg, Fachbereich Wirtschaftswissenschaften.
- Guillén, G., Badell, M., Espuña, A., Puigjaner, L., 2006. Simultaneous optimization of process operations and financial decisions to enhance the integrated planning/scheduling of chemical supply chains. *Computers Chemical Engineering* 30, 3, 421 – 436.
- Haberstock, L., 1971. *Zur Integrierung der Ertragsbesteuerung in die simultane Produktions-, Investitions- und Finanzierungsplanung mit Hilfe der linearen Programmierung*. Carl Heymanns, Köln et.al.
- Hahn, G., Kuhn, H., 2011. Value-based performance and risk management in supply chains: A robust optimization approach. *International Journal of Production Economics* 139, 1, 135 – 144.
- Hahn, G., Kuhn, H., 2012. Simultaneous investment, operations, and financial planning in supply chains: A value-based optimization approach. *International Journal of Production Economics* 140, 2, 559 – 569.
- Hax, H., 1964. Investitions- und Finanzplanung mit Hilfe der linearen Programmierung. *ZfbF* 16, 430–446.
- Hsu, C.I., Li, H.C., 2009. An integrated plant capacity and production planning model for high-tech manufacturing firms with economies of scale. *International Journal of Production Economics* 118, 2, 486 – 500.
- Jacob, H., 1964. *Neuere Entwicklungen in der Investitionsrechnung*. Gabler, Wiesbaden. Sonderdruck der ZfB.
- Jääskeläinen, V., 1966. *Optimal Financing and Tax Policy of the Corporation*. Helsinki Research Institute for Business Economics, Helsinki.
- Klein, C.M., 2009. Decision Analysis. In Ravindran, R.A. (ed.), *Operations Research Methodologies*. CRC Press, Boca Rato, London, New York, pp. 6–1 – 6–31.
- Majumdar, S., Chattopadhyay, D., 1999. A model for integrated analysis of generation capacity expansion and financial planning. *IEEE Transactions on Power Systems* 14, 2, 466–471.
- Mulvey, J.M., Vanderbei, R.J., Zenios, S.A., 1995. Robust optimization of large-scale systems. *Operations Research* 43, 2, 264–281.
- Neumann, J.v., Morgenstern, O., 1953. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.
- Park, J.I., 2008. *Simultane Planung von Investitions- und Finanzierungsprogrammen bei mehrfacher Entscheidungsfolge*. Dissertation, Georg-August-Universität Göttingen.
- Rajagopalan, S., Swaminathan, J.M., 2001. A coordinated production planning model with capacity expansion and inventory management. *Management Science* 47, 11, 1562–1580.
- Sabri, E.H., Beamon, B.M., 2000. A multi-objective approach to simultaneous strategic and operational planning in supply chain design. *Omega* 28, 5, 581 – 598.
- Steglich, M., 2016. CMPLServer - An open source approach for distributed and grid optimisation. *AKWI - Anwendungen und Konzepte der Wirtschaftsinformatik* 4, 9–21.
- Steglich, M., Schleiff, T., 2018. Cmpl - manual v1.12. http://www.coliop.org/_download/CMPL.pdf. Accessed: 27 Feb 2020.
- Weingartner, H.M., 1963. *Mathematical programming and the analysis of capital budgeting problems*. Prentice-Hall, Englewood Cliffs, N.J.
- Winston, W.L., 2004. *Operations Research* (4 edn.). Thomson Learning, Belmont, CA.

INFLUENCE OF COMPANY SIZES IN ADAPTED MASTER PRODUCTION SCHEDULING FOR IMPROVING HUMAN WORKING CONDITIONS

Marco Trost
Thorsten Claus
Technical University of Dresden,
Faculty of Business and Economics,
International Institute (IHI) Zittau,
Markt 23, 02763 Zittau, Germany
E-mail: marco.trost@tu-dresden.de

Frank Herrmann
Technical University of Applied Sciences Regensburg,
Faculty of Computer Science and Mathematics,
Innovation and Competence Centre for Production Logistics and Factory Planning (IPF),
Galgenbergstraße 32, 93053 Regensburg, Germany

KEYWORDS

exhaustion; workload; cost reduction; master production scheduling; linear optimisation; company size

ABSTRACT

Sustainability is an important topic in production planning and control. This article contributes in particular to further research on the social dimension. It presents a linear optimisation model for Master Production Scheduling in order to improve human working conditions. Existing approaches have already identified a considerable potential for improvements. Furthermore, this article analyses the influence of the company size on workload and costs using an application with a high proportion of manual activities. It is demonstrated that human working conditions can be improved independently from the company size without increasing costs. In addition, smaller companies tend to have a higher exhaustion and the workload affects the total costs more in smaller companies. Therefore, smaller companies might benefit more from an improvement in human working conditions.

INTRODUCTION

Sustainability is of rising relevance in research and practice, for example, due to various interest groups and other factors like resource shortages or a shortage of skilled workers. Production Planning and Control (PPC) enables corresponding improvements, since, for example, resources can be used more efficiently through an adapted planning approach. The PPC consists of a hierarchy of three levels: Master Production Scheduling (MPS), Material Requirements Planning and Scheduling; as introduced, for example, by Hax and Meal (1975) and Drexel et al (1994). However, existing approaches to sustainability focus on ecological aspects (see e.g. Grosse et al., 2017; Trost et al., 2019b). Therefore, we consider the integration of social aspects (i.e. exhaustion) in this paper. Due to exhaustion, performance deficits occur in the case of a high proportion of manual activities, for example, through higher error rates (Neumann and Dul, 2010). We found out that exhaustion effects have so far been integrated at the level of lot sizing and scheduling. At the MPS level, only the work of Trost et al (2019a) is known, who identified a high potential for improved human working conditions through an adapted MPS.

In our paper, we extend this approach by considering different company sizes. For example, Burgwal and Vieira (2014) identified a significant influence of the company size on sustainability reporting. We analyse the impact of improved human working conditions in terms of workload and costs. The company size is interpreted as the number of employees that is affected by different customer demands. The MPS is realised by a linear optimisation model and the exhaustion is integrated by a workload dependent capacity consumption.

For this, the paper is structured as follows. Section 2 presents a literature review. In section 3, the linear optimisation model is described. A test model is introduced in section 4 and the experimental design is presented in section 5. The numerical results and a discussion are outlined in section 6. The paper ends with a conclusion in section 7.

LITERATURE REVIEW

Due to exhaustion, performance deficits can be observed (Boenzi et al., 2015). For example, an increased error rate (Neumann and Dul, 2010; Yeow et al., 2014), lower productivity (Barker and Nussbaum, 2011) and decreasing motivation (Nijp et al., 2012) might occur. According to Nerdinger et al. (2014), physical diseases (e.g. musculoskeletal disorders) might occur as well. One reason for exhaustion, for example, are inadequate working conditions. In this regard, employees report, among other things, a high work intensity (DGB Index Gute Arbeit, 2014), the number of overtime hours (Ahlers, 2017) and deviations from regular working hours (Ahlers, 2017).

Exhaustion and performance deficits are not quantified for a long-term period (e.g. several years). Therefore we consider existing approaches to short-term muscle fatigue. In this context, there are methods for risk assessment of musculoskeletal disorders (e.g. OWAS, NIOSH, RULA, OCRA) to take into account the ergonomic risks of a particular task. These methods are considered, for example, to determine a maximum endurance time (MET) (e.g. Frey Law and Avin, 2010; Garg et al., 2002). In addition, an exponential increase in fatigue over time has been identified (see e.g. Ma et al., 2011).

Research on PPC has addressed this topic as follows. At the scheduling level, exhaustion is usually considered in the areas of assembly line balancing and job rotation. For

example, the ergonomic risks are minimised in Bautista et al. (2016) and Mossa et al. (2016) and constrained in Kara et al. (2014) as well as Nanthavanij et al. (2010). A comprehensive overview is given by Otto and Battaia (2017). In lot sizing, the already mentioned methods are used to avoid lot sizes with high ergonomic risks (e.g. Andriolo et al., 2016; Battini et al., 2017). Such considerations are also made in intra-logistics and warehouse management (Grosse et al., 2017). At the level of master production scheduling - to the best of our knowledge - there is only the work of Trost et al. (2019a), who identify the potential for improving human working conditions through an adapted MPS without increasing costs. So far, exhaustion has not been addressed at MPS widely. Therefore, we analyse the effect of different company sizes, based on the work of Trost et al (2019a).

MODEL FORMULATION

Our model is based on the extended linear optimisation model proposed by Trost (2018), who integrate an employee workload control. We use the following notation:

Sets

$EG = \{1, \dots, EG\}$ set of employee groups, indexed by eg
 $J = \{1, \dots, J\}$ set of production segments, indexed by j
 $K = \{1, \dots, K\}$ set of products, indexed by k
 $T = \{1, \dots, T\}$ set of time periods, indexed by t
 $Z = \{0, \dots, Z\}$ set of lead-time periods for capacity load, indexed by z

Parameters

$Capa_{eg}$ available capacity per period and an employee of employee group eg
 $d_{k,t}$ demand per product k in period t
 $f_{z,j,k}$ capacity load factors for lead-time period z , production segment j and product k
 h_k inventory holding costs per unit and period of product k
 I_k^{Init} initial inventory level for product k
 m_{eg}^{Cost} cost rate for hiring an employee of employee group eg
 n_{eg}^{Cost} cost rate for turnover of an employee of employee group eg
 R_j^{Max} maximum permitted employee utilisation per production segment j
 R_j^{Min} minimum permitted employee utilisation per production segment j
 $Staff_{eg}^{Cost}$ cost rate per employee of employee group eg
 $Staff_{eg,j}^{Init}$ initial number of employees per employee group eg and production segment j
 $Staff_{eg,j}^{Max}$ maximum number of employees per employee group eg in production segment j

$Staff_{eg,j}^{Min}$ minimum number of employees per employee group eg in production segment j
 $Staff_j^{TotalMax}$ maximum number of employees in production segment j
 V number of periods for overtime balancing
 w_{eg} lead-time periods for hiring employees of employee group eg
 wf_{eg} lead-time periods for employee turnover of employee group eg

Decision Variables

$a_{j,t}$ available capacity per production segment j in period t
 $b_{j,t}$ capacity requirement per production segment j in period t
 $I_{k,t}$ inventory level per product k in period t
 $m_{eg,j,t}$ number of hired employees of employee group eg in production segment j and period t
 $n_{eg,j,t}$ number of employee turnover of employee group eg in production segment j and period t
 $overtime_{j,t}$ used overtime per production segment j and period t
 $Staff_{eg,j,t}$ number of employees of employee group eg , production segment j and period t
 $x_{k,t}$ production quantity per product k in period t

Objective Function

The objective is to minimise the total costs from inventory holding, employment as well as employee hiring and turnover (equations (1)-(6)).

Objective Function: Minimise (TotalCosts) (1)

$$\begin{aligned} TotalCosts &= InventoryCosts \\ &+ StaffingCosts \\ &+ HiringCosts \\ &+ TurnoverCosts \end{aligned} \quad (2)$$

$$InventoryCosts = \sum_{t=1}^T \sum_{k=1}^K h_k \cdot I_{k,t} \quad (3)$$

$$StaffingCosts = \sum_{t=1}^T \sum_{j=1}^J \sum_{eg=1}^{EG} Staff_{eg}^{Cost} \cdot Staff_{eg,j,t} \quad (4)$$

$$HiringCosts = \sum_{t=1}^T \sum_{j=1}^J \sum_{eg=1}^{EG} m_{eg}^{Cost} \cdot m_{eg,j,t} \quad (5)$$

$$TurnoverCosts = \sum_{t=1}^T \sum_{j=1}^J \sum_{eg=1}^{EG} n_{eg}^{Cost} \cdot n_{eg,j,t} \quad (6)$$

Constraints

At the constraints, there are the inventory balance sheet (equation (7)), the definition of the initial inventory level (equation (8)) and equation (9) determine the capacity requirements.

$$x_{k,t} + I_{k,t-1} - I_{k,t} = d_{k,t} \quad \forall 1 \leq k \leq K; \forall 1 \leq t \leq T \quad (7)$$

$$I_{k,t=0} = I_k^{Init} \quad \forall 1 \leq k \leq K \quad (8)$$

$$\sum_{z=0}^Z \sum_{k=1}^K f_{z,j,k} \cdot x_{k,t+z} = b_{j,t} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq (T-Z) \quad (9)$$

Aspects of personnel requirements planning are considered as follows. We define the available capacity (equation (10)) and allow hiring and turnover of employees, which is integrated by the employee balance sheet (equation (11)) and the determination of the initial employee level (equation (12)). For this we distinguish between different employee groups (EG). Lead-times for hiring (we_{eg}) and turnover (wf_{eg}) are considered as well. With equation (13) and (14) we ensure that an adequate number of (skilled) employees are available and that only a limited number of employees can be employed. Equation (15) represents that the available number of skilled employees is limited on the labour market.

$$\sum_{eg=1}^{EG} Staff_{eg,j,t} \cdot Capa_{eg} = a_{j,t} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq T \quad (10)$$

$$Staff_{eg,j,t} = Staff_{eg,j,t-1} + m_{eg,j,t-we_{eg}} - n_{eg,j,t-wf_{eg}} \quad \forall 1 \leq eg \leq EG; \forall 1 \leq j \leq J; \forall 1 \leq t \leq T \quad (11)$$

$$Staff_{eg,j,t=0} = Staff_{eg,j}^{Init} \quad \forall 1 \leq eg \leq EG; \forall 1 \leq j \leq J \quad (12)$$

$$Staff_{eg,j,t} \geq Staff_{eg,j}^{Min} \quad \forall 1 \leq eg \leq EG; \forall 1 \leq j \leq J; \forall 1 \leq t \leq T \quad (13)$$

$$\sum_{eg=1}^{EG} Staff_{eg,j,t} \leq Staff_j^{TotalMax} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq T \quad (14)$$

$$Staff_{eg,j,t} \leq Staff_{eg,j}^{Max} \quad \forall 1 \leq eg \leq EG; \forall 1 \leq j \leq J; \forall 1 \leq t \leq T \quad (15)$$

The employee workload control is represented by equation (16) and (17). Thus, it is enabled to control the work intensity (average utilisation) as well as the deviations in regular working hours ($R_j^{Max} - R_j^{Min}$).

$$R_j^{Max} \cdot a_{j,t} \geq b_{j,t} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq (T-Z) \quad (16)$$

$$R_j^{Min} \cdot a_{j,t} \leq b_{j,t} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq (T-Z) \quad (17)$$

Overtime can occur if the maximum utilisation (R_j^{Max}) is over 100%. We control the use of overtime by equations (18)-(20). However, overtime do not result in additional costs because they have to be compensated within a specific time interval (by equation (19)) which meets legal restrictions. When the maximum utilisation is less than 100% these constraints are not restrictive.

$$b_{j,t} - a_{j,t} = overtime_{j,t} \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq (T-Z) \quad (18)$$

$$\sum_{t'=t-V}^t overtime_{j,t'} \leq 0 \quad \forall 1 \leq j \leq J; \forall 1 \leq t \leq (T-Z) \quad (19)$$

$$\sum_{t'=0-V}^{t=0} overtime_{j,t'} = 0 \quad \forall 1 \leq j \leq J \quad (20)$$

TEST PROBLEM

We consider an application from the railway industry in order to analyse the effects of different company sizes. Within the railway industry, there is a high human impact due to a low degree of automation (Neumann and Kripendorf, 2016). For good readability, this test problem is rather small. However, the results might not be dependent on the specific problem instance.

At first, general parameters are presented in Table 1. The different employee groups (EG) are interpreted as core employees ($eg = 1$) and temporary employees ($eg = 2$).

Table 1
General Parameters.

Parameter	Value
J	2
K	2
EG	2
Z	1

The employment of the temporary employees are outsourced to an employment agency as personnel leasing. For that reason, the costs for hiring (m_{eg}^{Cost}) and turnover (n_{eg}^{Cost}) as well as the lead-times for hiring (we_{eg}) and turnover (wf_{eg}) are higher for core employees than for temporary employees. The costs per employee of employee group and per period ($Staff_{eg}^{Cost}$) are taken from the IG Metall labour agreement for metal and electrical industries, Saxony, Germany from salary group five (additional level) (IG Metall, 2018) and 21.5% employer contribution are included as well. The staffing costs for the temporary employees are higher, due to the agency

service fees of 80%. Further, it is assumed that these employees have an experience gap compared to the core employees. This is taken into account by a lower available capacity per employee ($CAPA_{eg}$). Table 2 presents the concrete values.

Table 2

Parameters for core employees ($eg = 1$) and temporary employees ($eg = 2$) (abbr.: Money Units; seconds).

Parameter	$eg = 1$	$eg = 2$
$CAPA_{eg}$	524 400 s	393 300 s
m_{eg}^{Cost}	15 000 MU	1 500 MU
n_{eg}^{Cost}	60 000 MU	100 MU
$Staff_{eg}^{Cost}$	3 671 MU	5 435 MU
w_{eg}	3 months	1 months
wf_{eg}	3 months	0 months

Since different company sizes should be analysed, the minimum and maximum numbers of employees are set in such a way that they are not restrictive. Therefore, they are not presented.

Table 3 contains the cost rate for inventory holding and the initial inventory level.

Table 3

Parameters for inventory holding per product (K) (abbr.: Money Units; Quantity Units).

Parameter	$k = 1$	$k = 2$
h_k	115 MU	165 MU
I_k^{Init}	0 QU	0 QU

The remaining parameters: capacity load factors, demands and utilisation restrictions; are specific to our investigation and are explained in the next section.

EXPERIMENTAL DESIGN

This section explains the data for the individual experiments as well as the experiments itself. First, we determine a parameter setting so that a high work intensity, deviations in regular working time and overtime hours occur, which is described in the literature as reasons for exhaustion (see Ahlers, 2017; DGB-Index Gute Arbeit, 2014). We call it initial problem. Next, we modify the initial problem to improve the human working conditions and third, we consider different company sizes by defining a suitable customer demand.

Since the consequences of exhaustion analysed in this paper occur with a long-term overload, we consider a planning horizon of 84 months ($T = 84$). A 12-month warm up as well as run out phase are taken into account, so that the results from 60 months are analysed ($\hat{T} = 60$).

Initial problem

The employee utilisation is not further restricted. For this, the minimum utilisation is $U_j^{Min} = 0.00$ and overtime hours of 20% of the normal capacity are permitted

($U_j^{Max} = 1.20$). The overtime must be compensated by less working hours within 6 months ($V = 5$), so that in accordance with the working time law § 3 (in Germany) there are no additional costs. Table 4 presents the capacity load factors ($f_{z,j,k}$). They were set in such a way that similar to Trost et al. (2019a) a high average work intensity (average employee utilisation), high deviations from regular working time ($U_j^{Max} - U_j^{Min}$) and high overtime hours occur. Note that the capacity load occur only in lead-time period $z = 1$.

Table 4

Capacity load factors for the initial problem per product (K) and production segment (J) (abbr.: seconds).

Parameter		$j = 1$	$j = 2$
$f_{z=1,j,k}$	$k = 1$	3 867 s	13 976 s
	$k = 2$	4 092 s	10 184 s

Scenario with improved human working conditions

With the following setting, the human working conditions should be improved. Therefore, we limit the maximum utilisation to 95% ($U_j^{Max} \leq 0.95$). For a further reduction of the average workload, we reduce this maximum utilisation gradually (in steps of 5%). The deviations of the regular working time are limited to 10% ($U_j^{Max} - U_j^{Min} = 0.10$). In order to ensure that an arbitrarily low utilisation seems not to be economically, we consider a minimum utilisation of $U_j^{Min} \geq 0.65$. These lead to the following regarded utilisation intervals: 85-95%, 80-90%, 75-85%, 70-80% and 65-75%. The advantages of these improved working conditions are taken into account by adapted capacity load factors. For this, we assume an exponential decrease of capacity load with decreasing utilisation, similar to the course of muscle fatigue (see e.g. Ma et al., 2011). The new capacity load factors are based on the capacity load factors from the initial problem and both (the new and the old capacity load factors) are identical at an utilisation interval of 95-100%. The concrete values are given in Table 5.

Table 5

Capacity load factors for each utilisation interval of the scenario with improved human working conditions per product (K) and production segment (J) (abbr.: seconds).

Parameter	Utilisation		$j = 1$	$j = 2$
$f_{z=1,j,k}$	85-95%	$k = 1$	3 629 s	13 115 s
		$k = 2$	3 840 s	9 557 s
	80-90%	$k = 1$	3 405 s	12 306 s
		$k = 2$	3 603 s	8 967 s
	75-85%	$k = 1$	3 202 s	11 571 s
		$k = 2$	3 388 s	8 432 s
	70-80%	$k = 1$	3 017 s	10 902 s
		$k = 2$	3 192 s	7 944 s
	65-75%	$k = 1$	2 848 s	10 294 s
		$k = 2$	3 014 s	7 501 s

Company size

We define the company size as the number of employees (per production segment) and consider 16 different company sizes (see Table 6). Since the number of employees is a decision variable, we define the company size indirectly by the customer demand. For this, we consider that the demand is normally distributed with a standard deviation of 5%. The mean values are defined in such a way, that with respect to the capacity load factors from the initial problem and the available capacity per employee from employee group $eg = 1$ the corresponding company sizes (i.e. required number of employees) result. In Table 6 there are the concrete values for company size and mean value of demand series.

Table 6

Company size per production segment (J) and mean values per product (K) for determining demand series.

Company size		Demand mean value	
$j = 1$	$j = 2$	$k = 1$	$k = 2$
1 /	3	68	64
5 /	15	339	320
10 /	30	678	641
20 /	60	1 356	1 282
30 /	90	2 034	1 922
40 /	120	2 712	2 563
50 /	150	3 390	3 204
75 /	225	5 085	4 806
100 /	300	6 780	6 408
150 /	450	10 171	9 611
200 /	600	13 561	12 815
300 /	900	20 341	19 223
400 /	1 200	27 122	25 630
500 /	1 500	33 902	32 038
750 /	2 250	50 853	48 057
1 000 /	3 000	67 076	64 076

For each company size, we realise 10 individual demand series. For the total costs, we calculate confidence intervals with an error probability of $1 - \alpha = 0.95$ and a normal distribution. We calculate the relative deviation of

these confidence interval bounds to the mean value ($CI^{relative}$).

Each of these 4 160 individual planning problems (4 000 problems for the scenario with improved human working conditions from 25 utilisation combinations for segment 1 and 2, 16 company sizes and 10 demand series as well as 160 problems for the initial problem from 16 company sizes and 10 demand series) were solved with CPLEX from IBM-ILOG version 12.7.1 on a PC with a processor of 3.30 GHz and 192 GB of RAM.

NUMERICAL RESULTS and DISCUSSION

Of the 4 160 planning problems, 499 have such a long runtime that they are terminated after one hour. The resulting gap is on average 0.12% and on maximum 3.26% whereby 97.60% of the planning problems have a gap below 1%. The average runtime of all other planning problems is 160.65 seconds, while 92.90% have a runtime below 10 minutes and 1.23% above 50 minutes. For the total costs, the relative deviation of the confidence interval bounds to the mean value ($CI^{relative}$) are for all planning problems between 0.20% and 2.56% whereby the mean value is 0.47%. Thus, the confidence intervals are small and therefore they are not listed.

In the initial problem a high level of exhaustion should occur for all company sizes. In terms of work intensity, the average utilisation is 99.08% among all company sizes and production segments. The deviation of the regular working time (maximum amplitude of utilisation) is 16.14% on average. For example, this means a deviation of 6.13 hours for a 38-hours week. Overtime occur on average in 39.81% of periods, while the average overtime in these periods is 2.18% of the regular working time. However, exhaustion tends to be higher in smaller companies than in larger companies. This is illustrated exemplary in Figure 1 for the proportion of periods in which overtime occur in production segment 1. A comparable trend could be observed for production segment two, the average utilisation and the deviations of the regular working time.

As explained in section 'experimental design', the human working conditions from the initial problem are improved. For this, overtime is not permitted and the deviations from regular working time are limited to 10%. However, the permitted maximum amplitude of the utilisation is not

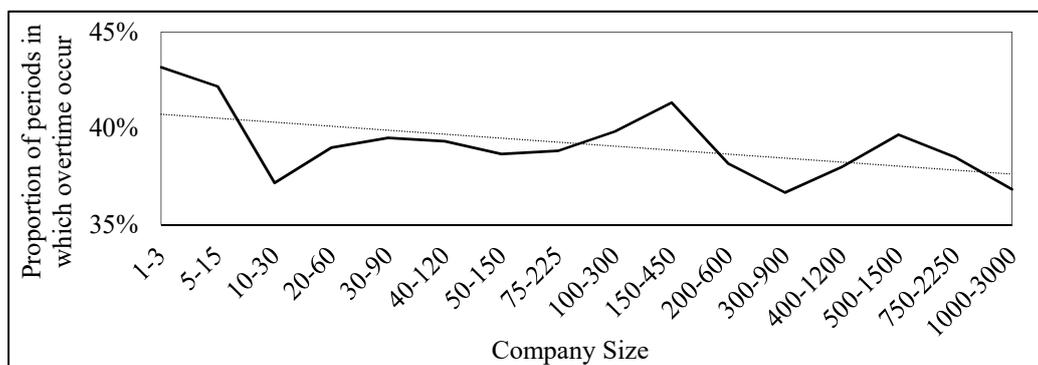


Figure 1 Proportion of periods in which overtime occur in production segment $j = 1$ and linear trend for each company size.

used to its limits. On average, the maximum utilisation amplitude is 6.74%. The work intensity is examined in the scenario with improved human working conditions by a (stepwise) reduced maximum utilisation. For the resulting utilisation intervals an average utilisation close to the upper interval bound occur.

In order to analyse these improvements, we also regard the resulting cost deviations compared to the initial problem. Figure 2 presents the range of cost deviations from all utilisation intervals in the scenario with improved human working conditions compared to the initial problem per company size. For all company sizes and production segments the lowest costs occur from the utilisation in-

terval 75-85%. Thus, the decision on the optimal employee treatment is not dependent on the company size. However, the range of cost deviations from all utilisation intervals is higher for smaller companies. This indicates that the utilisation interval affects the total costs more in smaller companies up to a certain company size and an inadequate treatment of the employees is more disadvantageous. With a suitable utilisation, smaller companies might benefit more than larger companies. Therefore, especially smaller companies should focus on the treatment of their employees. But also for larger companies a suitable employee treatment is advantageous.

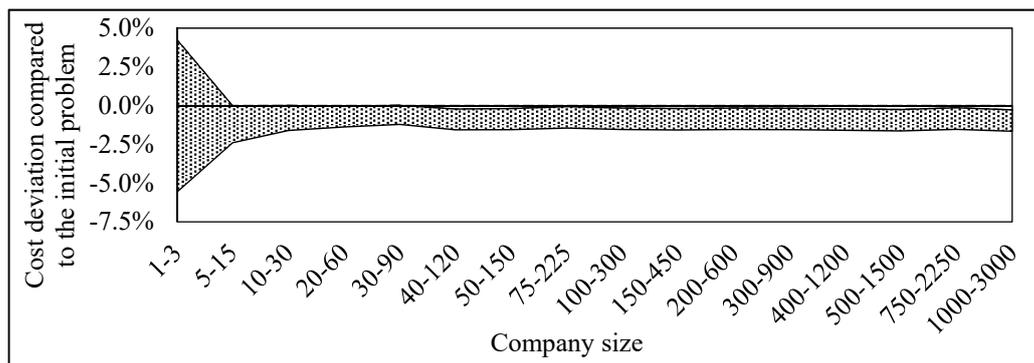


Figure 2 Range of cost deviations from all utilisation intervals per company size compared to the total costs from the initial problem.

CONCLUSION

In this paper, we adapt the MPS based on Trost (2018) by the consideration of exhaustion. We analyse its effects for different company sizes in terms of workload and costs through a sensitivity analysis. For this, we first identify exhaustion dependent performance deficits and their causes from the literature and consider the state of the art regarding the integration of exhaustion in PPC. Second the linear optimisation model for the MPS and the test problem is introduced. Thereby the exhaustion is modelled by utilisation dependent capacity load factors. The company size is interpreted as the number of employees and is defined by different normally distributed customer demands.

In the existing literature - to the best of our knowledge - there is only the work of Trost et al. (2019a), who identify the potential for improving human working conditions through an adapted MPS without necessarily increasing costs. We demonstrate that this occur independently of the company size und that the optimal utilisation is independent of the company size. According to this, the working conditions for smaller as well as for larger companies can be improved by an adapted MPS without an increase in costs. However, since in the initial problem smaller companies tend to be more exhaustive, its potential for improving working conditions is higher than in larger companies. Further the employee workload affects the total costs more in smaller companies and an adequate employee treatment is more advantageous. But also for

larger companies a suitable employee treatment is recommended.

Therefore, smaller companies might benefit more from an improvement in human working conditions. The results should also affect the next two levels of PPC. This extension is intended for future research.

REFERENCES

- Ahlers, E. 2017. Work and health in German companies. Findings from the WSI works councils survey 2015. WSI Report 33.
- Andriolo, A., D. Battini, A. Persona and F. Sgarbossa. 2016. A new bi-objective approach for including ergonomic principles into EOQ model. *International Journal of Production Research* 54 (9): 2610–2627.
- Barker, L. M., and M. A. Nussbaum. 2011. Fatigue, performance and the work environment: a survey of registered nurses. *Journal of advanced nursing* 67 (6): 1370–1382.
- Battini, D., C. H. Glock, E. H. Grosse, A. Persona, and F. Sgarbossa. 2017. Ergo-lot-sizing: An approach to integrate ergonomic and economic objectives in manual materials handling. *International Journal of Production Economics* 185: 230–239.
- Bautista, J., C. Batalla-García, and R. Alfaro-Pozo. 2016. Models for assembly line balancing by temporal, spatial and ergonomic risk attributes. *European Journal of Operational Research* 251 (3): 814–829.
- Boenzi, F., G. Mossa, G. Mummolo, and V. A. Romano. 2015. Workforce Aging in Production Systems: Modeling and Performance Evaluation. *Procedia Engineering* 100: 1108–1115.

- Burgwal, D. van de. and R. J. O. Vieira. 2014. Environmental disclosure determinants in Dutch listed companies. *Revista Contabilidade & Finanças -USP*, 25 (64): 60–78.
- DGB-Index Gute Arbeit. 2014. Der Report 2013: Wie die Beschäftigten die Arbeitsbedingungen in Deutschland beurteilen - Mit dem Themenschwerpunkt: Unbezahlte Arbeit. [online] <https://www.dgb.de/themen/++co++5f3b9a3c-bc06-11e3-a190-52540023ef1a>. Accessed 13th February 2019.
- Drexl, A., B. Fleischmann, H.-O. Günther, H. Stadler, and H. Tempelmeier. 1994. Konzeptionelle Grundlagen kapazitätsorientierter PPS-Systeme. *Zeitschrift für betriebswirtschaftliche Forschung*, 46 (12): 1022–1045.
- Frey Law, L. A., and K. G. Avin. 2010. Endurance time is joint-specific: a modelling and meta-analysis investigation. *Ergonomics* 53 (1): 109–129.
- Garg, A., K. T. Hegmann, B. J. Schwoerer, and J. M. Kapellusch. 2002. The effect of maximum voluntary contraction on endurance times for the shoulder girdle. *International Journal of Industrial Ergonomics* 30 (2): 103–113.
- Grosse, E. H., M. Calzavara, C. H. Glock, and F. Sgarbossa. 2017. Incorporating human factors into decision support models for production and logistics: current state of research. *IFAC-PapersOnLine* 50 (1): 6900–6905.
- Hax, A. C., and H. C. Meal. 1975. Hierarchical integration of production planning and scheduling. In M. A. Geisler (ed.). *TIMS Studies in Management Science, Vol. 1: Logistics*. North-Holland Publishing Company, 53–69.
- IG Metall. 2018. ERA-Monatsentgelte ab April 2018. [online] www.igmetall.de/download/docs_MuE_ERA_Entgelte_Juni2018_78d3e1848939887f53dcf9506907870bb637c493.pdf. Accessed 13th February 2019.
- Kara, Y., Y. Atasagun, H. Gökçen, S. Hezer, and N. Demirel. 2014. An integrated model to incorporate ergonomics and resource restrictions into assembly line balancing. *International Journal of Computer Integrated Manufacturing* 27 (11): 997–1007.
- Ma, L., D. Chablat, F. Bennis, W. Zhang, B. Hu, and F. Guillaume. 2011. A novel approach for determining fatigue resistances of different muscle groups in static cases. *International Journal of Industrial Ergonomics* 41 (1): 10–18.
- Mossa, G., F. Boenzi, S. Digiesi, G. Mummolo, and V. A. Romano. 2016. Productivity and ergonomic risk in human based production systems: A job-rotation scheduling model. *International Journal of Production Economics* 171: 471–477.
- Nanthavanij, S., S. Yaoyuenyong, and C. Jeenanunta. 2010. Heuristic approach to workforce scheduling with combined safety and productivity objective. *International Journal of Industrial Engineering* 17 (4): 319–333.
- Nerdinger, F. W., G. Blickle, and N. Schaper. 2014. *Arbeits- und Organisationspsychologie: Mit 51 Tabellen*, 3rd edn. Berlin: Springer.
- Neumann, L. and W. Krippendorf. 2016. *Branchenanalyse Bahnindustrie. Industrielle und betriebliche Herausforderungen und Entwicklungskorridore*. Düsseldorf: Hans-Böckler-Stiftung (Study / Hans-Böckler-Stiftung Reihe Praxiswissen Betriebsvereinbarungen).
- Neumann, P. W., and J. Dul. 2010. Human factors: spanning the gap between OM and HRM. *International Journal of Operations & Production Management* 30 (9): 923–950.
- Nijp, H. H., D. G. J. Beckers, S. A. E. Geurts, P. Tucker, and M. A. J. Kompier. 2012. Systematic review on the association between employee worktime control and work-non-work balance, health and well-being, and job-related outcomes. *Scandinavian journal of work, environment & health* 38 (4): 299–313.
- Otto, A., and O. Battaia. 2017. Reducing physical ergonomic risks at assembly lines by line balancing and job rotation: A survey. *Computers & Industrial Engineering* 111: 467–480.
- Trost, M. 2018. Master Production Scheduling With Integrated Aspects Of Personnel Planning And Consideration Of Employee Utilization Specific Processing Times. In *ECMS 2018 Proceedings. 32nd Conference on Modelling and Simulation*, Wilhelmshaven, Germany. May 22 - May 25, 329–335.
- Trost, M., T. Claus, and F. Herrmann. 2019a. Adapted Master Production Scheduling: Potential For Improving Human Working Conditions. In *ECMS 2019 Proceedings. 33rd International ECMS Conference on Modelling and Simulation*, Caserta, Italy. June 11 – June 14, 310–316.
- Trost, M., R. Forstner, T. Claus, F. Herrmann, I. Frank, and H. Terbrack. 2019b. Sustainable Production Planning And Control: A Systematic Literature Review. In *ECMS 2019 Proceedings. 33rd International ECMS Conference on Modelling and Simulation*, Caserta, Italy. June 11 – June 14, 303–309.
- Yeow, J. A., P. K. Ng, K. S. Tan, T. S. Chin, and W. Y. Lim. 2014. Effects of Stress, Repetition, Fatigue and Work Environment on Human Error in Manufacturing Industries. *Journal of Applied Sciences* 14 (24): 3464–3471.

AUTHORS BIOGRAPHY

MARCO TROST is a doctoral student and research associate at the professorship for Production and Information Technology at the International Institute (IHI) Zittau, a central academic unit of the Technical University of Dresden. His e-mail address is: *Marco.Trost@tu-dresden.de*.

PROFESSOR DR. THORSTEN CLAUS holds the professorship for Production and Information Technology at the International Institute (IHI) Zittau, a central academic unit of the Technical University of 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 the head of the Innovation and Competence Centre for Production Logistics and Factory Planning (IPF) at the Technical University of Applied Sciences Regensburg. His e-mail address is: *Frank.Herrmann@oth-regensburg.de*

SIMULATABLE REFERENCE MODELS TO TRANSFORM ENTERPRISES FOR THE DIGITAL AGE – A CASE STUDY –

Carlo Simon and Stefan Haag
Fachbereich Informatik
Hochschule Worms
Erenburgerstr. 19, 67549 Worms, Germany
E-Mail: {simon,haag}@hs-worms.de

KEYWORDS

Reference Models, Horizontal and Vertical Integration, Integration of Structural and Process Organization, Simulation, Petri nets

ABSTRACT

The digital transformation of enterprises forces organizations to handle increasingly complex models of their technical and information systems architectures. Both the existing and the desired architecture must be defined and analyzed in advance of the transformation itself. A purely static analysis of the architecture structure is not sufficient. Also, the dynamic behavior of horizontally and vertically integrated systems in and amongst organizations must be considered.

Reference models help modelers and organizations in finding inspiration to handle this complexity and to develop their own models. Though reference models exist, it was almost never the intention to immediately apply them to actual problems. Then, however, means are needed that help modelers to deeply understand the reference models, to support a decision finding whether a specific reference model can be applied to an organization or not, and to adapt the reference model to a given situation.

This paper explains an approach to use simulatable process models on the base of high-level Petri nets that help in this challenging situation. It describes a modelling and simulation environment called Process Simulation Center (P-S.C) in which processes of reference models can be executed with respect to the information objects needed to control these processes. Also, the structural organization can be defined with the aid of organigrams and can be linked to the process view. Finally, process maps are used to give an overview of the enterprise's processes and their interactions.

INTRODUCTION

Several years ago, (Morgan, 1998) discussed the influence the view on an organization has on its management. One of these views was the functional or mechanical one of an organization as a machine. The current discussion on the digital transformation of enterprises and the possibilities to automatize processes in a standardized way underline the importance of this view.

Today, more and more companies aspire after more automation. The upcoming internet of things and the ability to combine information systems and the real world in a cyber-physical system accelerate this development. But now, each enterprise that chooses this transformation path has to answer the following questions: What are the consequences of the transformation in practice? And which of the many options to automate processes in administration and production is the one which fits best?

Since the answers to these questions are so hard to find for a specific organization, modeling of the existing and the intended organizational structures in advance of the actual transformation seems to be reasonable. However, the next question immediately occurs: Where and how to start with developing a model of a modern organization?

Fortunately, the search for such a model does not have to start from scratch. Many reference models have been developed in the past, sometimes called best practices if they have been developed in a consulting context. These models can be adapted to fit to the according application situation (Becker, Delfmann, Knackstedt, 2007, p. 27).

Because of the long-ranging consequences of transformations, any such model should be regarded with suspicion and it should always be kept in mind that, according to (Stachowiak, 1973, pp. 131-133), any model follows three principles: it is a mapping of the reality, it is a reduction of the reality, and it has a pragmatic purpose. Since these principles might have been weighted differently by the modelers of the reference model than its applicators, a validation is always necessary.

The use of reference models in information systems is an established approach (Thomas, 2005, pp. 484-496). Especially in the German information systems research, this topic has been widely discussed. Exemplarily for this, the work of (Fettke and Loos, 2002) or (Scheer, 2011) is mentioned here. However, the use of simulatable reference models has not been investigated so far.

On this base, the following section explains the research design. Afterwards, different views on enterprises used in information systems development are discussed. The main part of the paper explains how these concepts have been realized in a modelling and simulation environment called Process Simulation Center (P-S.C) with the aid of an example. The advantages of using simulation for the evaluation of the reference model is discussed next. Further developments planned for the future are discussed at the very end.

RESEARCH METHOD

According to (Hevner et al., 2004), there are seven guidelines for Design Science Research. These and their implementation are briefly explained as follows:

Design as an Artifact: A web-based specification and simulation app for processes, the Process Simulation Center (P-S.C), has been developed which is extended by facilities to describe data structures, organizational structures and process maps.

Problem Relevance: In order to advance the quality of a reference model assessment, simulation can play an increasing impact. The applicability of a reference model to the concrete situation in an organization can be tested by simulation with respect to the given resources.

Design Evaluation: Companies already use the mentioned tool. Students of an integrated degree program in logistics adapt reference models of processes to a concrete problem in their company.

Research Contribution: The contribution consists of the creation of a practical application on the theoretical basis of high-level Petri nets combined with other views on organizations and an implementation example of a reference model. The tool, today, is able to map complex organizational structures such that new, simulatable reference models of organizations can be developed.

Research Rigor: The benefits of a simulation approach in opposite to pure visual methods is evaluated in bachelor and master courses and in cooperation with partner companies of integrated degree programs.

Design as a Search Process: The presented prototype is the latest in a series that starts from the initial implementation of the underlying principles and ends in a productive system. Each implementation step has been evaluated.

Communication of Research: The results achieved so far are relevant for both research and practice. They are presented on conferences but also, more illustrative, for practitioners.

VIEWS ON ORGANIZATIONS

Usually, the digital transformation of an enterprise leads to a total or partial replacement of analogue processes by digital, computer executable processes (Wolf and Strohschen, 2018). Figure 1 shows the dimensions of a digital transformation.

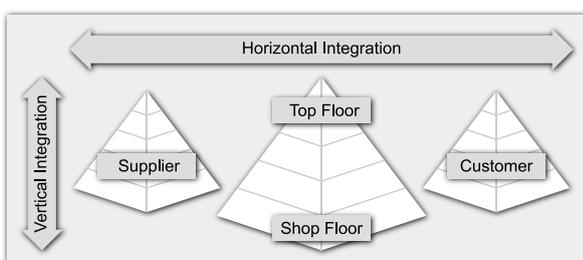


Figure 1: Integration of processes within and across companies adapted from (Simon and Haag, 2020, p. 106)

Horizontal integration occurs along the supply chain, connecting processes of both suppliers and customers with the company in focus.

Vertical integration enables seamless processes from the top floor (i.e. the business processes of the ERP system) to those of the shop floor (i.e. the IT systems for machine automation and in production).

The development of simulatable reference models for both vertically and horizontally integrated (information) systems needs either a shared semantical basis for different modelling languages or a unique modelling and simulation language such as Petri nets. The usefulness of Petri nets for the different tasks has been demonstrated in several publications. Exemplary for a large number in this field (Aalst and Stahl, 2011) can be referred for the application of Petri nets to business process and workflow management and (Zhou and Venkatesh, 1999) can be referred to the application of Petri nets to flexible manufacturing systems.

A Petri net-based approach to integrate processes across the different layers of the automation pyramid according to figure 2 has been demonstrated by (Haag and Simon, 2019). Hereby, also IoT components of the environment have been integrated into the process simulation and execution.

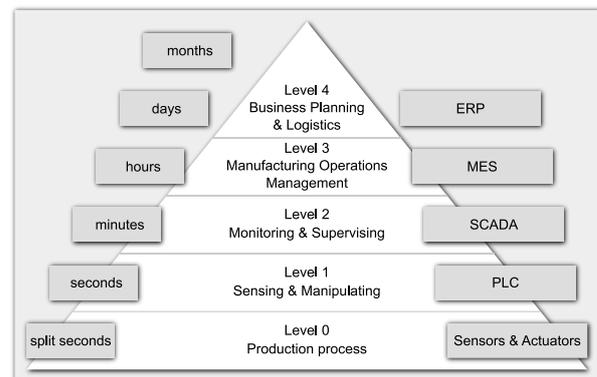


Figure 2: The automation pyramid according to the ISA 95 model (ANSI/ISA, 2005)

Actually, it needs more than processes to describe integrated information systems. Data structures typically defined with the aid of class diagrams (Hay, 2011) are needed to specify information classes that can be instantiated by objects. Combined with the structural organization of an enterprise represented in an organigram, the responsibility and obligation to conduct specific tasks and the right to use the relevant information for this completes the view on an integrated information system. (Scheer, 2000) described this in the architecture of integrated information systems (ARIS) several years ago as shown in figure 3.

The data models and organizational models exemplarily developed in the following sections can be located on the requirements level. Because of the use of Petri nets, the process models can be seen as examples of

all three levels: requirements definition, design specification, and implementation description.

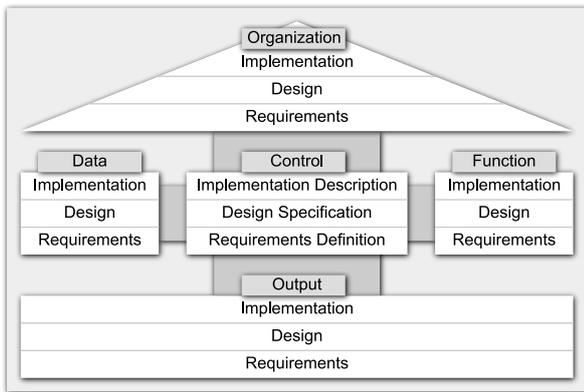


Figure 3: Architecture of integrated information systems adapted and summarized from (Scheer, 2000, p. 1)

In addition to this, process maps offer a further view on enterprises and visualize the interaction of its processes.

TOWARDS SIMULATABLE PROCESS MODELS

A short case which is introduced first might help to give a descriptive demonstration on how to integrate simulatable process models into reference models.

In our example (which has two real-world counterparts which must not be named here), the growing business volume of an online retailer for bicycle products causes an increased amount of returns. In the initial growing phase of the company the returns have been sufficiently handled with Excel. Now the process gets more expansive and complex at the same time. Mainly three variants have to be handled by the employees:

A wrong product has been delivered. Then the wrong item must return into the storage while a replacement is sent to the customer.

A defect product has been delivered. If the product exceeds a certain value and can be repaired, it is given to the own workshop and sent back to the customer after repairing has been finished. If the value is below this border, it is wrecked and a new product is sent to the customer. If it cannot be repaired and the critical value is exceeded, it is sent back to the producer and also in this case a new item is sent to the customer. In this case the claim must be settled with the producer in parallel to sending the product back.

Exceptional situations occur. Sometimes processes fail. A first example is that customers sometimes recognize wrong products after they have opened the packaging. In this case the product cannot return into the storage immediately but needs to be checked and repackaged first. If a damage is recognized throughout this check, the managing director decides on the further procedure. A second example occurs if employees believe that products can be repaired but realize after disassembling that this was wrong. In this case, resending the original product must be examined closely and again the managing director is involved.

A correct process execution depends highly on the distinct know how of the employees. Unfortunately, the growing business forces the company to deploy experienced personnel in core routines of the company like sales and marketing while handling the return process has been rated as a less important process in the past and, thus, is mainly staffed with temporary employees. For these employees, a clear and unambiguous process is needed. This was the initial reason why the company wanted simulatable business processes: the intention was to use them for trainings.

The case includes all elements of the architecture of an integrated information system with respect to figure 3:

Organizational roles inside the enterprise are the managing director, a clerk who decides on how to handle a return, service technicians in the workshop, and employees in the storage. Probably also employees of the purchasing department might be involved. External roles participating in the process are customer and producer.

Data objects are needed concerning the complaint and its current state in the return process, the customer, and the kind of product.

Important **functions** of the process are “check the returned good”, “repair the product”, “deliver a repaired product” or “deliver a substitute” to the customer. The wrecking of a return is omitted in order to reduce the complexity for this paper.

Output documents attended to the process are delivery receipts and supply notes attached to the product.

Although it is possible to model the entire behavior in a single but complex **process** model, it is more reasonable to divide it into the parts “classify a complaint”, “repair a product”, and “replace a product” which need to be reasonably chained. Like others, these processes are objectives of a continuous improvement process. And the two latter processes use standardized support processes like a centralized purchase.

On the base of this scenario, a reference model for returns has been developed with the aforementioned web-based modelling and simulation environment Process Simulation Center (P-S.C). With the aid of a tiny mockup language, models for the different views are specified and can be linked to each other. The following concentrates on the actual models that have been built with the aid of the tool without discussing the detailed specification. The reference model focusses on the mentioned facts and is an excerpt of the entire model which would blow up the scope.

The initial model for the described case is a process map of the mentioned processes as depicted in figure 4. The primary processes that immediately address a customer problem are centered in the middle segment. As can be seen, repair and replacement process follow the classification process. Which one occurs is chosen in the classification process as demonstrated later on. The upper segment is typically reserved for the management processes. The lower segment is reserved for so called secondary or support processes.

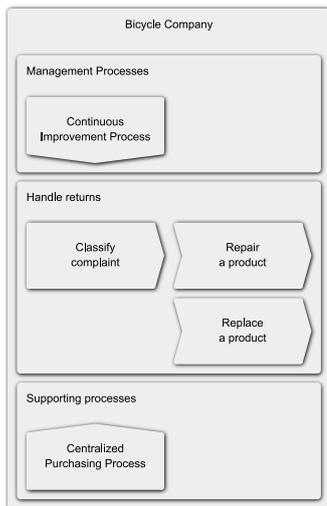


Figure 4: Process map of the exemplary company (rendered with P-S.C)

The second important view is on the organizational structure of the enterprise as shown in figure 5. With respect to the limited space, central departments like purchasing and central storage are bundled graphically into a single unit. The managing director for the supply chain is seen as part of a general management team which is not further described here. External roles involved in the process are not shown in the (internal) organigram.

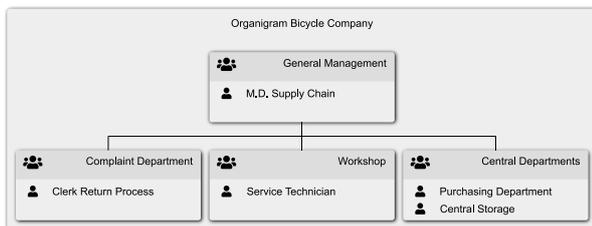


Figure 5: Organizational structure of the exemplary company (rendered with P-S.C)

As part of the specification language for processes it is possible to define datatypes and associate them with places – a Petri net element to store information. Process and data flow can be combined in a single model with the aid of this concept as explained in the next two sections.

Figure 6 depicts a simple data structure for complaints, customers and products. Each return is uniquely associated with a customer and may contain one or more products of given product types. The same product type might occur in several returns caused by the same or by different users. Currently, the company does not distinguish between the individual products but is completely satisfied with the product type level.

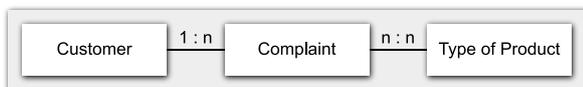


Figure 6: Simple data structure for handling returns

A more detailed view on the attributes used for the distinct data objects will be given in the next section when this information is used to control the business process.

SIMULATABLE PROCESS MODELS

In this section, exemplary processes of the case are modelled as Petri nets which we assume to be known (Reisig, 2013). In a stepwise approach, the core process structure is defined first and extended by data-flow concepts afterwards.

The authors have chosen Petri nets as modeling language for several reasons:

An **execution semantic** is formally defined on base of the so-called firing rule for transitions. This semantic can be seen as a dominant reason for the influence of Petri nets on the development of workflow management systems in the past (Aalst and Hee, 2002).

Distinctive information objects, named tokens, that flow through Petri nets are crucial for process control and to predict the quantitative process behavior. The best-known approaches are Predicate/Transition Nets (Pr/T-nets) as defined by (Genrich and Lautenbach, 1981) and Coloured Petri Nets as defined by (Jensen, 1992).

As tokens may also contain **time information** schedules can be examined or planned.

The authors chose to use Pr/T-Nets in the following.

Figure 7 shows a first version of process “Classify complaint” and needs to be explained since the Process Simulation Center draws Petri nets in a modern, compared to classic Petri nets unusual way.

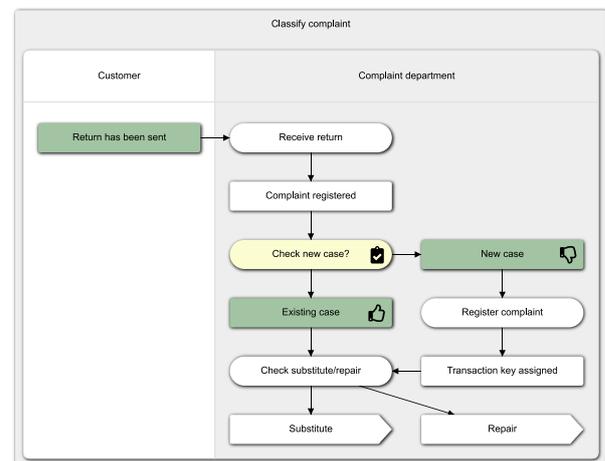


Figure 7: Initial version of process “Classify complaint” (rendered with P-S.C)

Places and transitions are stretched such that the label can be put into the nodes. Places have round left and right borders (like “Receive return”) while transitions are rectangles (like “Complaint registered”) with one exception: models can be chained such that the firing of a transition opens another net where simulation can be continued. These special transitions are drawn as process arrows (like “Repair”).

Enabled transitions and marked places are graphically emphasized: An enabled transition is drawn green and a marked place is drawn either green or yellow. The latter is chosen if the place is the reason for a conflict situation where two transitions compete against each other concerning tokens. Moreover, transitions and places can be supplemented by icons dependent on their enabling status or the number of tokens on the place.

Places are chosen to describe activities that happen or the coming to a decision. Important events or the results of decisions are depicted as transitions. Swim lanes are used to express responsibilities.

The model depicted in figure 7 describes the following process: after the customer has sent the return, it is received by the complaint department and registered first. Due to some mishaps in the past, it is checked whether this really is a new case or whether it has been registered already (the situation indicated by the shown marking). After a new complaint has possibly been registered, the clerk of the complaint department decides on whether the return must be substituted or repaired. Dependent on this decision, the next process begins.

In the next section, this model is enriched by information concerning the returned goods and their processing. This is done for two reasons: it can be checked how to automatize the process and it can be the base for the simulation of different process behavior after a digital transformation process has occurred.

INFORMATION CONTROLLED SIMULATION

In Pr/T-Nets, places are typed and can be marked with tuples that match this type structure. Hence, places correspond to tables of a database and their tokens to the records within a database table.

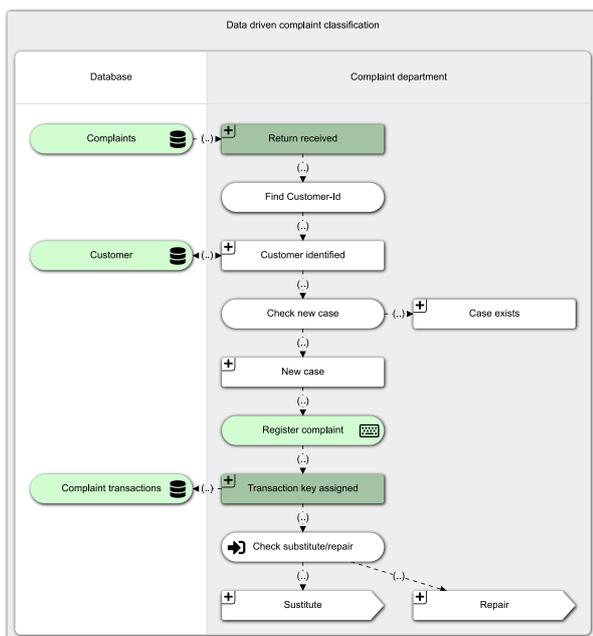


Figure 8: Data driven complaint classification process (rendered with P-S.C)

In figure 8, all places are typed. The places in swim lane “Database” contain the information concerning the incoming complaints, the customers (in order to identify the right customer-id), and complaint transactions (used for the documentation of all handled complaints). Figure 9 is a screenshot showing the content of latter place.

Ticket	Timestamp	CustomerID	Classification	Priority	Description	Status
27	2020-02-19T18:45:36	1265	substitute	2	Wrong color	finished
28	2020-02-19T19:35:29	2124	repair	1	Front light defect	finished
29	2020-02-20T04:43:15	4711	repair	1	Brake defect	finished
30	2020-02-21T11:18:58	3486	substitute	2	Wrong seat	open
31	2020-02-21T11:36:48	4513	repair	3	Scratches	finished
32	2020-02-21T14:14:34	2124	substitute	2	Wrong color	open
33	2020-02-24T13:34:43	3781	substitute	3	Wrong seat	open
34	2020-02-24T17:19:31	1934	substitute	3	Kickstand to short	open
35	2020-02-24T22:42:57	1969	repair	1	Back light defect	open

Figure 9: Screenshot of the marking of place “Complaint transactions”

Also, the places of swim lane “Complaint department” are typed and carry the information concerning the current complaint handled. Clicking the plus-symbol enlarges the transitions in order to show transition conditions and select-criteria that control the sequence in which the complaints are handled. Different strategies can be followed: a first-come-first-serve strategy handles the complaints corresponding to their automatically generated ticket number. Another strategy could be to follow the priority criterium instead.

Finally, place “Check substitute/repair” has an extra symbol. It indicates that the information concerning the current complaint is exported to the next net opened after transition “Substitute” or “Repair” are fired. The idea behind this concept has been explained in (Simon, 2018) already. The selection which of these two transitions fires depends on the value of attribute “Classification”.

The nets “Repair a product” and “Replace a product” are defined in a similar way.

FROM SIMULATION TO EXECUTION

The definition of unambiguous business rules that standardize the process execution is crucial for process automation and hence for the digital transformation of enterprises. The conflict situation shown in figure 7 is a good example for a process model which is still incomplete and the Process Simulation Center helps to find such situations. Whenever the simulation must stop because a conflict occurs which can only be solved by the operator of the software, the same would occur in the real-world process.

To find out how well the process is automated so far, the P-S.C can work with real world data which can be imported with the aid of a CSV interface.

In the concrete case of the demonstrated return process, students of a logistics company had to take the reference process (which – in practice – is of course more elaborated) and adapt it stepwise to a real-world return process in their company. For this they had to adapt the model on all levels concerning the architecture of integrated information systems shown in figure 3:

Organizational roles are labeled differently and in addition to the clerks who decide on how to proceed with the returned goods there are also employees who transport the goods from the loading zone to the complaint department and further on to the main storage. A supervisor coordinates the work of the team.

Data objects are used differently in several aspects since in the real-world example the logistics company works as a sub-contractor for a producer. Hence, the internal data must always be combined with the purchaser's data. As a further consequence, returns are typically bundled into larger lots that must be distinguished in the information perspective.

Furthermore, the logistics company is interested in the consequences of applying different strategies to the return process which is currently reorganized. Therefore, each step within the process is assessed concerning the average processing time. This time information is also considered in the complaint transaction record.

More **functions** exist in the real world than shown in figures 7 and 8. As described in the beginning, it is also possible that some returned goods are wrecked since it is not reasonable to restore low-grade goods. In order to increase the readability of this paper, these details have been omitted.

Output documents of the real-world process have been used as a source to complete the data objects needed in the simulation of the said process.

Although the same **processes** have been considered as described here, the management was also interested in measuring the process behavior. For this reason, the students established key performance indicators (KPI) to measure the process and defined them with the aid of a standardized KPI-sheet.

The application of the simulatable reference model in the modelling environment demonstrated the advantage of this approach. The stepwise refinement approach enabled the students to define not only a simulatable but also executable process definition within four weeks. Moreover, they could predict the behavior of optional reorganizations in advance.

At the beginning of this paper, the following questions have been formulated: What are the consequences of a digital transformation in practice? And which of the many options to automate processes in administration and production is the one which fits best? The simulatable models helped the students and their company to precisely answer these questions for a concrete field of application in advance. The application to further processes is planned for the future.

FURTHER RESEARCH

The current state of research opens manifold further research questions:

The **number of reference models** is increased constantly in order to apply this approach to a larger number of cases. Especially branches of industry partners of the university are considered first. The former section also showed a demand for inter-organizational reference models.

Another approach is to increase the **number of reference models** with respect to established ERP systems and to develop simulatable processes for them.

In a combination of the research presented here and research conducted in the past, the **automation of manufacturing processes** is also an aim of the future work. The purpose of this is to make vertical integration processes more illustrative for students and partner companies.

Finally, it is the aim to **combine the research and teaching of modelling and simulation techniques**. Modern pedagogic approaches called problem-based learning and research-oriented learning try to substitute classical ways of fact-based learning and to develop and train the scrutiny of students. Simulatable models help students to explore the world they have to work in and to derive an appropriate solution for a given problem from an existing one (Simon and Haag, 2020).

REFERENCES

- van der Aalst, W. and K. van Hee. 2002. *Workflow Management – Models, Methods, and Systems*. MIT Press, Cambridge, MA.
- van der Aalst, W. and C. Stahl. 2011. *Modeling Business Processes: A Petri net - Oriented Approach*. MIT-Press, Massachusetts.
- ANSI/ISA. 2005. *Enterprise Control System Integration, Part 3: Activity Models of Manufacturing Operations Management*.
- Becker, J.; Delfmann, P. and R. Knackstedt. 2007. "Adaptive Reference Modeling: Integrating Configurative and Generic Adaptation Techniques for Information Models". In: *Reference Modeling*. Becker, J. and Delfmann, P (Eds.). Springer, 27-58.
- Fettke, P. and P. Loos. 2002. "Der Referenzmodellkatalog als Instrument des Wissensmanagements: Methodik und Anwendung". In: *Wissensmanagement mit Referenzmodellen*. Becker J.; Knackstedt R. (Eds.). Physica, Heidelberg, 3-24.
- Genrich, H. J. and K. Lautenbach. 1981. "System Modelling with High-Level Petri Nets". *Theoretical Computer Science* 13, 1, 109-135.
- Haag, S and C. Simon. 2019: "Simulation of Horizontal and Vertical Integration in Digital Twins". In: *33rd International ECMS Conference on Modelling and Simulation*, 284-289.
- Hay, D. C. 2011. *UML and Data Modeling: A Reconciliation*. Dorset House Publishing, New York.
- Hevner, A. R.; March, S. T.; Park, J. and S. Ram. 2004. "Design Science in Information Systems Research". *MIS Quarterly* 28, 1 (Mar), 75-105.
- Jensen, K. and L. M. Kristensen. 2009. *Coloured Petri Nets*. Springer, Berlin, Heidelberg, Germany.

- Morgan, G. 1998. *Images of Organization - The Executive Edition*. Berrett-Koehler, San Francisco.
- Reisig, W. 2013. *Understanding Petri Nets*. Springer, Wiesbaden.
- Scheer, A.-W. 2000. *ARIS – Business Process Modeling*. 3rd edition, Springer, Berlin.
- Scheer, A.-W. 2011. *Wirtschaftsinformatik: Referenzmodelle für industrielle Geschäftsprozesse*. 7th edition, Springer, Berlin.
- Simon, C. 2018. “Web-based Simulation of Production Schedules with High-level Petri Nets”. In: *Proceedings of the 32nd European Conference on Modeling and Simulation* (Wilhelmshaven, May 22-25), 275-281.
- Simon, C. and S. Haag. 2020. “Digitale Zwillinge modellieren und verstehen“. In *Joint Proceedings of Modellierung 2020* (Vienna, Feb 19), 101-112.
- Stachowiak, H. 1973. *Allgemeine Modelltheorie*. Springer, Wien. Cited after (Thomas, 2005, 8-10)
- Thomas, O. 2005. “Understanding the Term Reference Model in Information Systems Research: History, Literature Analysis and Explanation”. In *Business Process Management Workshops*. LNCS 3812. Springer, Nancy, France, 484-496.
- Thomas, O. 2005. “Das Modellverständnis in der Wirtschaftsinformatik: Historie, Literaturanalyse und Begriffsexplikation“. Techn. Ber. 184, Universität des Saarlandes, Institut für Wirtschaftsinformatik, Saarbrücken.
- Wolf, T. and J.-H. Strohschen. 2018. “Digitalisierung: Definition und Reife“. In *Informatik-Spektrum* 41, 1 (Feb), 56-64.
- Zhou, M. C. and K. Venkatesh. 1999. *Modeling, Simulation, and Control of Flexible Manufacturing Systems – A Petri net Approach*. World-Scientific, Singapore.

AUTHOR BIOGRAPHIES



CARLO SIMON studied Informatics and Information Systems at the University of Koblenz-Landau. For his PhD-Thesis, 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 Provdadis 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 modelling 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.

SIMULATION-BASED EVALUATION OF RESERVATION MECHANISMS FOR THE TIME WINDOW ROUTING METHOD

Thomas Lienert

Florian Wenzler

Johannes Fottner

Chair of Materials Handling, Material Flow, Logistics

Department of Mechanical Engineering

Technical University of Munich

Boltzmannstrasse 15, 85748 Garching, Germany

Email: thomas.lienert@tum.de, florian.wenzler@tum.de, j.fottner@tum.de

KEYWORDS

Automated Warehouses, Mobile Robots, Time Window Routing Method, Discrete Event Simulation

ABSTRACT

Automated warehouses operated by a fleet of robots offer great flexibility, since fleet size can be adjusted easily to throughput requirements. Furthermore, they provide higher redundancy compared to common solutions for automated storage and retrieval systems.

On the other hand, these systems require more complex control strategies to run robustly and efficiently. Special routing and deadlock handling strategies are necessary to avoid blocking and collisions among the robots.

In this contribution, we focus on the time window routing method, an approach for avoiding deadlocks by reserving routes in advance. We present and discuss different reservation mechanisms that are evaluated by the means of simulation.

INTRODUCTION

Automated warehouses that are run with a fleet of mobile robots for the part-to-picker order picking have become the subject of intensive research since different decision problems need to be resolved during operation of these systems.

Basically, these systems consist of a rack system containing storage items and a fleet of robots moving within the storage area. The robots use a rectangular grid of paths to fulfil storage and retrieval requests. We refer to these systems as mobile-robot-based warehouses. There are several types of mobile-robot-based warehouses that differ in regard to their storage systems.

Robotic mobile fulfilment systems (RMFS) consist of a single storage tier, where items are stored on shelves on the ground. Robots travel underneath these shelves, lift them and bring them to the picking zone that is located somewhere near the storage area (Azadeh et al. 2018a).

In contrast, shuttle-systems consist of several tiers that are connected by lifts. These lifts link the storage system to the picking zone and enable vertical movements of the robots that are denominated as shuttles in this context (Tappia et al. 2018).

In another type, robots move horizontally as well as vertically within an aisle placed between two single-deep storage racks. Picking zones are located at one or at both ends of each aisle (Azadeh et al. 2018b).

Although these types differ slightly, they provide the same benefits relative to common stacker-crane-based storage systems. They are easily scalable. The whole system can theoretically be run with a single robot and if a higher throughput is needed, more and more robots can be added. The layout can be changed flexibly and the system can be enlarged easily. Furthermore, a required sequence can be established within the storage system (Lienert and Fottner 2018) and high redundancy can be provided as long as suitable failure-handling strategies are used (Lienert et al. 2019). In the literature, mobile-robot-based warehouses are widely discussed. Among others, storage assignment (Boysen et al. 2019), order batching (Boysen et al. 2017), dispatching (Yuan and Gong 2017), battery charging and swapping (Zou et al. 2017) as well as dwelling strategies for idle robots (Roy et al. 2016) are addressed.

In this contribution, we focus on the routing and deadlock-handling, more precisely on the time window routing method. This approach enables conflict-free routing of robots by reserving the path to be travelled in advance. Acceleration and deceleration processes are usually neglected when the time window routing method is applied. We present different reservation mechanisms that include acceleration and deceleration processes and fit different requirements regarding communication between the robots and the material flow control. The remainder of this paper is organized as follows. In the next section, we briefly introduce the time window routing method. We subsequently describe and discuss three different reservation mechanisms that are compared with a simulation and taking into account an RMFS before we conclude our work.

TIME WINDOW ROUTING METHOD

In mobile-robot-based warehouses, robots move using the layout given by storage-aisles and cross-aisles. Since several robots are operating in the system at the same time, traffic must somehow be controlled and deadlocks or even worse, collisions must be avoided. In general a deadlock describes a situation where one or more

processes are blocked forever because the requests for resources by the processes can never be satisfied (Kim et al. 1997). In the context of routing robots, the processes correspond to the execution of the routes and the resources for the layout of segments along these routes. For example, a deadlock occurs if two robots driving in opposite directions meet each other within a storage aisle.

The approach of the time window routing method avoids deadlocks by reserving the path for a robot from its current situation to the destination in advance. On each layout segment that needs to be travelled along this path, a time window is blocked, during which the layout segment is claimed exclusively by a robot and during which the layout is not available for the movement of any other robot. Since time windows on neighbouring layout segments of a route overlap each other, robots can move safely through the layout.

To apply this method, the layout is represented by a graph. Each node corresponds to a layout segment, whereas the edges give information about predecessors and successors of the nodes. For each node, there is a time line with reserved and free time windows (figure 1). $f_{i,l}$ denominates the l^{th} free time window on node i .

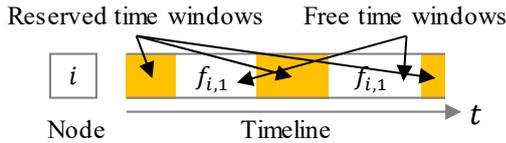


Figure 1: Reserved and free time windows on a node i

In the case that a robot has to be routed from its current position to a given destination, the method searches for a route through the free time windows on the nodes using an A*-algorithm. Once a conflict-free route is found, the corresponding time windows are reserved and the robot can start moving. From the point of time that the idea of this method was introduced first by Kim and Tanchocco (Kim and Tanchocco 1991), it has been applied in different contexts as the routing of automated guided vehicles in container terminals (Stenzel 2008) for the organization of taxi traffic at airports (Bussacker 2005) and for managing a fleet of robots in an RMFS (Hvězda et al. 2018). For more detailed insights into the modelling of the layout as a graph we refer to (Lienert and Fottner 2017).

The time window routing method can be applied in systems that use centralized material flow control as well as in systems that are operated by a multi agent system.

When it comes to the execution of a reserved route, deadlocks or collisions among the robots threaten to occur even though the routes are theoretically conflict-free. Robots might be delayed – due to several reasons – and not match their reserved time windows. Therefore it is essential that the node’s crossing order of the robots based on the conflict-free schedule is maintained (Maza and Castagna 2005). Hence, a robot is only allowed to travel the next node along its reserved route if it has reserved the next time window on that node.

Figure 2 shows an example that clarifies this approach. At timestamp T the reserved time window of robot r_2 begins on node j . Robot r_1 is delayed. According to the planning, it should reside on node i , but it has not yet passed node j , k and l .

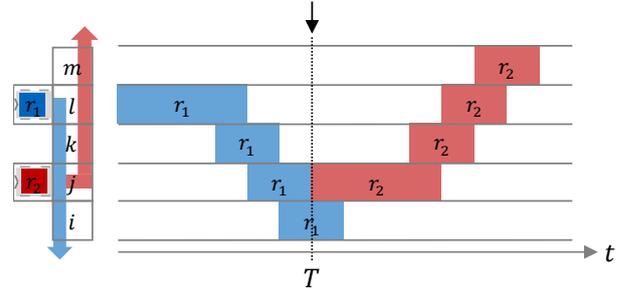


Figure 2: Robot r_1 is delayed and a deadlock might occur

If robot r_2 nevertheless continues with the execution of its route, both robots will face each other somewhere in between the nodes j and l and a deadlock will occur.

Note that robots must also not enter a node before their reserved time window has started. In the example in figure 3, robot r_2 has just entered node j at timestamp T and before the corresponding time window has started. Routing robot r_1 will lead to a feasible route traversing nodes l , k and j before the reservation of robot r_2 begins (as in figure 2). Once again, a deadlock is likely to happen.



Figure 3: Robot r_2 is early and a deadlock might occur

Maintaining the node’s crossing order is easily realizable if acceleration and deceleration processes are neglected since a robot can stop immediately in case it is not allowed to enter the next node. However, if acceleration and deceleration processes are taken into account, the robust execution of a route must be implemented with some sort of lookahead.

In a previous work, we modified the time window routing method to incorporate acceleration and deceleration processes. During the planning, so-called “segments” are created that describe movement of a robot over several nodes in a straight line. The computed route is executed segment by segment, respecting the node’s crossing order (Lienert et al. 2018a). Figure 4 shows the creation of a segment during the planning phase. Starting with a free time window on the node i , the algorithm checks whether the free time window on node j is reachable. In that case, the segment is – if possible – extended node by node.

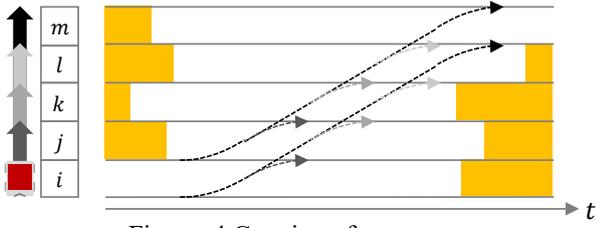


Figure 4: Creation of a segment.

To extend a segment, none of the time windows within the segment can violate any existing reservation. In the example in figure 4, all free time windows are reachable since no trajectory touches any existing reservation. Each trajectory is described by a pair of arrows that represent the movements of the front and rear of a robot. Free time windows that are reachable are candidates for the first node of a new segment in a later iteration. For a comprehensive description of the overall routing procedure, we refer to (Lienert et al. 2018a).

RESERVATION MECHANISMS

In this section, we focus on the part of the algorithm that examines the reachability of free time windows from a free time window at a specific time stamp, taking into account three different reservation mechanisms. These mechanisms differ on the one hand in regard to the requirements of the communication between robots and centralized material flow control. On the other hand, the required length of the reserved time windows of a segment differ, which leads to varying resource utilizations.

Mechanism 1: Triangle

First, we assume that once a robot starts with the execution of a segment, it has to be ensured that the robot can finish that segment without any interference due to other robots that are late. In settings where communication between robots and central material flow control cannot be guaranteed at any given time in real-time, this procedure is necessary to avoid collisions and deadlocks.

The reservations of a segment are preponed so they start with the departure of the robot from the first node of the segment. Time windows are deleted as soon as a robot has left a node completely. As a result, reservations of a segment form a triangle (see figure 5).

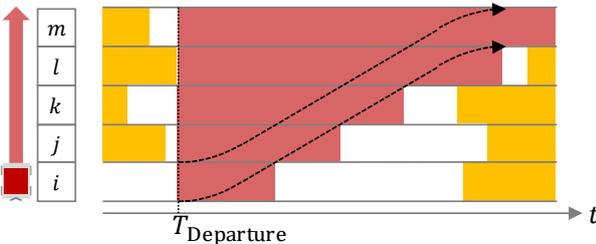


Figure 5: Reservations of a segment display the form of a triangle

We assume free time window $f_{i,l}$ to be the first time window of a segment and free time window $f_{k,p}$ the free time window whose reachability is analyzed (figure 6).

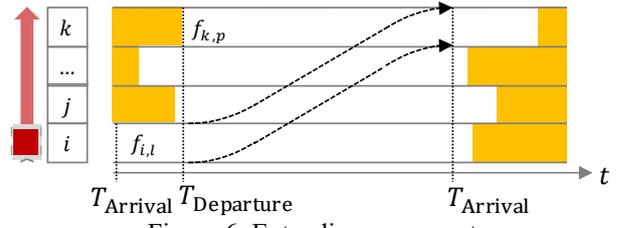


Figure 6: Extending a segment

The flowchart in figure 7 shows the procedure of checking the reachability of a free time window (FTW) in detail.

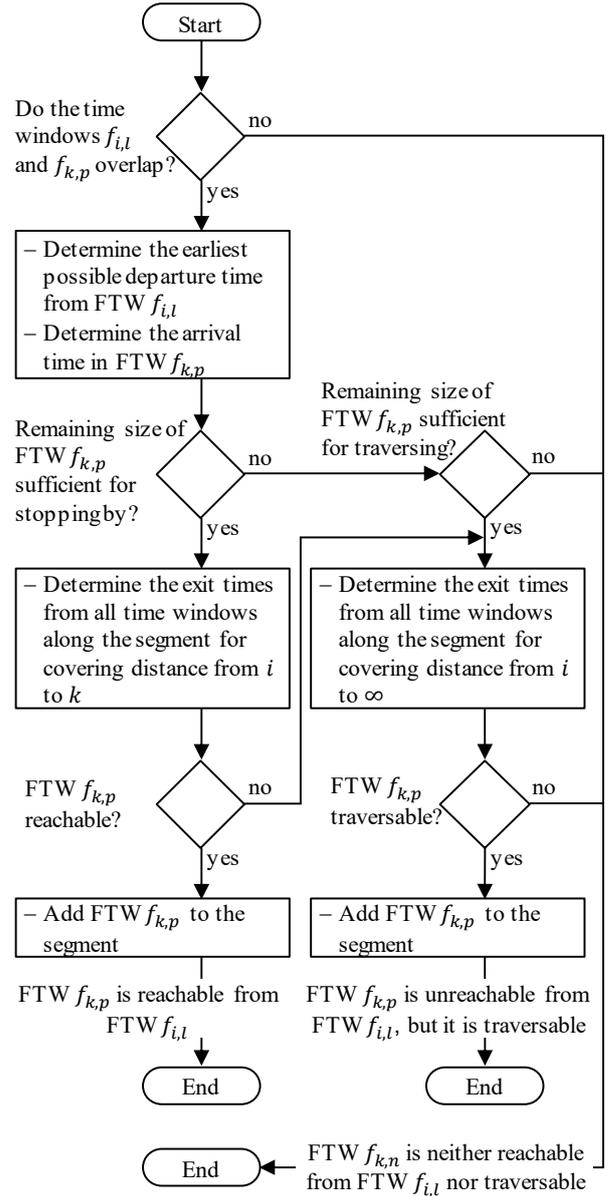


Figure 7: Reachability check of free time window $f_{k,p}$

First of all, it is mandatory that both of the free time windows $f_{i,l}$ and $f_{k,p}$ overlap each other. In that case, the earliest departure time from free time window $f_{i,l}$ can be determined. The departure cannot take place before the robot arrives on node i and not before any of the free time windows of the segment including the free

time window $f_{k,p}$ start. Next, the arrival time in free time window $f_{k,p}$ can be calculated taking into account the distance as well as the robot's parameters.

The remaining size of free time window $f_{k,p}$ after the arrival must be sufficient to leave the node completely before the next reservation starts. In that case, all of the required reservations on all nodes of the segment can be determined. If no existing reservation on any node of the segment is violated, the free time window $f_{k,p}$ is reachable and the segment can be extended. In a later iteration, free time window $f_{k,p}$ serves as a starting time window of another segment.

If the remaining size of free time window $f_{k,p}$ is not sufficient for stopping by, node k might be traversed by the robot before the next reservation starts. In that case, all required time windows on all nodes of the segment can once again be determined assuming a movement of infinite length. If no existing reservation on any node of the segment is violated, free time window $f_{k,p}$ is traversable and the segment can be extended (as in figure 8).

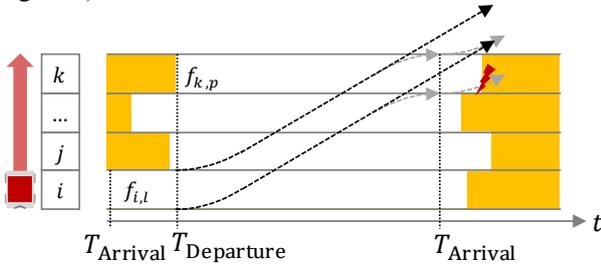


Figure 8: Remaining size of free time window $f_{k,p}$ does not allow an intermediate stop on node k , but traversing the node.

In case the remaining size of the free time window is sufficient for stopping by but a previous reservation is violated, traversing the node without violating any reservation might once again be possible (as in figure 9).

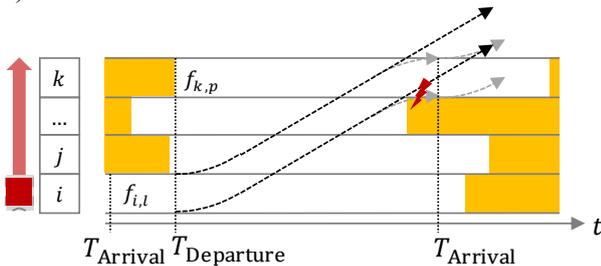


Figure 9: An existing reservation does not allow an intermediate stop on node k , but does allow traversing the node.

Note that in both cases (figure 8 and figure 9), free time window $f_{k,p}$ is not a potential candidate for the first time window of another segment.

During execution of a reserved route, a robot is only allowed to start a segment if the robot is not early and if the robot is to travel all nodes of the segment next, taking into account the node's crossing order.

Mechanism 2: Rectangle

The second reservation mechanism is even more restrictive. Time windows are deleted only after a segment is finished and the robot comes to a standstill. As a result, reservations of a segment form a rectangle (see figure 10).

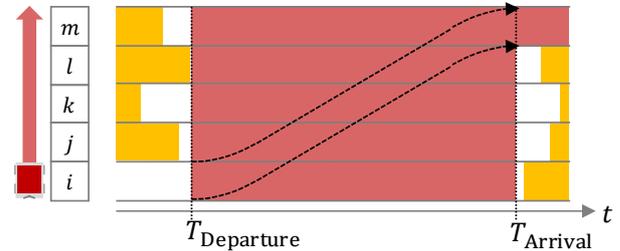


Figure 10: Reservations of a segment display the form of a rectangle

This approach is suitable for applications that do not allow the deletion of time windows in real-time. Once a robot starts with the execution of a segment, no information of the robot's position is available until the robot finishes the segment.

The procedure to check reachability corresponds to the flow chart in figure 7 apart from determining the necessary reservations. These once again begin with the departure, but end with the arrival on the last node of the segment.

Mechanism 3: Stairs

The third reservation mechanism enables a more efficient use of resources but requires communication between robots and centralized material flow control more often. Time windows do not begin at the departure, but there is a buffer before each reservation that enables a safety deceleration in case a robot with a previous reservation is delayed. As a result, reservations of a segment form a stairs pattern (see figure 11).

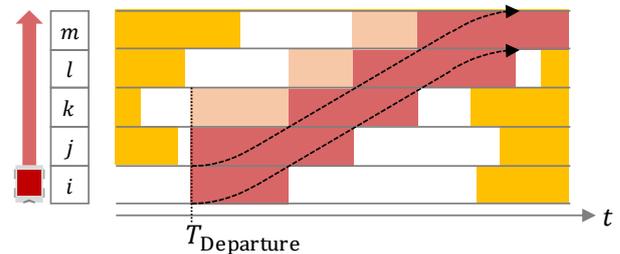


Figure 11: Reservations of a segment display the form of a stairs pattern.

To analyze reachability, the procedure differs slightly and is shown in the flow chart in figure 14. Note that it is not necessary that both free time windows $f_{i,j}$ and $f_{k,p}$ overlap each other.

First of all, the start of the reservation of the free time window $f_{k,p}$ has to be determined. Starting from the penultimate node of the segment on which the robot has to stop in case a preceding robot is delayed, the

distances between the nodes are added up until the deceleration distance is reached. Let the node a be the node in which this summed distance reaches or exceeds the deceleration distance. The reservation of the free time window $f_{k,p}$ must begin as soon as the stopping position of the node a has been passed (figure 12).

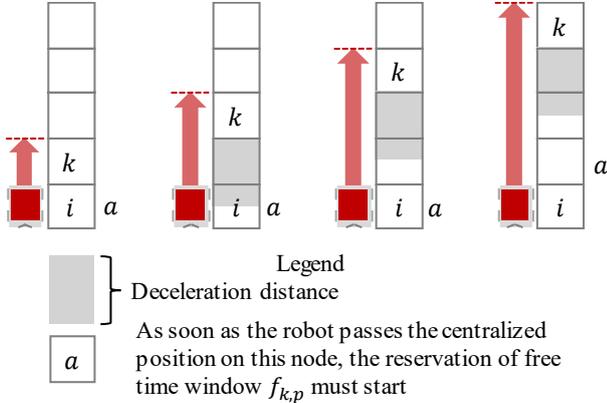


Figure 12: Determination of the start of the reservation of the free time-window $f_{k,p}$

If the calculated start does not fall in the free time-window, it is checked whether it is possible to postpone the departure, so that the start coincides with the start of the free time window. Otherwise the free time window $f_{k,p}$ is not reachable.

The arrival time can be determined next. The remainder of the procedure corresponds to the one described in the flowchart in figure 7 apart from the determination of the necessary reservations. For each free time window in the segment, the start of the reservations needs to be recalculated (as described above) since the departure from the first free time window $f_{i,j}$ of the segment might have been shifted. Reservations end as soon as a robot has completely left a node.

In case a robot that reserved a preceding time window on any node of a segment is delayed, the additional buffer allows the robot to decelerate and to stop safely. In the example in figure 13, robot r_1 is delayed at timestamp T and its reserved time window on node m is not yet deleted, robot r_2 is not allowed to enter node m and starts decelerating, coming to a standstill on node l .

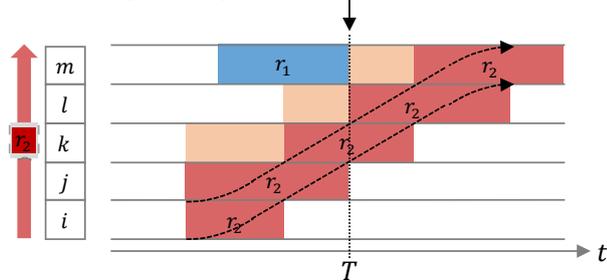


Figure 13: Robot r_1 being late leads to an unplanned intermediate stop on node l for robot r_2 .

Note that if acceleration and deceleration processes are neglected, additional buffers disappear.

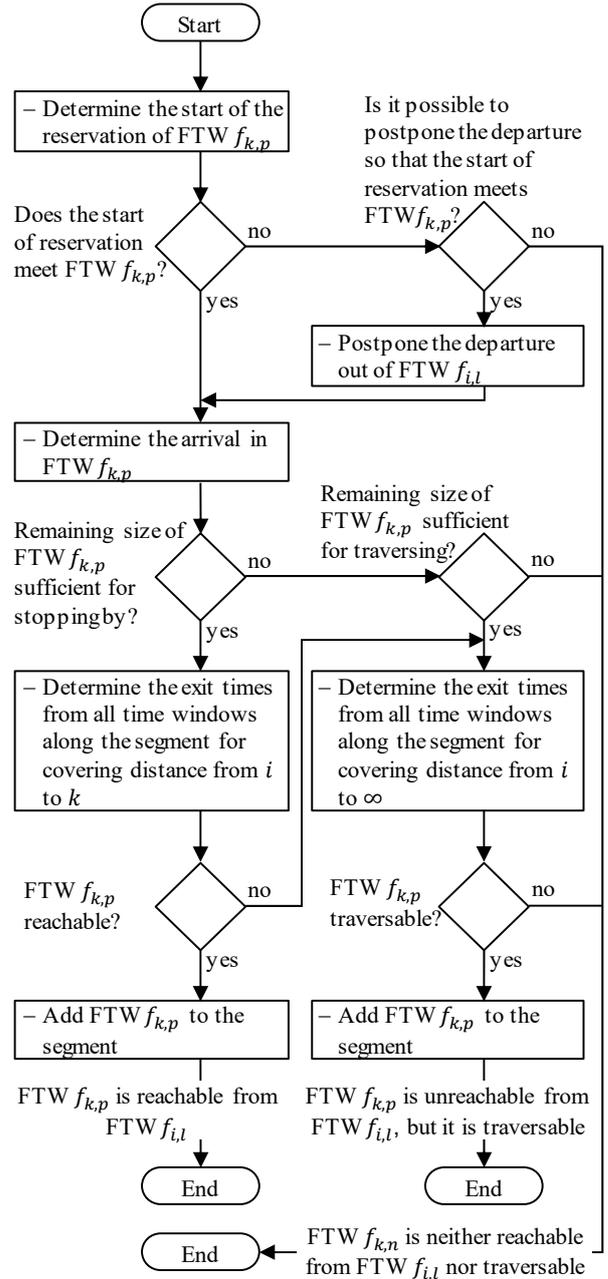


Figure 14: Reachability check of free time window $f_{k,p}$

Theoretically there is another reservation mechanism where reservations start with an additional buffer but end with the arrival on the last node of the segment. However, we assume that if real-time communication is enabled (which is necessary for the stairs mechanism), time windows can be deleted immediately after a node has been left.

SIMULATION STUDY

In this section, we compare the previously described reservation mechanisms by performing a simulation study, considering an RMFS.

There is a trade-off regarding the segment length using the triangle and rectangle reservation mechanisms. Short segments entail frequent stops and accelerations of the robots. Longer segments enable robots to achieve

maximum speed and reach their destinations with fewer intermediate stops. But in this case, the reserved time windows on the nodes are larger and the nodes are blocked longer for other vehicles. As a consequence, the optimal maximal segment length has to be determined for each number of robots. This is done first before the mechanisms are compared to each other.

Considered System

We apply the strategies to a fleet of robots moving within an RMFS with 336 storage locations that are arranged in seven double rows divided by storage aisles. There are two cross-aisles located at one third and at two thirds of the aisle length. All aisles can be used for bi-directional traffic. There are four picking areas with five picking places, with each arranged in front of the storage system. In front of these places, there are two unidirectional cross-aisles. A replenishment area, where empty racks are refilled, is located on the opposite side of the storage area. Robots are dedicated to a picking zone and perform three different cycles to maintain the material flow between storage locations, picking area and replenishment area. For a more detailed description of the system, we refer to (Lienert et al. 2018b).

We implemented the RMFS using the Tecnomatix Plant Simulation discrete event simulation environment.

Parameter Settings

We vary the number of robots, starting with four robots (one for each picking zone) and increasing this up to 60 robots working in the system in steps of four, and repeat the experiments for each reservation mechanism. With both the triangle and rectangle reservation mechanisms, we limit the segment length to a certain number of nodes. We start with a maximum segment length of only one node, meaning that robots only move node by node, stopping at every single node along their routes. We vary the maximum segment length increasing it up to 30 nodes. All of the remaining parameters, such as the robot's acceleration and maximum speed, remain the same. Simulation time is set to 24 hours. No warm-up time is taken into account, as the goal of the simulation study is to compare the reservation mechanisms. We conduct five replications for each parameter setting.

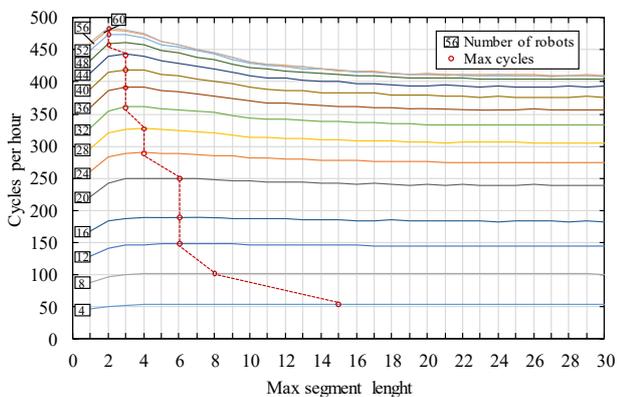


Figure 15: Throughput using the triangle reservation mechanism

Results

First we analyse the trade-off between shorter and longer segments. Figure 15 shows the throughput reached with the corresponding parameter setting using the triangle reservation mechanism.

As can be seen, the more robots that are operating in the system, the shorter the optimal segment length becomes. Figure 16 shows the throughput reached using the rectangle reservation mechanism.

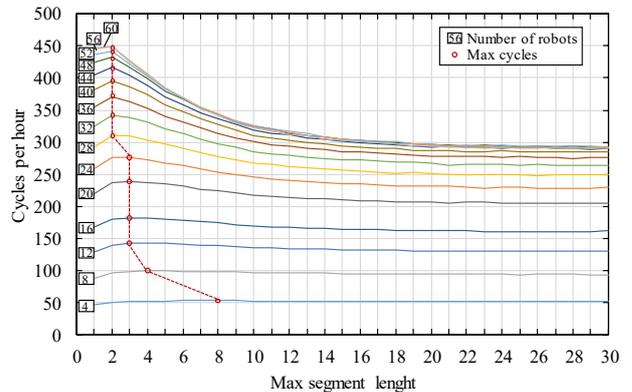


Figure 16: Throughput using the rectangle reservation mechanism

A similar behavior can be observed. However, the optimal maximum segment length is even shorter. That is expectable, since nodes are reserved for a longer time periods than using the triangle mechanisms.

Finally we compare these maximums to the stairs reservation mechanisms. Figure 17 shows the throughput reached.

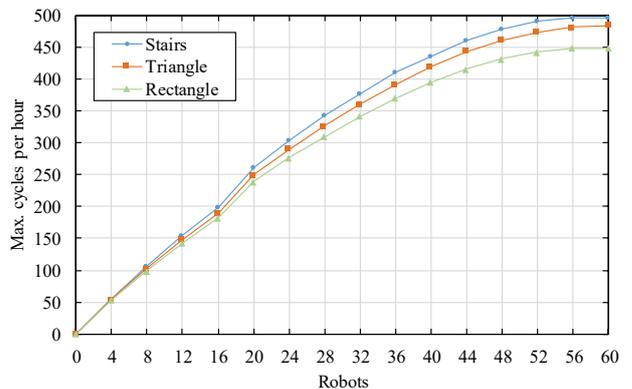


Figure 17: Throughput for varying number of robots

All curves look similar with a linear increase and a small knee between 16 and 20 vehicles before reaching saturation with 56 robots. Since each robot is assigned to a certain picking zone, the number of robots per picking zone with 20 robots equals the number of picking places, and a different strategy is used for the supply of the picking zone (see Lienert et al. 2018b).

For a small number of robots, the throughput of the different mechanisms is equal. In fact, using a single robot in the whole system yields the same throughput, no matter which reservation mechanism is used.

However, the more robots that are operating in the system, the more evident the difference becomes. As expected, the stairs mechanism reaches the highest throughput followed by the triangle one. Reaching a throughput of 400 cycles per hour using the stairs mechanism requires 36 robots, 40 robots using the triangle mechanism and as many as 44 robots using the rectangle mechanism.

Note that saturation is reached at different levels. Using the stairs reservation mechanism yields the highest throughput, whereas the triangle mechanism reaches 97.6% and the rectangle mechanism 90.4% of that throughput.

CONCLUSION

In this contribution, we considered mobile-robot-based warehouses. We presented three different reservation mechanisms for the time window routing method. These mechanisms require different levels of communication and differ regarding the resource utilization while executing a calculated route. We conducted a simulation study to compare the performance of these mechanisms considering an RMFS. As expected, the stairs mechanism enables the highest throughput followed by the rectangle mechanism.

In our consideration, robots accelerate, move with constant maximum speed or decelerate. For future work, we suggest enabling robots to move with a constant, but reduced speed to avoid intermediate stops. This behavior must be modelled and taken into account for the analysis of reachability of free time windows.

REFERENCES

- Azadeh, K.; de Koster, R. and Roy, D., 2018 (a), "Robotized and Automated Warehouse Systems: Review and Recent Developments." *Transportation Science* 53 No.4, 917–945.
- Azadeh, K.; Roy, D.; De Koster, R., 2018 (b): "Design, Modeling, and Analysis of Vertical Robotic Storage and Retrieval Systems." *Transportation Science* 53 No.5, 1–22.
- Boysen, N.; Briskorn, D. and Emde, S., 2017, "Parts-to-picker based order processing in a rack-moving mobile robots environment." *European Journal of Operational Research* 262, No.2, 550-562.
- Boysen, N.; de Koster, R. and Weidinger, F., 2019, "Warehousing in the e-commerce era: A survey." *European Journal of Operational Research* 227 No.2, 396–411.
- Busacker, T. 2005. *Steigerung der Flughafen-Kapazität durch Modellierung und Optimierung von Flughafen-Boden-Rollverkehr – Ein Beitrag zu einem künftigen Rollführungssystem*. Dissertation. Technische Universität Berlin.
- Havězda, J.; Rybecký, T.; Kulich, M. and Přeučil, L., 2018, "Context-Aware Route Planning for Automated Warehouses." In *Proceedings of the 21st International Conference on Intelligent Transportation Systems*, 2955-2960.
- Kim C. W. and Tanchoco J. M. A., 1991, "Conflict-free shortest-time bi-directional AGV routing." *International Journal of Production Research* 29, No.12, 2377-2391.
- Kim C. W., Tanchoco J. M. A. and Koo P., 1997 "Deadlock Prevention in Manufacturing Systems with AGV Systems: Banker's Algorithm Approach." *Journal of Manufacturing Science and Engineering* 119, No.4, 849-854.
- Lienert, T. and 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. and Fottner, J., 2018 (a), "Robust integration of acceleration and deceleration processes into the time window routing method." In *Proceedings of the 9th International Scientific Symposium on Logistics*, 66-86.
- Lienert, T.; Staab, T.; Ludwig, C. and Fottner, J., 2018 (b), "Simulation-based Performance Analysis in Robotic Mobile Fulfilment Systems." In *Proceedings of the 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, 383-390.
- Lienert, T. and Fottner, J., 2018, "Routing-based Sequencing Applied to Shuttle Systems." In *Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2949-2954.
- Lienert, T., Stigler, L. and Fottner, J., 2019, "Failure-handling strategies for mobile robots in automated warehouses." In *Proceedings of the 33rd European Conference on Modelling and Simulation*, 199-205.
- Maza, S. and Castagna, P., 2005, "A performance-based structural policy for conflict-free routing of bi-directional automated guided vehicles." *Computers in Industry* 56, No.7, 719-733.
- Roy, D.; Krishnamurthy, A.; Heragu, S. and Malmborg, C., 2015, "Queuing models to analyze dwell-point and cross-aisle location in autonomous vehicle-based warehouse systems", *European Journal of Operational Research* 242, No. 1, 72-87.
- Stenzel, B. 2008. *Online Disjoint Vehicle Routing with Application to AGV Routing*. Dissertation. Technische Universität Berlin.
- Tappia, E.; Roy, D. de Koster, R. and Melacini, M., 2018, "Modeling, Analysis, and Design Insights for Shuttle-Based Compact Storage Systems", *Transportation Science* 51, No.1.
- Yuan, Z.; Gong, Y.Y., 2017, "Bot-In-Time Delivery for Robotic Mobile Fulfilment Systems." *IEEE Transactions on Engineering and Management* 64, No.1, 83-93.
- Zou, B.; Xu, X.; Gong, Y.Y., de Koster, R., 2017, "Evaluating battery charging and swapping strategies in a robotic mobile fulfillment system." *European Journal of Operational Research* 267, No. 2, 733-753.

THOMAS LIENERT has been working as a research assistant at the Chair of Materials Handling, Material Flow and Logistics, Technical University of Munich, since 2014. His research deals with the simulation of mobile-robot-based warehouses. His email address is: thomas.lienert@tum.de.

FLORIAN WENZLER has been working as a research assistant at the Chair of Materials Handling, Material Flow and Logistics, Technical University of Munich, since 2013. His research deals with the optimization of project plans in construction. His email address is: florian.wenzler@tum.de.

JOHANNES FOTTNER is professor and head of the Chair of Materials Handling, Material flow, Logistics at the Technical University of Munich. His email address is: j.fottner@tum.de.

MATHEMATICAL SIMULATION OF ADJACENT-COUPLING AMMONIA ABSORPTIVE REACTOR.

Wenchan Qi
René Bañares-Alcántara
Department of Engineering Science
University of Oxford
Parks Road, Oxford OX1 3PJ, UK
E-mail: wenchan.qi@eng.ox.ac.uk

KEYWORDS

Ammonia synthesis, mathematical model, absorption, Backflow Cell Model, absorptive reactor, adjacent-coupling

ABSTRACT

The development of an efficient process for ammonia synthesis is a goal that has been long sought after; therefore, the application of an absorptive reactor for ammonia synthesis is important since it allows the reaction to occur under milder conditions. In the adjacent-coupling absorptive reactor, absorbent particles are positioned downstream the fixed ammonia synthesis catalyst bed. This kind of absorptive reactor leads to the enhanced conversion of ammonia synthesis under milder conditions, compared to the equivalent reactor used without absorbent. Here, we present the transient backflow cell model (BCM) to explain and analyse the phenomenon of absorption-enhanced reaction. The transient BCM, based on the first principle of mass balances, is developed to simulate that the backflow existing through the absorptive reactor. As a reference, the transient cell model (CM) is also implemented to simulate the absorptive reactor when assuming no existing backflow existing. These two models demonstrated that backflow through the absorptive reactor promotes the ammonia reaction conversion via two mechanisms: longer residence time for reaction and faster reaction rate due to the absorption of ammonia absorbed.

1. INTRODUCTION

Ammonia synthesis through the Haber-Bosch process is widely recognised as one of the most significant industrial applications (Bruce and Faunce 2015). Gas reactants hydrogen and nitrogen are converted to ammonia, operating in a catalytic packed bed reactor under high pressure (150-250 bar) and high temperature (400-500°C) (Appl 1999). Due to the problems of low single-pass conversion and energy-intensive operating conditions, the development of an ammonia synthesis reactor to enhance single-pass product yield and to adapt technology to and a less energy-intensive operation has attracted much attention.

A small number of research studies have investigated using the absorptive reactor to improve ammonia synthesis (Huberty et al. 2012; Wagner et al. 2017; Smith et al. 2019). The ammonia synthesis was enhanced by magnesium chloride absorption (Huberty et al. 2012), as the composite absorbents based on chlorides of alkaline-earth metals have a high ammonia absorption capacity (Zhu et al. 2009). Column absorption for reproducible cyclic separation in small scale ammonia synthesis was introduced in (Wagner et al. 2017). C. Smith et al. even ran the reaction and absorption at one temperature in the same piece of process equipment (Smith et al. 2019). However, no research focuses on the case when the catalyst bed and absorbent beds are sequentially packed in one column. In this paper, we focused on this adjacent-coupling configuration of the absorptive reactor for ammonia synthesis. As an increase in conversion was measured experimentally, it can be concluded that the flow regime is not an ideal plug flow, but backflow exists through the reactor. In this paper, the backflow is proposed to explain the absorption-enhanced conversion of ammonia synthesis. The reason for the improvement of the conversion will be identified in this paper. The transient backflow cell model (BCM) for the absorptive reactor was created to specify or verify the extent of absorption effect and determine the most efficient absorbent amount. The backflow cell model has actually been adopted in many types of research before (McSwain and Durbin 1966; Sinkule et al. 1976). The backflow cell model is of convenient because the mathematical treatment is simple, as shown by a numerical investigation for multiple stages. The transient backflow cell model (BCM), based on the first principle of mass balances, was developed and implemented in MATLAB®. Meanwhile, the transient cell model (CM) was implemented as a reference to simulate the absorptive reactor when assuming no backflow and, therefore, no absorption effect. The sensitivity on the number of cells, as one parameter used in the model, is also considered.

2. EXPERIMENTAL SECTION

The synthesis of Cs-Ru/MgO followed the incipient wetness impregnation procedure. The effect of the absorbent on ammonia production was tested in a fixed-bed continuous-flow reactor using Cs-Ru/MgO catalyst. A glass reactor tube (cross-sectional area $5.024 \times 10^{-5} \text{ m}^2$) was packed with 0.1 g Cs-Ru/MgO catalyst and 1 g absorbent (MgCl_2) with quartz wool in between. A 3:1

mixture of H₂ and N₂ was passed through the reactor at a pressure of 10 bar and a weight hourly space velocity of 36000 mL g⁻¹ h⁻¹. The reactor was equipped with three independent furnaces that allowed different sections to be heated to different temperatures. Each furnace was controlled by an independent thermocouple. The middle thermocouple was directly inserted into the catalyst bed and was set to synthesis temperature. During the reaction step, the catalyst was at 400 °C and the absorbent at 150 °C. Finally, any ammonia absorbed on the absorbent was desorbed by heating the absorbent to 400°C under flowing nitrogen at ambient pressure. The reaction rate of the catalyst was 7884 μmolg⁻¹h⁻¹ with the absorbent loaded behind the catalyst and 5082 μmolg⁻¹h⁻¹ without the absorbent.

3. MATHEMATICAL MODEL

The backflow process leads to two possible mechanisms that affect the single-pass conversion of the reaction: increased residence time and ammonia removal due to the absorption effect. In this section, we identify and compare the extent of each mechanism via two models: the transient BCM and the transient CM for the absorptive reactor (catalyst bed adjacent to an absorbent bed) and the conventional reactor (catalyst bed only). Figure 1 represents a schematic of the transient BCM in an absorptive ammonia synthesis reactor.

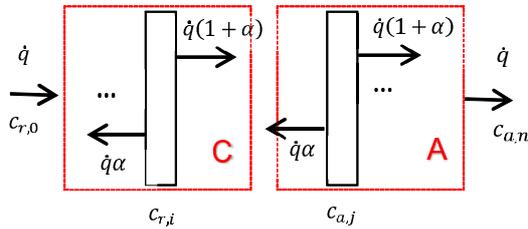


Figure 1: Schematic Diagram of the BCM in the Ammonia Synthesis Absorptive Reactor ('C' refers to the catalyst bed while 'A' refers to the absorbent bed).

The BCM hypothesises a stable unchanging backflow, expressed by $\dot{q}\alpha$ between each cell, to characterise the backflow mechanism in the reactor. There are m cells in the catalyst bed and n cells in the absorbent bed. In this case, the backflow ratio α represents the backflow amount. An α of zero is the plug flow limit; while when α equals infinity, the equations model approaches the stirred tank limit. The assumptions that govern the development of transient BCM are as follows:

- Due to symmetry in the radial direction of the packed bed, the governing equations are independent of this direction. Thus, a one-dimensional model is considered.
- Constant actual (interstitial) velocity is assumed.
- Isothermal condition is assumed in the catalyst bed.
- Each cell is assumed to be perfectly mixed.

- The volumetric flow of gas is assumed to be constant throughout the axial position of the bed.
- Instantaneous equilibrium of the solute in the bulk fluid with the absorbate.
- Catalyst/absorbent packing has uniform voidage and particle size.

According to these assumptions, the differential equations of the transient BCM were developed along the axial direction based on the principle of mass balance. The mathematical relationship for these cells can be written as follows. For the first cell of the catalyst bed:

$$V_{r,f} \cdot \frac{dc_{r,1}}{dt} = \dot{q}c_{r,0} + \dot{q}\alpha c_{r,2} - \dot{q}(1+\alpha)c_{r,1} + V_{r,c} \cdot R_r \quad (1)$$

Similarly, the modelling equations for the intermediate cells of the catalyst bed ($i = 2, 3, \dots, m-1$) can also be written.

$$V_{r,f} \cdot \frac{dc_{r,i}}{dt} = \dot{q}(1+\alpha)c_{r,i-1} + \dot{q}\alpha c_{r,i+1} - \dot{q}(1+2\alpha)c_{r,i} + V_{r,c} \cdot R_r \quad (2)$$

For the last cell of the catalyst bed:

$$V_{r,f} \cdot \frac{dc_{r,m}}{dt} = \dot{q}(1+\alpha)c_{r,m-1} + \dot{q}\alpha c_{r,m+1} - \dot{q}(1+2\alpha)c_{r,m} + V_{r,c} \cdot R_r \quad (3)$$

For the first cell of the absorbent bed:

$$V_{a,f} \cdot \frac{dc_{a,1}}{dt} = \dot{q}(1+\alpha)c_{a,0} + \dot{q}\alpha c_{a,2} - \dot{q}(1+2\alpha)c_{a,1} + V_{a,c} \cdot R_a \quad (4)$$

For the intermediate cells of the absorbent bed, $j = 2, 3, \dots, n-1$:

$$V_{a,f} \cdot \frac{dc_{a,j}}{dt} = \dot{q}(1+\alpha)c_{a,j-1} + \dot{q}\alpha c_{a,j+1} - \dot{q}(1+2\alpha)c_{a,j} + V_{a,c} \cdot R_a \quad (5)$$

For the last cell of the absorbent bed:

$$V_{a,f} \cdot \frac{dc_{a,n}}{dt} = \dot{q}(1+\alpha)c_{a,n-1} - \dot{q}(1+\alpha)c_{a,n} + V_{a,c} \cdot R_a \quad (6)$$

where \dot{q} is the inlet gas flowrate (in m³/s) and $\dot{q}\alpha$ is the backflow. The forward flow increases to $\dot{q}(1+\alpha)$. The mass balances for hydrogen, nitrogen, and ammonia are respectively calculated by Equations (1) through (6), with different reaction rates and absorption rates.

Dyson and Simon modified the Temkin expression to calculate the intrinsic rate of reaction in kmol/m^3 (Dyson & Simon 1968), given in Equation (7):

$$R_{r,NH_3} = 2k_2 \left[K_a^2 a_{N_2} \left(\frac{a_{H_2}^2}{a_{NH_3}^3} \right)^\beta - \left(\frac{a_{NH_3}^2}{a_{H_2}^3} \right)^{1-\beta} \right] \quad (7)$$

The parameter β is a constant, taking values between 0.5 and 0.75, and the expression of the activity of the components, a_k , is shown in Equation (8):

$$a_k = \frac{f_k}{P^0} = \frac{\phi_k y_k P}{P^0} \quad (8)$$

P^0 is the reference pressure and is assumed to be atmospheric (1 bar). Gillespie and Beattie calculated the equilibrium constant, K_a , and proposed the correlation in Equation (9) (Gillespie and Beattie 1930):

$$\log_{10} K_a = -2.691122 \log_{10}(T) - 5.519265 \times 10^{-5} T + 1.848863 \times 10^{-7} T^2 + \frac{2001.6}{T} + 2.689 \quad (9)$$

In turn, k_2 is estimated by the Arrhenius equation (Gillespie & Beattie 1930):

$$k_2 = 8.849 \times 10^{14} e^{\left(\frac{-40765}{1.9877} \right)} \quad (10)$$

The absorbents only absorb ammonia. Therefore, the absorption rate for hydrogen and nitrogen is zero.

In this mathematical model, the absorbent bed is first assumed to be in the non-saturated condition. Then, the saturation condition of the absorbent is additionally considered to determine the influence of the absorbent amount. While a detailed mechanism for absorption is not known, kinetic experiments conducted previously (Smith et al. 2018) allow the fitting of an analytical equation:

$$r_{abs} = k_{a1} e^{k_{a2}(P_{NH_3} - P_{NH_3}^+)} \text{ when } (P_{NH_3} - P_{NH_3}^+) < 1 \text{ bar} \quad (11)$$

$$r_{abs} = k_{a3} (P_{NH_3} - P_{NH_3}^+)^2 \text{ when } (P_{NH_3} - P_{NH_3}^+) > 1 \text{ bar}$$

where r_{abs} is the rate in micromoles per second per gram of absorbent, k_{a1-a3} are experimentally determined constants (Smith et al. 2019), P_{NH_3} is the pressure of ammonia in kPa, and $P_{NH_3}^+$ is the equilibrium pressure of ammonia at a given temperature in kPa, as illustrated in (Smith et al. 2019). Based on these analytical equations, the temperature profile of the absorbent bed could be

ignored as the parameters and equations were assumed to be independent of temperature-irrelevant.

The governing equation for the conventional reactor in BCM is Equation (12). To some extent, the existing of backflow can be assumed as that the flow goes through the catalyst bed with longer residence time. Therefore, the cell number of the catalyst bed is m' , which is higher than m .

$$V_{r,f} \cdot \frac{dc_{r,i'}}{dt} = \dot{q}c_{r,i'-1} - \dot{q}c_{r,i'} + V_{r,c} \cdot R_r \quad (12)$$

In the CM of the absorptive and the conventional reactors, the reactants flow in one direction. The modelling equations for the cells of the catalyst bed were consistent with Equations (1) to (6) but ignored the backflow.

4. SIMULATION AND DISCUSSION

The governing equations of the transient BCM and transient CM are ordinary differential equations which were solved using the ode15s solver from MATLAB[®]. The numerical results and discussion are presented in this section. Nitrogen is taken as the reference to calculate the reaction conversion.

4.1 The Conventional Reactor

Figure 2 depicts the concentration of these three components in the catalyst bed at different times (arrow points in the direction of an increasing number of cells). The steady-state concentration of hydrogen and nitrogen for the downstream position is lower than upstream positions as the reaction progresses.

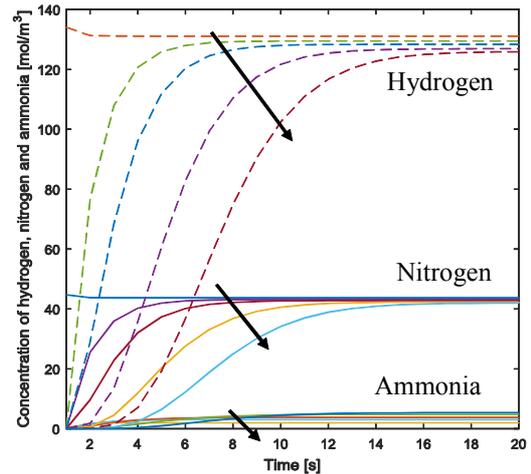


Figure 2: Concentration of hydrogen, nitrogen, and ammonia in the conventional reactor.

As illustrated in Figure 3, the concentration of ammonia increases from the first cell to the end as a result of the superposition of the produced ammonia (arrow points in the direction of an increasing number of cells). The

conversion in the catalyst bed is 6.25% according to the simulation results, which depend on the parameters used in the model, such as the length of the catalyst bed, temperature, and pressure.

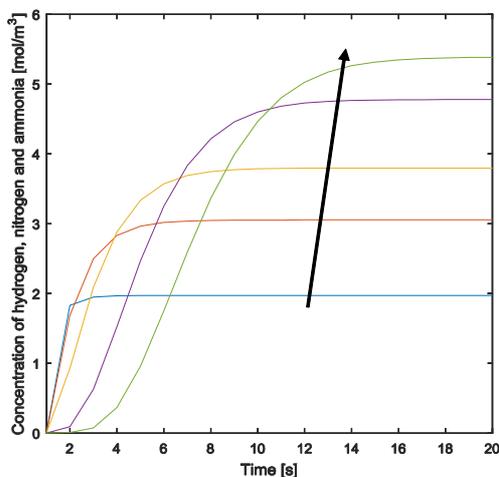


Figure 3: Concentration of ammonia in the conventional reactor.

Therefore, the reaction equilibrium has not been reached, as the thermodynamic equilibrium conversion of ammonia synthesis from $H_2: N_2 = 3:1$ at $T = 400\text{ }^\circ\text{C}$ and $P = 10\text{ bar}$ is 7.45% (Ogura et al. 2018). Then, the effect of the catalyst bed length on the reaction conversion was considered here. As demonstrated in Figure 4, the conversion of the ammonia synthesis rises sharply at first with the increasing length of the catalyst bed, then reaches a stable value: the equilibrium point of this reversible reaction. This balance cannot be broken unless the reaction conditions change. In Figure 4, the ammonia synthesis reaches the equilibrium when the catalyst bed is longer than 0.04 m. The synthesis conversion is independent of the residence time if the equilibrium has been reached.

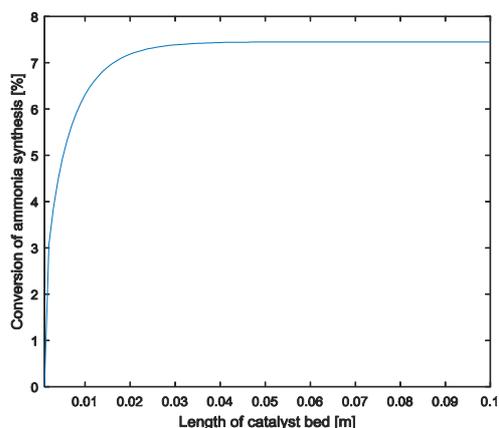


Figure 4: Variation of the conversion of ammonia synthesis with the length of the catalyst bed.

Figure 5 presents the conversion of ammonia synthesis under various backflow extents (0% to 100%) when the

length of the catalyst bed is 0.01 m. An increase in ammonia production conversion from 6.25% to 7.25% can be observed. In the transient BCM of the conventional reactor, the extent of backflow was assumed to be related to the number of cells in the catalyst bed, which replace the increase in residence time. When the extent of backflow is 100% the catalyst bed is assumed to be extended by about 10 cells, which means the flow has a longer residence time for reaction.

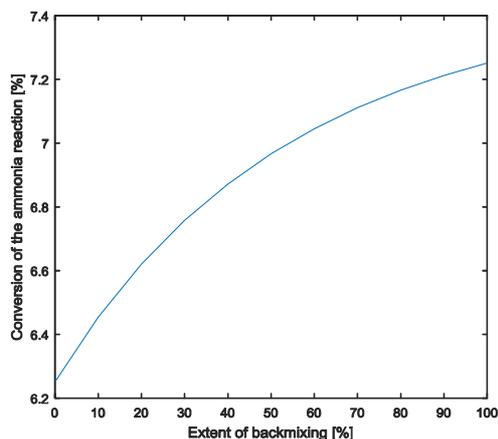


Figure 5: Effect of backflow on the ammonia synthesis conversion.

The number of cells is an important parameter in the BCM and CM. Therefore, it is necessary to analyse the sensitivity of the number of cells at a fixed-length catalyst bed, which is displayed in Figure 6 (arrow points in the direction of an increasing bed length). The cell number sensitivity study is a significant aspect to avoid inaccurate results due to an ill-conditioned cell number. In Figure 6, the cell number should be larger than 8 when the length of the catalyst bed is longer than the equilibrium length, and then the results stop changing with cell number. When the length of the catalyst bed is shorter than 0.04 m, increasing the number of cells will slightly increase the reaction conversion. Therefore, in this case, a cell number of 12 is taken to minimise the deviation of cell number dependence.

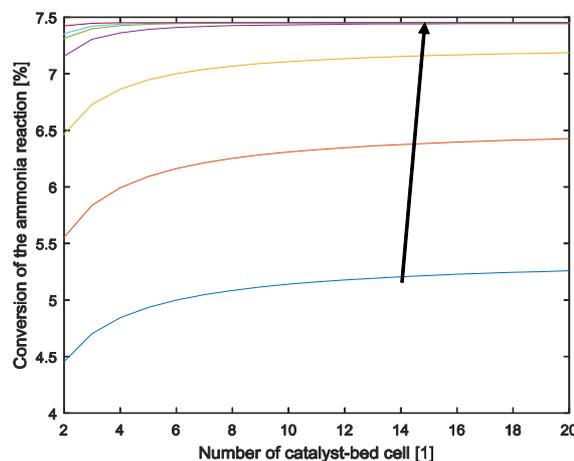


Figure 6: Sensitivity on the number of cells in the catalyst bed.

4.2 The Absorptive Reactor

Based on the results in Section 4.1 for the simulation of the absorptive reactor, the length of the catalyst bed was fixed to 0.1 m to ensure that the reaction equilibrium was reached and the cell number was set to be 8 to decrease computation time. The single-pass conversion of the absorptive reactor is 7.45%, calculated by the transient CM, which is the same as the conversion of the conventional reactor (0.1 m length). Therefore, the absorption does not affect the ammonia synthesis conversion when using the transient CM, which is not consistent with the experimental results.

Based on the experimental results, an absorption bed following the catalyst bed has an impact on the single-pass reaction conversion. Therefore, the transient CM does not apply to the absorptive reactor and the simulation can be switched to the transient BCM with a fixed 50% backflow ratio. The solid lines in Figures 7 and 8 display the concentration profiles of nitrogen and ammonia. Unlike the transient CM results (dashed lines), the ammonia concentration decreases at the end of the catalyst bed, when the flow has yet to reach to absorbent bed. The nitrogen concentration has a second drop at the junction of the catalyst bed and absorbent bed, shown in the red square of Figure 7.

The comparison of the transient BCM and CM simulations is illustrated in Figures 7 to 8 for a 50% backflow ratio. The arrows point in the direction of increasing time. As illustrated in Figures 7 and 8, the dashed curves and solid curves are noticeably different, which means that simulated results were influenced greatly by the backflow setting. Under the same cell number and same time conditions, a lower nitrogen concentration with higher ammonia concentration is revealed for the absorptive reactor with the BCM simulation in comparison with that in the CM simulation. This implies that the single-pass synthesis conversion in the BCM simulation is higher than the one predicted by the CM simulation. Undoubtedly, the transient BCM matched experimental results better than transient CM.

Based on the BCM simulation, the absorption capacity of the absorptive reactor, which promotes the reaction conversion, can be further analysed. In this BCM simulation, the equilibrium of the reaction in the catalyst bed has been reached as its length is set at 0.1 m with 8 cells. Therefore, the single-pass conversion of ammonia synthesis does not change when there is the only backflow without absorption or increased residence time, based on the results of Section 4.1. Therefore, the reason for improved conversion is due to the absorption capacity of the absorptive reactor, which speeds up the reaction rate and drives the reaction equilibrium as the partial pressure of ammonia decreases. Then, the backflow ratio,

length and cell number of absorbent bed, as well as the saturation degree of absorbent bed, will be discussed.

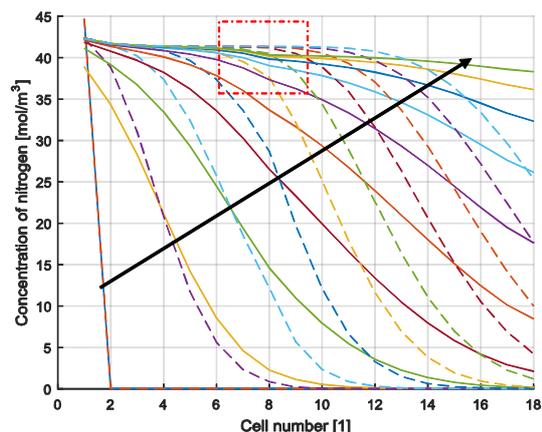


Figure 7: Concentration of nitrogen with respect to the position in the absorptive reactor.

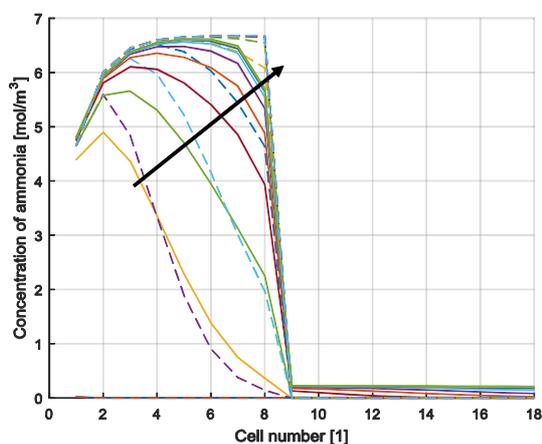


Figure 8: Concentration of ammonia with respect to the position in the absorptive reactor.

Figure 9 reveals a relationship between the synthesis conversion of the absorptive reactor and the backflow ratio. The reaction conversion has a linear growth as the backflow ratio increases when the absorbent bed has not been saturated. The reaction conversion increases from 7.45% (the equilibrium conversion) to 11.2% in Figure 9. Therefore, the absorption in the absorptive reactor can drive the synthesis equilibrium in the right direction when the thermodynamic equilibrium of the reaction has already been reached.

Then, the length and cell number of the absorbent bed were considered for their impact on the conversion improvement. Based on the simulation results, it is illustrated that increasing infinitely the length or the cell number of the absorbent bed does not affect the synthesis conversion when the absorbent bed was assumed to be unsaturated.

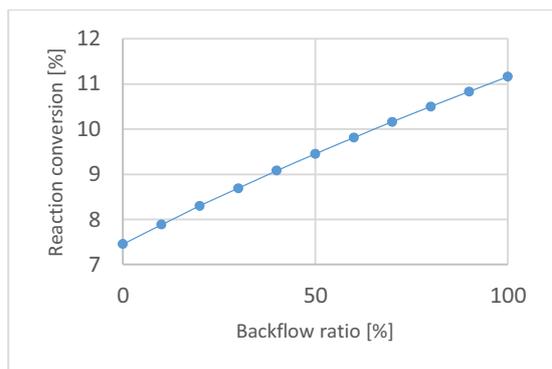


Figure 9: Relationship between reaction conversion and backflow ratio in the absorptive reactor.

However, in reality, the absorbent bed is gradually saturated as the synthesis take place, and the effect of absorption on the synthesis conversion begins to wane as the absorbent pellets stop absorbing ammonia. For example, Figure 10 displays a relationship between the absorptive reactor conversion and the saturated cells in the absorbent bed. In this case, the backflow ratio is set at 50% with a fixed-absorbent-bed length of 0.1 m. A saturated cell number equal to six means that the first six cells of the absorbent bed have been saturated and the absorption occurs at the third cell. Figure 10 reveals that the saturation state of only the first six cells in the absorbent bed can affect the synthesis conversion. Once the first six cells are saturated, the conversion no longer increases, and there is no difference between an absorptive reactor and a conventional reactor.

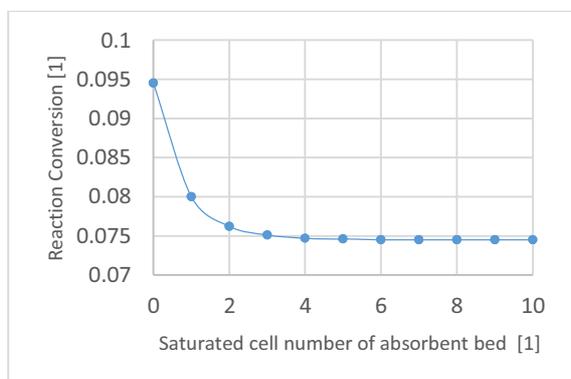


Figure 10: Reaction conversion of ammonia synthesis with a gradually saturated absorbent bed.

When the ammonia synthesis in the catalyst bed has not reached the thermodynamic equilibrium, the results indicate that, under the kinetically controlled regime, both the longer residence time and the absorption of ammonia can yield higher ammonia conversion to approach the thermodynamic value as shown in our model. If the packing after the catalyst bed contains solid pellets but no absorbent, the reaction conversion can be improved only by the residence time. However, once the equilibrium is reached, the increase of the reaction conversion is due to the absorption. The backflow ratio

can be regarded as the scale of the fluid disorder in the absorptive reactor, which will directly relate the impact extent of absorption on synthesis. If the backflow ratio is zero, the decrease in ammonia concentration or partial pressure due to absorption will not affect the synthesis reaction. Furthermore, when the backflow ratio is one hundred, the reactants with the reduced partial pressure of ammonia (ammonia was absorbed) will flow to upstream cells, speeding up the reaction and driving the equilibrium.

It can be summarized that the reaction conversion is entirely independent of the cell number of the absorbent bed. Once the absorbent bed length is long enough, e.g. at least 0.1 m, the reaction conversion will not increase even if the absorbent bed is longer. Moreover, if the first several cells (about 0.06 m) of the absorbent bed are saturated, the absorption will not affect the synthesis conversion any more.

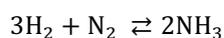
The reaction conversions for experimental results were 1.09% and 0.71%. These were much lower than the simulation conversion of a stoichiometric feed, which was due to the activity of the catalyst being calculated based on an iron catalyst (Dyson and Simon 1968) in the simulation, not on the Ru catalyst that was used in the experiment as the lack of rate expression for Ru catalyst. Furthermore, the data at mild conditions were not a good comparison for the reaction rates under ideal simulation conditions. However, as the backflow effects on the single-pass conversion in the simulation have the same increasing trend as in reality, this model of the ammonia absorptive reactor can, to some extent, explain the experimental results.

There are potential limitations to this simulation. In reality, both ammonia synthesis and absorption are highly exothermic. However, in this simulation, we assumed that the temperature of the catalyst bed was constant at 400 °C, while it declined linearly at the beginning of the absorbent bed from 400 °C to 150 °C. Also, the ammonia synthesis reaction is not equimolar (with 4 mol of reactants being converted to 2 mol of products), and absorption will only remove ammonia from the gas phase. In this simulation, the single-pass conversion of the reaction is quite low at about 5 – 7 %. Therefore, the volume change during the reaction and absorption processes was low enough to be ignored in this paper. Nevertheless, these are future simulation we would like to perform.

5. CONCLUSION

In this paper, a transient BCM was performed to simulate an absorptive reactor for ammonia synthesis. Meanwhile, the transient CM was taken as a reference to identify what role the absorbent plays in the enhancement of synthesis conversion. Modelling results indicate that the absorbent packed adjacently after the catalyst significantly affects the single-pass synthesis conversion through increased

residence time and faster reaction rate, driving equilibrium to the right. The backflow increases the residence time, while the faster reaction rate and equilibrium-shift are due to the reduced ammonia partial pressure (absorbed in the absorbent). The bench-scale absorptive ammonia synthesis reactor was operated to obtain experimental data to underline this point. Furthermore, the absorption has more impact on the conversion improvement than the increased residence time, as the proportional increase in Figure 9 is more significant than that in Figure 5. These two models could demonstrate that backflow in the absorptive reactor promotes the ammonia reaction conversion via two mechanisms: longer residence time for reaction and a faster reaction rate, driving the equilibrium of the reaction to the right (reversible reaction shown as below).



REFERENCES

- Appl, M., 1999. *Ammonia: principles and industrial practice*, Weinheim: Wiley-VCH.
- Bruce, A. and Faunce, T., 2015. Sustainable fuel, food, fertilizer and ecosystems through a global artificial photosynthetic system: Overcoming anticompetitive barriers. *Interface Focus*, 5(3), pp.1–9.
- Dyson, D.C. and Simon, J.M., 1968. A kinetic expression with diffusion correction for ammonia synthesis on industrial catalyst. *Industrial & Engineering Chemistry Fundamentals*, 7(4), pp.605–610.
- Gillespie, L.J. and Beattie, J.A., 1930. The Thermodynamic Treatment of Chemical Equilibria in Systems Composed of Real Gases. I. An Approximate Equation for the Mass Action Function Applied to the Existing Data on the Haber Equilibrium. *Physical Review*, 36(4), pp.743–753.
- Huberty, M.S. et al., 2012. Ammonia absorption at haber process conditions. *AIChE Journal*, 58(11), pp.3526–3532.
- McSwain, C. V and Durbin, L.D., 1966. The Backflow-Cell Model for Continuous Two-Phase Nonlinear Mass-Transfer Operations Including Nonlinear Axial Holdup and Mixing Effects. *Separation Science*, 1(6), pp.677–700.
- Ogura, Y. et al., 2018. Efficient ammonia synthesis over a Ru/La_{0.5}Ce_{0.5}O_{1.75} catalyst pre-reduced at high temperature. *Chem. Sci.*, 9(8), pp.2230–2237.
- Sinkule, J., Hlaváček, V. and Votruba, J., 1976. Modeling of chemical reactors-XXXI. The one-phase backflow cell model used for simulation of tubular adiabatic reactors. *Chemical Engineering Science*, 31(1), pp.31–36.
- Smith, C. et al., 2018. Rates of Ammonia Absorption and Release in Calcium Chloride. *ACS Sustainable Chemistry & Engineering*, 6(9), pp.11827–11835.
- Smith, C., McCormick, A. V and Cussler, E.L., 2019. Optimizing the Conditions for Ammonia Production Using Absorption. *ACS Sustainable Chemistry and Engineering*, 7(4), pp.4019–4029.
- Wagner, K. et al., 2017. Column absorption for reproducible cyclic separation in small scale ammonia synthesis. *AIChE Journal*, 63(7), pp.3058–3068.
- Zhu, H. et al., 2009. Large-Scale Synthesis of MgCl₂ · 6NH₃ as an Ammonia Storage Material. , pp.5317–5320.

AUTHOR BIOGRAPHIES

WENCHAN QI was born in Shandong, China and went to the University of Birmingham, where she studied chemical engineering and obtained her MEng degree in 2016. She then moved to the University of Oxford for DPhil degree study, and she is now working on a research of the optimisation of the Haber-Bosch reactor. Her e-mail address is wenchan.qi@eng.ox.ac.uk.

RENE BANARES-ALCANTARA has an MEng from UNAM (Mexico), and an MSc and PhD from Carnegie Mellon University (USA), all of these degrees in Chemical Engineering. He has worked in the Department of Engineering Science at Oxford since 2003 and is a Fellow of New College. His research interests are in the area of Process Systems Engineering, mainly process design, synthesis and simulation. Since 2014 he has been involved in projects related to long-term (chemical) storage of renewable energy and the production of ‘green’ ammonia.

Implementation of the optimizer of SOA system deployment architecture

A. P. Woźniak

Military University of Technology

Institute of Computer and Information Systems

Ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland

E-mail: adrian.wozniak@wat.edu.pl

KEYWORDS

Service-Oriented Architecture, SOA, business process, optimization, simulation

ABSTRACT

Optimization of business processes in SOA systems has been done using three separate types of methods: Resource Allocation, Service Scheduling and Service Composition. All three may influence each other, so the new method has been proposed to find an optimal combination of those three. It is based on a genetic algorithm that uses a simulator of the SOA system to evaluate solutions. The article describes a model for the optimization criteria for such solutions. Subsequently, some basic concepts used to implement the simulator and optimizer have been presented. Finally, the performance results of the optimizer have been described, including the conclusions on how they might be improved.

INTRODUCTION

The optimization of the system performance has always been important. It is due to many reasons, but first and foremost due to the limitation of resources or drive to increase system performance. It is no different in the case of the Service-Oriented Architecture (SOA) systems. Nonetheless, there are a few differences in how the systems might be optimized. The differences are mainly caused by fragmentation of the SOA systems. To add value to such systems, many components must cooperate. Each component is a software module that may be implemented in a technology different than other components and deployed independently. The most important part of the SOA components is that they deliver services. A service is a function of a component that is usually provided through a www. Users get value out of the SOA system by invoking the so-called composite services or business processes which are sequences of services realized by components.

The literature includes 3 types of methods for optimizing business processes in the SOA systems: Service Composition, Service Scheduling and Resource Allocation. Each of them is focused on a different stage of the SOA system implementation or execution. The first type is the Resource Allocation. It consists in determining which components should be deployed on which servers. Each component may be deployed simultaneously on many servers. Therefore, during Resource Allocation, it is also decided how many component instances should be running. Example of such a method uses Quality of Service (QoS) constrains and

resource usage cost as an input (Almeida et al. 2006; Huang et al. 2016; Mennes et al. 2016). Then it searches for optimal allocation using Fixed Point Iteration technique. The second way of optimizing business processes in SOA is to use the Service Composition method, which is the most popular in the literature (e.g. Ebrahim 2011; da Silva et al. 2015; Zhao et al. 2017; Wang et al. 2011; Xianwen et al. 2009). This type of optimization method is used when a given service is available on multiple servers. Usually, it is because the component is deployed on many servers. The Service Composition method is about deciding which server should execute a service instance. Usage of genetic algorithm is very common in solving this problem. Example of such approach is presented by Ebrahim (Ebrahim 2011). He suggests using a genetic algorithm where the chromosome has a number of genes equal to the number of services that must be called in the process. Each gene indicates an instance of the service that should be called in the process. The best chromosomes are those that provide the best QoS with minimal cost of service and minimal diversity of suppliers. The third type of methods is Service Scheduling (Dyachuk and Deters 2008). It is executed last and it is least popular in the literature. It may be used when multiple service invocations are organized in a queue of one component. Then it is possible to determine the order of their execution. For example, the Service Scheduling method presented in (Dyachuk and Deters 2007) finds services on a critical path of a business process and prioritizes them in the component queue.

All of these three methods are considered independently in the literature, even though they influence each other. Different service composition methods may give best results on different allocations and service scheduling algorithms. It means that we should strive to optimize all three aspects. Such optimization concept is proposed in (Woźniak and Nowicki 2019). It is based on a SOA system simulator that is used to evaluate solutions. The simulator takes, as an input, the SOA system model, which includes: services, components, execution environments, servers, business processes, etc. Each of the above-mentioned elements should be described with attributes, such as a random variable resources (CPU and RAM) used by each service invocation. In addition to the model, the simulator takes, as an input, the matrix of resource allocation and selected algorithms of the Service Composition and Service Scheduling methods. During simulation, output values that constitute criteria for selecting the best solutions (out of those that were simulated) may be obtained. To search through solutions,

a genetic algorithm was used in combination with a brute force approach. It is a unique feature of the SOA system allowing to define the optimization criteria from the business process point of view, which will be subsequently translated into the infrastructure. The reason behind this is that in SOA, business processes may be mapped to services. The following optimization model is an extension of one presented in (Woźniak and Nowicki 2019). Its main difference from the three SOA optimization types of methods is that it takes into account that all of those three influences each other so it finds optimal three: resource allocation, service composition algorithm and service scheduling algorithm.

OPTIMIZATION MODEL

Several optimization criteria for business processes in the SOA system may be defined. They are focused on two aspects of the process: service quality for the user and costs for the company. The first two criteria are the following:

1. Average execution time of business processes weighted by the expected number of instances of business process.

$$\bar{k}_1(t, X) = E(k_1(t, X))$$

where:

$$k_1(t, X) = \sum_{b=1}^B \left(\frac{H_b(t)}{\sum_{i=1}^B H_i(t)} \cdot \frac{\sum_{i=1}^{LR_b} CR_{i,b}(t, X)}{LR_b(t)} \right)$$

B – number of business processes,

t – simulation time,

X – analysed solution which consists: boolean matrix of allocation of components to servers (genotype), selected Service Scheduling and Service Composition algorithms, $H_x(t)$ – expected value for the number of x-type business process instances running,

$CR_{i,b}(t, X)$ – random variable denoting the time of implementation of the i-th instance of the b-th business process during t,

$LR_b(t)$ – random variable denoting the number of completed instances of the bth business process during time t.

2. Average variance of execution time of business processes

$$k_2(t, X) = \sum_{b=1}^B \frac{H_b(t)}{\sum_{i=1}^B H_i(t)} \cdot \frac{\left(\sum_{i=1}^{LR_b} \left(\frac{\sum_{i=1}^{LR_b} CR_{i,b}(t, X)}{LR_b(t)} - CR_{i,b}(t, X) \right)^2 \right)}{LR_b(t)}$$

The costs criteria are included in the form of resources that are used by the system to calculate how they should be minimized. They are defined in the following manner:

1. The expected amount of processor resources used to provide services over a given time t.

$$\bar{k}_3(t, X) = E(k_3(t, X))$$

where:

$$k_3(t, X) = \frac{\sum_{s=1}^S m_s tu_s(t, X)}{\sum_{s=1}^S (m_s \cdot Y(s)) \cdot t}$$

S – number of servers,

m_s – computational power of the s-th server,

$tu_s(t, X)$ – random variable denoting the time spent by s-th server on processing services

$Y(s)$ – function Y (s) takes the value 1 if any component is assigned to server s.

2. expected utilization rate of allocated memory resources for the provision of services during t

$$\bar{k}_4(t, X) = E(k_4(t, X))$$

where:

$$k_4(t, X) = \sum_{s=1}^S (p_s \cdot Y(s)) \cdot t$$

p_s – amount of RAM on s-th server.

However, during the tests of optimizers, the \bar{k}_3 criterion had two effects. The first one led to maximizing the use of a processor, which was beneficial. The second effect promoted the solutions that had long queues leading to longer execution times of business processes, which was unintended. The effect was partly nullified by the first criterion. However, to increase the convergence of the method, it was decided to replace it with a simpler one:

The expected degree of utilization of allocated processor resources for the provision of services in a given time t

$$\bar{k}_5(t, X) = E(k_5(t, X))$$

where:

$$k_5(t, X) = \sum_{s=1}^S (m_s \cdot Y(s)) \cdot t$$

Furthermore, during the experiments, it was noticeable that some solutions with high evaluation scores, according to the above criteria, had many unrealized business processes. The process may be abandoned if a server does not have enough resources to realize it or is damaged. To solve that, the another criterion was added:

$$\bar{k}_6(t, X) = E(k_6(t, X))$$

where

$$k_6(t, X) = \sum_{b=1}^B r_b(t, X)$$

$r_b(t, X)$ – random variable denoting the number of unrealized instances of a b-th type process.

Therefore, the target function of such optimization may be defined as:

$$\bar{k}(t, X) = (\bar{k}_1(t, X), \bar{k}_2(t, X), \bar{k}_4(t, X), \bar{k}_5(t, X), \bar{k}_6(t, X))$$

Not all component allocations are acceptable solutions. At least three restrictions on the solution should be defined:

Restriction 1. Processor. The server processor power should exceed its consumption as resulting from the operations of the components and execution environments.

$$\forall_{s \in (1, S)} \sum_{k=1}^K K_{k,s} M_k + \sum_{su=1}^{SU} S_{su,s} MU_{su} < m_s$$

where:

K – number of components

$K_{k,s}$ – binary value whether the k component has been assigned to the server s

M_k – amount of processing power consumed by k-th component

SU – number of execution environments,

$S_{su,s}$ – binary value whether the su-th execution environment has been assigned to the server s,

MU_k – amount of processing power consumed by su-th execution environment.

Restriction 2. RAM. The server RAM resources should exceed its consumption as resulting from the operations of the components and execution environments.

$$\forall_{s \in (1, S)} \sum_{k=1}^K K_{k,s} P_k + \sum_{su=1}^{SU} S_{su,s} PU_{su} < p_s$$

where:

P_k – amount of RAM consumed by k-th component

PU_{su} – amount of RAM consumed by su-th execution environment

Restriction 3. At least one instance of each component should be deployed.

$$\forall_{k \in (1, K)} \exists_{s \in (1, S)} K_{k,s} = 1$$

SIMULATOR IMPLEMENTATION

To evaluate the solutions using the above criteria, the simulator of the SOA system was implemented in Java

language with the DISSim library. DISSim is a toolset for developers that allows to perform discrete event simulation. It uses BasicSimObject class for every element that is simulated and generates events in simulation time. The second class is BasicSimEvent that is put on simulation calendar. A simulation engine searches through the calendar and picks up the nearest event. Every event has a code attached thereto that is executed when the simulation time comes.

Class Model of the Simulator

The BasicSimEvent and BasicSimObject classes are abstract. The latter is a parent object to the Organization and Server classes (Figure 1). The organization starts and contains instances of business processes that are specified by the Business Process Definition. The business process is defined as a graph, in which each step is a service. All services in the process are interconnected by arches described with the probability of choosing each path (which reflects the operation of the gates in BPMN). The services are described by the processor power, the RAM they need for their operations and the volume of data that should be sent through the network to provide the service. The services are associated with the components that execute them. The components contain a list of execution environments on which they can run. The components and execution environments are run on a server, which is a simulation object. They are described by the processor power and RAM memory necessary for their operations. The servers have specified amount of the processing power and RAM needed to run the components and execution environments as well as to provide the services. In addition, the servers are described with a matrix of network bandwidth between them. What is more, each server class object contains random variables that indicate the time of its damage and repair. Server damage events are created after starting the simulator in random time according to the random variable assigned to the server. When a server damage event occurs, the server's status is changed to inoperative and a server repair event is generated at a random time from the time of failure. The repair event generates the damage event, etc.

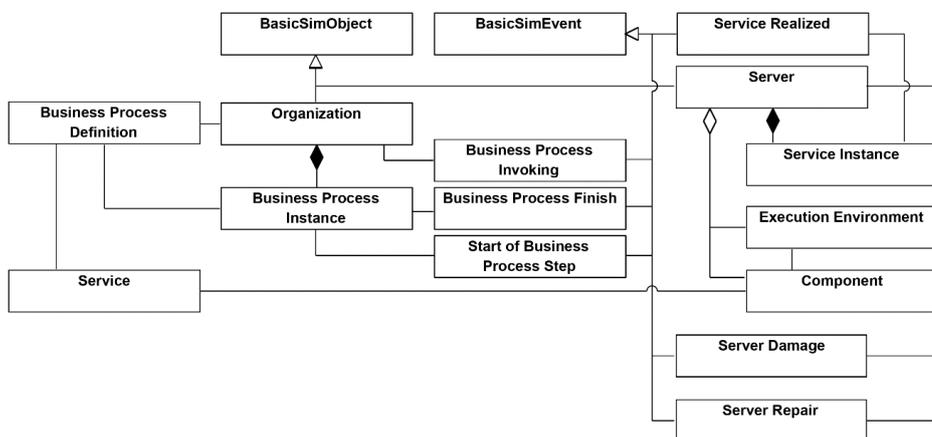


Figure 1 Class Model of the Simulator

Simulation Process

Creating the Business Process Instance

The events in the simulator are interdependent, and the logic of their occurrence is presented in Figure 2. After running the simulation, the "cyclic invocation of business process instance" events are generated. There is one event for every business process definition. It represents the creation of a new business process instance.

Once the simulation time reaches the event time, a new process instance and a new calendar event are created and will occur for the randomly generated simulation time. The time between the successive process start events is established according to the random variable specified in the business process definition.

Business Process Realization

When the business process instance is created, its first step with the current simulation time is generated. Each event representing a step in the business process aims at invoking the services to accomplish such step. First of all, a server is appointed to execute the service. It is the operation of the load balancer, which consists in the selection of the correct instance of the component, i.e. implementation of the Service composition algorithm. Two Service Selection strategies have been implemented in the simulator:

- select least loaded server,
- select server with the shortest expected response time.

It is possible to add further Service Composition algorithms to the simulator. If a server capable of performing the step in the process is not found, then the process is terminated and the information about the

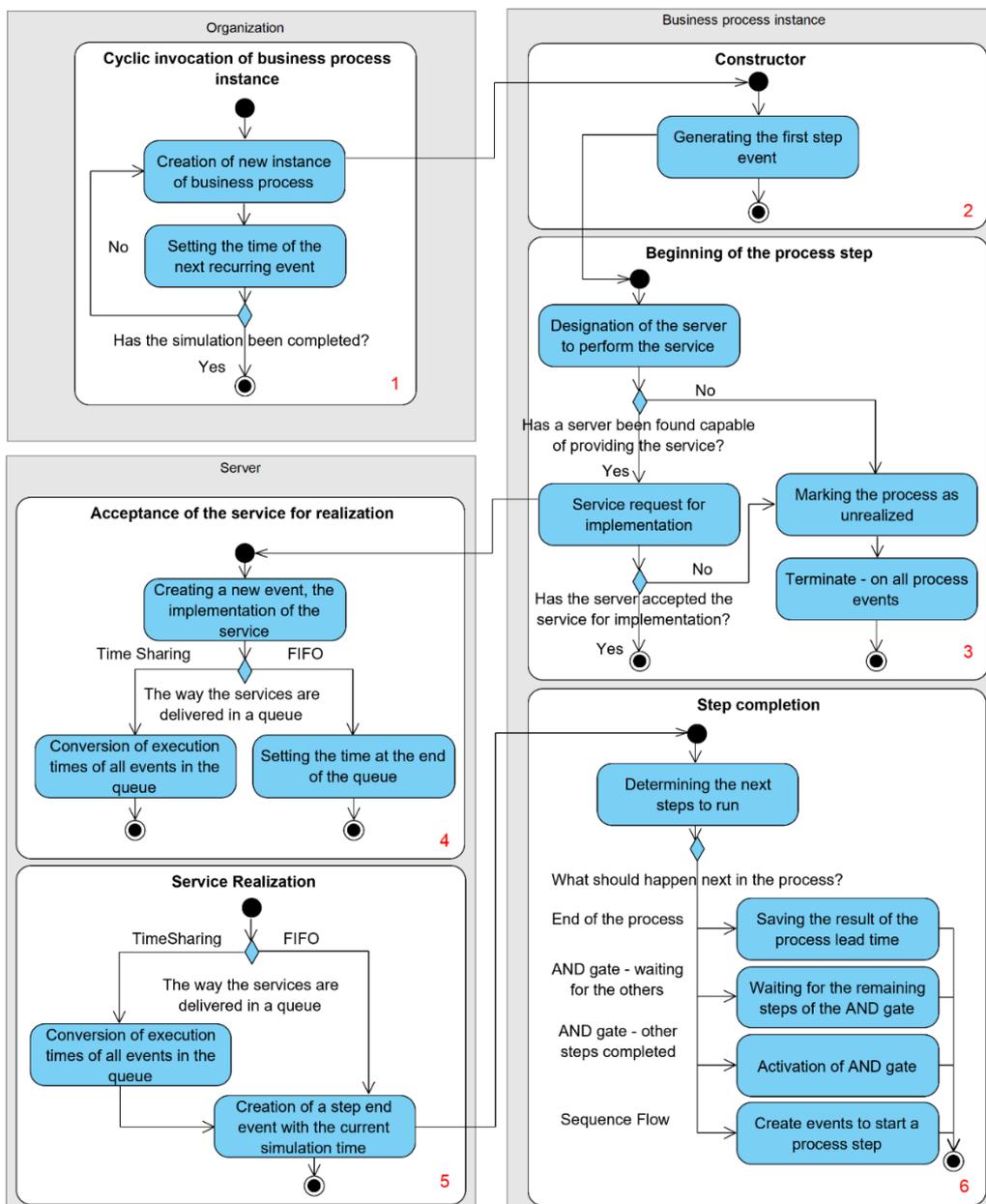


Figure 2 Simulation Process of the SOA System

system's inability to implement the process is included in the simulation results.

Service Execution

If the load balancer has found the server capable of providing the service, i.e. the server that:

- has enough free RAM memory,
 - has a component capable of providing the service,
- a service order is created. The time necessary for the service to be completed is the sum of:
- data transfer through the network,
 - service execution time on the server.

The transfer time depends on the volume of data to be transferred, as defined in the service, and the network bandwidth. The execution time depends on:

- the processor power allocated by the server to execute the services,
- the number of services invoking the orders,
- the power needed to perform the services,
- the component operation model (FIFO or Time Sharing).

If the component operates in the time-sharing mode, then each appearance of a new service to be executed and each termination of the service require recalculation of the expected service realization times.

Business Process Step Execution

The completion of the service creates a step completion event in the business process. Its aim is to determine the next steps. It may cause the termination of the process or creation of start events for one or more steps. If one of the XOR, OR or AND gateways was used after the step in the business process, then such step is interrelated with many other subsequent steps. Each relationship is described by the probability of path selection for XOR and OR gates. In case of the XOR gate, exactly one next step in the process is selected within the probability of the relations that add up to 1. In case of the OR gate, the probability of each path is calculated independently. Their sum may be greater than 1, hence, one or more of the following steps may be chosen. For the AND gate, all of the following steps are always run. The same is true for connecting gates (before the step) - in this case, it is essential to first complete one or more of the previous steps before starting the next step.

The above-described process is executed for every instance of the business process that can occur in large numbers. Simulation is performed for a fixed simulation time defined as a parameter before its beginning.

OPTIMIZER IMPLEMENTATION

Population Initiation

The optimizer is based on multiple simulations organized in a genetic algorithm and the brute force algorithm. The brute force algorithm is a loop that executes the genetic algorithm for each combination of the Service Composition and Service Scheduling methods. The genetic algorithm is used to find optimal resource allocation for the Service Composition and Service

Scheduling methods. The genetic algorithm starts with the generation of the population of genotypes. A genotype is a Boolean matrix that shows allocation of components to servers, where 1 indicates that the component has been allocated to the server. To generate an initial population, two layers of randomness are applied. The first one is the generation of numbers between 0 and 1 for each genotype. The number represents the probability of allocation. The second layer is randomization of 0 or 1, which shows whether the component is allocated to the server. This randomization is done with the probability of allocation from the first layer. In this way, not only different allocations are analyzed, but also solutions with different density of allocation.

After the generation of the initial population, it is necessary to adjust the solutions so that all restrictions are met. To that end, the genotypes that do not meet the first two restrictions are selected. Subsequently, the loop is executed for each server that is overloaded. Within each iteration of the loop, the randomly selected components and their execution environments are removed from the server allocation. The process continues until the server is not overloaded anymore. To guarantee that the third restriction is met, all genotypes that have at least one component not allocated to any server are selected. Thereupon, in case of each such component, a randomly selected server having enough resources to handle it is allocated.

Solution Evaluation and Selection

To evaluate the solutions, each genotype is simulated. Each genotype is simulated multiple times to minimize the influence of randomness. During the simulation, the values of every decision criterium are subject to measurement. The criterion evaluation is averaged across multiple simulations of the same solution. After all solutions within a given population have been simulated, the ideal solution with the best values in each criterion in the population is formulated. The ideal solution is hypothetical. It is used as reference to evaluate other solutions. The values of all solutions in all criteria are normalized, where 0 constitutes the criterion value of the ideal solution and 1 - the worst value ever found. Subsequently, the distance from a given solution to the ideal one is calculated using the Euclidean metric. Best solutions are those that are closest to the ideal solution. The next step is the selection process. To do that, genotypes are sorted from best to worst. Survival chances are allocated to all solutions linearly, where the best solution has the probability of survival to the next population equal to 1. The worst solution has the probability of 0. The solutions are eliminated from the population according to the probability assigned thereto.

Crossover and Mutation

The final steps of the genetic algorithm are crossover and mutation. During the crossover, new solutions are generated. Each new genotype has two parents randomly

selected. The probability of being selected is higher in case of better solutions. The weight of being picked as a parent is the same as the probability of survival in the previous step. Each pair of parents has two children. The genes of the children are randomly picked from one of the parents. If the first child inherited a gene from one parent, then the second child inherits it from the other parent. Next, new genotypes are undergoing mutation. There is a small probability that each gene will be mutated. Mutation is changing the value of a gene from 0 to 1 or from 1 to 0. This is to broaden the spectrum of the solutions searched. Finally, to ensure that all restrictions on solutions are met, the same algorithm as in case of the population initiation is performed.

PERFORMANCE

The presented optimizer was used to find solutions to multiple problems. To test convergence, the same problem was optimized with multiple times using different seeds. Convergence depends on the following parameters:

- number of genetic algorithm iterations,
- size of population,
- variance of the simulation output data (which may be minimized by increasing the simulation time and number of repetitions),
- size of the problem (number of: servers, components, business processes, etc.).

The values of such parameters may be increased to achieve better convergence, but it would also make the optimization time longer. In the end it all comes down to the processing power and time. The more we have of those, the better convergence may be achieved.

The simulation time is not only dependent on how long it has to be processed, but also on how many events must be executed in the environment. The number of events depends on the number of business processes and their two attributes:

- expected value of a random variable of time between business process invocations,
- expected number of steps in a process to complete it.

Furthermore, there is one more value that has great impact on the simulation time. It is the ratio of load generated by the processes to available resources. The more load generated by the processes in comparison with the available resources, the longer service queues on the components. When the service instance is executed, the estimated execution times of all other service instances are updated. In case of a long queue, a lot of services have to be updated. According to the data gathered by a Java profiler, the service instance updating the process may consume up to 90% of the computation power provided to the optimizer. In case of a very short queue (shorter than 1 on average), it does not consume so many resources and the simulation process may be performed up to 10 times faster. Additionally, during experiments, it turned out that with short queues, the Service Scheduling algorithm of the optimal solution had a very low convergence. It is almost as if it was selected randomly. The reason for this is that when the queue is

short, then the Service Scheduling algorithm has nothing to optimize.

Table 1 shows the execution times on different parameters for the problem that comprises: 100 business processes and an average of 17.5 steps needed for their execution. Each business process definition had an average time between launches in a range of 1 to 500. This average time was a parameter for calling subsequent instances of business processes with exponential distribution. However this average time between business process instances was subject to changes in different simulation variants. The services were assigned an average execution time on a standard processor from 5 to 120 (note that one server has multiple processors so real execution time can be much shorter). This value was an input parameter for the execution time of individual service instances that were randomized according to the normal distribution (both the average time and standard deviation). The simulation length was 2000 time units. Every solution was simulated 10 times to evaluate the values of its decision criteria. There was $3 \cdot 10^6$ solutions searched. Each solution had resource allocation problem size of 50 components allocated to 30 servers. The ratio between the resources required by the processes and the resources available on servers were from 2 to 1 (note that not all servers are used to minimize the resources consumed in the optimal solution). Optimization has been done on virtual machine with 3 Intel Xeon E5-2640 2.60 GHz virtual cores.

Table 1. Optimization times of problems with different amount of business process instances

Average time between invocations	Average number of events in simulation	Total optimization time (days)	Average time of evaluation of 1 solution (seconds)
10	350000	15,202	0,438
20	175000	8,445	0,243
40	87500	6,110	0,176

The differences in values of decision criteria between the variants varied between 1% and 15%, except for the variance of the business process execution time, which differed between 18% and 47%. Best convergence (1% difference between best solutions of different seeds) had been achieved when optimization was performing for 60 days for a problem of same size as presented above.

Another experiment was also carried out on the same problem, but with different population sizes and the number of iterations as shown in table 2. Each experiment was repeated five times, and the results of the time of each repetition are shown in Figure 3. The results show that the number of iterations in the proposed solution has a much greater impact on simulation time than the size of the population. Experiment 4 was carried out much longer than experiment 3 despite the need to review the same number of solutions. The reason for this may be greater optimization convergence with the

parameters of experiment 3 and thus a greater proportion of simulations closer to optimal, which require less computing power. Concept to combine all three types of optimization methods of business processes in SOA is new so it is impossible to compare those results to others.

Table 2. Parameters of experiments

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Iterations	5 000	5 000	5 000	10 000
Population size	100	150	200	100

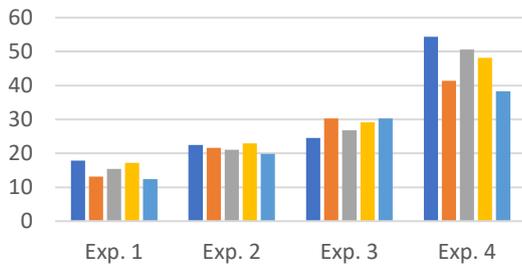


Figure 3 Execution time (days) of each experiment repetition.

SUMMARY

The results of the proposed method may be interpreted dually. On one hand, it may give optimal solutions with high convergence, but on the other, it greatly depends on the resources and the problem size. To optimize it, greater utilization of multi-threading should be implemented in the simulator. Each simulation should be run as a separate thread. If further increase of its efficiency is needed, additional changes could be made. For example, each solution could be simulated on a different virtual machine. It is up to the user to decide if such efficiency of the optimizer is satisfactory. Still, the core of the optimizer would be same as described in the paper. What is more, with proper amount of time, it may have high convergence. As this method is new, it is impossible to compare it to other optimizers.

REFERENCES

- Almeida J.; V. Almeida; D. Ardagna; C. Francalanci; and M. Trubian. 2006. "Resource Management in the Autonomic Service-Oriented Architecture". *IEEE International Conference on Autonomic Computing*. Dublin, Ireland.
- BPMN Specification documents. Accessed 31.01.2020. <https://www.omg.org/spec/BPMN/2.0/About-BPMN/>.
- da Silva A. S.; H. Ma; and M. Zhang. 2015. "A GP approach to QoS-aware web service composition including conditional constraints". *IEEE Congress on Evolutionary Computation (CEC)*. Sendai, Japan.
- Dyachuk D.; and R. Deters. 2008. "Ensuring Service Level Agreements for Service Workflows". *IEEE International Conference on Services Computing*. Honolulu, USA.
- Dyachuk D.; and R. Deters. 2007. "Service Level Agreement Aware Workflow Scheduling". *IEEE International Conference on Services Computing*. Salt Lake City, USA.
- Ebrahim G. A. 2011. "Intelligent Composition of Dynamic-Cost Services in Service-Oriented Architectures". *Fifth UKSim European Symposium*. Madrid, Spain.
- Huang K. C.; Y. C. Lu; M. H. Tsai; Y. J. Wu; and H. Y. Chang. 2016. "Performance-Efficient Service Deployment and Scheduling Methods for Composite Cloud Services". *IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC)*. Shanghai, China.
- Mennes R.; B. Spinnewyn; S. Latre; and J. F. Botero. 2016. "GRECO: A Distributed Genetic Algorithm for Reliable Application Placement in Hybrid Clouds". *5th IEEE International Conference on Cloud Networking (Cloudnet)*. Pisa, Italy.
- Schmid M. 2011. "An approach for autonomic performance management in SOA workflows". *12th IFIP/IEEE International Symposium on Integrated Network Management and Workshops*. Dublin, Ireland.
- Wang Z. J.; Z. Z. Liu; X. F. Zhou; and Y. S. Lou. 2011. "An approach for composite web service selection based on DGQoS". *The International Journal of Advanced Manufacturing Technology*, 56 (9). London, Great Britain 1167-1179.
- Woźniak A.; and T. Nowicki. 2019. "The Problem of Effective Deployment Architecture in SOA". *Computer Science and Mathematical Modelling* (9). Warsaw, Poland. 33-44.
- Xianwen F.; F. Xiaoqin; and C. Jiang. 2009. "An Efficient Approach to Web Service Selection". *Web Information Systems and Mining: International Conference*. Shanghai, China. 271-280.
- Xie L.; J. Luo; J. Qiu; J. A. Pershing; Y. Li; and Y. Chen. 2008. "Availability "weak point" analysis over an SOA deployment framework". *IEEE Network Operations and Management Symposium*. Salvador, Bahia, Brazil.
- Zhang C.; R. N. Chang; C. S. Perng; E. So; C. Tang; and T. Tao. 2009. "An Optimal Capacity Planning Algorithm for Provisioning Cluster-Based Failure-Resilient Composite Services". *IEEE International Conference on Services Computing*. Bangalore, India.
- Zhao Y.; W. Tan; and T. Jin. 2017. "QoS-aware Web Service Composition Considering the Constraints between Services". *12th Chinese Conference on Computer Supported Cooperative Work and Social Computing*. Chongqing, China.



ADRIAN P. WOŹNIAK studied at the Military University of Technology in Warsaw, where he obtained a degree in Information Technology in 2012. Subsequently, he worked as a system and business analyst. In this role, he designed many SOA systems. After that, his position changed to Architect and then Main IT Architect in the largest retail company in Poland, which allowed him to gain extensive experience in the field of the SOA system optimization problem. At the Military University of Technology, he is a lecturer in software engineering, system design and system integration. The subjects he teaches are also his main areas of research. His e-mail address is: adrian.wozniak@wat.edu.pl

EFFICIENT TASK PRIORITISATION FOR AUTONOMOUS TRANSPORT SYSTEMS

Maximilian Selmair
Vincent Pankratz
BMW Group
80788 Munich, Germany
Email: maximilian.selmair@bmw.de

Klaus-Jürgen Meier
University of Applied Sciences Munich
80335 Munich, Germany

KEYWORDS

Automated Guided Vehicle; autonomous Automated Guided Vehicle, Self-Driving Vehicle; Task Allocation; Pick-up and Delivery; Prioritisation

ABSTRACT

The efficient distribution of scarce resources has been a challenge in many different fields of research. This paper focuses on the area of operations research, more specifically, Automated Guided Vehicles intended for pick-up and delivery tasks. In time delivery in general and flexibility in particular are important KPIs for such systems. In order to meet in time requirements and maximising flexibility, three prioritisation methods embedded in a task allocation system for autonomous transport vehicles are introduced. A case study within the BMW Group aims to evaluate all three methods by means of simulation. The simulation results have revealed differences between the three methods regarding the quality of their solutions as well as their calculation performance. Here, the *Flexible Prioritisation Window* was found to be superior.

LIST OF ABBREVIATIONS

AGV Automated Guided Vehicle
FMS Flexible Manufacturing System
GAP Generalised Assignment Problem
HM Hungarian Method
ILP Integer Linear Programming
KPI Key Performance Indicator
NVA-share non-value-adding share
POI Point of Interest
SDV Self-driving Vehicle
STR Smart Transport Robot
VAM Vogel's Approximation Method
VAM-nq Vogel's Approximation Method for non-quadratic Matrices
aAGV Autonomous Automated Guided Vehicle

INTRODUCTION

Constantly changing external conditions associated with global competition, individualised customer demands and more complex production structures force companies to reorganise their production systems to become more flexible (Braunisch 2015). Amongst others, Flexible Manufacturing Systems (FMSs) rely on so-called Autonomous Automated Guided Vehicles (aAGVs), which are able to find the fastest way from a source to a sink by themselves and drive around obstacles, something that Automated Guided Vehicle (AGV) are unable to achieve. The function of the superordinate fleet management system is to allocate the transportation tasks to agents in form of vehicles (Wagner 2018). According to Le-Anh and Koster (2006), there are three different main criteria that ought to be considered when allocating tasks: time, utilisation and distance. The dispatching can be either static or dynamic. When operating with static dispatching, a task allocation decision is made and carried out without considering system changes that occur during the execution of said task (Gudehus 2012). These system changes (for instance task urgency, deadline changes, machine malfunctions) can make the planned sequence inoperative (Ouelhadj and Petrovic 2009). Dynamic dispatching, in contrast, adapts the planned sequence of the transportation tasks to changes either in certain time intervals or when a certain event happens. This increases the flexibility of the task allocation process (Gudehus 2012). After a task is allocated to a vehicle, the vehicle drives to the source of the task and collects the material. Subsequently, it moves to the source and delivers the container at the designated place (sink). During this process, delays can occur due to pedestrians, other vehicles or obstacles that have to be circumvented (Clausen et al. 2013).

RELATED LITERATURE

The issue of transportation is an extensively studied topic in operational research (Díaz-Parra et al. 2014).

The methods for solving the assignment problem aim to minimise the total transportation costs while bringing goods from several supply points (e. g. warehouses) to demand locations (e. g. customers). In general, each point of departure features a prearranged amount of goods that can be distributed. Correspondingly, every destination requires a certain amount of units (Shore 1970). The underlying use case, where tasks have to be assigned to Self-driving Vehicles (SDVs), differs in some regards from the classical transportation problem. In our case, each vehicle has a capacity restriction of one, i. e. a maximum of one load carrier can be transported at a time. Furthermore, each task corresponds to a demand of one. This means that every task can only be allocated to one single vehicle. In the research domain, the just mentioned case is known as the Generalised Assignment Problem (GAP). In an effort to minimise it, researchers aim for minimal costs Z between n tasks and m agents while each task is assigned to one agent (Kundakcioglu and Alizamir 2009). According to Srinivasan and Thompson (1973) the formulation of the GAP is:

$$\text{minimize } Z = \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \quad (1)$$

In subject to the constraints:

$$\sum_{j \in J} r_{ij} x_{ij} \leq b_i, \text{ for } i \in I \quad (2)$$

$$\sum_{i \in I} x_{ij} = 1, \text{ for } j \in J \quad (3)$$

$$x_{ij} \in \{0, 1\}, \text{ for } i \in I \text{ and } j \in J \quad (4)$$

A set of agents $I = \{1, 2, \dots, i, \dots\}$ has to be assigned to a set of tasks $J = \{1, 2, \dots, j, \dots\}$. Variable c is representing the costs accrued when the task j is allocated to the agent i . r_{ij} represents the capacity which is needed when the task j is allocated to the agent i . Variable b represents the available capacity of agent i in total. The binary variable x_{ij} equals 1 if task j is assigned to agent i , otherwise it remains 0.

There are several methods that solve the GAP. Beside optimisation methods like Integer Linear Programming (ILP), there are, for example, algorithms like the Hungarian Method (HM) proposed by Kuhn (1955), Vogel's Approximation Method (VAM) proposed by Reinfeld and Vogel (1958) or Vogel's Approximation Method for non-quadratic Matrices (VAM-nq) proposed by Selmaier et al. (2019). The last method was used in the simulation study of this paper due to its superior calculation performance and satisfactory results in terms of non-quadratic matrices.

DEVELOPED PRIORITISATION METHODS

Solving the GAP minimises the transport costs, in this case operationalised as meters driven, between

tasks and agents. For the presented use case, transport costs can be equated to the transport distances or – for prioritised orders – the transportation duration. Minimising transport distances is an important goal as it leads to faster task processing and less traffic on the routes. As the transport route between source and sink is the same for every vehicle, in this scenario only the empty drive distance from the current location of each vehicle to the source of a task is minimised.

If there are consistently more tasks than free vehicles in a system, waiting times can occur for tasks that could not be allocated in former allocation cycles due to high transport costs or long pick-up distances. These waiting times can cause the performance of the overall system. In automotive intralogistics, which oversees the supply of parts to assembly lines, delays can cause high costs and must be reduced as much as possible. Consequently, the following prioritisation methods were developed to unite efficiency and in time delivery in the task allocation of autonomous transport systems.

Fix Prioritisation Window

The first prioritisation method presented here is the *Fix Prioritisation Window*. This method was inspired by the backwards calculation in dynamic task dispatching, which can be found in the article contributed by Gudehus (2012). The idea behind this method is to force the allocation of tasks that are at risk of becoming delayed. When a task exceeds the point of time where a in time delivery is critical, it is marked “prioritised” and is thus allocated in the next allocation cycle. In this manner, any delays should be avoided. The prioritisation window is the period between the deadline of a task and time of prioritisation. During this time the task must be allocated to a vehicle, the vehicle must drive to the source and collect the material, drive to the sink and deposit the material. This method requires that the prioritisation window has the same length for each task and is determined by analysing the data of prior simulations.

The prioritisation window must allow enough time for the transport vehicles to fulfil tasks with long distances between the source and the sink as well as include a buffer for unexpected delays. However, a too generously planned prioritisation window will reduce the system's flexibility and thus impact negatively on its efficiency. Furthermore, if too many tasks were to become prioritised and had to compete for vehicles, delays might become inevitable.

Flexible Prioritisation Window

The *Flexible Prioritisation Window* is quite similar to the *Fix Prioritisation Window*, yet they differ in that the *Flexible Prioritisation Window* assigns an individual time window for each task, depending on the distance between source and sink and the time required to cover this distance. In order to determine the duration of this time window, an analysis of statistical data is performed. In this case, the time between the allocation of the task and the delivery is measured,

<i>Parameter</i>	<i>Characteristics</i>			
Prioritisation Methods	No Prioritisation	Fix Prioritisation Window	Flexible Prioritisation Window	Bidding Approach
Number of Vehicles	10	25	50	
Ratio Vehicles / Tasks	1 / 0.5	1 / 1	1 / 1.5	1 / 2

TABLE I: Experimental Plan illustrated as Morphological Box

Combination (vehicles / mission pool)	Fix Time Window	Flexible Time Window	Vehicle delivers in time? Yes and more than two others	Bidding Factor
10 / 5	468 s	309 s	Yes and two others	1
10 / 10	468 s	305 s	Yes and one other	0.7
10 / 15	519 s	407 s	Yes – the only one	0.2
10 / 20	475 s	298 s	No	0.1
25 / 13	485 s	309 s		3
25 / 25	437 s	279 s		
25 / 38	437 s	346 s		
25 / 50	485 s	344 s		
50 / 25	452 s	280 s		
50 / 50	439 s	244 s		
50 / 75	442 s	279 s		
50 / 100	444 s	291 s		

TABLE II: Fix and Flex Time Window Values for the Scenarios of the Simulation-Study

but without the theoretical travel time between source and sink (distance divided by the average speed of the vehicle). This theoretical travel time is different for each task and is added to the determined time window value. This total duration represents the *Flexible Prioritisation Window*, which is calculated individually by backward-scheduling from the deadline of each task.

This approach can be considered an extension of the *Fix Prioritisation Window* method. It is expected that by means of the *Flexible Prioritisation Window*, fewer tasks are going to be prioritised, which will result in advantages in terms of efficiency and in time delivery.

Bidding Approach

The bidding approach adds a further strategy for prioritising tasks increased in complexity. Its operation method is different from the former presented approaches. The main difference being that this approach does not allocate prioritised tasks ahead of non-prioritised tasks, but adjusts the transportation costs when tasks become critical, ensuring these tasks are allocated and are dispatched on time. This approach was inspired by multi-attribute dispatching rules like Lampe and Clausen (2006) or Klein and Kim (1996) which take into account multiple criteria to determine the priority / costs of a task. These multi-attribute costs are processed by the GAP method which is solved

TABLE III: Bidding Factors for different Situations

by the VAM-nq heuristic developed by Selmaier et al. (2019).

The bidding factors that lower or raise the costs are calculated by including the distance of pick-up and transport from source to sink in a formula, to establish whether a vehicle is able to carry out the task in time. The prioritisation window w is compared with the period between the deadline of the task and the current time. If w is smaller than the period between the deadline and current time the task can be executed on time.

$$w = \frac{d_P + d_S}{v} * (1 + u) + t_P + t_D \quad (5)$$

Where d_P is the distance from the current location of the vehicle to the source, d_S is the distance between the source and sink of the task, v is the average speed of the vehicle, u is the uncertainty factor that increases the travel time to compensate for unforeseen events, t_P is the time for collecting the material at the source, t_D is the time to deposit the material at the sink.

After ascertaining if the vehicle is able to fulfil the task in time, the vehicle checks if it is the only one to do so and recalculates the costs for the task with the bidding factor in Table III. These factors were determined by means of a parameter simulation.

SIMULATION STUDY

The following study built in the software *AnyLogic* aims to examine whether the presented methods of prioritisation achieve results that are similar in terms of efficiency to those obtained when solving the GAP without prioritisation. The main KPI used to proof the efficiency of the system will be the non-value-adding share of vehicles. Vehicles are not adding value when driving without any load – accordingly they are adding value when delivering a load from a source to its sink. The non-value-adding share of movement per task will

be referred simply to NVA-share in the further course. Additionally, it is proposed that, by means of these measures, better results in terms of deliveries in time will be yielded. Therefore, the different methods were tested in a simulation study with different scenarios (see Table I).

The scenarios are simulated in a homogeneous grid-structured production environment with a total length of 4,072 meters in which a number of 5,000 tasks are processed for each scenario. The ratio between vehicles and tasks remains constant during every simulation run. That is, every time a task is completed, a new one appears – until the limit of 5,000 tasks is reached. The vehicles move with a maximum speed of $1.5 \frac{m}{s}$. Their acceleration is parametrised to $1 \frac{m}{s^2}$, the deceleration is set to $5 \frac{m}{s^2}$. The tasks must be performed at an randomly selected point in time that lies between 9 to 15 minutes after a task has been generated. Additionally, tasks are more likely to be allocated at the beginning of said time interval. A solution is calculated in a cycle-time of 20 seconds to react to changes in the system, such as the appearance of new tasks or changes of the vehicles' status. While picking up or depositing material, the vehicles lower their speed so that the corresponding path is blocked for 25 seconds. In actual industrial applications, during this time period, the vehicles calculate the exact angle in which to drive to the Point of Interest (POI), by means of information gathered by 3D cameras.

The parameters of the *Fix* and *Flexible Prioritisation Windows* are determined by analysing the times between allocation and delivery of a task in a simulation run performed without prioritisation by using the GAP for allocating the tasks. The final parameter was set to the longest time that was measured for each vehicle / utilisation (v/u) combination (equals the 100%-percentile) plus a 20-second-cycle time. Table II shows the parameters for the *Fix* and *Flexible Prioritisation Windows*.

For the uncertainty variable u of the bidding approach, a value of 30 % was used. This was based on the average travel uncertainty of the combinations, which lies between 20 % and 25 %. The set value includes a small buffer for accidental uncertainties.

RESULTS

The following section provides a summary of the results with particular focus on the two main Key Performance Indicators (KPIs): in time delivery and NVA-share of the vehicles. Moreover, it was decided to only include the results of simulation runs with 50 vehicles, as all other scenarios yielded very similar behaviour.

In order to explore the lower range of values associated with the NVA-share of vehicles, a scenario free of any prioritisation rules was utilised. This means that for the task allocation determined by the GAP, all tasks were assigned by following the fewest possible NVA-share (from the overall system perspective) – without imposing any deadlines and thus time pressure. In Figure 1, this value is represented by the first bar, from

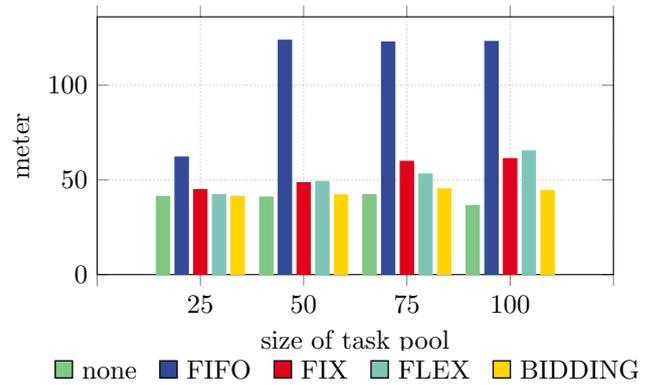


Fig. 1: Mean NVA-share per task for a scenario with 50 vehicles (5,000 Tasks)

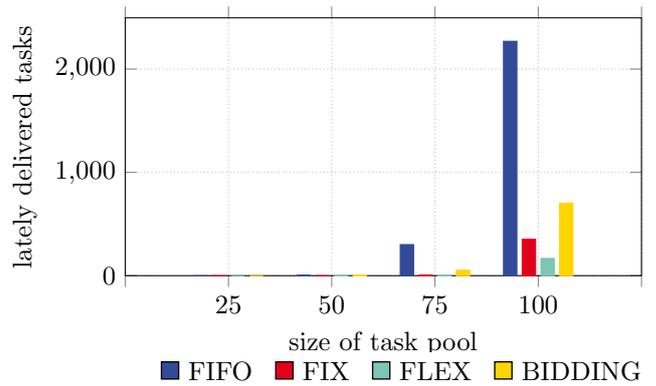


Fig. 2: Late Tasks of the original FIFO method and the three developed prioritisation approaches (5,000 Tasks)

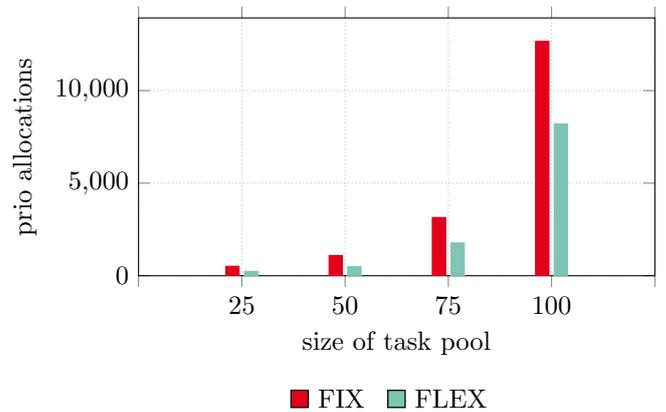


Fig. 3: Total Allocations with Priority for the *fix* and *flex* approach (5,000 Tasks) – Each task can be assigned more than once if decisions are frequently changed

left to right, of every group, where each group represents one scenario of utilisation. The different levels of utilisation are achieved by keeping the simultaneous task pool sizes constant (25, 50, 75 and 100 tasks).

The second bar in Figure 1 represents the originally applied task allocation strategy: first-in-first-out in combination with the nearest-agent-first policy. For

the first scenario with a pool of 25 tasks, the system can freely select the closest of two vehicles, as the task to vehicle ratio is favourable at 1 : 2. As the number of tasks exceeds 50, substantially changing the allocation ratio, the system’s choice is restricted to only one idle vehicle, which may not be the closest. It follows that the pick-up distances, measured in meters, are thus more likely to exceed 120 m.

For low utilisation scenarios (25 and 50 tasks), the *fix* and the *flex* approaches yield results close to those of the *bidding* approach with a maximum difference of 18 %. For a high utilisation (100 tasks), the NVA-share per task is 40 % higher than for the *bidding* approach. This will be elucidated further in the discussion.

The next three bars in the same Figure, bars 3 to 5, represent the pick-up meters for the three developed prioritisation methods in combination with the assignment by means of the GAP. For low utilisation levels, e.g. 25 and 50 tasks in the pool, all three prioritisation methods are quite similar to the “optimal” first bar which is not involving any prioritisation. The differences are only between 1 % and 8 %.

To evaluate all three methods, a closer look at the respective late tasks is deemed essential. Figure 2 illustrates the number of late orders for each scenario of utilisation. For the first scenarios with task pools of 25 and 50, all methods resulted in late orders below 0.3 %. However, the initially applied FIFO method resulted in 6 % late orders for the scenario with 75 tasks, that is, as soon as the content of the task pool exceeded the number of available vehicles. The high-pressure scenario with a task pool of 100, yielded late orders for every applied method. 55 % were on time when applying the FIFO rule, the *fix* approach delivered 93 % of all orders on time, while the *flex* approach resulted in 97 % of tasks being delivered on schedule. The results of the *bidding* approach in combination with the high utilisation scenario showed that only 86 % of all orders were able to be delivered on time.

DISCUSSION

The previously presented simulation results support the notion that the three methods, developed within this scope of research, are quite similar in terms of efficiency (see Figure 1). Furthermore, in high utilisation scenarios (75 and 100 tasks), all three methods yielded superior results in comparison to the currently applied FIFO method (see also Figure 1). Based on the concept that the *fix* and *flex* approach will prioritise more tasks in a high utilisation than in a low utilisation scenario (see Figure 3), it was expected that for high utilisation scenarios, the pick-up meters would exceed those of the scenarios without prioritisation. That is, all three approaches reduce flexibility and force the system to allocate urgent tasks immediately and these priority-tasks are delivered as fast as possible instead of way-efficient.

That means, that the *bidding* approach is able to prioritise each scenario with only little impact on NVA-share of vehicles in terms of meters. Nonetheless, it is presumed that the late tasks contribute to this out-



Fig. 4: The Smart Transport Robot (STR) of the BMW Group in its natural habitat

come: starting with the scenario of 75 tasks, the *bidding* approach begins to deliver undesirable results. In order to ascertain the reason behind this, a close look at the calculation method of the *bidding* approach was necessary. In this approach, the calculation effort for urgent tasks is scaled down to support an urgent allocation. When handling more tasks with fewer available vehicles, the calculation efforts are scaled down in the calculation matrix for formulating the GAP. This leads to a reduction in the relative differences between the already scaled down efforts. Subsequently, tasks which are located peripherally are more likely to be processed than others. Especially these tasks contribute to the large number of delayed deliveries. Furthermore, the *bidding* approach is the approach with the highest calculation effort of all three introduced methods due to its complexity.

The *fix* and the *flex* approach yielded comparatively good results, both in terms of the distance of NVA-share as well as number of late tasks. The disadvantage of the *fix* and *flex* approach is associated with their independence regarding their possible vehicle choice. This independence is assessed through a time buffer that leads to lower flexibility, which in turn ensures that tasks are processed in time. The advantage of both methods is the low calculation effort: only one statistical value has to be determined in advance and applied continuously to each task. As described in the Section *Simulation Study*, this value is determined by using the 100 %-percentile. Due to this fact, most of the values are ascertained with a long buffer (see Table II). Nevertheless, this technique is necessary to ensure a minimal number of delayed tasks.

CONCLUSION

This paper has introduced three prioritisation approaches for a task allocation system of a transport vehicle system. All approaches differ in their complexity and calculation efforts. To evaluate their suitability for industrial application, all approaches were compared by means of simulation. Three different fleet sizes and four levels of utilisation were combined in order to as-

sess performance under a variety of conditions. For the evaluation, two KPIs were focused on: the number of late tasks and the NVA-share in the system, mapped as average distance per task. While the significance of the first KPI is presumed to be self-evident, the importance of the second KPI is associated with resource conservation. That is, by minimising the NVA-share, fewer traffic conflicts occur, which allows traffic to flow more smoothly.

On the whole, all three approaches have qualified for regular utilised transport vehicle systems due to their ability to ensure that tasks are delivered in time. For performance and robustness reasons, the *fix* and *flex* approach are suggested to be more economically attractive within actual industrial use-cases. At the BMW Group, the *flex* approach has been implemented in a self-developed and self-built transport vehicle system (see Figure 4). The critical factor for this decision were the readily comprehensible calculation effort and the consistently positive results.

Further research might evolve the methods in order to provide a higher flexibility to the system whenever possible. Currently, the statistical determined values of Table II are overstated for the most tasks. This issue might be tackled in the next level of development.

REFERENCES

- Le-Anh, Tuan and M. B. M. De Koster (2006). “A review of design and control of automated guided vehicle systems”. In: *European Journal of Operational Research* 171.1, pp. 1–23.
- Braunisch, Dirk (2015). *Multiagentensysteme in der rückführenden Logistik: Entwurf einer Systemarchitektur zur Steigerung der Prozesseffizienz durch dynamische Disposition der Sekundärrohstofflogistik*. Berlin.
- Clausen, Uwe, Peiman Dabidian, Daniel Diekmann, Ina Goedicke, and Moritz Pötting (2013). “Entwicklung und Analyse einer multikriteriellen Einsatzsteuerung von Staplern in einem manuell bedienten Distributionslager.” In: *Simulation in Produktion und Logistik*, pp. 207–216.
- Díaz-Parra, Ocotlán, Jorge A. Ruiz-Vanoye, Beatriz Bernábe Loranca, Alejandro Fuentes-Penna, and Ricardo A. Barrera-Cámara (2014). “A Survey of Transportation Problems”. In: *Journal of Applied Mathematics* 2014.3, pp. 1–17.
- Gudehus, Timm (2012). *Dynamische Disposition: Strategien, Algorithmen und Werkzeuge zur optimalen Auftrags-, Bestands- und Fertigungsdisposition*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.
- Klein, C. M. and J. Kim (1996). “AGV dispatching”. In: *International Journal of Production Research* 34.1, pp. 95–110.
- Kuhn, H. W. (1955). “The Hungarian method for the assignment problem”. In: *Naval Research Logistics Quarterly* 2.1-2, pp. 83–97.
- Kundakcioglu, O. Erhun and Saed Alizamir (2009). “Generalized Assignment Problem”. In: *Encyclopedia of Optimization*. Ed. by Christodoulos A. Floudas and Panos M. Pardalos. Boston, MA: Springer US, pp. 1153–1162.
- Lampe, Heiko and Uwe Clausen, eds. (2006). *Untersuchung von Dispositionsentscheidungen in Umschlagterminals des kombinierten Verkehrs Schiene-Straße*. Logistik, Verkehr und Umwelt. Dortmund: Verl. Praxiswissen.
- Ouelhadj, Djamilia and Sanja Petrovic (2009). “A survey of dynamic scheduling in manufacturing systems”. In: *Journal of Scheduling* 12.4, pp. 417–431.
- Reinfeld, N. V. and W. R. Vogel (1958). *Mathematical Programming*. Prentice-Hall, Englewood Cliffs.
- Selmair, Maximilian, Klaus-Jürgen Meier, and Yi Wang (2019). “Solving non-quadratic Matrices in assignment Problems with an improved Version of Vogel’s Approximation Method”. In: *European Conference of Modelling and Simulation*.
- Shore, Harvey H. (1970). “The Transportation problem and the Vogel Approximation Method”. In: *Decision Sciences* 1.3-4, pp. 441–457.
- Srinivasan, V. and G. L. Thompson (1973). “An Algorithm for Assigning Uses to Sources in a Special Class of Transportation Problems”. In: *Operations Research* 21.1, pp. 284–295.
- Wagner, Rainer Maria (2018). *Industrie 4.0 für die Praxis: Mit realen Fallbeispielen aus mittelständischen Unternehmen und vielen umsetzbaren Tipps*. Wiesbaden: Springer Fachmedien Wiesbaden.



MAXIMILIAN SELMAIR is doctoral student at the University of Plymouth. Recently employed at the SimPlan AG, he was in charge of projects in the area of material flow simulation. Currently, he is working on his doctoral thesis with a fellowship of the BMW Group. His email address is: maximilian.selmair@bmw.de and his website can be found at maximilian.selmair.de.



VINCENT PANKRATZ contributed to the paper during his master thesis which was written in collaboration with the BMW Group. With this thesis, he finished his logistics studies at the Regensburg University of Applied Sciences and works now for a renowned automotive company. His email address is: vincent.pankratz@gmx.de.



KLAUS-JÜRGEN MEIER holds the professorship for production planning and logistic systems in the Department of Engineering and Management at the University of Applied Sciences Munich and he is the head of the Institute for Production Management and Logistics (IPL). His email address is: klaus-juergen.meier@hm.edu.

AN APPROACH TO CREATING A SIMPLE DIGITAL TWIN FOR OPTIMIZING A SMALL ELECTRIC CONCEPT VEHICLE DRIVETRAIN

Tamás Dóka

Péter Horák, PhD

Department of Machine and Product Design

Faculty of Mechanical Engineering

Budapest University of Technology and Economics

1111, Műegyetem rkp. 3, Budapest, Hungary

Email: doka.tamas@gt3.bme.hu

KEYWORDS

3D Model-Based Simulation; Digital Twin; Electric Vehicle Drivetrain

ABSTRACT

Since modeling and simulation are integral tools in engineering, the question is not *if* they should be used in a design process, but rather *how* they should be used to deliver the best solutions. The objective of this paper is to outline an approach to creating a simple Digital Twin for a small electric vehicle drivetrain utilizing only parametric 3D CAD models, widely used simulation tools and some programming libraries. First, the concept of the Digital Twin, its benefits, then the possibilities of using Generative Design are briefly introduced, afterwards electric vehicles' advantages are reviewed. In an example project the properties and opportunities of the 3D CAD- and simulation models are demonstrated. Finally, future improvements and automated optimization opportunities are discussed.

INTRODUCTION

Before today's modeling and simulation technologies have been emerged, designing a system and ensure its proper behavior was expensive and time-consuming. The only way to test a system in operation was to build it physically and to subject it to effects and impacts that the designers thought would be necessary (Grieves and Vickers 2017). In the second half of the 20th-century, Computer-Aided Design (CAD) software solutions enabled to create different variations of a system relatively easy. Finally, with a Digital Twin, in theory, we could analyze any product's behavior in different environments without having the physical representation itself.

With the recently available computational power and cloud-based services, the use of complex simulations and detailed virtual prototypes are no longer privileges of the biggest companies, but they can be created and also run on personal computers. In this paper, an example is shown how a simple Digital Twin of a product

can be set up and prepared to be used for optimization, using only open-source or student license software solutions. The subject of this example is a small electric concept car's drivetrain; therefore, the properties of a battery electric vehicle are briefly introduced.

Digital Twins

The term Digital Twin (DT) has many slightly different definitions which are slowly changing through the years. "The Digital Twin is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level. At its optimum, any information that could be obtained from inspecting a physical manufactured product can be obtained from its Digital Twin." (Grieves and Vickers 2017). In that way, a DT could exist without physical representation and could serve as a substitute for real prototypes.

In applications like aerospace or automotive industries, any modification to the product could generate unforeseen consequences to the whole system, so these effects should be adequately modeled and simulated before applying them to the system in operation (Goossens 2017). Not long ago, Digital Twins have been developed in a bottom-up philosophy after the real system was specified. The Digital Twin concept was used to create a virtual environment where a detailed simulation model is running. Based on that model and sensor information from the real physical world, the behavior of the real-world twin could be estimated. This approach has many case studies where the existing system is monitored, the Digital Twin could easily detect occurring problems and practical solutions could be calculated to solve these issues.

Traditionally in the conceptual phase, possible design alternatives were explored by engineers, which required a lot of experience and time. CAD models and simulations were only involved in the process after the design space was narrow enough to analyze, validate, and fabricate the design (Khan and Awan 2018). A detailed model for a product or process is not only capable of examining the system based on real-life data, but it can be leveraged in the concept design phase to

define, test, and evaluate different variants of the system. In this method, the virtual model is not only used for diagnosis but to find out which version of the system should be built in the first place. It is more common today to utilize methods that offer a more standardized description of the models (Rodič 2017), allowing to use optimization algorithms acting directly on the digital model by modifying its parameters and comparing the simulation results. The Digital Twin technology, combined with novel optimization algorithms using artificial intelligence, could generate feasible systems that not only correspond with the requirements but are optimal in the prescribed aspects.

Generative Design

It is usually hard to say how a specific parameter will affect the whole system without knowing the system itself. In many fields during the concept design phase, even experts are using best practices to set up the initial boundaries of the product. This method could lead to sub-optimal solutions even if very precise optimization takes place in later stages.

Generative Design systems are using parametric design, optimization and simulation techniques, which allow engineers to iterate through a large number of design alternatives. Taking a problem definition as input Generative Design systems could create a set of optimal solutions for the given problem (Khan and Awan 2018). An example is shown in Figure 1.

Commercially available Generative Design systems are promising tools for creating complex mechanical parts for a given load- and constraint set using artificial intelligence and topology optimization. This method assumes that the surroundings and functions of the element are known. However, the constraints and loads acting on the part usually depend on other members of the system, so these generated solutions are only suitable for situations where all other components of the system remain the same. Still, the idea that engineers should only carefully define the functions and coarse boundaries of a system and artificial intelligence could do the rest of the work could create previously unimaginable new inventions.



Fig. 1. Generated Design Variations of a Motorcycle (Khan and Awan 2018)

Generative Design combined with Digital Twin technology could allow the system-level optimization where components are heavily co-dependent, such as in electric vehicles.

Small Battery Electric Vehicle Concept

As the present trend suggests, electric vehicles are likely to replace internal combustion engine (ICE) vehicles in the near future (Un-Noor et al. 2017). This trend could be explained by electric vehicles (EVs) being more environmentally friendly, quiet, easy to operate, require less money for fuel and also, they provide instant torque from the startup. EVs are the unquestionably better choice for urban transport, but for longer journeys, two more factors come into question: power and range. Providing a bigger range needs more batteries; therefore, the overall mass of the vehicle grows; thus, the power consumption increases, limiting the achievable range of the vehicle. Finding the right size and arrangement of the electric powertrain components is not as evident as it is in ICE vehicles, because power can be transmitted through electrical wires, enabling to create very different configurations. Moreover, most of the components' parameters are determined by other parts (Un-Noor et al. 2017); thus, optimization becomes even more complicated.

In this paper, an approach for creating a Digital Twin is demonstrated on a small electric vehicle concept. The powertrain of an electric vehicle consists of an electrical and a mechanical subsystem; thus, it is necessary to accurately model and analyze them together to get a real insight into the vehicle's dynamic behavior (Park et al. 2014).

APPROACH TO CREATING A SIMPLE DIGITAL TWIN

As mentioned before, a Digital Twin is basically a set of information about an entity permitting to analyze it from different aspects accurately in a virtual environment. Although every area (mechanical engineering, electrical engineering, etc.) deals with the same product, each of these areas approach the parts that make up these components in a different way (Grieves and Vickers 2017). At present, computers with relatively high computing capacity are affordable, and many modeling, simulating, and optimizing tools and programs are available. Thus the opportunity of using Digital Twins is at hand for smaller businesses and smaller projects too. The design process could be even more improved if we could integrate the available software solutions which we are using, allowing the data exchange between them. Usually, the connection within these software products is not provided, so technically, it requires effort to implement such complex systems. The key to creating a consistent Digital Twin is to persist a homogeneous perspective of the information across functional boundaries. This can be realized by having an application that controls and manages data between different platforms and areas, as shown in Figure 2.

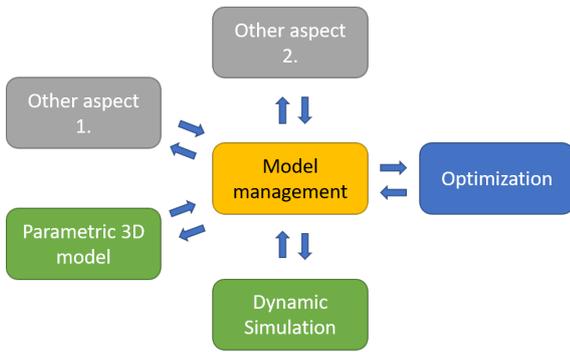


Fig. 2. Block Diagram of the Digital Twin Environment

In the example project, the vehicle was approximately modeled, focusing more on its drivetrain. A detailed, skeleton based top-down 3D model of the drive module (electric motor, fixed gear speed reducer and differential) was modeled and built (see in Figure 3) for further measurements, parameter identification and testing. A simple dynamic simulation was created, using parameters from the CAD models to analyze the vehicle's longitudinal behavior and to provide a basis for optimization in the future.

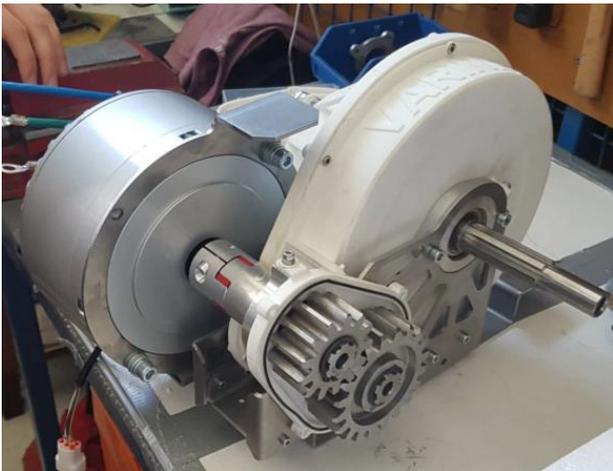


Fig. 3. Drive Module of the Small Electric Vehicle Concept

Parametric 3D model

A Computer-Aided Design (CAD) model is typically used to visualize the entity, how it will look like in its physical form. Almost every product design starts with an approximate 3D model after the main concept was laid out. These 3D models represent the product's mechanical and physical properties, such as total mass, material properties, or geometric boundaries. Today's advanced CAD systems are capable of performing many different tasks, such as FEM analysis, topology optimization, motion- and dynamic simulations. These integrated CAD systems provide pervasive solutions, but dynamic simulations usually do not require the detailed 3D representation of a product to deliver the desired results or, on the contrary only the aggregated properties of a body (mass, moment of inertia, center of gravity) should be taken into consider-

ation to avoid long computational time.

In the project the vehicle was modeled in PTC Creo 4.0, because it is a high-level CAD system equipped with a wide variety of embedded modules, which can be automated. Make use of the skeleton-based top-down approach, all of the key parameters could be modified, and the assembly regenerated through the managing algorithm.

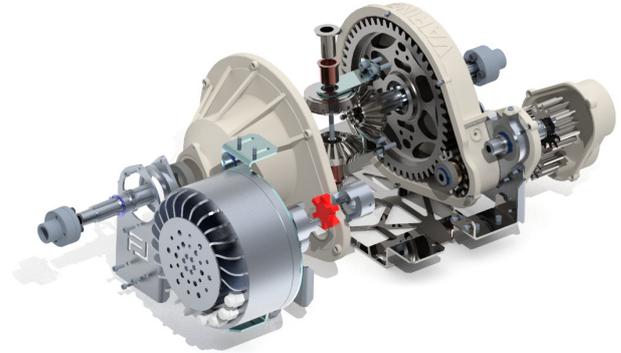


Fig. 4. Exploded View of the Drive Module

Simulation model

Using only aggregated parameters from the 3D CAD models (not the actual 3D mesh) enables us to simulate the system efficiently. Even without a detailed 3D model, simulation can be processed with approximate parameters for estimations. For example, in the early stages, the total mass of the vehicle is unknown, but the drive module could be tested with different scenarios. Later, when the actual vehicle model is available, approximate parameters could be replaced with precise ones. In the simulation model, a simplified, single mass point representation of the vehicle was implemented, which analyses only the longitudinal motion of the car. In general, there are two main approaches towards the longitudinal motion simulations of vehicles, kinematic and dynamic simulations. In case of a kinematic simulation, the actual state of the simulation components is calculated backward from a given driving condition, using the gear ratios and efficiencies to determine the required input values of the propulsion unit. In a dynamic simulation, where the calculations are forward-directed, a driver model calculates the torque demand of the vehicle from the current vehicle speed and provides the corresponding input to the propulsion unit. In the following blocks, the states of the drivetrain components are calculated, resulting in the vehicle's actual speed, which is connected to the driver model to close the simulation loop. Using MATLAB Simulink in the project, a dynamic simulation was implemented because this kind of model ensures a more realistic and accurate simulation of the vehicle's drivetrain than the kinematic approach (Winke and Bargende 2013). The simulation model is divided into three main blocks (see in Figure 5):

- *Driver module*
- *Drivetrain module*
- *Vehicle module*

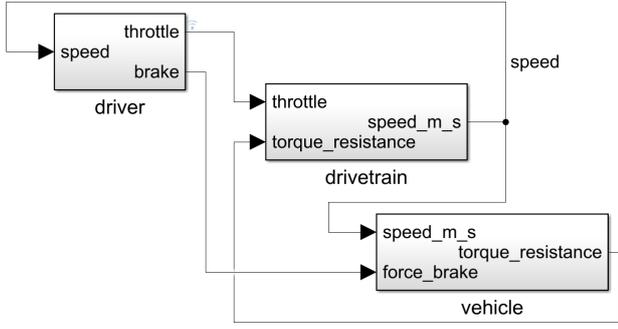


Fig. 5. The Basic Structure of the Dynamic Simulation Model

The *driver module* controls the electric motor's throttle based on the New European Driving Cycle (NEDC). If the current speed of the vehicle is lower than the desired, the module raises the throttle signal, if the opposite is true, it lowers the throttle signal, or even initiates braking.

The *drivetrain module* simulates the dynamics of the electric motor, the fixed-gear drive module, and the vehicle's wheel (see on Figure 6), using the throttle signal and the resistances acting on the vehicle as inputs to calculate the vehicle's speed. Each block calculates its inner state according to the input and output torques and angular accelerations, assuming that the system is totally stiff, and does not contain any non-linearity such as backlashes.

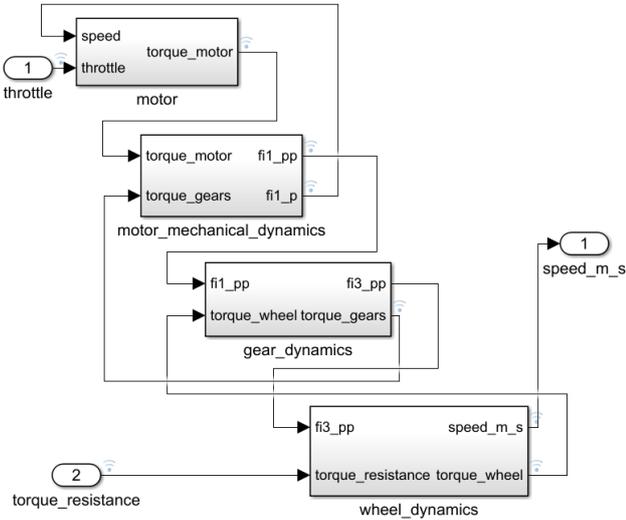


Fig. 6. Drivetrain Simulation Module's Block Diagram

Inside the *drivetrain module*, the *motor* block provides torque input to the *motor mechanical dynamics* block based on the current motor speed, the throttle signal (which can be between 0 and 1), and a look-up-table (LUT) which contains the motor's torque-speed characteristics. The LUT determines the maximal motor torque at the given speed; finally, this value is multiplied by the throttle signal value. This method enables to create the needed motor torque between zero and the nominal maximum torque of the motor for each motor speed. In the *motor mechanical dynamics* block, the angular acceleration of the motor is calculated as

shown in Equation (1):

$$\ddot{\varphi}_m(t) = \frac{\tau_{motor}(t) - \tau_{gears}(t)}{J_{motor}} \quad (1)$$

where,

- φ_m is the motor's angular position,
- τ_{motor} is the motor torque,
- τ_{gears} is the input torque for the gearbox,
- J_{motor} is the moment of inertia of the motor's rotating parts

Integral of the angular acceleration over time gives the motor speed, which is connected to the *motor* block.

The *gear dynamics* block consists of the input-, intermediate- and differential shaft assemblies, which are connected together by gear pairs (see on Figure 7).

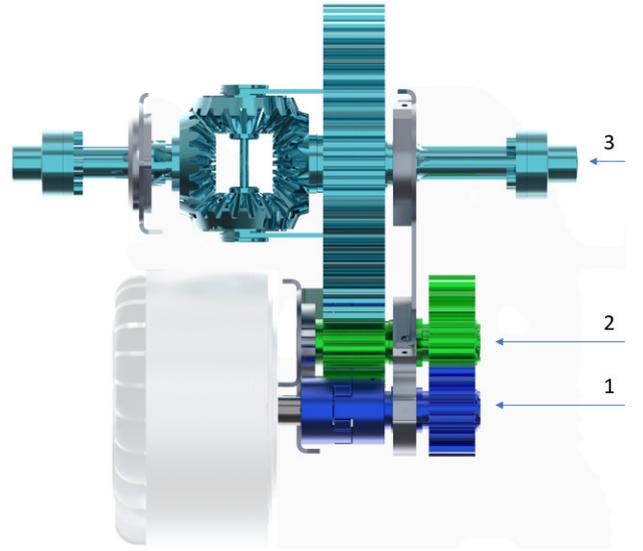


Fig. 7. 3D Model of the Fixed-gear Drive Module: Input- (1), Intermediate- (2), Differential Shaft (3) Assemblies

Since the motor is connected to the input shaft assembly, the angular acceleration of these two is equal. To ensure this the input torque for the gearbox (τ_{gears}) is calculated as shown in Equation (2):

$$\tau_{gears}(t) = \tau_{l1} + J_1 \cdot \ddot{\varphi}_m(t) + \frac{\tau_{l2} + J_2 \cdot \ddot{\varphi}_m(t) + \frac{\tau_{l3} + J_3 \cdot \ddot{\varphi}_m(t) + \tau_{wheels}(t)}{i_{2,3}}}{i_{1,2}} \quad (2)$$

where,

- τ_{wheels} is the input torque for the wheels
- τ_{l_j} is an estimated torque loss (e.g. from bearings) for the j-th shaft assembly,
- J_j is the moment of inertia of the j-th shaft assembly,
- $i_{j,k}$ is the gear ratio of the gear pair between the j-th and the k-th shaft assemblies.

The *wheel dynamics* block represents the connection between the drivetrain's rotational- and the vehicle's longitudinal movement. The torque transferred

through the gearbox to the wheels is compensated by static friction on the ground; thus, the vehicle's center of gravity starts to accelerate in a longitudinal direction. Assuming that the contact points of the wheels are not sliding on the ground, the relation between the longitudinal and rotational measures can be expressed as shown in Equation (3) - (5):

$$a(t) = \ddot{\varphi}_{wheels}(t) \cdot r \quad (3)$$

$$m = \frac{J_{red}}{r^2} \quad (4)$$

$$F_p(t) = \frac{\tau_{red}(t)}{r} \quad (5)$$

where,

- a is the vehicle's longitudinal acceleration,
- φ_{wheels} is the wheels' angular position,
- r is the wheel radius,
- m is the vehicle's mass,
- J_{red} is the reduced moment of inertia of the vehicle's mass,
- F_p is the pulling force,
- τ_{red} is the torque that effectively accelerates the vehicle in longitudinal direction.

When the vehicle is moving straight, the wheels rotating together at the same speed as the differential shaft assembly, which angular acceleration can be calculated from the motor's angular acceleration (see in Equation (6)). Thus the torque acting on the wheels (τ_{wheels}) is responsible for the wheels' angular acceleration, the vehicle's longitudinal acceleration, and the compensation of resistance torques (see in Equation (7)).

$$\ddot{\varphi}_{wheels}(t) = \frac{\ddot{\varphi}_m(t)}{i_{1,2} \cdot i_{2,3}} \quad (6)$$

$$\tau_{wheels}(t) = (J_{wheels} + J_{red}) \cdot \ddot{\varphi}_{wheels}(t) + \tau_{res} \quad (7)$$

where,

- J_{wheels} is the moment of inertia of the wheels,
- τ_{res} is the resultant torque from resistance forces acting on the vehicle.

Integral of Equation (3) over time gives the vehicle's longitudinal speed, which serves as an input in the *vehicle module* to calculate resistances, and also in the *driver module* to determine the throttle and brake signals.

The separation of the blocks allows us to change the current model into more detailed versions, add more gear stages, and to read out the inner states of each component throughout the simulation.

In the *vehicle module*, the resistance forces acting on the vehicle are simulated (see in Figure 8).

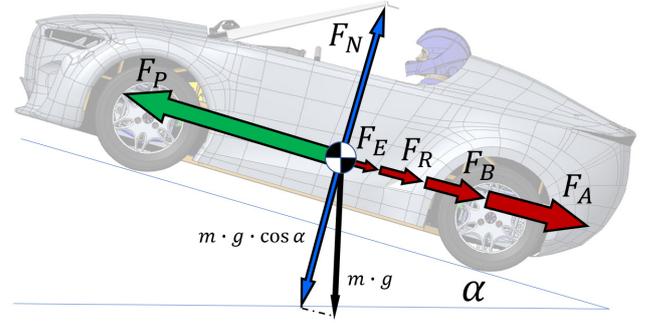


Fig. 8. Forces Acting on the Vehicle, at a Constant Velocity Equilibrium State

- F_p : pulling force from τ_{red} ,
- F_A : force from air resistance,
- F_R : force from rolling resistance,
- F_E : force from elevation on the slope,
- F_B : force from braking,
- F_N : normal force,
- g : gravitational acceleration,
- α : angle of slope.

For every equilibrium speed, the pulling force is equal to the summation of the resistance and elevation forces. If the pulling force is greater than the resistance forces, the vehicle is accelerating; else, it is decelerating.

Using this model a wide variety of valuable information can be extracted from the simulation results, such as the required motor torque to meet the prescribed speed profile (see in Figure 9) or the energy consumption during the examined time. This information can serve as input to other applications, which can check how good the final concept is. Modification of the parameters of mechanical parts is essential in the design process. Changes could be propagated through the simulation, and based on the results, optimal values could be calculated for the initial parameters. If we manipulate data in 3D models, a managing algorithm should keep track of the evaluated (or shared) parameters after the models are updated, to keep the Digital Twin consistent across CAD systems, simulation programs, or any other platforms.

Managing algorithm

Usually, data exchange between different applications from different fields (e.g. 3D modeling, simulation, optimization) is not provided. It requires some programming skills to extract data from one in a form that is useful to the other. Each of these software products rely heavily on graphic interfaces, however, if the models are appropriately set up, both of them can be managed from a third program. This managing program could enable the parameter optimization of the product; in our case, the drivetrain parameters can be optimized to minimize energy consumption.

If access to the model and the simulation results is provided, an optimizer algorithm could tune parameters in order to reach optimal properties of the modeled

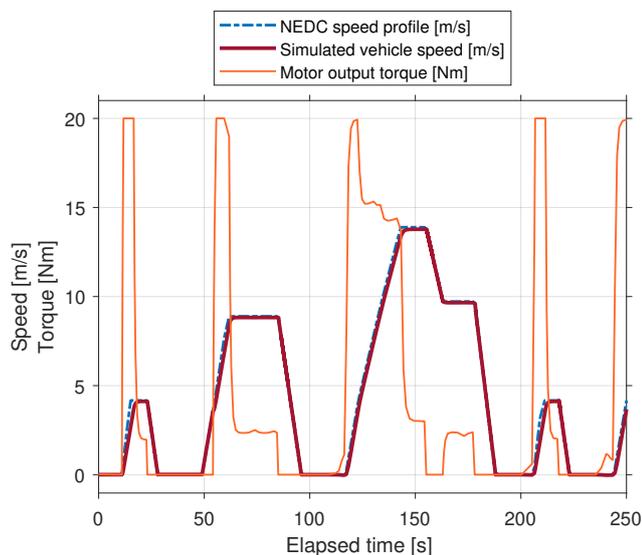


Fig. 9. Visualization of Simulation Results: Required Motor Output Torque to ensure the Prescribed Speed Profile for the Vehicle

system, however such an optimizer algorithm needs to be implemented in this project in the future. In the project, a simple Python script is used to ensure communication between Creo and Simulink.

CONCLUSIONS

The concept of the Digital Twin and Generative Design was introduced with its benefits and possibilities. Then small electric vehicles' advantages and disadvantages were reviewed. In the example project, the 3D CAD model, the simulation model, and the drive module's mathematical model were detailed.

FURTHER DEVELOPMENT

The detailed 3D model of the drive module and a simple simulation model of the powertrain is set up and ready to be utilized for measurements and tests. The model parameters and results could be accessed from a Python script, through which other different programs could be used to analyze the product from different aspects. As this managing algorithm could handle parameters consistently, it is capable of implementing an optimization algorithm to generate various optimal solutions for a defined problem. In the future, implementing a sensor network on the drive module to measure the actual torques and velocities, to identify the estimated loss parameters accurately, to test different scenarios and validate the simulation model is necessary. Based on these corrected parameters, the drivetrain module could be tested in dangerous situations virtually without damaging or breaking the physical twin. With a proper optimization algorithm (e.g. genetic algorithm), geometrical parameters could be optimized. Later, the model completed with the vehicle's suspension system could be used in a 3D physical engine for further optimization and testing.

ACKNOWLEDGEMENT

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI), and by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

REFERENCES

- Goossens, P. (2017). Industry 4.0 and the power of the digital twin. Maplesoft Engineering Solutions.
- Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. https://doi.org/10.1007/978-3-319-38756-7_4
- Khan, S., & Awan, M. (2018). A generative design technique for exploring shape variations. *Advanced Engineering Informatics*, 38, 712–724. <https://doi.org/10.1016/j.aei.2018.10.005>
- Park, G., Lee, S., Jin, S., & Kwak, S. (2014). Integrated modeling and analysis of dynamics for electric vehicle powertrains. *Expert Systems with Applications: An International Journal*, 41, 2595–2607. <https://doi.org/10.1016/j.eswa.2013.10.007>
- Rodič, B. (2017). Industry 4.0 and the new simulation modelling paradigm. *Organizacija*, 50, 193–207. <https://doi.org/10.1515/orga-2017-0017>
- Un-Noor, F., Sanjeevikumar, P., Mihet-Popa, L., Molah, M., & Hossain, E. (2017). A comprehensive study of key electric vehicle (ev) components, technologies, challenges, impacts, and future direction of development. *Energies*, 10. <https://doi.org/10.3390/en10081217>
- Winke, F., & Bargende, M. (2013). Dynamic Simulation of Urban Hybrid Electric Vehicles. *MTZ worldwide*, 74, 56–63. <https://doi.org/10.1007/s38313-013-0089-1>



TAMÁS DÓKA is a PhD student at the Department of Machine and Product Design of Budapest University of Technology and Economics (BME), Hungary. He had received masters degree in Mechatronic Engineering in 2019. He is actively taking part in an Electric Vehicle development university project called BME Fuse, focusing on concept creating, 3D modeling, simulation and optimization. Email address: doka.tamas@gt3.bme.hu



PÉTER HORÁK, PhD is an associate professor at the Department of Machine and Product Design of Budapest University of Technology and Economics (BME), Hungary. He had received PhD degree in Mechanical Engineering in 2004. His research fields are tribology, kinematics and geometry of gears, design theory and methodology. Email address: horak.peter@gt3.bme.hu

DEVIATION IN ENERGY CONSUMPTION ON AGGREGATE PRODUCTION PLANNING LEVEL IN INDUSTRIAL PRACTICE

Hajo Terbrack
Thorsten Claus
Technical University of Dresden
Faculty of Business and Economics
International Institute (IHI) Zittau
Markt 23, 02763 Zittau, Germany
E-Mail: hajo.terbrack@mailbox.tu-dresden.de

Frank Herrmann
Ostbayerische Technische Hochschule Regensburg
Faculty of Computer Science and Mathematics
Innovation and Competence Centre for Production
Logistics and Factory Planning (IPF)
Prüfening Str. 58, 93049 Regensburg, Germany

KEYWORDS

Energy Consumption, Aggregate Production Planning

ABSTRACT

In this article, we discuss energy consumption of producing firms on aggregate production planning. While almost constant energy consumption can be the case for a producing firm, highly fluctuating energy demand can occur as well. Together with volatile energy supply, e.g. due to renewable energy sources, this combination of fluctuating energy supply and demand can result in planning uncertainty and high energy costs. We propose different case studies in which such high deviation in the electricity consumption of a producing firm occurs due to aggregate production planning without appropriate consideration of energy consumption.

INTRODUCTION

Global energy consumption and respectively its costs are rising and therefore, the need to integrate energy consumption in industrial production planning is given. Without consideration of energy usage in production planning, highly fluctuating energy demand can arise. While energy suppliers already face problems in controlling energy supply, potential deviation in energy demand can strengthen planning uncertainty. As a result, high costs occur for energy suppliers to stabilize the power grid, leading to higher energy costs for the demand side, e.g. producing firms.

Within industrial production planning, a common concept is the hierarchical production planning as proposed by Hax and Meal in 1975 and many others (e.g. Claus et al. 2015). In this concept, the aim is to harmonise decisions that are necessary in the long term and that span multiple production sites, right up to short-term, to-the-second decisions at the plant level, across three planning levels: Production Program Planning (consisting of Aggregate Production Planning and Master Production Scheduling), Lot Sizing and Scheduling. Thereby, aggregate production planning (APP), as the upper level within production program planning, aims to smooth employment over a planning horizon of usually one to two years. Seasonally varying demand or possible economic fluctuation are taken into account in this mid-term planning. Based on product types and a typical period length of one month, capacity planning for one or

more production sites is fulfilled. Besides an optimization of inventory level costs and costs for the usage of additional capacity, further optimization goals within APP can include transports between production sites, external procurement or multi-level supply processes.

To improve the planning situation for energy suppliers, mid-term production planning should consider energy consumption. Therefore, this article discusses energy consumption on aggregate production planning.

The article is structured as follows. Chapter 2 gives an overview of relevant literature on industrial energy usage and demand side management approaches on production program planning. Chapter 3 shortly introduces the APP optimization model used for the case studies, which are presented and discussed in Chapter 4. In Chapter 5, the paper concludes with a brief summary and an outlook for further research.

PROBLEM DEFINITION AND LITERATURE

With the ongoing integration of renewable energy sources, which are characterised by a very volatile supply, in the power grid (see Kabelitz et al. 2014, Simon 2017), corresponding fluctuations in the generation and feed-in of electricity from renewable energy sources are increasingly putting a strain on the electricity grids of electricity suppliers, who have to balance out irregular load distributions accordingly (see Paulus & Borggreffe 2011). In addition to this volatile electricity supply, strong fluctuations in the amount of energy purchased by the demand side can occur, as it can be the case with producing firms (see U.S. Department of Energy 2006). Together with increasing and strongly varying energy prices (see Rösch et al. 2019, Simon 2017), both, the energy supplier and the producing firm have to face planning uncertainty resulting in grid overloads and high costs for energy as part of total production costs. To improve power grid stability and to reduce the resulting costs, the energy demand side is more and more willing to involve energy costs and consumption into its planning (see Paterakis et al. 2017, Paulus & Borggreffe 2011) – as it is also the case for producing firms by corresponding production planning and control. Various so-called "Demand Side Management" (DSM) or "Demand Response" (DR) approaches pursue the goal of leading the demand side in the electricity market to change its

consumption in order to, among other things, take into account the strong fluctuations in electricity supply or to shift the use of electricity to periods with lower demand (see Paterakis et al. 2017, Paulus & Borggreffe 2011).

Within hierarchical production planning, different DSM approaches can be assigned to the respective planning levels, a large part of which are assigned to the scheduling level (see Biel & Glock 2016). However, since production quantities and available production capacities are already defined in lot sizing and scheduling where the planning horizon usually corresponds to one week, there is limited room for action at these lower levels of hierarchical production planning (see Claus et al. 2015, Günther & Tempelmeier 2016). In order to also achieve improved energy-oriented planning in the medium term, appropriate measures are necessary within production program planning.

Within aggregate production planning, there are only a few articles that focus on sustainability and in specific, on energy consumption. Cheraghalikhani et al. (2019), who present a literature review on aggregate production planning models, point out the open research issue of adding sustainability as well as green concepts to aggregate production planning. A DR approach within aggregate production planning is pursued by Latifoğlu et al. (2013) in their article. The authors present an approach that considers the effects of "Interruptible Load Contracts" (ILCs) at the level of aggregate production planning. A robust production plan is set up to meet all customer needs despite supply-side interruptions in the electricity supply, while minimizing electricity costs by taking the economic incentives of ILCs into account. Modarres and Izadpanahi (2016) present a robust optimization approach to minimize operational costs, energy costs and carbon emission in aggregate production planning. Under consideration of uncertain costs, energy and carbon parameters and uncertainty in demand and maximum capacity, the robust optimization approach is applied to a smelting manufacturer and the relationship between budgets of uncertainty and optimal values is analysed. In the article of Chaturvedi (2017), different production facilities are assumed with different facility-specific energy consumption rates per produced unit. By determining the optimal production capacity of each production facility through linear programming, the overall annual energy consumption is minimized while the demand is satisfied. An approach to maximize profit in aggregate production planning by consideration of labor costs, inventory costs, production costs, shortage costs and electricity costs is presented by Nour et al. (2017). The effect of electricity price changes on a porcelain manufacturer is analysed and total costs can be reduced by 23% compared to the former production planning.

However, none of the approaches presented can sufficiently reduce the deviations in the energy consumption of a manufacturing company in the medium-term production programme planning. Furthermore, due to the lack of a possibility to predict the existing energy supply with satisfactory accuracy in the

medium term (see Kabelitz et al. 2014), the demand for electricity cannot be adjusted to supply forecasts in the medium term.

ENERGY-ORIENTED AGGREGATE PRODUCTION PLANNING

To integrate energy consumption into a common production program planning model, the aggregate production planning model AGGRPLAN, a linear optimization model, is expanded (for a detailed description see Claus et al. 2015). The basic model aims at smoothing employment over a planning horizon of several months by optimizing production quantities for product types in order to minimize costs for inventories and costs for additional capacity usage. We introduce product-type-specific energy consumption coefficients. These coefficients represent the electricity consumed in production of one unit of the corresponding product type. Energy consumption is considered in terms of production quantities multiplied with energy consumption per produced unit.

CASE STUDIES

To point out the relevance of high deviation in energy consumption of a producing firm, three case studies are presented. In these case studies, aggregate production planning is fulfilled resulting in fluctuating energy consumption. Additionally, a fourth scenario is described in which almost constant energy consumption occurs in aggregate production planning.

Case study (1) deals with the production of agricultural machines. A company in Germany produces two product types, ploughs, and hay rakes. Ploughs are necessary on the field to loosen and turn the soil before seeds can be sown. Hay rakes are frequently used to gather harvested material such as hay or straw for later collection. These agricultural product types are usually ordered by agricultural cooperatives and by individual farmers. In the manufacturing company, the production of these two product types is organised as a job-shop with six production segments and a total of 13 production machines. The layout of the shop floor is illustrated in Figure 1.

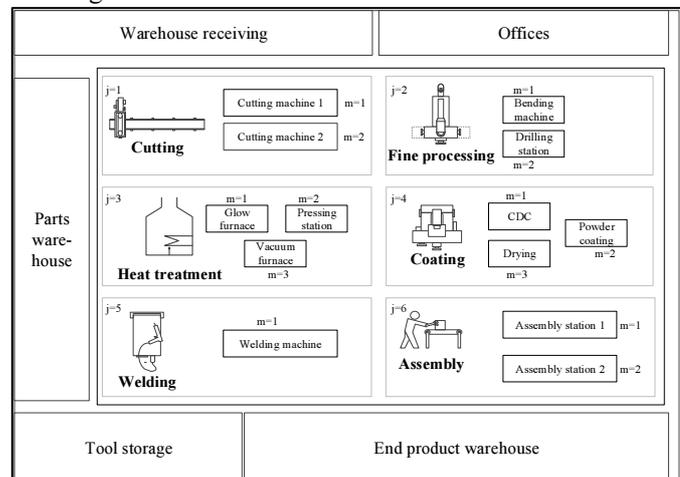


Figure 1: Shopfloor layout – case study (1).

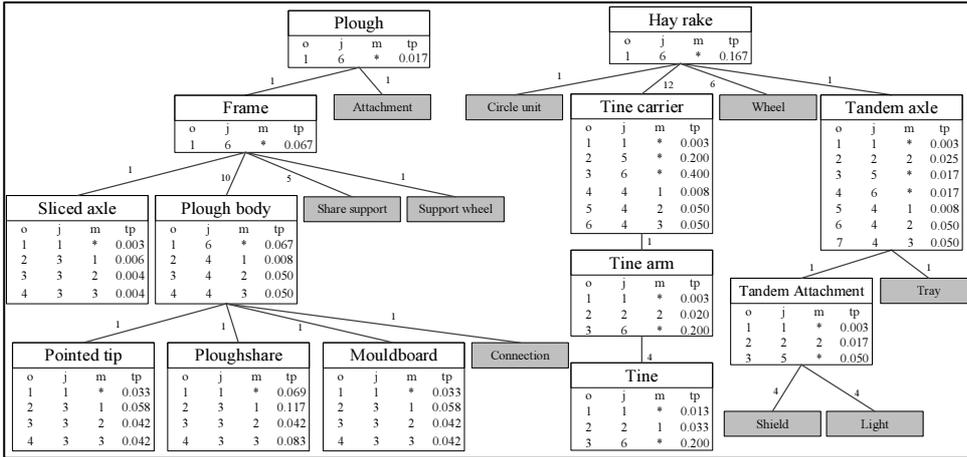


Figure 2: Gozintograph for product type 1 (plough) and product type 2 (hay rake) - case study (1).

A two-shift model is used whereby ten workers are available each shift. Gozintograph and corresponding processing times (tp) are stated in Figure 2. Column o represents the sequence of processing steps, j and m the production segments and machines as named in Figure 1. tr equals processing time on the machines in time units [TU]. Processing steps that can be carried out by all machines within a production segment are labelled with *. Purchase parts are shaded grey.

Throughput time of each product type was used as technical production coefficients in aggregate production planning. Therefore, by simulating material requirement planning and scheduling for 720 periods (representing 24 APP periods with an aggregation factor of 30) based on same demand data later used in aggregate production planning, throughput time was determined for each product type. Mean value and deviation is shown in Table 1. To determine personnel production coefficients, net processing times of each product type were analysed and referred to personnel and technical capacity usage. Based on the proportion of personnel and technical capacity usage, personnel production coefficients were calculated by multiplying the relation of personnel and technical net processing time with the mean value of throughput time. Summed up personnel and technical net processing times as well as energy consumption for producing one unit of each product type are shown in Table 2.

Case study (1) – Throughput time			
[in time units]	Mean value	Deviation	
Plough	182.22	31.41	
Hay rake	202.10	39.71	

Table 1: Throughput time [in TU] – case study (1).

Case study (1) – Net processing times [in TU] and energy consumption [in EU]			
	Personnel	Technical	Energy Consumption
Plough	0.55	0.93	73.01
Hay rake	1.60	1.66	18.24

Table 2: Net processing times [in TU] and energy consumption [in EU] – case study (1).

Energy consumption in energy units [EU] per quantity unit of each product type was determined by calculating the electricity consumption of the corresponding production processes. Note that the production of

ploughs is by far more energy-intensive than the production of hay rakes since the main components (ploughshare, sliced axle, pointed tip, and mouldboard) get heated, pressed, and hardened due to their need of being very robust.

In Table 3, the relevant product type parameters for case study (1) are stated. f_k^P represents personnel capacity usage in time units, f_k^T technical capacity usage in time units and f_k^E the consumed energy in producing one unit of the product type, in energy units. Inventory holding costs per quantity unit and APP period, h_k , are 140 money units (MU) for product type 1, ploughs, and 80 MU for product type 2, hay rakes. The initial inventory level is zero for both product types k ($I_{k0} = 0$).

Case study (1) – Product type parameters				
	f_k^P	f_k^T	f_k^E	h_k
	[in TU]	[in TU]	[in EU]	[in MU]
Plough	106.67	182.22	73.01	140
Hay rake	195.52	202.10	18.24	80

Table 3: Product type parameters – case study (1).

Available capacity per APP period (personnel capacity b_t^P , maximum additional personnel capacity U_t^{max} , technical capacity b_t^T) and cost rate for additional capacity usage (u_t) is shown in Table 4.

Case study (1) – Capacity parameters				
	b_t^P	U_t^{max}	b_t^T	u_t
	[in TU]	[in TU]	[in TU]	[in MU]
per APP period	3200	800	6240	36

Table 4: Capacity parameters – case study (1).

Aggregated demand quantities for each APP period are given in table 5. While different demand scenarios can occur, the assumed quantities represent a typical demand trend of the analysed company.

Fulfilling the aggregate production planning model AGGRPLAN leads to an optimization of inventory holding costs and costs for additional capacity usage. Figure 3 shows inventory levels per period, personnel capacity usage as well as the energy consumption for a planning horizon of 24 APP periods. Capacity usage is almost constant over the planning horizon and inventory level is fluctuating depending on the amount of pre-production. A deviation in energy consumption per APP

APP Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Plough	8	9	6	5	13	9	15	4	9	7	17	12	13	10	8	7	10	3	12	15	14	13	8	8
Hay rake	14	10	15	15	5	9	6	12	10	13	9	7	10	18	12	15	11	14	10	8	7	6	14	15

Table 5: APP demand quantities – case study (1), scenario (1).

period of 26% was measured, the difference between maximum and minimum energy consumption equals 823 EU and the average energy consumption is 914 EU.

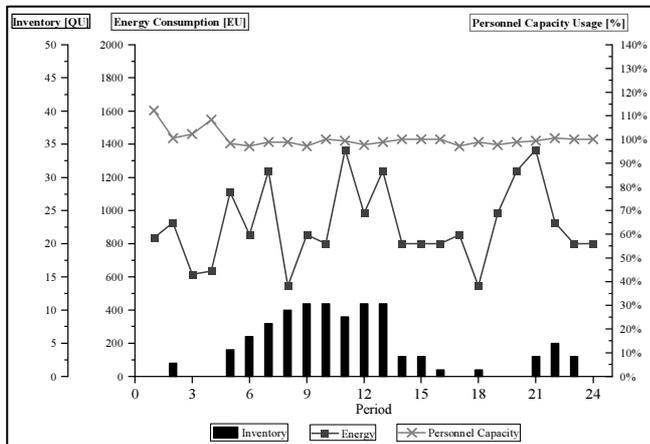


Figure 3: APP resulting in high deviation in energy consumption – case study (1), scenario (1).

In a second scenario, assuming a different demand structure, aggregate production planning was carried out for twelve APP periods. The new demand situation is given in Table 6. In this second scenario, customer demands and its mix are quite stable along the planning horizon. As a result, the corresponding energy consumption per period is less fluctuating (see Figure 4) with a deviation equal to 8%.

APP Period	1	2	3	4	5	6	7	8	9	10	11	12
Plough	8	9	8	7	8	9	8	7	7	7	8	7
Hay rake	14	12	13	14	14	9	13	12	10	16	9	12

Table 6: APP demand quantities – case study (1), scenario (2).

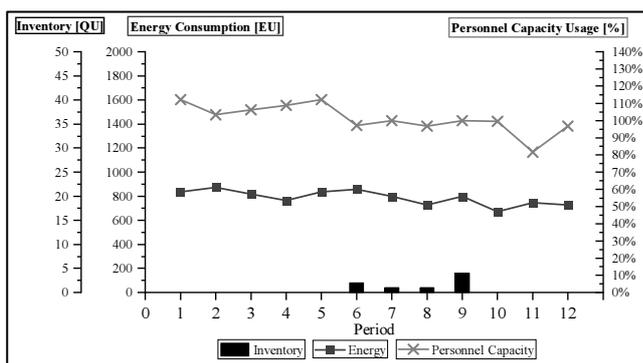


Figure 4: APP resulting in almost constant energy consumption – case study (1), scenario (2).

This case study (1) shows, that the combination of two product types with considerable differences in energy consumption and a varying demand structure can lead to high deviations in energy demand. As already mentioned, such huge differences in energy consumption can strengthen the planning uncertainty for the electricity provider and lead to high energy costs for a producing firm.

Case study (2) discusses the manufacturing of three components for plant construction in a company in Poland. 22 workers produce pipings, conveyor lines and carousels in a two-shift model, whereby eleven machines are ordered as a job shop with five production segments. Shop floor layout is outlined in Figure 5. Corresponding gozintographs and work schedules are stated in Figure 6 and 7.

The finished components are later installed in larger machines, such as bottling machines. Usually, multiple units of pipings and conveyor lines get installed in such machines and represent the main part of customer demand for the company. While production of pipings and conveyor lines do not differ much in energy usage due to almost similar process steps like cutting, sawing, drilling and bending, milling and welding of the inner and outer ring of the carousel lead to a high energy usage in production of carousels.

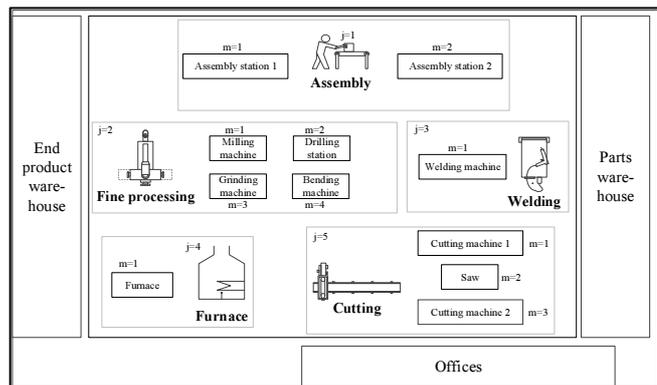


Figure 5: Shop floor layout – case study (2).

Production coefficients were determined by simulation equivalent to case study (1). MRP and scheduling was simulated for 360 periods (equal to 12 APP periods with an aggregation factor of 30) on the basis of demand data later used for APP. Mean value and deviation of throughput times are shown in Table 7, net processing time and energy consumption for each product type is listed in Table 8.

Case study (2) – Throughput time		
[in time units]	Mean value	Deviation
Piping	180.78	32.64
Conveyor line	178.97	35.71
Carousel	183.60	28.25

Table 7: Throughput time [in TU] – case study (2).

Case study (2) – Net processing times [in TU] and energy consumption [in EU]			
	Personnel	Technical	Energy Consumption
Piping	2.63	2.63	13.70
Conveyor line	2.04	2.04	4.90
Carousel	14.54	20.16	443.82

Table 8: Net processing times [in TU] and energy consumption [in EU] – case study (2).

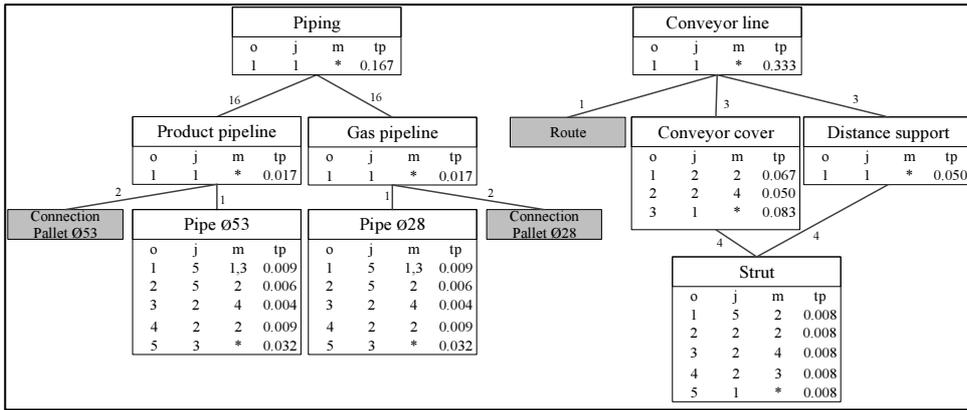


Figure 6: Gozintograph for product type 1 (piping) and product type 2 (conveyor line) – case study (2).

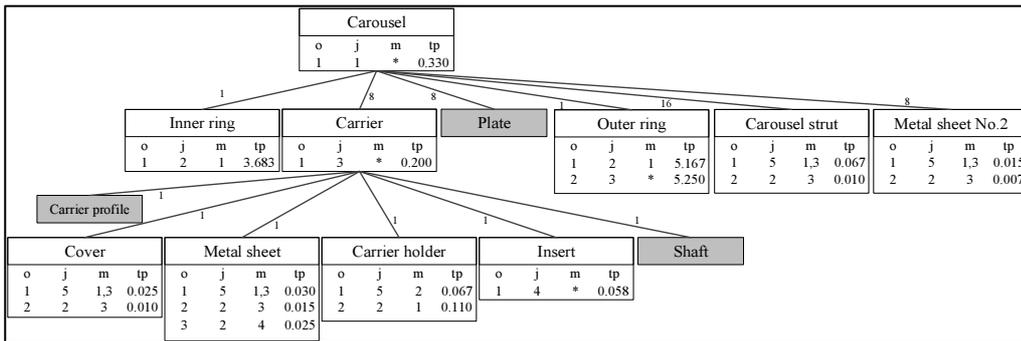


Figure 7: Gozintograph for product type 3 (carousel) – case study (2).

	Case study (2) – Product type parameters			
	f_k^P [in TU]	f_k^T [in TU]	f_k^E [in EU]	h_k [in MU]
Piping	180.78	180.78	13.70	50
Conveyor line	178.97	178.97	4.90	20
Carousel	132.42	186.30	443.82	130

Table 9: Product type parameters – case study (2).

	Case study (2) – Capacity parameters			
	b_t^P [in TU]	U_t^{max} [in TU]	b_t^T [in TU]	u_t [in MU]
per APP period	3520	880	5280	20

Table 10: Capacity parameters – case study (2).

Product type parameters are stated in Table 9. Table 10 shows the capacity parameters for case study (2). The initial inventory level is zero for all product types k ($I_{k0} = 0$). Customer demand for 360 periods was aggregated for APP. Demand quantities for each APP period is given in Table 11.

APP Period	1	2	3	4	5	6	7	8	9	10	11	12
Piping	13	9	6	6	4	11	12	4	9	7	5	16
Conveyor line	6	8	15	13	5	7	10	12	14	7	12	11
Carousel	3	1	2	3	3	1	1	2	3	1	3	1

Table 11: APP demand quantities – case study (2).

Aggregate production planning was carried out for twelve APP periods. Figure 8 shows inventory levels per period, personnel capacity usage as well as the energy consumption for the optimal solution of AGGRPLAN. Average energy consumption equals 1057 EU per APP period. Due to the peak in period 9 (2381 EU), difference between maximum and minimum energy usage per period is 1057 EU while deviation in energy usage per period equals 54%.

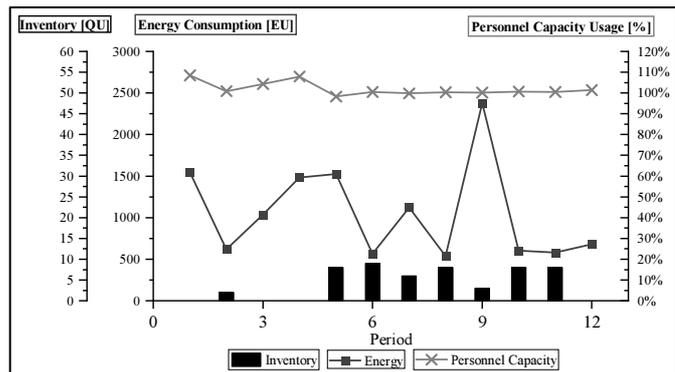


Figure 8: APP resulting in high deviation in energy consumption – case study (2).

In Period 9, five units of product type 3, carousels, are produced, leading to very high energy consumption in this period. Also, due to the changing demand structure along the planning horizon, in other periods energy consumption is varying while personnel capacity usage remains almost stable.

Case study (3) deals with the production of steel bridges and windtowers in a German steel working company in Northern Bavaria. Due to the long service life and individual layout of steel-fabricated bridges and windtowers, this case represents a make-to-order production approach.

27 employees work in production, producing steel parts for bridges and wind towers in a three-shift model. In total, seven machines are installed for polishing, cutting, edging, bending, welding, milling and coating the steel products. The shopfloor layout is shown in Figure 9.

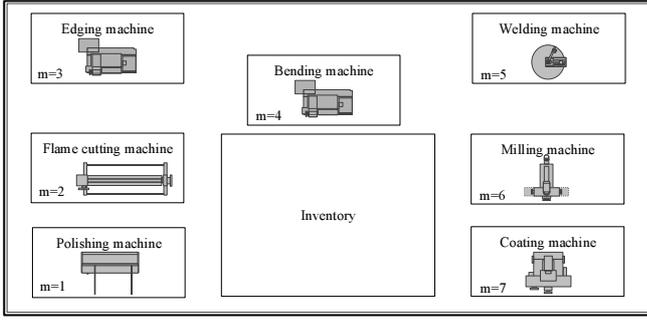


Figure 9: Shopfloor layout – case study (3).

Processing times are very long since both products are large in size and must meet specific safety regulations. The corresponding work schedules are given in Figure 10.

Bridge			Windtower		
o	m	tp	o	m	tp
1	1	48	1	1	48
2	2	120	2	2	96
3	3	48	3	4	120
4	5	144	4	5	120
5	6	120	5	6	96
6	7	96	6	7	72

Figure 10: Work schedules for product types – case study (3).

Summed up net processing times as well as energy consumption for both product types are listed in Table 12.

Case study (3) – Net processing times [in TU] and energy consumption [in EU]			
	Personnel	Technical	Energy Consumption
Bridge	864.00	576.00	15720
Windtower	828.00	552.00	9696

Table 12: Net processing times [in TU] and energy consumption [in EU] – case study (3).

While net processing times are similar for both product types, the production of the bridge part is more energy-intensive in the processes of polishing, welding and milling. The reason for the higher energy consumption lays in the higher steel thickness of the bridge sections compared to the wind tower sections, resulting in a higher total energy consumption in the production of bridge parts.

Due to the long processing times of both product types, for aggregate production planning, the aggregation factor was increased to 90 simulation periods equal one APP period. Simulating 540 periods (equal to six APP periods in this case study) produced the following results for throughput time (Table 13).

Case study (3) – Throughput time		
[in time units]	Mean value	Deviation
Bridge	665.78	59.01
Windtower	607.11	48.37

Table 13: Throughput time [in TU] – case study (3).

For aggregate production planning, the following parameters in Tables 14 – 16 were used. The initial inventory level is zero for all product types k ($I_{k0} = 0$). Based on the varying demand quantities and the differences in energy consumption of bridges and

Case study (3) – Product type parameters				
	f_k^P [in TU]	f_k^T [in TU]	f_k^E [in EU]	h_k [in MU]
Bridge	998.67	665.78	15720	4860
Windtower	910.67	607.11	9696	7830

Table 14: Product type parameters – case study (3).

Case study (3) – Capacity parameters				
	b_t^P [in TU]	U_t^{max} [in TU]	b_t^T [in TU]	u_t [in MU]
per APP period	11520	2880	10080	40

Table 15: Capacity parameters – case study (3).

APP Period	1	2	3	4	5	6
Bridge	2	0	12	1	9	3
Windtower	13	13	2	12	3	11

Table 16: APP demand quantities – case study (3).

windtowers, the optimal solution of AGGRPLAN leads to a fluctuating energy consumption per period, as shown in Figure 11. With a minimum energy consumption equal to 132072 EU and a maximum energy consumption equal to 192312 EU, a deviation of 14% occurs.

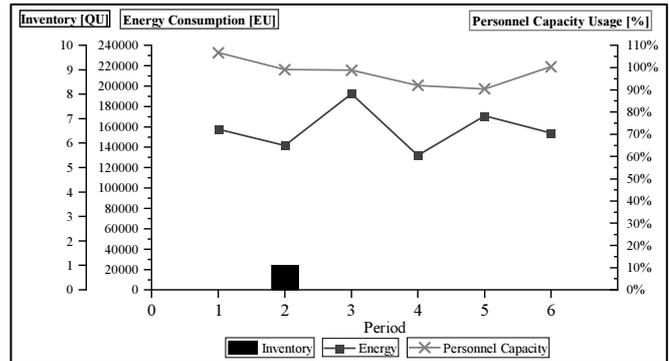


Figure 11: APP resulting in medium high deviation in energy consumption – case study (3).

CONCLUSION AND OUTLOOK

While energy-oriented production planning is widely known in short-term planning, existing literature lacks on mid-term planning that focus on energy consumption, especially on possible deviation in energy usage within a planning horizon.

Different product types having remarkable differences in energy consumption can be the case in production, as shown for manufacturing of two agricultural machines, three components for plant construction as well as for two steel parts, leading to high deviation in energy consumption per period as long as demand quantities and demand mix are varying. Those differences in energy consumption can occur in more companies and industries, as long as there is a product range that differs in the usage of energy-intensive processes and demand quantities are not constant. Thereby, such volatile energy consumption of a producing firm can result in planning uncertainty for energy suppliers and therefore in high costs for the energy supplier and its customers. By reducing such high deviation in energy consumption, both sides could benefit. The energy supplier gains

planning certainty and therefore could offer a favourable price to the energy demand side, leading to lower energy costs. However, when a producing firm reduces this deviation in energy consumption, one moves away from the former optimal production program in terms of costs, e.g. inventory costs and costs for additional capacity usage. Additionally, there can be situations, in which companies lack on the flexibility to change production quantities and corresponding energy consumption, as it might be the case for the make-to-order example in case study (3).

Therefore, future research needs to integrate energy consumption, respectively deviation in consumption, and its costs to find an optimal solution for both objectives, costs and deviation in energy consumption.

REFERENCES

- Biel, K.; Glock, C.H. (2016): Systematic literature review of decision support models for energy-efficient production planning, in: *Computers & Industrial Engineering*, Vol. 101, pp. 243–259.
- Chaturvedi, N.D. (2017): Minimizing energy consumption via multiple installations aggregate production planning, in: *Clean Technologies and Environmental Policy*, Vol. 19, No. 7, pp. 1977–1984.
- Cheraghalikhani, A.; Khoshalhan, F.; Mokhtari, H. (2019): Aggregate production planning: A literature review and future research directions, in: *International Journal of Industrial Engineering Computations*, Vol. 10, No. 2, pp. 309–330.
- Claus, T.; Herrmann, F.; Manitz, M. (2015): *Produktionsplanung und -steuerung: Forschungsansätze, Methoden und deren Anwendungen*, Springer, Berlin, Heidelberg.
- Günther, H.O.; Tempelmeier, H. (2016): *Produktion und Logistik: Supply Chain und Operations Management*, BoD–Books on Demand.
- Hax, A. C.; Meal, H.C. (1975): Hierarchical integration of production planning and scheduling, in M. A. Geisler (Ed.): *Logistics*. North Holland, Amsterdam, S. 53–69.
- Kabelitz, S.; Streckfuß, U.; Gujjula, R. (2014): Einsatz von mathematischen Optimierungsverfahren zur energieorientierten Produktionsplanung, in: *TBI2014*.
- Latifoğlu, Ç.; Belotti, P.; Snyder, L.V. (2013): Models for production planning under power interruptions, in: *Naval Research Logistics (NRL)*, Vol. 60, No. 5, pp. 413–431.
- Modarres, M.; Izadpanahi, E. (2016): Aggregate production planning by focusing on energy saving: A robust optimization approach, in: *Journal of Cleaner Production*, Vol. 133, pp. 1074–1085.
- Nour, A.; Galal, N.M.; El-Kilany, K.S. (2017): Energy-Based Aggregate Production Planning For Porcelain Tableware Manufacturer in Egypt, in: *Proceedings of the International Conference on Industrial Engineering and Operations Management*, Rabat, Morocco.
- Paterakis, N.G.; Erdinç, O.; Catalão, J.P. (2017): An overview of Demand Response: Key-elements and international experience, in: *Renewable and Sustainable Energy Reviews*, Vol. 69, pp. 871–891.
- Paulus, M.; Borggrefe, F. (2011): The potential of demand-side management in energy-intensive industries for electricity markets in Germany, in: *Applied Energy*, Vol. 88, No. 2, pp. 432–441.
- Rösch, M.; Lukas, M.; Schultz, C.; Braunreuther, S.; Reinhart, G. (2019): An approach towards a cost-based production control for energy flexibility, in: *Procedia CIRP*, Vol. 79, pp. 227–232.
- Simon, R. (2017): *Nachfrageseitige Flexibilitätsoptionen: Demand-Side-Management, Energiespeicher und Regelenergie*, in: Matzen F., Tesch R. (Hrsg.): *Industrielle Energiestrategie*, Springer Gabler, Wiesbaden.
- US Department of Energy (2006): Benefits of demand response in electricity markets and recommendations for achieving them – a report to the United States congress pursuant to section 1252 of the Energy Policy Act of 2005, February 2006.

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 Dresden Technical University. 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 Dresden Technical University, 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.

Modeling and Simulation for Performance Evaluation of Computer-based Systems

Modelling and Simulation of Data Intensive Systems

-

Special Session

Computing resilience of interconnected systems by piecewise linear Lyapunov functions

Alberto Tacchella and Armando Tacchella

KEYWORDS

Simulation of Control Systems, Piecewise-Linear Switched Systems, Cyber-Security and Critical Infrastructure Protection

ABSTRACT

Resilience, i.e., the ability to withstand and recover from disruption, is a fundamental requirement for mission-critical automation systems. Interconnection among systems makes the analytical determination of resilience zones harder than in isolated systems, because the failure of a system usually has an impact on those connected to it. In this paper we propose an algorithm to determine resilience zones of interconnected systems based on the computation of piecewise linear Lyapunov functions. The algorithm is based on a model introduced previously that abstracts from specific system dynamics and enables analysis in the space of performances. Experiments with an interconnected system inspired to a real wastewater treatment facility show the feasibility and the accountability of our approach.

I. INTRODUCTION

The number of security incidents affecting industrial automation systems has been steadily increasing over the past few years — see, e.g., [Lou15]. The main problem is that more and more such systems do not work in isolation, but are routinely connected between them, often relying on wide-area networks including the Internet. Through such connections, cyber-attacks can be brought to the systems and cause disruption in services, damage to equipment or severe impairment of human activities. Detecting weaknesses, fixing them and monitoring critical events in industrial automation systems are compelling and heavily investigated matters, but we must also acknowledge that, in spite of all the efforts made to secure them, connected systems may never be fully secure. In this scenario, the concept of *resilience* emerges as an additional target, complementary to prevention and protection from attacks, but not less important. This line of thought is pervasive in the Presidential Policy Directive 21 [Oba13] about the security of critical infrastructure, which defines resilience as “[...] *the ability to [...] withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or in-*

idents”. More recently, the term *cyber-resilience* has been coined to identify specifically “*the ability to continuously deliver the intended outcome despite adverse cyber events*” [BHSZ15] and this is the interpretation whereto we adhere in the following.

As mentioned in [AF13], the analysis of interconnected automation systems involves a number of challenging issues. One such issue is that they are built as hierarchies of subsystems which makes their design, implementation and maintenance feasible from an engineering point of view. However, the analysis based on decomposition in systems components is either hardly feasible or of poor utility. If we consider resilience, one problem is the difficulty of inferring the resilience of the overall system considering the resilience of its components. The well-known phenomenon of *cascading failures* — see, e.g., [CLM04] — is an example of how individually healthy subsystems might become unstable and fail because other connected subsystems cease to function properly, leading to failures that potentially affect the whole system. In case of resilience it is important to understand how variations in the behavior of one system can affect the others, and in particular whether outside the nominal operating range there exist *resilience regions*, i.e., zones in the system state space wherein a failure of some component can be “absorbed” by the system which remains functional, albeit possibly at a degraded level.

The problem of devising models to compute resilience in interconnected systems has been studied previously. In [AF13], the authors introduce a model where systems are treated as abstract entities, decomposed into a set of elements interconnected by functional dependencies. The failures considered are of two types: internal operation drifts and external cascade effects. The modeling approach is quite general, yet simple enough to be applied on a variety of systems. While we use the contribution of [AF13] as a basis for ours, the topology analyzed in that paper applies only to small systems consisting of two components connected through a “feedback” loop. A slight generalization of the results in [AF13] is presented in [LPZ14] where the authors consider also “loop structures” for the analytical determination of resilience. Finally, in [LFZ17], the application of the framework introduced in [AF13] is considered in the context of critical infrastructure.

In this paper we contribute an extension of the methodology proposed in [AF13] which in principle allows to infer automatically Lyapunov functions to study the stability, and thus the resilience of interconnected systems. To the extent of our knowledge, this is the first time that a computational approach is inves-

Alberto Tacchella is with Fondazione Bruno Kessler, Trento, Italy — E-mail: atacchella@fbk.eu. Armando Tacchella is with DIBRIS, University of Genoa, Genoa, Italy — E-mail: armando.tacchella@unige.it

tigated to attack this problem, providing also some experimental evidence of its feasibility and accountability on a case study extracted from a real critical infrastructure. More in detail, our contribution can be fleshed out as follows:

- An iterative method to compute piecewise linear Lyapunov functions characterizing the resilience of specific topologies of interconnected systems;
- A case study about the model of an urban wastewater treatment facility;
- Experimental results on the case study showing promise about the method.

The rest of the paper is structured as follows. In Section II we introduce basic terminology, notation and definitions related to Lyapunov stability, piecewise linear switching systems, and the resilience model for interconnected systems that we consider as a basis for our investigation. In Section III we extend the basic resilience model to systems having a specific topology and, for systems having such topology, in Section IV we present an approach to compute piecewise linear Lyapunov functions to tackle their stability. Finally, in Section V, we introduce our case study related to wastewater treatment, show its simulation model, how we extracted relevant parameters from it, and the results of applying our approach to the system under consideration. We conclude in Section VI with some final remarks and an agenda for future developments.

II. BACKGROUND

We recall briefly the method of Lyapunov functions for proving the stability of an equilibrium point of a nonlinear system; see e.g. Khalil [Kha92] for a textbook introduction. We also refer the reader to [GH15] for a comprehensive review on Lyapunov functions and related computational methods.

Let us consider an autonomous, first-order system of ordinary differential equations (ODEs)

$$\dot{\mathbf{x}} = f(\mathbf{x}),$$

where f is a C^1 function $\mathbb{R}^n \rightarrow \mathbb{R}^n$. We assume that the corresponding dynamical system has (at least) an equilibrium point, which can be taken to be at $0 \in \mathbb{R}^n$ without loss of generality. We denote by $\Phi_t(\mathbf{x}_0)$ the corresponding flow at time t with initial value \mathbf{x}_0 .

The equilibrium point at 0 is called:

- *stable* if for all $\varepsilon > 0$ there exists $\delta > 0$ such that for all \mathbf{x}_0 with $\|\mathbf{x}_0\| < \delta$ we have that $\|\Phi_t(\mathbf{x}_0)\| < \varepsilon$ for every $t \geq 0$;
- *asymptotically stable* if it is stable and there exists a $\delta' > 0$ such that for every \mathbf{x}_0 with $\|\mathbf{x}_0\| < \delta'$ we have that $\lim_{t \rightarrow \infty} \|\Phi_t(\mathbf{x}_0)\| = 0$;
- *exponentially stable* if it is asymptotically stable and the rate of convergence to zero of $\|\Phi_t(\mathbf{x}_0)\|$ is exponential.

Classically, a *Lyapunov function* for the system (at the given equilibrium point) is a C^1 function $V: U \rightarrow \mathbb{R}$ defined on a neighborhood U of 0 and such that:

- V is positive away from 0 and $V(0) = 0$ (in other words, V is a positive function with a global minimum at 0), and

- V is strictly decreasing along the solution curves of the ODE (outside of the equilibrium point). If we denote by $\dot{V}(\mathbf{x})$ the derivative of V along the orbit passing through \mathbf{x} , then a sufficient condition for this to happen is that $\dot{V}(\mathbf{x}) < 0$ for every $\mathbf{x} \in U \setminus \{0\}$.

Lyapunov's theorem asserts that the existence of a Lyapunov function is sufficient to guarantee the asymptotic stability of the equilibrium. Moreover, any sublevel set of V which is contained in U is also a subset of the basin of attraction of the equilibrium point.

For the purposes of the present paper it is important to note that the above formulation of Lyapunov's theorem can be considerably weakened, for instance by removing the hypothesis that the Lyapunov function be C^1 or even continuous. The only important requirement is that V must decrease in a suitable sense along the trajectories of the system. In the sequel we shall always speak of Lyapunov functions in this more general sense.

A *switched system* is a hybrid system in which a continuous-time evolution law is coupled to a set of discrete "switching" events. Typically, the continuous state space is partitioned into a finite number of *operating regions* by means of a family of *switching surfaces*. In each operating region, the evolution of the system is described by a given system of ODEs. Whenever the trajectory of the system hits a switching surface, the continuous state jumps instantaneously to a new value, specified by a *reset map*. When the reset map is trivial, that is the switching events only involve a change in the continuous evolution law, one speaks of an *autonomous* switched system, or *switching system*. For a general introduction to switched (and switching) systems we refer the reader to the books by Liberzon [Lib03], or Sun and Ge [SG11].

We are interested in a particular subclass of switching systems, namely the ones in which the state space is a finite-dimensional linear space (say \mathbb{R}^n for some $n \in \mathbb{N}$) and in each operating region the continuous evolution law is either linear or affine. Following Johansson [Joh03], we shall call such systems *piecewise linear (PWL) switching systems*¹.

A piecewise linear switching system is then defined by:

- a partition $\{X_i\}_{i \in I}$ of the state space \mathbb{R}^n in a set of operating regions, and
- for each $i \in I$ an affine evolution law of the form

$$\dot{\mathbf{x}} = A_i \mathbf{x} + b_i$$

where A_i is a $n \times n$ matrix and b_i is a column vector. Let us recall from [Joh03] how to reinterpret this kind of dynamics in matrix terms. Given $\mathbf{x} \in \mathbb{R}^n$, we define an extended state vector \bar{x} as

$$\bar{x} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \quad (1)$$

¹In the literature one often finds also the equivalent term *piecewise affine*.

Similarly, given an $n \times n$ matrix A and a vector $\mathbf{b} \in \mathbb{R}^n$ we define \bar{A} to be the $(n+1) \times (n+1)$ block matrix

$$\bar{A} = \begin{pmatrix} A & \mathbf{b} \\ 0_{1 \times n} & 0 \end{pmatrix} \quad (2)$$

In this way the affine evolution equation $\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}$ can be formulated more compactly as $\dot{\hat{x}} = \bar{A}\hat{x}$.

In this paper we are going to study the stability of a certain class of PWL switched systems using *piecewise-linear Lyapunov functions*, as described for instance in [Joh03], Section 4.10. The advantage of piecewise-linear functions over piecewise-quadratic ones is that the former are typically the outcome of a *linear programming* (LP) problem, whereas piecewise-quadratic Lyapunov functions are usually computed by solving a system of *linear matrix inequalities* (LMI), which is computationally more difficult (although faster methods are sometimes available, see e.g. [AGS⁺16]).

A difference with respect to the approach in [Joh03] is that we do not require our Lyapunov functions to be continuous on the switching surfaces. (In particular, this means that we can do without the complex machinery of “continuity matrices”.) We shall only require that V decreases on every switch, possibly in a discontinuous manner. The reasons for this choice will become clear later on.

We now briefly describe the model introduced in [AF13] to describe the resilience of interconnected systems.

The starting point is a set of n linear dynamical systems which are interconnected according to the edges of a directed graph G . The state of each system $k \in \{1, \dots, n\}$ is abstracted into a single variable $x_k \in [0, 1] \subseteq \mathbb{R}$ (“percentage of service loss”), with 0 indicating regular functioning and 1 indicating total failure.

Each system has a corresponding *service recovery rate* $\mu_k > 0$, which quantifies its recovery capability when the state is perturbed away from zero. Moreover, for each pair of systems (i, k) connected by an arc $i \rightarrow k$ in the graph G the following parameters are defined:

- a *coupling coefficient* $\alpha_{ik} > 0$ between the state variables of the two systems;
- a *state threshold* $\sigma_{ik} \in (0, 1)$ representing the maximum percentage of service loss beyond which the system i is considered to be failed for the descendant system k ;
- a *service loss rate* $\lambda_{ik} > 0$ representing the rate of approaching failure of system k when the ancestor system i is malfunctioning.

Let I_k be the set of incoming neighbours of node k . System k is in *nominal operation mode* when $x_i \leq \sigma_{ik}$ for every $i \in I_k$. The evolution equation for the system is then

$$\dot{x}_k = -\mu_k(x_k - \sum_{i \in I_k} \alpha_{ik}x_i) + d_k, \quad (3)$$

where d_k is a real function modeling the (time-varying) external disturbance experienced by the k -th system.

If any of the input dependencies are malfunctioning the systems enters a failure mode. Denoting by m_k the cardinality of I_k , there are $2^{m_k} - 1$ such modes for system k : m_k where a single ancestor system is malfunctioning, $\binom{m_k}{2}$ where two ancestors are malfunctioning, and so on. For each $i \in I_k$, if $x_i > \sigma_{ik}$ is the only failed input then the evolution equation becomes

$$\dot{x}_k = \lambda_{ik}(1 - x_k), \quad (4)$$

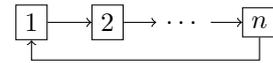
and more generally if $J_k \subseteq I_k$ is the subset of failed inputs we put²

$$\dot{x}_k = \sum_{j \in J_k} \lambda_{jk}(1 - x_k). \quad (5)$$

This system of evolution laws defines a piecewise-linear switching system whose state space is the n -dimensional hypercube $[0, 1]^n \subset \mathbb{R}^n$. The basin of attraction of the equilibrium point located at the origin will be called the *resilience region* of the system.

III. SYSTEM CONFIGURATION

In the sequel we shall only consider the models that correspond to the family of graphs G_n defined as follows. For each $n \in \mathbb{N}$, the graph G_n has n nodes; for each $i = 1 \dots n-1$, node i node is connected by an edge to node $i+1$; finally, node n is connected to node 1. In other words, we consider a chain of systems with a final “feedback” connection:



For $n = 2$, this is exactly the model that was considered originally in [AF13]. As in the latter paper, we are only interested in evaluating the resilience against “instantaneous” disturbances that shift the state of the system away from the equilibrium, so from now on we shall assume that the noise term d_k in the evolution equation (3) is zero.

To each edge³ $i \rightarrow i+1$ in the graph G_n there corresponds a threshold $\sigma_{i,i+1} \in [0, 1]$ that we are going to denote more briefly by σ_i . This notation is quite natural in this setting, since the value σ_i pertains to the state variable of the i -th subsystem.

Each threshold divides the segment $[0, 1]$ along the i -th axis of the state space in two parts, $[0, 1] = [0, \sigma_i] \cup [\sigma_i, 1]$. It follows that the state space is partitioned into a total of 2^n cells. Each cell is a n -dimensional (rectangular) hypercuboid; the switching surfaces are the hypersurfaces $x_i = \sigma_i$ for $i = 1 \dots n$.

The remaining parameters defining the systems are:

- n service recovery rates μ_1, \dots, μ_n ;
- n service loss rates $\lambda_1, \dots, \lambda_n$, where (compared to the notations of Section II) λ_i is a shorthand for $\lambda_{i-1,i}$;
- n coupling coefficients $\alpha_1, \dots, \alpha_n$, where α_i is a shorthand for $\alpha_{i-1,i}$.

²This rule for combining failures is not explicitly stated in [AF13].

³Here and in the sequel every index running from 1 to n is understood to be taken modulo n , so that $n+1 = 1$.

It is important to note that the model is *positive* (all state variables are constrained to assume only positive values) and *bounded*.

Let us introduce a more systematic labeling for the cells of this switching system. For any subset S of $\{1, \dots, n\}$ we shall denote by R_S the region where every system whose index belongs to S is over the threshold, and all the remaining systems are under the threshold. In particular R_\emptyset denotes the region corresponding to the normal operation mode, i.e. the hypercuboid

$$R_\emptyset = [0, \sigma_1] \times \dots \times [0, \sigma_n],$$

whereas for instance

$$R_{\{1\}} = [\sigma_1, 1] \times [0, \sigma_2] \times \dots \times [0, \sigma_n]$$

represents the cell in which system 1 is failed for system 2, and so on, up to

$$R_{\{1\dots n\}} = [\sigma_1, 1] \times \dots \times [\sigma_n, 1]$$

which denotes the region of total failure.

The resilience region obviously includes the whole of R_\emptyset and is disjoint from region $R_{\{1\dots n\}}$. We are interested in delineating its boundary in the remaining $2^n - 2$ regions where some, but not all, of the subsystems are failing.

Let us reformulate the dynamics of the system in matrix form, using the linear embedding of affine dynamics described by equations (1-2). In region R_\emptyset the system is linear with evolution matrix

$$A_\emptyset = \begin{pmatrix} -\mu_1 & 0 & 0 & \dots & \mu_1 \alpha_1 \\ \mu_2 \alpha_2 & -\mu_2 & 0 & \dots & 0 \\ 0 & \mu_3 \alpha_3 & -\mu_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \mu_n \alpha_n & -\mu_n \end{pmatrix} \quad (6)$$

An easy proof by induction shows that

$$\det A_\emptyset = (-1)^n \mu_1 \dots \mu_n (1 - \alpha_1 \dots \alpha_n)$$

It follows that the dynamics is nonsingular if and only if $\alpha_1 \dots \alpha_n \neq 1$; in the following we are going to suppose this is the case.

We also assume that in this region the system is stable. Using the well-known Routh-Hurwitz criterion for positive systems it is easy to check that this happens if and only if $\alpha_1 \dots \alpha_n < 1$.

When some system fails, for instance when system n fails for system 1, the dynamics switches from linear to affine and the new (extended) evolution matrix reads

$$\bar{A}_{\{1\}} = \begin{pmatrix} -\lambda_1 & 0 & 0 & \dots & 0 & \lambda_1 \\ \mu_2 \alpha_2 & -\mu_2 & 0 & \dots & 0 & 0 \\ 0 & \mu_3 \alpha_3 & -\mu_3 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \mu_n \alpha_n & -\mu_n & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 \end{pmatrix}$$

Similarly, the failure of the i -th system causes the i -th row of the matrix (6) to be replaced with a row of the form

$$0 \quad \dots \quad 0 \quad -\lambda_i \quad 0 \quad \dots \quad 0 \quad \lambda_i$$

In particular for the region number $2^n - 1$ where all subsystems have failed the evolution matrix is

$$\bar{A}_{\{1\dots n\}} = \begin{pmatrix} -\lambda_1 & 0 & \dots & 0 & \lambda_1 \\ 0 & -\lambda_2 & 0 & \dots & \lambda_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda_n & \lambda_n \\ 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

It is interesting to note that these models are actually exactly solvable: the shape of the resilience region can (at least in principle) be determined analytically. This was done in [AF13] for the case $n = 2$ with no coupling between the two subsystems and no external disturbances. The boundary of the resilience region turns out to be the part of the square $[0, 1]^2$ which lies below the two curves

$$x_2 = 1 - (1 - \sigma_2) \left(\frac{x_1}{\sigma_1} \right)^{\lambda_2/\mu_1}$$

and

$$x_2 = \sigma_2 \left(\frac{1 - x_1}{1 - \sigma_1} \right)^{\mu_2/\lambda_1}$$

which meet (with a corner) at the point (σ_1, σ_2) .

A similar analysis can be performed for a system with $n > 2$ nodes; however, the explicit description of the boundary gets more and more complex as the dimension grows. Notice also that the functions defining the boundary are transcendental, so they are quite expensive to compute with arbitrary precision.

For this reason it is advantageous to have a quickly computable (e.g. piecewise-linear) approximation of the boundary. In the next Section we turn to this goal.

IV. FROM STABILITY TO RESILIENCE

The general summary of our approach is the following. We look for a Lyapunov function whose sublevel sets approximate the basin of attraction of the equilibrium at zero. This function must be computationally cheap to work with; the simplest choice is to consider a piecewise-linear function on a polyhedral partition of the state space. The partition is computed starting from the cells of the switching system by successive refinements. Since the models we use are positive and have no limit cycles, we expect piecewise-linear Lyapunov functions to be a good fit.

Let us describe in detail our approach to construct a piecewise-linear Lyapunov function for the family of switching models described in Section III.

We start by looking for a piecewise-linear Lyapunov function defined on the same polyhedral partition of the state space given by the switching surfaces, namely the 2^n cells R_S for $S \subseteq \{1, \dots, n\}$. Actually, we already know that the cell corresponding to the choice $S = \{1, \dots, n\}$ is surely outside of the resilience region, so

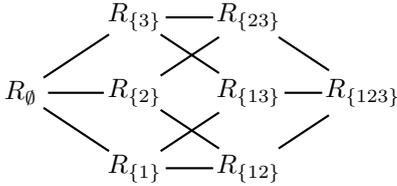
we can safely ignore this cell in what follows. Thus, we consider a Lyapunov function candidate of the form

$$V(x) = \begin{cases} \mathbf{k}_0 \cdot \mathbf{x} & \text{if } \mathbf{x} \in R_\emptyset, \\ \bar{k}_S \cdot \bar{x} = \mathbf{k}_S \cdot \mathbf{x} + k_{S,n+1} & \text{if } \mathbf{x} \in R_S. \end{cases} \quad (7)$$

Here $\mathbf{k}_0 \in \mathbb{R}^n$, $\bar{k}_S = (\mathbf{k}_S, k_{S,n+1}) \in \mathbb{R}^{n+1}$ and the index S runs over the proper nonempty subsets of the set $\{1, \dots, n\}$. The total number of parameters to be determined is then $n + (2^n - 2)(n + 1)$.

If we now require V to be continuous on *every* switching surface we see that as soon as $n > 2$ the continuity constraints eat up all the degrees of freedom in the multiple-failure regions. To clarify this point, let us start by noting that every cell in the decomposition $\{R_S\}$ can be put in one-to-one correspondence with a vertex of the hypercube $[0, 1]^n$: for any subset $S \subseteq \{1, \dots, n\}$ the region R_S corresponds to the vertex $V(S)$ whose coordinates are 1 for each axis x_i with $i \in S$ and 0 for each axis x_i with $i \notin S$.

The incidence relations between the vertices and the edges of an n -dimensional hypercube are described by the *hypercube graph* Q_n [HHW88]. Using the correspondence defined above, we can use this graph to describe the relationship between different regions in the partition $\{R_S\}$. For instance when $n = 3$ we have



The dimension of the boundary between two regions can be recovered from this graph: indeed $R_S \cap R_{S'}$ is a cuboid of dimension $n - k$, where k is the (geodesic) distance between R_S and $R_{S'}$ in the graph Q_n . For instance in the $n = 3$ case R_\emptyset and $R_{\{1\}}$ intersect in a rectangle (dimension 2), R_\emptyset and $R_{\{12\}}$ intersect in a segment (dimension 1) and finally R_\emptyset and $R_{\{123\}}$ intersect in a single point (dimension 0).

We can now cook up a recipe to impose some reasonable continuity constraints on the candidate Lyapunov function (7). Our strategy is the following:

- for each $i \in \{1, \dots, n\}$ we require $V_{\{i\}}$ to match V_\emptyset on the whole $(n - 1)$ -dimensional cuboid $R_\emptyset \cap R_{\{i\}}$;
- for each $i < j \in \{1, \dots, n\}$ we require $V_{\{ij\}}$ to match V_\emptyset on the $(n - 2)$ -dimensional cuboid $R_\emptyset \cap R_{\{ij\}}$;
- for each $i < j < k \in \{1, \dots, n\}$ we require $V_{\{ijk\}}$ to match V_\emptyset on the $(n - 3)$ -dimensional cuboid $R_\emptyset \cap R_{\{ijk\}}$; and so on. We remark that on each $(n - k)$ -dimensional cuboid it is enough to select a subset of vertices forming a $(n - k)$ -dimensional simplex, since two affine functions coinciding on (the vertices of) a simplex coincide everywhere.

The number of constraints generated by this rule is given by

$$\begin{aligned} n_c &= \binom{n}{1}n + \binom{n}{2}(n - 1) + \dots + \binom{n}{n-1} \cdot 2 \\ &= \sum_{i=1}^{n-1} \binom{n}{i}(n - i + 1). \end{aligned}$$

Using some well-known properties of the binomial coefficients we see that

$$n_c = (2^n - 2)\left(\frac{n}{2} + 1\right),$$

so that the number of free parameters in (7) is reduced to

$$n + (2^n - 2)(n + 1) - n_c = 2^{n-1}n.$$

We shall also require that $V_\emptyset(\sigma_1, \dots, \sigma_n) = 1$. This can be seen as a normalization condition for V ; it is particularly convenient because the trajectories that hit the critical point $(\sigma_1, \dots, \sigma_n)$ lie exactly on the boundary of the resilience region.

We now formulate the conditions which guarantee that the function (7) is a Lyapunov function for the system. We need two kinds of constraints:

- we should ensure that V is positive on each vertex of region R_\emptyset and decreasing on every vertex in a boundary between two cells (it is of course sufficient to consider the vertices where no continuity conditions are imposed);
- we should ensure that V is strictly decreasing along the trajectories of the system in the interior of each cell. Since the dynamics there is linear, it suffices to check that

$$\begin{cases} \mathbf{k}_0 \cdot A_\emptyset \cdot \mathbf{v} < 0 & \text{for every vertex } \mathbf{v} \text{ of } R_\emptyset, \text{ and} \\ \bar{k}_S \cdot \bar{A}_S \cdot \bar{v} < 0 & \text{for every vertex } \bar{v} \text{ of } R_S. \end{cases}$$

If we group the entries of the vectors \mathbf{k}_0 and \bar{k}_S in a single column matrix K , both these conditions can be expressed as a system of inequalities of the form $MK < 0$ for a suitable block matrix M .

Finally, we should select a suitable function of the coefficients of V to minimize in order to capture the largest possible subset of the resilience region. As V is everywhere positive, it suffices to minimize the sum of the values attained by V on a suitable set of $2^n - 1$ points (one for each region of the partition). Since (as explained above) each region is in one to one correspondence with a vertex of the cube $[0, 1]^n$, it is natural to minimize the value of V on such points. Thus we define

$$F(K) := V(1, 0, \dots, 0) + V(0, 1, 0, \dots, 0) + \dots + V(1, 1, 0, \dots, 0) + \dots$$

and look for the matrix $K = (\mathbf{k}_0, \bar{k}_{\{1\}}, \bar{k}_{\{2\}}, \dots)$ which minimizes this function.

Summing up, we are led to the problem of minimizing $F(K)$ subject to the constraints:

$$\begin{cases} MK \leq 0 \\ EK = 0 \\ V(\sigma_1, \dots, \sigma_n) = 1 \end{cases} \quad (8)$$

where E is a block matrix capturing the continuity constraints expressed earlier.

The solution of this linear programming problem (which we expect to be feasible for generic values of the parameters) determines a piecewise-linear Lyapunov

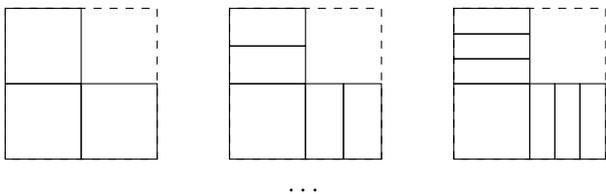
function for the system. Then, by construction, the sublevel set

$$\{x \in [0, 1]^n \mid V(x) \leq 1\}$$

gives a piecewise-linear approximation to the resilience region.

The cell partition introduced above is rather coarse. As we saw in Section III, the boundary of the resilience region is given by a system of transcendental equations. In general, a single hyperplane will give a rather bad approximation for this boundary. It is then natural to seek a refinement of the partition $\{R_S\}$ in order to better capture the shape of the resilience region for the system.

For this purpose we can try to exploit again the special geometry of the state space of the model. The simplest idea is to consider refinements built according to the following rule, illustrated here for the $n = 2$ case:



In other words, we divide each single-failure region $R_{\{i\}}$ in $m > 1$ subregions and let the coefficients $\bar{k}_{\{i\}}$ in the expression (7) vary between those subregions.

Of course, things get quickly more complicated when $n > 2$: clearly, the subdivision described above automatically induces a subdivision of each double-failure region $R_{\{ij\}}$ in m^2 subregions, and so on. In general, one ends up with a total of

$$1 + nm + \binom{n}{2}m^2 + \dots + \binom{n}{n-1}m^{n-1} = (m+1)^n - m^n$$

subregions. Notice that for $m = 1$ the above formula recovers the original number of regions ($2^n - 1$).

The new candidate Lyapunov function now reads

$$V(x) = \begin{cases} \mathbf{k}_0 \cdot \mathbf{x} & \text{if } \mathbf{x} \in R_\emptyset, \\ \bar{k}_{S,j_S} \cdot \bar{x} & \text{if } \mathbf{x} \in R_{S,j_S} \end{cases} \quad (9)$$

where, as before, the index S runs over the proper nonempty subsets of the set $\{1, \dots, n\}$ and the additional index j_S runs over the set $\{1, \dots, m^{|S|}\}$, where $|S|$ denotes the cardinality of the set S . The total number of parameters to be determined is in this case $n + (n+1)((m+1)^n - m^n - 1)$.

Generalizing the problem (8) to the new situation is, at least in principle, straightforward; the only difficulty has to do with the bookkeeping needed to manage the various constraints. In particular the continuity constraints adopted above for the case $m = 1$ can be generalized in multiple ways to the case $m > 1$, giving rise to linear programming problems which are more or less constrained. We shall leave a detailed analysis of such issues to future work.

V. EXPERIMENTS

To test our approach we consider a model inspired by a real urban wastewater treatment facility which ensures depollution and dumping at sea of urban wastewater produced by domestic and economic activities in an international tourist area encompassing a marine reserve. The facility handles an estimated maximum of 36,000 people, roughly equivalent to a waste-water supply of 250 liters per person, per day. The plant is heavily automatized: all biological, chemical and mechanical processes are supervised by a control system connected through the Internet with a remote monitoring center. The plant consists of several compartments characterized by interconnected tanks, making it an ideal test bench for our framework. In particular, we focus on a series of three interconnected subsystems inside the facility: a balancing (BA) tank, a denitrification tank (DE) and a nitrification-oxidification (NO) tank. The BA tank receives sewage from previous compartments and feeds the DE tank with enough liquor to maintain the level of the NO tank at a desired level. The DE tank regulates its own level, dumping excess fluid to the NO tank. Liquor from the NO tank simply falls by gravity to other compartments for further treatment. As it can be observed in Figure 1, the topology of the interconnected systems respects the one considered in Section III. The input flow is a single-step function at $0.08 \text{ m}^3/\text{s}$, corresponding to the average inlet flow of the real facility (approximately $280 \text{ m}^3/\text{h}$). Gaussian white noise is added to the single-step function in order to simulate hourly variation of the inlet flow. The absolute value of the variation is within 20% of the average flow, consistently with data recorded at the facility⁴.

For each tank $i \in \{1, 2, 3\}$ of Figure 1, the performance loss x_i is obtained from the corresponding output flow u_i through a *figure-of-merit function*, i.e., a mapping from the space of state variables to the space of performances. This is the standard way in which state variables are mapped to performance indicators — see, e.g., [HRM12] — and our specific mapping can be observed in Figure 2. Here we posit that an output flow of $0.08 \text{ m}^3/\text{s}$ corresponds to maximum functionality, whereas flows outside the interval $[0.06; 0.10] \text{ m}^3/\text{s}$ correspond to total loss of functionality. The state threshold σ is placed at $x_i = 0.8$, corresponding to flows in the interval $[0.064; 0.096] \text{ m}^3/\text{s}$, i.e., a variation of $\pm 20\%$ consistent with the actual input data. The choice of these parameters is motivated by the fact that the average input flow to the facility is $0.08 \text{ m}^3/\text{s}$ with a variation of $\pm 20\%$. Outside this range, flows are considered to be abnormal, whereas variations above $\pm 25\%$ are considered total loss of functionality. In our experiments $\sigma_{ij} = \sigma$ for all $i, j \in \{1, 2, 3\}$ with $i \neq j$.

In order to analyze the resilience of the overall system according to the framework presented in Section II, we need to compute the values of the service loss rates,

⁴The complete model, as well as instructions to run it, are available on the companion web site of the paper at <https://gitlab.sagelab.it/armtac/ifm2019companion.git>

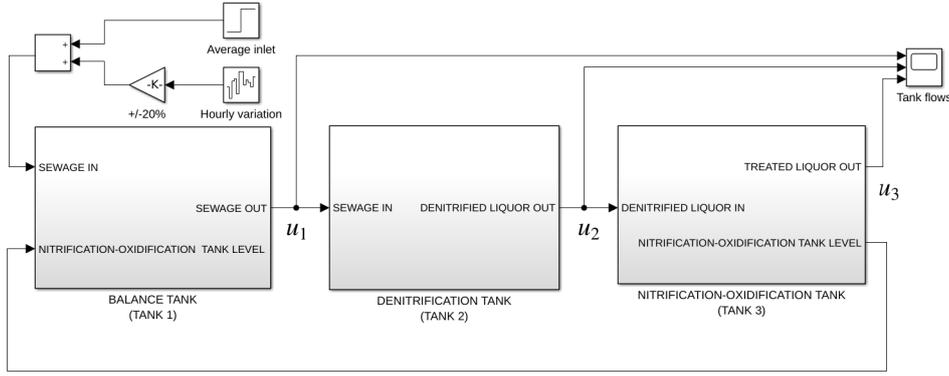


Fig. 1. Matlab/Simulink[®] model of three interconnected subsystems inside a wastewater treatment facility.

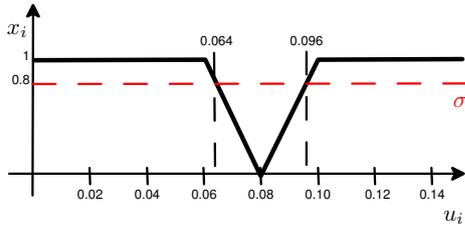


Fig. 2. Figure-of-merit (FOM) function to map tank outlet flow u_i to performance loss x_i . The red dotted line is the threshold σ beyond which the system is considered to be failed. Black dotted lines mark the points on the u_i axis within which the system is considered to be working properly.

recovery rates and coupling coefficients. To this end, we consider the definitions presented in [AF13] and, given two interconnected systems $i, j \in \{1, 2, 3\}$ with $i \neq j$, we compute μ_{ij} and λ_{ij} as follows:

$$\mu_{ij} = -\frac{\ln(\sigma_{ij})}{TTR_{ij}} \quad \lambda_{ij} = -\frac{\ln(1 - \sigma_{ij})}{TTF_{ij}} \quad (10)$$

where

- TTR_{ij} (Time To Recover) is the time that system j takes to recover from a failure of system i and return within the state threshold σ_{ij} from the internal state $x_j = 1$, and
- TTF_{ij} (Time To Fail) is the time that system j takes to cross the state threshold σ_{ij} from the state $x_j = 0$.

Estimation of TTR , TTF and the coupling coefficient α can be done on the system model by introducing faults and then observing the behavior of relevant process variables. In particular we proceed as follows:

- To estimate TTF_{12} and TTR_{12} we introduce a stuck-at-0 fault in the pump that drains liquor from the BA tank into the DE tank. This causes the flow u_1 to decrease abruptly to 0. Consequently, the flow u_2 also decreases to 0 since the DE tank regulator tries to maintain a desired level in the tank, and stops sending fluid to the NO tank eventually. This effect is shown graphically in Figure 3 (top).
- To estimate TTF_{23} and TTR_{23} we introduce a stuck-at-0 fault in the pump that keeps the level of the DE tank stable. This causes the flow u_2 to decrease abruptly to 0. Consequently, the NO tank is not feeded

TABLE I: System parameters computed by injecting faults in the system.

i	j	TTF	TTR	λ	μ	α
1	2	110.35	261.50	0.01458	0.00085	0.9
2	3	2251.00	1714.00	0.00071	0.00013	0.1
3	1	2127.00	1540.00	0.00076	0.00014	0.7

anymore and starts emptying: reduction in the volume of the NO tank causes also reduction in its output flow which is due to gravity only.

- to estimate TTF_{31} and TTR_{31} we introduce a fault in the NO tank outlet by simulating a partial obstruction, which causes the NO tank to reduce its outlet flow below the average system input flow. Since the pump in the BA tank is regulated on the level of the NO tank, this will cause the output flow of the BA tank to decrease as well. The recovery from this fault is shown graphically in Figure 3 (bottom).

To estimate the coupling coefficient we proceed similarly, but instead of stuck-at faults, we consider drifting faults from nominal values and evaluate the ratio between the slopes of the overall flow variations. For instance, in order to estimate α_{12} we decrease the efficiency of the pump that drains liquor from the BA tank into the DE tank. Both u_1 and u_2 start to decrease until the normal functionality is resumed. The value of α_{12} is computed as the ratio between the slopes of u_1 and u_2 considering the onset of the failure as initial point and the end of the failure as final one. In all the cases above, faults are injected when the system reaches a regime condition (after 6 hours from the start) and last for a specified amount of time (1 hour).

Using Matlab we solved the linear programming problem (8) for the three-dimensional model with the coefficients listed in Table I. The complete matlab scripts and instructions on how to run them can be found on the companion web site mentioned above. We found the following values for the 27 parameters describing the Lyapunov function V :

$$\mathbf{k}_0 = \begin{pmatrix} 0.7081 \\ 0.0082 \\ 0.5338 \end{pmatrix} \quad \bar{k}_{\{1\}} = \begin{pmatrix} 1.8903 \\ 0.0082 \\ 0.5338 \\ -0.9458 \end{pmatrix}$$

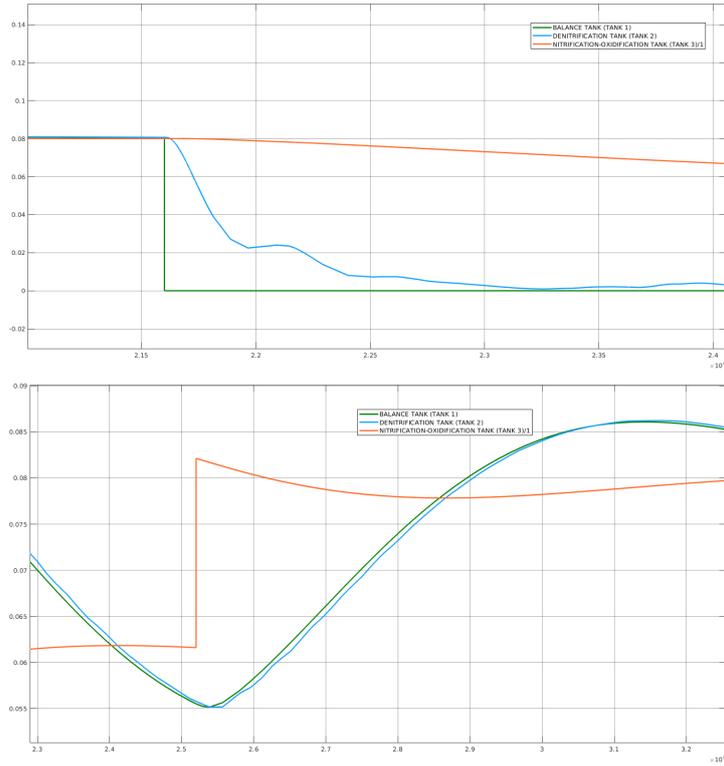


Fig. 3. Effects of injected faults on output flows. In each plot the lines represent the evolution of the tank outlet flows over time (x -axis in seconds, y -axis in cubic meters per second). The plot on the top shows the cascading effect of a “stuck-at-0” fault in the pump that drains liquor from the BA tank into the DE tank. The plot on the bottom shows the recovery from the effects of a partial obstruction in the NO tank on the flows of the BA and DE tanks.

$$\bar{k}_{\{2\}} = \begin{pmatrix} 0.7081 \\ 4.4070 \\ 0.5338 \\ -3.5191 \end{pmatrix} \quad \bar{k}_{\{3\}} = \begin{pmatrix} 0.7081 \\ 0.0082 \\ 5.6899 \\ -4.1249 \end{pmatrix}$$

$$\bar{k}_{\{12\}} = \begin{pmatrix} 384.7189 \\ 4.4070 \\ 0.5338 \\ -310.7277 \end{pmatrix} \quad \bar{k}_{\{13\}} = \begin{pmatrix} 1.8903 \\ 0.0082 \\ 5.6899 \\ -5.0707 \end{pmatrix}$$

$$\bar{k}_{\{23\}} = \begin{pmatrix} 0.7081 \\ 13.4647 \\ 5.6899 \\ -14.8901 \end{pmatrix}$$

Some graphs of the resulting piecewise-linear Lyapunov function are represented in figure 4 and figure 5 as slices in the (x_1, x_2) plane for various values of x_3 .

We notice that the resulting Lyapunov function has a very steep slope in the region $\{12\}$, i.e., when system 3 is the only subsystem operating normally. This is presumably due to both the (comparatively) higher value of $\lambda_2 = \lambda_{12}$ and the strong coupling between the two failed subsystems. On the contrary, V is quite flat in the three single-failure regions and also in the region $R_{\{13\}}$, thanks to the good recovery capability of subsystem 1. Running the system model with faults injected confirms the picture obtained with the analysis.

VI. CONCLUSIONS

In this paper we have introduced a new method to compute piecewise linear Lyapunov functions to study

the stability of interconnected systems, and thus the related resilience regions. We have shown that the method allows to analyze automatically the model of a real wastewater treatment facility by providing a computation approach and experimental results.

For a future development of this work it would be interesting to generalize our stability analysis to the full resilience model, with a nonzero time-varying external disturbance term in the evolution equation (3). Since the disturbance is not known a priori the model becomes a so-called *uncertain system* in this case, and different techniques need to be applied.

We would also like to consider interconnected systems with different (but still computationally tractable) topologies. In this regard, let us remark that the family of graphs described in Section III has the property that the number of interconnections grows only *linearly* as a function of the number of nodes n (as opposed to, e.g., a complete graph, where the number of edges grows quadratically in n). Other families of graphs with this property look relevant for the modeling of real-world systems.

In general, when some subsystem has more than one connection to other subsystems we expect to see more complex dynamical phenomena emerge (e.g. limit cycles). In these cases piecewise-linear Lyapunov functions are no longer a good fit, and it may become necessary to use a more general class of Lyapunov function candidates (piecewise quadratic, sum of squares, etc.).

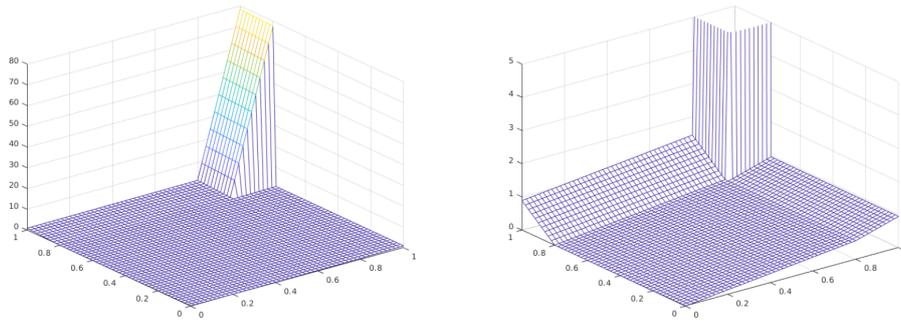


Fig. 4. Values of the Lyapunov function for $x_3 = 0$; the right-hand pane shows the same graph clipped to the region $V \leq 5$.

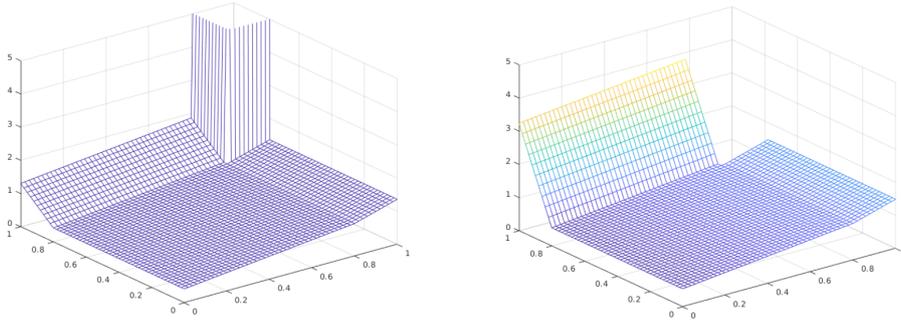


Fig. 5. Values of the Lyapunov function for $x_3 = 0.78$ and $x_3 = 0.82$ (threshold is $\sigma_3 = 0.8$).

REFERENCES

- [AF13] Angelo Alessandri and Roberto Filippini. Evaluation of resilience of interconnected systems based on stability analysis. In *Critical Information Infrastructures Security*, pages 180–190. Springer, 2013.
- [AGS⁺16] Xavier Allamigeon, Stéphane Gaubert, Nikolas Stott, Eric Goubault, and Sylvie Putot. A scalable algebraic method to infer quadratic invariants of switched systems. *ACM Trans. Embedded Comput. Syst.*, 15(4):69:1–69:20, 2016.
- [BHSZ15] Fredrik Björck, Martin Henkel, Janis Stirna, and Jelena Zdravkovic. Cyber resilience—fundamentals for a definition. In *New contributions in information systems and technologies*, pages 311–316. Springer, 2015.
- [CLM04] Paolo Crucitti, Vito Latora, and Massimo Marchiori. Model for cascading failures in complex networks. *Physical Review E*, 69(4):045104, 2004.
- [GH15] Peter Giesl and Sigurdur Hafstein. Review on computational methods for Lyapunov functions. *Discrete Contin. Dyn. Syst. Ser. B*, 20(8):2291–2331, 2015.
- [HHW88] Frank Harary, John P. Hayes, and Horng-Jyh Wu. A survey of the theory of hypercube graphs. *Comput. Math. Appl.*, 15(4):277–289, 1988.
- [HRM12] Devanandham Henry and Jose Emmanuel Ramirez-Marquez. Generic metrics and quantitative approaches for system resilience as a function of time. *Reliability Engineering & System Safety*, 99:114–122, 2012.
- [Joh03] Mikael Johansson. *Piecewise linear control systems*, volume 284 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin, 2003. A computational approach.
- [Kha92] Hassan K. Khalil. *Nonlinear systems*. Macmillan Publishing Company, New York, 1992.
- [LFZ17] X Liu, E Ferrario, and Enrico Zio. Resilience analysis framework for interconnected critical infrastructures. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 3(2):021001, 2017.
- [Lib03] Daniel Liberzon. *Switching in systems and control*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2003.
- [Lou15] George Loukas. *Cyber-physical attacks: A growing invisible threat*. Butterworth-Heinemann, 2015.
- [LPZ14] Xing Liu, Ionela Prodan, and Enrico Zio. On the resilience analysis of interconnected systems by a set-theoretic approach. *Safety and Reliability: Methodology and Applications, CRC Press, Leiden, The Netherlands*, pages 197–205, 2014.
- [Oba13] Barack Obama. Presidential Policy Directive 21 (PPD21): Critical infrastructure security and resilience. *Washington, DC*, 2013.
- [SG11] Zhendong Sun and Shuzhi Sam Ge. *Stability theory of switched dynamical systems*. Communications and Control Engineering Series. Springer, London, 2011.

Towards Artificial Neural Network Hashing with Strange Attractors Usage

Jacek Tchórzewski
AGH University of Science and Technology
30-059 Cracow, Poland
Cracow University of Technology
31-155 Cracow, Poland

Agnieszka Jakóbik
Cracow University of Technology
31-155 Cracow, Poland

KEYWORDS

Chaotic Attractor; Strange Attractor; Artificial Neural Networks; Intelligent Security System; Hash Functions; Lightweight Hash Functions;

ABSTRACT

A broad variety of methods ensuring the integrity of data in the mobile and IoT equipment is very important nowadays. Hash functions are used for detecting the unauthorized modification of data and for digital signatures generation. Traditional hash functions like SHA-2 or SHA-3 have relatively high computational power requirements, therefore are not always suitable (or optimal) for devices with limited computational capacity or battery capacity.

Instead, light cryptography hash functions may be used. They are processing data strings of the shorter length and offers simpler mathematical models as the basis of hash calculation.

In this paper Artificial Neural Network (ANN)-based model hashing is proposed. Instead of using s-boxes or complicated compression function, a simple two-layered non-recurrent ANNs are used for hash calculation. In order to provide a very high quality of the randomization of the output, several different chaotic attractors were incorporated into ANNs training phase. ANNs output was tested with appropriate statistical tests and compared with hashes returned by traditional hashing methods. Using shorter hash length enables implementing those methods in the mobile and IoT equipment. Our approach allows merging the low complexity of ANN processing with the high-quality standards of cryptography hash functions.

I. INTRODUCTION

Cryptographic services for data processed by low computing capacity machines are more challenging than traditional cryptography methods. From among them, light cryptography hash functions are commonly used for verifying the data set integrity. Methods used for Cloud or Big Data processing like SHA-2 and SHA-3 may be in some cases too computationally demanding for smartphones or tablets. Instead, dedicated hashing methods are used, like DM-PRESENT-80, PHOTON-80/20/16, Parallel Keccak-f[200], Spongent-88/80/8 or ARMADILLO, [6].

The hash length that they are calculating varies from 48 to 256 bits. Their construction is based on many cycles of computation for the single hashing block. For example, DM-PRESENT-80 needs 4547, PHOTON-80/20/16 uses 708 cycles and ARMADILLO 176 cycles for single block processing. Their construction is based on light sponge algorithms or short 4-bit S-boxes. Instead of long permutation series, the shorter ones are implemented in order to design more hardware-friendly procedures.

The paper presents simpler model that is based on two layers ANN. Weighed sums of the input signal are transferred by the sigmoid activation function of the neuron unit. The quality of hashing providing the sufficient level of randomization of the hash output is coded inside the memory of the ANN trained by using chosen chaotic dynamical nonlinear systems. Three concurrent strange attractors were used [9], [5], [17]. Results will be compared to the original hashing standards, like SHA-2 and SHA-3.

The paper is organized as follows. Section (II) is describing concurrent hashing models that involve the usage of chaotic systems. Section (III) is presenting our model which consists of Artificial Neural Networks and three different attractors: Lorenz, Rossler, and Henon. Definitions of all attractors are introduced as well as the main idea of our hashing system. In section (IV) we are defining statistical tests that were applied on produced hashes. We gave details about testing procedures, illustrated results, and compared them with the same tests applied on certificated hashing functions. We are also describing how ANNs input and target data were prepared. 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

A hash function is a method of transforming the given input bit string into a fixed size output bit string, [23]. The cryptographic hash function may be used for data integrity verification and digital signature generation. They also must fulfill three additional requirements:

1. calculation of hash from long messages may involve multiple usages of hashing procedure.
2. it is computationally infeasible to calculate input string from a given output string (hash).
3. the probability that two different input strings will have the same hash is very low.

A cryptographic hash function is like a random transformation of the input string. Classical algorithms used nowadays (SHA-2 and SHA-3) enables to obtain 256, 386 or 512 bits of output, [1], [2]. Light cryptography hash functions, [6], output strings are equal to 60, 80, 128, or 160 bits. Those numbers are determined by algorithms, which implicates the fact that no in-between hash sizes are available.

Different alternative methods were introduced for obtaining hash functions based on chaotic time series and chaotic dynamical systems. In [15] a Lorenz system was used for obtaining the secret key, used for the next step of the computation. An output hash is generated in four different iterations. Each of them feeds with an intermediate hash value of the previous one, part of the input and the secret keys. Iterations involve: dividing intermediate results into parts, padding them, performing logical XOR and AND operations with usage of a secret key, and some procedures known from calculating SHA-2 hash. Moreover, the rotations are also used like in SHA-2. The time of calculation was proposed as a computational complexity measure and was compared to the outdated standard Secure Hash Algorithm 1 (SHA-1). The final comparison with SHA-1 was proved to be satisfactory, however next-generation hashing methods (SHA-2 or SHA-3) were not examined by authors.

In [14] authors presented a hash-alike algorithm based on Lorenz's attractor that may be used also for checking the integrity of the data. The algorithm incorporates a series of powering operations using large numbers and calculating the modulus of the results with base equal to large powers of 2. Those operations may be considered as similar to the RSA ciphering scheme. This algorithm does not outperform the AES algorithm as far as the time of computation is considered. It was also not tested as classical cryptography hash function.

In [12] authors presented hash algorithm using the hyper-chaotic Lorenz system. The algorithm is based on a sponge function that absorbs the input message. The function is quite complicated which results in some time-varying perturbations. Basic operations were similar to the [14]: multiplying by large numbers and calculating modulus. The algorithm was tested for producing 256-bit and 512-bit hash values, and compared with SHA-2 and SHA-3 hashing standards. This solution, however, can generate 256, 512, 1024 bits or longer hash values. It was not tested for the light cryptography hash output strings.

In [3] hash function based on the 3D chaotic map was proposed. The algorithm consumes chunks of the message one by one with the usage of the equation presented in the article. XOR and modulus operations are also incorporated. Authors tested their solution from

many perspectives but results were compared only to the traditional hash functions such as SHA-1 and MD5. This method offers the following digests sizes (in bits): 128, 160, 256, or 512.

In [7] authors presented a color image encryption scheme with the usage of one-time keys, based on crossover operation and the pseudo-random sequences generated by the Lorenz system. The 256-bit hash value is used as the one-time key and then applied to calculate the initial values of the Lorenz system. The permutation-diffusion process used after is based on the crossover operation and XOR operator. This solution is not a hashing method but the experiment results indicate that incorporating a chaotic system may result in a high-quality encryption scheme, even when assuming only one round of the process.

More pieces of information about hashing functions and hashing strategies can be found in [19]. Different approach assuming usage of evolutionary algorithms and genetic programming for hashing purposes was presented in [11].

All methods described above were not authorized by international security agencies like The International Organization for Standardization (ISO), The National Institute of Standards and Technology (NIST), or The British Standards Institution (BSI) so far.

In contrast to the solutions described above our model assumes only light hashing with usage of Artificial Neural Networks and three different chaotic attractors, and can be compared with [22] where Mackey-Glass [4] chaotic time series was used for hash creation. Our idea assumes preparing a light hashing algorithm that may be used in cryptographic procedures and offers the possibility to change the bit length of the fingerprints (hashes) of messages. To achieve this goal, we compared three chaotic series to decide which one will offer the best hash quality.

III. CHAOTIC MODELS USED FOR ANN TRAINING

The algorithm of hash calculation is based on the usage of Artificial Neural Networks. Trained ANNs were operating on strings of data (ANN inputs), that had fixed length. For the simplicity of the presentation, input data were considered as bit strings. All three chaotic models presented in this section were used for ANNs target data preparation. Trained ANNs were treated as potential hash creators. The detailed report describing how inputs/targets were generated, how long were hashes and what was ANNs structure is presented in Section (IV).

The first presented chaotic model is the Rössler attractor. It is a system of three differential equations presented in eq. (1 - 3), that solved create a strange attractor visualized in fig. (2).

$$\frac{dx}{dt} = -(x + z) \quad (1)$$

$$\frac{dy}{dt} = x + ay \quad (2)$$

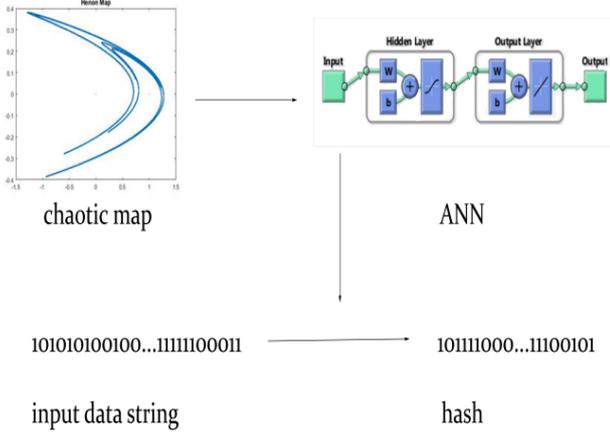


Fig. 1: The model of the proposed hashing method

$$\frac{dz}{dt} = b + xz - cz \quad (3)$$

where a , b and c are real parameters. The dynamics exhibit chaotic behavior when $c = 5.7$ and a, b are fixed at $a = 0.2$, and $b = 0.2$, [20].

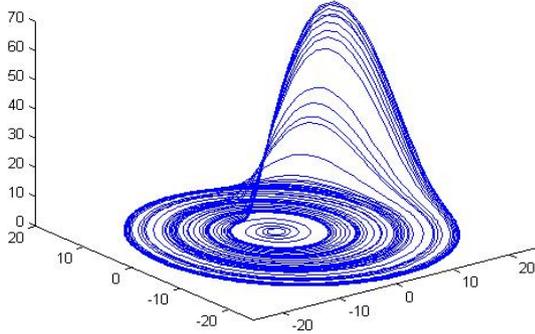


Fig. 2: The behaviour of the Rossler model presented in [18]

Second chaotic system used for ANN training was Chaotic Lorenz System presented in eq. (4 - 6) and visualized in fig. (3).

$$\frac{dx}{dt} = -a(y - x) \quad (4)$$

$$\frac{dy}{dt} = -cx - y - xz \quad (5)$$

$$\frac{dz}{dt} = -xy - bz \quad (6)$$

where a, b, c are also real parameters, and when $a = 10$, $b = 8/3$, $c = 28$, the system is chaotic.

Third system used was generalized Henon map, a two-dimensional map (see eq. (7 - 8)).

$$\frac{dx}{dt} = a + by - x^2 \quad (7)$$

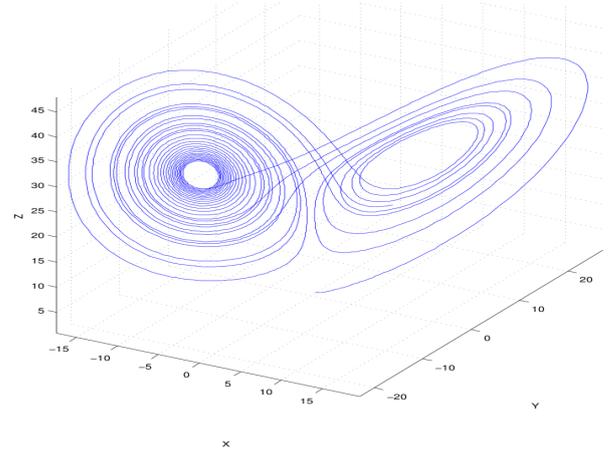


Fig. 3: The behaviour of the Lorenz model presented in [10]

$$\frac{dy}{dt} = x \quad (8)$$

which appears to have chaotic behaviour when $a = 1.4$, $b = 0.3$, see fig. (4).

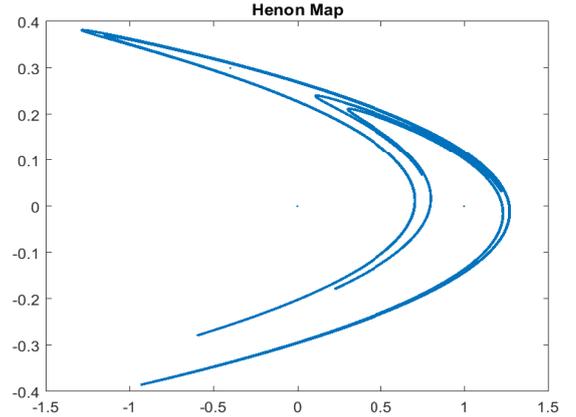


Fig. 4: The behaviour of the Henon model presented in [16]

IV. NUMERICAL TESTS

In this section, we will describe the process of Feed-Forward ANNs learning, testing, hashes creation and hashes evaluation in 5 steps algorithm. Two hash lengths were tested: 50 and 100 bits. Those values belong to the light cryptography family solutions.

STEP 1: Input data preparation. Input data structure is presented in eq. (9).

$$INPUT_{\{L,R,H\}}^{train}[i] = [b_1, b_2, \dots, b_n] \quad (9)$$

Where b_x is representing a single bit, i number of generated bit strings, n length of the hash and L, R, H is indicating the attractor used for target generation (Lorenz, Rossler, Henon). Those distinct bit strings

were considered as random messages which should be hashed in the future. Parameter n was equal to 50 or 100 because all attractors were tested for 50 and 100 bits hash. Parameter i was equal to 10000 in Lorenz and Rossler case (both, 50 and 100 bits hash length), 8276 in the Henon 50 bits hashes and 6831 for Henon 100 bits hashes. The values of parameter i were determined by the time of computation, accuracy of attractors (note, that in each case there were differences in attractors' input parameters), and effectiveness of ANNs after many empirical tests. As a result, six different inputs for all possible combinations of chosen attractors and chosen hash lengths were obtained.

STEP 2: Target data preparation. Models presented in eq. (1 - 8) were used for target generation. Lorenz and Rossler's equations were transformed into its' discrete form and Runge - Knutta Fourth Order method (RK4) was applied to solve them. Henon map was solved by the explicit formula:

$$x_{t+1} = 1 - 1.4 * x_t^2 + y_t \quad (10)$$

$$y_{t+1} = 0.3 * x_t, t = 0, \dots, 40000 \quad (11)$$

To connect random message (input binary string from $INPUT^{train}$) with a corresponding attractor, initial conditions were used. It was done in the same way for all attractors:

$$x_0 = [b_1, b_2, \dots, b_{\frac{n}{2}}] \Rightarrow [0, 1] \in \mathbb{R} \quad (12)$$

$$y_0 = [b_{\frac{n}{2} + 1}, \dots, b_n] \Rightarrow [0, 1] \in \mathbb{R} \quad (13)$$

Each message was divided into two substrings of the same length (half of the hash size). The first substring was mapped into a real number from range $[0, 1]$ and assigned to the x_0 . The second was mapped in the same way and assigned to the y_0 . In Lorenz and Rossler's case, z_0 was chosen randomly for each input. During solving the equations, the only parameters which were not constant were x_0, y_0 and (if applicable) z_0 , for all attractors. In all cases, 40 000 samples (in all dimensions, three in the case of Lorenz and Rossler attractors, two in the case of Henon map) were generated for further processing. Initial statistical tests (the same tests as described in STEP 4, but done on pure results after solving chosen equation) indicated that the highest potential for hash creation was in the second vector of solution (y) in all three attractors.

The final procedure can be represented in six stages:

Stage 1: Choose hash length (50 or 100) and an attractor (L, R, H).

Stage 2: Generate input data (random distinct bit strings). Input data is a matrix containing i binary words, each with a fixed length.

Stage 3: For $j = 1, \dots, i$ do Stages 4, 5, 6

Stage 4: Map message j into attractor initial values x_0, y_0 .

Stage 5: Solve equation with appropriate method (RK4, explicit), generate 40 000 samples of a solution. 40 000 is a value checked empirically, samples are far enough from each other which enables to create a hash.

Stage 6: Take samples from a chosen dimension (in all cases it was vector y).

The target after generation of samples is presented in eq. (14).

$$TARGET_{\{L,R,H\}}^{input}[i] = [s_1, s_2, \dots, s_{40000}] \quad (14)$$

Only n samples were necessary to generate a hash of length n . To cover the whole space of solutions and avoid usage of neighboring samples in the final target generation, appropriate step k was calculated. The step k is determined by the number of samples and by the hash size n (see eq. (15)).

$$k = \frac{40000 - 1000}{n} \quad (15)$$

First 1000 generated samples were skipped in all cases (the distance between them were too small). The final targets for each attractor and each ANN are presented in eq. (16).

$$TARGET_{\{L,R,H\}}^{train}[i] = [s_{1*k}, s_{2*k}, \dots, s_{n*k}] \quad (16)$$

where i is denoting target bit string corresponding to i -th input string, and s_h is denoting a value returned by chosen attractor in time h .

STEP 3: ANNs used for hash generation. Classical Two-layered ANNs were considered, see fig. (5). Each artificial neural network had M sigmoid neurons in the hidden layer and n (hash size) linear output neurons in the output layer. Different numbers of neurons M were tested for the ANNs configurations 50-M-50-50 and 100-M-100-100. All networks' inputs, targets, and testing data were generated with the usage of the related chaotic model. The scaled conjugate gradient back-propagation learning method was used, [8] due to the large size of data sets. All ANNs were implemented in MATLAB.

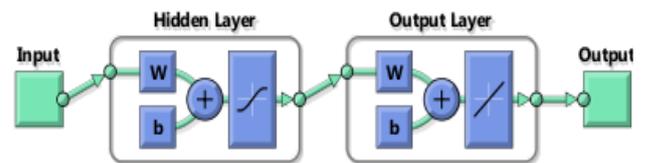


Fig. 5: ANN structure with M sigmoid neurons in the hidden layer and n (hash size) linear output neurons in the output layer

Results of statistical tests, number of neurons in hidden layers and expected values are presented in tab. (1) and discussed in steps 4 and 5 of the algorithm.

STEP 4: Statistical tests. After the learning phase the hashing procedure itself was tested. The $INPUT^{test}$ data was produced in the same way as $INPUT^{train}$, however, the size of $INPUT^{test}$ was

equal to 5000. It means that $INPUT^{test}$ contained 5000 random bit strings of length 100 or 50 bits for all three attractors.

ANNs producing 50 bits were tested for 15, 25, 50, 75 and 100 neurons in a hidden layer (M parameter). ANNs producing 100 bits hash were tested for 50, 75, 100, 125, 150, 175 and 200 neurons in a hidden layer. Four statistical tests on ANNs outputs were performed. The Z statistics (see eq. (17)):

$$|Z| = \left| \frac{AVG - \mu}{std} * \sqrt{p} \right|. \quad (17)$$

where AVG is the average value, μ is the expected value, std is the standard deviation and p is the number of elements in the considered data set. A significance level $\alpha = 5\%$ was used for deciding if the particular hashing model passed the hamming distance test or not. If $|Z| > 1.96$ the test was not passed. The same statistics and restrictions were used for the bits prediction test. In a series test statistics was calculated for each hash separately and differently than in hamming distance and bits prediction tests. To calculate Z, Wald-Wolfowitz formula was used. The series test was passed if less than 5% of all hashes failed it. The collision test was failed if there were at least one collision in the output hashes set.

- **Collision Test** was testing whether any of the chosen ANN produced at least once two same hash values. Not all ANNs (6 out of 36) were free from collisions (see tab. (1)). From among all configurations, collisions appeared only during producing 50 bits hash.

- **Series Test** designed by Wald Wolfowitz, also called runs test, was calculated for every hash independently. If the test was passed hash was produced randomly. For each ANN output set, a 5000 element vector containing Z statistics values was produced. Each Z statistic value is coding the fact if a chosen hash passed the test. If more than 5% of hashes failed the test, it can be assumed that hashes contain internal dependencies and particular ANN failed it.

- **Hamming Distance Test** was measuring the hamming distance between hashes and its' corresponding messages for each ANN. Hamming distance is calculated as a number of ones in vector given by formula: $INPUT^{test}[i] \text{ XOR } OUTPUT_o[i]$ (i -th output from ANN, potential hash). The result of the test for a particular ANN is a vector containing 5000 elements from range $[0, 50]$ or $[0, 100]$ (it is determined by the hash length). The expected value μ is equal to the half of the hash size. This value is taken from the characteristics of certified hashing methods like SHA-2 or SHA-3, [21].

- **Bit Prediction Test** was measuring whether a particular bit of output hash can be predicted, or not. The probability of '1' on a particular hash position was calculated according to the formula eq. (18).

$$P_j^o(1) = \frac{\sum_{i=1, \dots, 5000} OUTPUT_o[i][j]}{5000}, \quad (18)$$

$$j = 1, \dots, n,$$

where o denotes chosen ANN and n denotes hash length. The vectors of n elements indicating the probability of '1' on the particular position of the hash was examined. The expected value for this test is equal to 0.5 on each hash position.

The best results of each test are presented in fig. (6), fig. (7), and fig. (8). Parameters of ANNs are described in figures captions. Horizontal lines are representing expected values. In fig. (7) the expected value is equal to 25 (half of the hash size) and the closer particular hash (in Hamming Distance sense) is to this value, the better. In fig. (6) expected value is equal to 0.5. Values representing the probability of '1' on a particular hash position should be also as close to this value as possible. In fig. (8) expected value should be lower than 1.96. The more hashes were below the horizontal line ($Z = 1.96$), the better. Note, that even the best result in one particular test doesn't implicate the positive results in the rest of the tests. A comprehensive comparison is presented in the tab. (1).

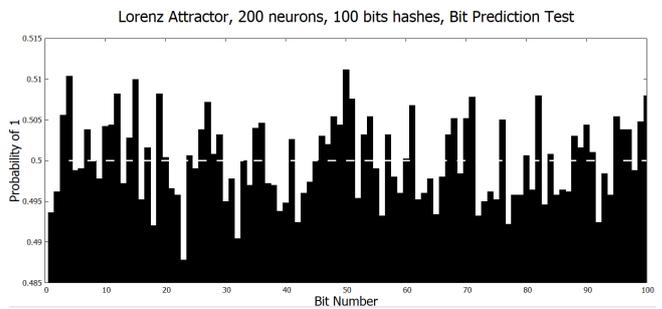


Fig. 6: The best result from Bits Prediction Test

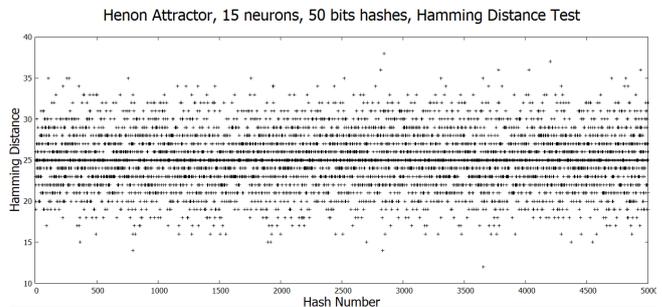


Fig. 7: The best result from Hamming Distance Test

STEP 5: Tests summary. The results of the tests are presented in the tab. (1). M denotes the number of neurons in the ANN hidden layer, B.P.T. is a Bits Prediction Test (expected value less than 1.96), H.D.T. is a Hamming Distance Test (expected value less than 1.96), S.T. is a Series Test (expected value less than 5%), and C.T. is a Collision Test (expected value equal to 0).

Results from the tab. (1) can be compared with results of the same tests done on three certificated hashing functions: SHA-1, SHA-512 and SHA3-512 presented in the tab. (2). Results from the tab. (2) are broadly discussed in the [21].

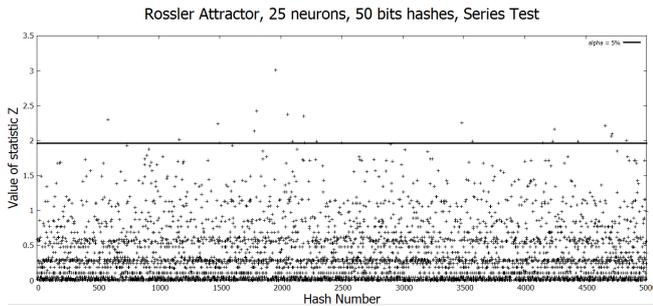


Fig. 8: The best result from Series Test

Usage of Lorenz attractor for creating 100 bits hash produced the best average results in bits prediction test. The worst was in case of generating 100 bits hash with the usage of Henon map, however, most of ANNs passed this test. It means, that bit position did not determine bit value in case of all three attractors.

Almost all ANNs failed the Hamming distance test. Statistically, the best average Z statistic values were obtained in 100 bits hashes generated by ANNs trained by Rossler attractor (3.95), and the worst by 50 bits hash generated by ANNs trained by the same attractor (average value of Z statistic equal to 9.97). Results indicate that generated hashes are either too similar to original messages or too chaotic.

ANNs producing 50 bit hash trained by Rossler attractor gained the best results in Series Test (average percent of failed tests equal to 0,82). The worst results were in the case of ANNs producing 100 bits hash also trained by Rossler attractor (65%). The rest of average values were relatively close to 5%, which indicates that in most cases there were no internal dependencies in hashes (hashes were generated randomly).

Collision Test was failed for all ANNs producing 50 bits hash with the usage of Rossler attractor, and for ANN producing 50 bit hash with the usage of Lorenz attractor with 15 neurons in a hidden layer. Results indicate, that Rossler attractor may demand more generated samples or different configurations of neural networks for smaller hashes production.

Two ANNs which passed all tests were: ANN trained by Henon attractor with 15 neurons in a hidden layer producing 50 bits hash, and ANN trained by Lorenz attractor with 125 neurons in a hidden layer producing 100 bits hash.

V. CONCLUSIONS

In this paper, we presented an intelligent ANN-based model of generation of light cryptographic hashes. The model was based on the usage of three different chaotic systems: Lorenz attractor, Rossler attractor, and Henon map as the base for simple two-layered ANNs learning. Results indicated that some configurations of attractors and ANNs' structures allow producing light hashes (all statistical tests were passed). There were also set of ANNs which didn't pass all tests but may be used for light hash production under restrictions described below.

TABLE 1: The results of ANN learning and testing procedure

M	B.P.T. (Z)	H.D.T. (Z)	S.T. (%)	C.T. (No)
Lorenz Attractor, $n = 50$				
15	2.00	2.93	7.48	3
25	1.63	10.66	6.24	0
50	0.53	4.82	5.72	0
75	0.38	7.21	6.04	0
100	1.94	3.27	5.90	0
Lorenz Attractor, $n = 100$				
50	0.08	8.80	5.10	0
75	0.59	4.48	5.52	0
100	0.87	5.14	4.82	0
125	1.64	0.22	4.8	0
150	0.34	3.39	5.48	0
175	0.13	6.29	5.00	0
200	0.06	6.20	4.78	0
Rossler Attractor, $n = 50$				
15	0.53	11.48	1.16	102
25	0.83	7.13	0.46	385
50	1.40	5.16	0.82	58
75	1.60	12.80	0.82	26
100	0.76	13.30	0.84	53
Rossler Attractor, $n = 100$				
50	2.18	7.02	42.38	0
75	1.58	2.06	52.98	0
100	0.73	4.45	74.02	0
125	0.92	2.97	78.26	0
150	0.44	3.55	75.78	0
175	0.30	3.27	73.04	0
200	2.55	4.33	65.36	0
Henon Attractor, $n = 50$				
15	0.10	0.21	4.06	0
25	1.96	13.67	5.20	0
50	0.70	2.97	6.24	0
75	1.03	1.05	5.06	0
100	0.90	2.37	5.66	0
Henon Attractor, $n = 100$				
50	1.87	6.72	5.46	0
75	1.45	9.72	4.92	0
100	0.93	3.01	5.16	0
125	1.51	7.69	5.48	0
150	0.33	3.69	4.92	0
175	1.51	6.73	5.20	0
200	1.77	5.05	4.94	0

TABLE 2: The results of statistical test for certificated hashing functions

SHA	B.P.T. (Z)	H.D.T. (Z)	S.T. (%)	C.T. (No)
1	1.04	1.17	4.80	0
512	0.86	0.15	4.98	0
3-512	0.32	0.44	4.17	0

Results indicate that:

1. Not all combinations were strong enough against statistical tests. ANNs which failed the collision test

shouldn't be used for hashing purposes.

2. Two combinations out of 36 passed all tests and can be considered as potential light hashing functions.
3. The same statistical tests were performed on certificated hashing functions which enabled to obtain expected values and expected results.
4. Rossler attractor should not be used for hash generation in the proposed configuration. Hashes produced with the usage of this attractor were full of collisions in a shorter version (50 bits) and full of internal dependencies in a longer version (100 bits).
5. ANNs trained with the usage of Lorenz attractor or with the usage of the Henon map, which failed the hamming distance test but passed the rest of tests, may be useful under some assumptions. Hash produced by them is truly random but is also too chaotic or too close (in hamming distance sense) to the original message. Both situations may result in collisions in the future, however, the probability of such situation may be acceptable if hashes won't be stored for a long time. It means that those ANNs may be used for creating short term hashes (e.g. fingerprints of tasks for checking data integrity in the cloud) but are not appropriate for long term hashes (e.g. storing passwords in databases).

Further investigations may involve: usage of different dimensions of solutions produced by the same attractors (or combinations of solutions from different dimensions), usage of more sophisticated ANNs types or structures which may result in increased computation time but also may be beneficial in chaotic behavior mapping, usage of more sophisticated chaotic structures, such as Triangular Chaotic map (TCM) [13].

REFERENCES

- [1] SECURE HASH STANDARD . <https://csrc.nist.gov>.
- [2] SHA-3 Standard . <https://nvlpubs.nist.gov>.
- [3] A. Akhavan Masoumi, A. Samsudin, and A. Akhshani. A novel parallel hash function based on a 3d chaotic map. *EURASIP Journal on Advances in Signal Processing*, 2013:126, 12 2013.
- [4] M. Cococcioni. Mackey-glass time series generator. <https://www.mathworks.com/matlabcentral/fileexchange/24390-mackey-glass-time-series-generator>. Accessed: 04.2019.
- [5] Z. Fan, T. Xiao-Jian, L. Xue-Yan, and W. Bin. Hash function based on the generalized henon map. *Chinese Physics B*, 17:1685, 05 2008.
- [6] Z. Gong. Survey on lightweight hash functions. *Journal of Cryptologic Research*, 3(1):1 – 11, 2016.
- [7] R. Guesmi, M. Ben Farah, A. Kachouri, and M. Samet. Hash key-based image encryption using crossover operator and chaos. *Multimedia Tools and Applications*, pages 1–17, 02 2015.
- [8] S. Haykin. *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Prentice-Hall, Inc., USA, 2007.
- [9] K. Ibrahim, R. Jamal, and F. Hasan. Chaotic behaviour of the rossler model and its analysis by using bifurcations of limit cycles and chaotic attractors. *Journal of Physics: Conference Series*, 1003:012099, 05 2018.
- [10] M. Igor. Lorenz attractor plot. <https://de.mathworks.com/matlabcentral/fileexchange/30066-lorenz-attaractor-plot>. Accessed: 02.2020.
- [11] M. Kidon and R. Dobai. Evolutionary design of hash functions for ip address hashing using genetic programming. In *2017 IEEE Congress on Evolutionary Computation (CEC)*, pages 1720–1727, June 2017.
- [12] H. Liu, A. Kadir, and J. Liu. Keyed hash function using hyper chaotic system with time-varying parameters perturbation. *IEEE Access*, 7:37211–37219, 2019.

- [13] M. Maqableh. A novel triangular chaotic map (tcm) with full intensive chaotic population based on logistic map. *Journal of Software Engineering and Applications*, 8:635–659, 12 2015.
- [14] A. Marco, A. Martinez, and O. Bruno. Fast, parallel and secure cryptography algorithm using lorenz's attractor. *International Journal of Modern Physics C - IJMPC*, 21, 01 2012.
- [15] H. Medini, M. Sheikh, D. Murthy, S. Sathyanarayana, and G. Patra. Identical chaotic synchronization for hash generation. *ACCENTS Transactions on Information Security*, 2:16–21, 12 2016.
- [16] L. Moysis. The henon map. <https://www.mathworks.com/matlabcentral/fileexchange/46600-the-henon-map>. Accessed: 02.2020.
- [17] J. Peng, S. Jin, H. Liu, and W. Zhang. A novel hash function based on hyperchaotic lorenz system. *Fuzzy Inform. Eng*, 2:1529–1536, 01 2009.
- [18] U. Prajapati. The rossler attractor, chaotic simulation. <https://www.mathworks.com/matlabcentral/fileexchange/56600-the-rossler-attractor-chaotic-simulation>. Accessed: 02.2020.
- [19] M. Singh and D. Garg. Choosing best hashing strategies and hash functions. In *2009 IEEE International Advance Computing Conference*, pages 50–55, March 2009.
- [20] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, 2000.
- [21] J. Tchórzewski and A. Jakóbiak. Theoretical and experimental analysis of cryptographic hash functions. *Journal of Telecommunications and Information Technology*, 1/2019.
- [22] J. Tchórzewski, A. Jakóbiak, and D. Grzonka. Towards ann-based scalable hashing algorithm for secure task processing in computational clouds. In *33rd European Conference on Modelling and Simulation*, pages 421–427, 2019.
- [23] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):769–790, April 2018.

AUTHOR BIOGRAPHIES

JACEK TCHÓRZEWSKI received his B.Sc. and M.Sc. degrees with distinctions in Computer Science at the Cracow University of Technology, Poland, in 2016 and 2017 respectively. Currently he is a Research and Teaching Assistant at the Cracow University of Technology and Ph.D. student at the AGH Cracow University of Science and Technology. His e-mail address is: jacek.tchorzewski@onet.pl



AGNIESZKA JAKÓBIK (KROK)

received her M.Sc. in the field of Stochastic Processes at the Jagiellonian University, Poland and a Ph.D. 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. Her e-mail address is: ajakobik@pk.edu.pl



Towards a multiparadigm approach to model energy management in WSN for IoT based edge computing applications

Lucilla De Arcangelis
Dipartimento di Ingegneria Industriale e
dell'Informazione
Università degli Studi della Campania
"L. Vanvitelli"
via Roma 29
81030, Aversa, Italy

Mauro Iacono
Eugenio Lippiello
Dipartimento di Matematica e Fisica
Università degli Studi della Campania
"L. Vanvitelli"
viale Lincoln 5
81100, Caserta, Italy

KEYWORDS

Performance evaluation; energy management; wireless sensor networks; simulation; ns-3; Internet of Things; edge computing

ABSTRACT

Energy management in Wireless Sensor Networks (WSN) is a vastly analyzed, yet still open issue in the scientific literature. Managing energy is of paramount importance when sensors are battery-powered and distributed in large, hardly (or expensively) accessible sites, such as a forest, for environmental or safety monitoring, or a stretch of sea, to counter contraband or drug traffic or other illegal activities, or dangerous scenarios, such as extended fires or areas on which toxic or dangerous chemical agents are insisting as the result of an incident or an attack. As an extended WSN presents a significant complexity in terms of the number, type, heterogeneity and workload of nodes, communications, interactions with the environment and information propagation and management. To deal with this issue, we present here a preliminary feasibility study for a parameterizable modeling framework aiming to investigate energy balancing and management in WSN. To cope with complexity and with the diversity of the problems to be analyzed in the small and in the global scale of a WSN system, and to provide a tool to evaluate and take into account the actual technological stacks available and arbitrary software workloads, we chose a customizable multiparadigm approach, that has been designed within the research agenda of VALERE research project ePassion.

I. INTRODUCTION

The WSN technology has become a commodity and has been adopted in uncountable different contexts to easily deploy monitoring devices over existing scenarios. WSN are used in small scale environments, both indoor spaces like apartments, warehouses, museums or industrial shelters and outdoor spaces like campuses or parking lots, and large scale environments, such as a forest, a stretch of sea, farmland, a forest, a city.

Applications span from surveillance, security, smartification of existing facilities, emergency management, Industry 4.0 solutions, smart cities, military support, environmental protection, with a continuous proliferation due to the increasing availability of more and more affordable WSN nodes and sensors, increasing in both capabilities and computing power.

In some scenarios, WSN nodes have to be battery-powered, e.g. if installed in locations that do not allow wiring for historical, practical or cost reasons, or because they must be relocatable, self-moving or just deployed by throwing them in a scenario that is not otherwise practically reachable. In some cases, it is suitable and convenient to adopt technologies that can recharge batteries by harvesting energy from the environment, e.g. by solar panels, tidal waves, mechanical vibrations. However, even when harvesting is possible, a proper and careful design is needed to ensure the correct engineering of both the nodes and the network as deployed. Indeed, to accomplish the mission, a WSN needs to keep undamaged both the environmental coverage and the connection between all nodes: this requires the definition of a proper number of nodes and a location strategy to ensure survivability, also by facing node loss using reconfigurations that, in turn, need tools to model and evaluate with sufficient accuracy energy consumption on nodes as deployed in the scenario.

When dealing with large scale heterogeneous WSN, complexity is significantly increased for various reasons, such as the number of nodes, scenario dimensions, increased complexity in scheduling and routing, diversity of node workloads and battery level dynamics.

In this paper we sketch a modeling strategy designed to evaluate and manage energy usage and balancing in large WSN applications. This modeling framework is still in its preliminary development. This notwithstanding, here we discuss the approach as a case of multiparadigm modeling, with special attention to the heterogeneity of the involved modeling techniques, the different modeling planes, the parameterization of the adoptable modeling techniques and the advantages of

multiparadigm approaches for dealing with problems characterized by different scales and detail levels.

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 an instantiation of the methodology with specific modeling formalisms; Section VI points at some relevant literature; conclusions close the paper.

II. ANALYSIS OF WSN CHARACTERISTICS AND REQUIREMENTS

WSN nodes can be considered embedded computers equipped with a radio connection, used to support networking, and a number of different sensors, whose type and characteristics depend on the application. Nodes may be battery operated or connected to the common electrical grid: in the following, we only focus on battery-operated sensors. The software stack running on the nodes can consist of a proprietary solution or open software, either derived from other segments (e.g. Linux distributions or analogous projects) or natively design for WSN. Network support is generally based on standard general-purpose (e.g. based on IPv4 or IPv6 and the TCP/IP protocol suite) or specialized (e.g. ZigBee) network protocols, to ensure interoperability. Costs are one of the main drivers in the design of commodity WSN solutions: this notwithstanding, a node may have enough resources to be able to run applications locally or as part of a distributed environment, also including edge computing solutions.

The nature of the problem, the equipment and the Hardware/Software (HW/SW) architecture are not the only key elements to understand WSN: the other relevant factor is the management of the network and communications. A shared practice is adopting infrastructure-less solutions, that allow WSN to be deployed and self-sufficient by dedicated routing and management techniques (e.g. mesh, ad-hoc networks, MANET for mobile WSN), also providing reconfiguration features. This approach also inherently allows the extension of a WSN to grow up to hundred or thousand of nodes, in principle, introducing scaling problems for the characterization of the WSN as a whole, due to propagation and lack of centralization of management and state of the network.

When dealing with battery operated nodes, both node activities and network dynamics impact on energy consumption.

Modeling energy in WSN nodes specifically targets the general and local optimization of the available energy level to reorganize, periodically or dynamically, the workloads in terms of local computing, the delegation of tasks (if possible) and preferred routing paths, acting simultaneously and synergistically on two aspects: energy supply and energy consumption. Therefore, the modeling approach is intended to support the development of such systems, both guiding the design or the choice of the electronics and the integration of a node and the identification of optimal operating param-

eters, and analyzing the effects of active use of single nodes and the whole network. The aim is achieving the maximization of the energy available at the sensor nodes and the ratio between service quality and energy consumption, with reference to a specific application scenario. The complexity of the system and the number of possible scenarios and applications suggest the use of multiple coordinated modeling approaches to cope both with the network aspects, by abstract techniques, and the node features with a high level of detail. This allows to minimize the complexity and the analysis time and to keep a high degree of fidelity when dealing with the actual behavior of the used technological stack, and, finally, helps validation. While both analytical and numeric evaluation techniques are precious to model the system on a small scale, soft computing or approximated techniques may help in coping with global analysis while keeping the complexity of the evaluation manageable. For this reason, we chose a multiparadigm approach as evaluating framework, to leverage the benefits of very different modeling techniques inside the same model for different levels of the system.

Modeling a node requires taking into account: a model of the actual hardware configuration with the actual activation scheme of each sensor; a model of the software environment in terms of background activity of the operating system and the activation/sleeping state of the node; a model of the tasks that run on the node and its schedule, that might be non-deterministic; a model of the activities that the node performs on the network according to the generated traffic and the routed traffic depending to its role in the network; and, finally, a model of the harvesting process, if any, depending on the conditions of a specific node in the network and the environment. Using an analytical modeling technique (e.g. a stochastic technique based on Markov chains or Petri nets or queuing networks) provides a good first approximation model but cannot capture all details that would allow a realistic validation. Simulation offers specialized tools that actually mimic the technological stack faithfully, but only allows a limited number of possible configurations and behaviors of the system to be actually simulated to keep feasibility, and has a high level of time and memory complexity if the system has a very high scale.

At the network level, global characteristics of the system are strongly dependent on the specific topology: the topology, even neglecting the technological aspects, affects routing decision and more abstract characteristics of the network, such as robustness and resilience. These two characteristics are an important target, because they are connected to keeping the coverage of the WSN or to a smart degradation of the coverage when nodes start exhausting their batteries. They may be evaluated on an abstract level by means of advanced mathematical descriptions, e.g. borrowed from the statistical physics corpus of knowledge, from the artificial intelligence domain, from the soft computing field. Of course, this abstraction step must be intended as a way

to provide a guide for parameter space exploration and high-level dynamics evaluation that must then be validated, e.g. by properly simulating the real scenario.

III. A PROPOSAL OF MODELING METHODOLOGY

The proposed methodology aims at joining the advantages of high-level models and approaches with the realism that may be obtained by using a specific simulation platform, capable of capturing actual events at the desired detail level. The methodology is based on considering the WSN on three different levels: the global level, the simulation level and the node environment level. At the global level, optimization is performed considering a synthetic representation of nodes or node groups, to abstract details and to allow the use of different methods; at the simulation level, a faithful model of nodes and network mechanisms is exploited by means of a specialized, extensible, modular network simulator, to evaluate the actual costs of communications, network organization and configurations in terms of energy, both to obtain synthetic parameters for the global level and to validate the results of the optimization; at the node environment level, a faithful description of the HW/SW configuration of each node is integrated by a model of the harvesting mechanism and a model of the applications to be executed on the node and of their scheduling and reactive behavior.

The methodology leverages a multiparadigm approach [25] to allow maximum flexibility in the choice of the optimization strategy. This allows defining an interface representation of parameters, topology and configuration to integrate the tools used at the global level and the simulation level.

The modeling and evaluation process is articulated in 8 phases, as in Fig. 1. The starting phase (*Requirements elicitation*) is meant to understand the problem the WSN should solve, the general characteristics of the environment, the constraints on available hardware and possible topology. This leads to the *Definition of node architecture*, that includes hardware, configuration, devices and harvesting model, and provides as output the node simulation model. The phase of *Definition of the scenario* allows defining the global information needed to produce the optimization model and the contexts for the generation of the testbeds. The phase of *Definition of exploration testbed* produces a local simulation testbed around single nodes, typical because of their configuration or their position or role in the WSN, or node groups, that is used to generate parameters for the global level model in the *Estimation by exploration testbed* phase. In the *Optimization* phase, the global level model is evaluated, to generate the optimal configuration for the WSN, that is used to generate the simulation model evaluated in the *Validation* phase. If the simulation of the configuration of the optimal configuration confirms the results of the Optimization phase, the overall final configuration for the system is generated; otherwise, in the *Parameter refinement* phase a new set of parameters is generated, and two possible it-

erations are taken until the Validation phase produces a satisfactory result.

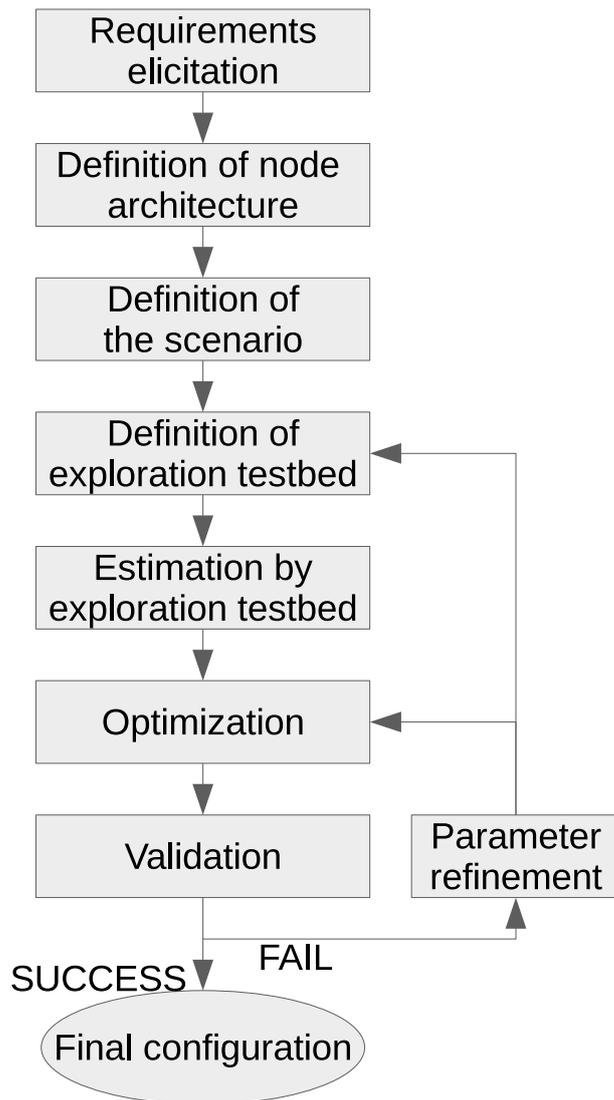


Fig. 1. The simulation methodology

IV. AN INSTANTIATION OF THE METHODOLOGY

To provide an application of the methodology, we defined a framework suitable for the analysis of energy management in large WSN. In this case, the global level is implemented by a modeling and evaluation technique borrowed from statistical physics, namely *percolation theory*; the simulation level is implemented by leveraging a well established network simulator, namely *ns-3*; the node environment level is implemented by a *multiformalism* modeling approach [10] that exploits *Fluid Stochastic Petri Nets* (FSPN), Ordinary Differential Equations (ODE) and again *ns-3*.

A. The global layer

This level of implementation is focused on the topology of the WSN. More precisely, it assumes that a given number of sensors are distributed in space in fixed space positions. Each sensor dissipates energy to perform

different operations, data acquisition, elaboration and transfer and is supplied by a harvesting device. As a consequence, there is a finite probability that temporal periods exist in which one or more nodes are inactive. The aim is to find the topology of the links among the different nodes, which maximizes the probability that, at each time step, each node is connected to the external receiver sink. It requires as input information the average energy balance of a single sensor which is provided by the previous analysis for a typical scenario. More precisely, the central quantity is the average value of the dissipated rate for data transfer, as a function of the distance between transmitter and receiver, for data elaboration and data acquisition. At the same time, it depends on the average rate of energy accumulation.

It is quite evident that the larger the number of links incoming and/or outgoing from a single node, the larger is the energy dissipated by that node and, therefore, the larger is the probability for that node to become inactive. As a consequence, the optimization of the link topography corresponds to the minimization of the number of links. This is a typical problem of statistical physics, known as “percolation theory” [23], [1]. Within this framework indicating with p the probability that a link is present between two arbitrary nodes, there exists a critical value p_c such that for $p < p_c$ the network is not connected, i.e. information cannot be conveyed to the external sink. In the case of the WSN, the situation is more complicated for two reasons: i) since the cost for energy transfer depends on the distance between the connected nodes, links are not equivalent, but there are links that are less energy-consuming than others; ii) nodes are not always active, and this information must be included in the optimization procedure. Accordingly, the WSN must be treated as a weighted graph where each node is a vertex to which is assigned a random variable assuming value 1 or 0 if the node is active or inactive, respectively. The problem can be faced by means of advanced mathematical tools developed in the field of graph and information theory, focusing for instance on quantities like entropy, joint entropy, and mutual information. The entropy of a random variable is a function which attempts to characterize its intrinsic “unpredictability”. The Joint entropy is the entropy of a joint probability distribution or a multi-valued random variable, and the mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. A precise definition of these quantities can be found in the literature [7].

For our specific problem, a key role can be played by the betweenness centrality, a widely used measure in graph theory that captures the role of a node in allowing information to pass from one part of the network to the other [3], [18]. More precisely, to evaluate the betweenness of a node J , for all pairs of nodes, one must identify the shortest paths between those nodes; then one evaluates the fraction of those shortest paths that pass through node J , which provides the betweenness centrality for node J . A node with higher betweenness

centrality would have more control over the network because more information will pass through that node. Nodes with the largest betweenness do not automatically correspond to most connected nodes, and a possible optimization procedure corresponds to minimizing the energy consumption of node with large betweenness. This would allow us to stabilize those nodes which are more central in transferring information. The optimization procedure will be implemented by means of typical methods of statistical mechanics, such as Monte Carlo sampling [15] or simulated annealing [13].

B. The simulation level

For the simulation layer ns-3 (www.nsnam.org/) has been chosen, a discrete event simulator that is specialized for computer networks simulation. The choice is motivated by some of its key characteristics. As first, it is designed as open and to be extended according to a modular architecture, that allows the generation of additional components to support our methodology. As second, it is designed to faithfully reproduce all aspects of network protocols, management, hardware, routing, and other details, so that simulation can closely reproduce realistic situations for a given scenario and execution trace. Finally, it is widely used in the scientific community and by practitioners for WSN [5]. The logic of the simulation is based on the generation of events to be scheduled for execution in the temporal order in which they would happen in the real system [20]. ns-3 is supported by a vast community that develops and validates new features and components, and existing components also support energy management. Finally, ns-3 also has a real-time scheduler to implement simulation-in-the-loop and can execute real workloads on the nodes, including virtual machines, thus allowing a very close tracking of the software layer of each node as well. As the authors declare that they aim at enabling the reuse of real-world protocols without specific reimplementations, the level of details possible is only limited by the need of longer wall-clock simulation execution times as much as the simulation is close to reality and the dimension of the simulated network grows. As ns-3 is very popular and widely accepted, we suggest interested readers a systematic literature review of papers dealing with the use of ns-3 in WSN in [5] for further details.

C. The node environment level

Modeling the node environment allows the determination of average node behavior for typical nodes to obtain the initial parameters. As in the considered node model, each node executes software tasks when needed besides normal activities related to data sensing and processing and communications to implement an edge computing platform. A stochastic oriented modeling approach is consequently needed. In addition, harvesting has external influencing factors depending on the environment and exhibits a continuous behavior of battery energy level, that varies for the two opposing factors (usage and harvesting), and differently accord-

ing to the power state of each node (idle, sensing, processing, high performances). In this isolated model, the contribution of interactions between nodes, as well as routing problems and connected influences on energy levels and contributions of software tasks that are launched as a reaction to events detected on the field in the scenario are not known, but can be taken into account on a probabilistic basis as a first approximation.

Due to these requirements, the chosen modeling technique is a multiformalism approach based on FSPN and ODE, that are used together to evaluate node behavior and parameters and to partially generate the Node simulation model. FSPN are a variant of stochastic Petri nets in which additional places, transitions and arcs exist that deal with continuous marking (i.e., a continuous place is not marked by an integer number of tokens, but by a continuous level, and incoming and outgoing arcs influence its marking by a rate when the enabling conditions hold). They are thus suitable to easily represent energy levels and the power drain due to node activities that happen for a given duration but are activated with a discrete event logic described by the ordinary (non-fluid) elements of the formalism (e.g. a scheduling scheme, reactions to events, guards or activities with stochastic duration): for formalization and details on modeling and analysis with FSPN, interested readers may refer to [11], for an example of application to [8]. ODE well describe the mathematical model of harvesting-related phenomena and are easily comparable with the fluid part of FSPN, so that they can be modeled by a domain expert that is not familiar with the FSPN. However, the reduced semantic distance between the two modeling formalisms may be easily bridged within multiformalism frameworks, like SIMTHESys [2].

V. MODELING AND EVALUATION PROCESS

The modeling and evaluation process is depicted in Fig. 2. The inputs for the process, generated in the Requirements elicitation phase and Definition of node architecture phase, are the Scenario model, the Harvesting analytical model, the Node HW configuration and the Node SW configuration.

The *Scenario model* is an abstract description of the geometry, the characteristics and the configuration of the scenario in which the WSN energy management evaluation has to be performed. This model considers the requirements of the application to be implemented and is used to generate¹ the *Global analytical optimization model*, that is the base of the Global level in this implementation, and the information needed to define the *Parameters exploration testbed*, namely the *Scenario characterization*.

The *Harvesting analytical model* is provided in terms

¹In the prototype implementation for the experimental campaign that will be carried on in the next phase of the ePassion project, the generation is performed manually. However, a transformation based automatic generation is possible leveraging the semantic description of the interface: the same holds for other parts of the process, but this is out of the scope of this paper.

of a parametric ODE description of the harvesting model of a node. It is used to generate the *Harvesting numerical model* as a component of the ns-3 Node simulation model, and, together with the Node HW configuration and Node SW configuration, the *Node environment model*, in turn used to instantiate parameters in the Node simulation model and the Parameters exploration testbed.

The *Node HW configuration* and the *Node SW configuration* are used to generate the Node configuration, that in turn is used to generate the ns-3 Node simulation model.

The *Node simulation model* is used to produce the *Node instances* that are used to compose both the Parameters exploration testbed and the Global WSN model, to be used during the Estimation by exploration testbed phase and the Validation phase, respectively.

Finally, the *Global WSN model*, as configured after the execution of the Global analytical optimization model, is used to simulate the WSN by ns-3 in detail, and the results are compared with the results of the Global analytical optimization model results to decide if an iteration is needed or if the final result is satisfactory.

VI. RELATED WORK

Energy management in WSN has been widely discussed in the literature, focusing the attention on one or more aspects of the problem. For an introduction and a general description of relevant issues, we suggest [29], [16] and [4], that also introduce the main elements needed to define a modeling framework. Modeling approaches can be found in [17] and [27]. Power management methods are examined in [12], [19], while energy saving in protocols and network management is surveyed in [28] and [24]. A survey about energy harvesting and energy management is provided in [26]. Different energy optimization techniques have been applied to WSN, including neural networks [21][14], reinforcement learning [22], game theory [6] and genetic algorithms [9].

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented an experimental modeling methodology for the evaluation and the optimization of energy use in WSN, that may scale to extensive configurations and allows realistic verification by simulation. Future works include the application to real-world case studies and the implementation of automatic mechanisms, where possible, for the evaluation process.

VIII. 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”.

REFERENCES

- [1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

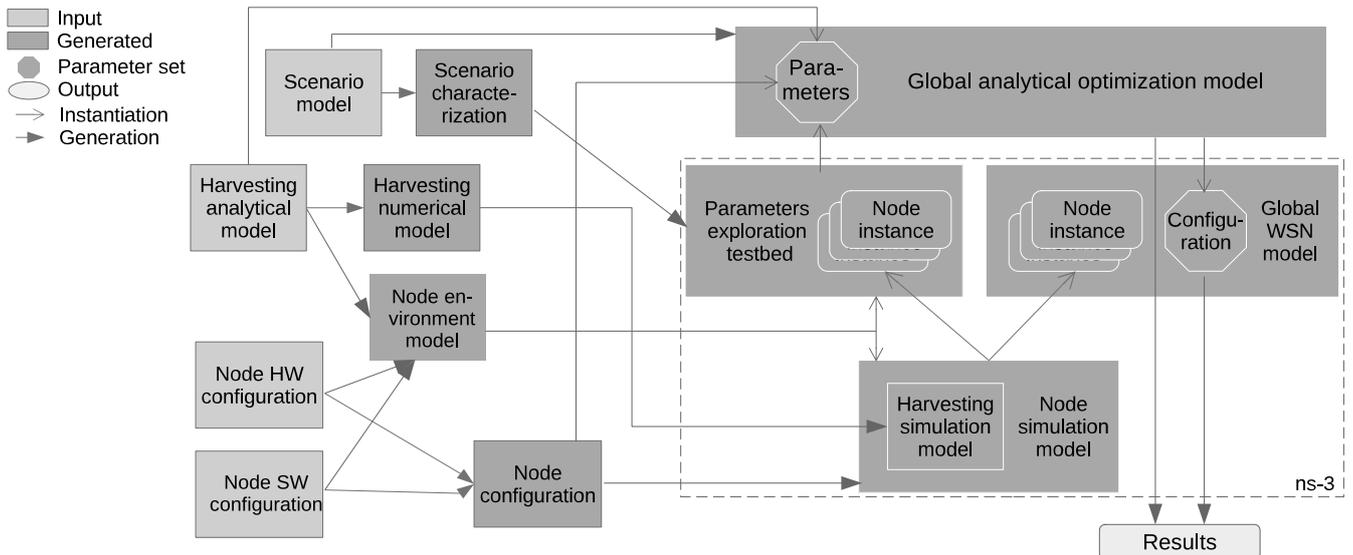


Fig. 2. The modeling and evaluation process

- [2] E. Barbierato, M. Gribaudo, and M. Iacono. Modeling hybrid systems in SIMTHESys. *Electronic Notes in Theoretical Computer Science*, 327:5 – 25, 2016.
- [3] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.
- [4] E. Bitar, E. Baeyens, and K. Poolla. *Energy management in wireless sensor networks*. 2012.
- [5] L. Campanile, M. Gribaudo, M. Iacono, F. Marulli, and M. Mastroianni. Computer network simulation with ns-3: A systematic literature review. *Electronics*, 9(2):272, Feb 2020.
- [6] 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.
- [7] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Ed., 1991.
- [8] 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.
- [9] K. Ferentinos and T. Tsiligiridis. Adaptive design optimization of wireless sensor networks using genetic algorithms. *Computer Networks*, 51(4):1031–1051, 2007.
- [10] M. Gribaudo and M. Iacono. An introduction to multi-formalism modeling. In M. Gribaudo and M. Iacono, editors, *Theory and Application of Multi-Formalism Modeling*, pages 1–16. IGI Global, Hershey, 2014.
- [11] M. Gribaudo, M. Sereno, A. Horváth, and A. Bobbio. Fluid Stochastic Petri Nets augmented with flush-out arcs: Modelling and analysis. *Discrete Event Dynamic Systems: Theory and Applications*, 11(1-2):97–117, 2001.
- [12] 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.
- [13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [14] U. Kulkarni, D. Kulkarni, and H. Kenchannavar. Neural network based energy conservation for wireless sensor network. pages 1312–1316, 2018.
- [15] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- [16] R. Mini and A. Loureiro. Energy-efficient design of wireless sensor networks based on finite energy budget. *Computer Communications*, 35(14):1736–1748, 2012.
- [17] P. Murali, A. Challa, M. Kasyap, and C. Hota. A generalized energy consumption model for wireless sensor networks. pages 210–213, 2010.
- [18] M. Piraveenan, M. Prokopenko, and L. Hossain. Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PLOS ONE*, 8(1):1–14, 01 2013.
- [19] E. Popovici, M. Magno, and S. Marinkovic. Power management techniques for wireless sensor networks: A review. pages 194–198, 2013.
- [20] G. F. Riley and T. R. Henderson. *The ns-3 Network Simulator*, pages 15–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [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] D. Stauffer and A. Aharony. *Introduction to percolation theory*. Taylor and Francis ed., 1994.
- [24] 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.
- [25] H. Vangheluwe, J. Lara, and P. Mosterman. An introduction to multi-paradigm modelling and simulation. *Proceedings of the AIS’2002 Conference*, 01 2002.
- [26] Z. Wan, Y. Tan, and C. Yuen. Review on energy harvesting and energy management for sustainable wireless sensor networks. pages 362–367, 2011.
- [27] Q. Wang and W. Yang. Energy consumption model for power management in wireless sensor networks. pages 142–151, 2007.
- [28] 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.
- [29] 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

LUCILLA DE ARCANGELIS



is Full Professor at Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi della Campania "L. Vanvitelli", Italy. She was visiting scientist at the University of Cologne and the CEA in Saclay. In 1990 she was awarded a CNRS (CR1) position at the ESPCI in Paris and in 1993 a Faculty position in Italy. Her research interests span from percolation, fractals, cellular automata to spin glass, models for fracture and gelation. Recently, she has focused her research on statistical properties of earthquake and solar flare occurrence and on the critical features of spontaneous neuronal activity. She is Associate Editor of JSTAT, Physical Review E, Physica A and Frontiers in Physiology. Since 2018 she is secretary of the C3 IUPAP Commission. Her email is lucilla.dearcangelis@unicampania.it.

Her research interests span from percolation, fractals, cellular automata to spin glass, models for fracture and gelation. Recently, she has focused her research on statistical properties of earthquake and solar flare occurrence and on the critical features of spontaneous neuronal activity. She is Associate Editor of JSTAT, Physical Review E, Physica A and Frontiers in Physiology. Since 2018 she is secretary of the C3 IUPAP Commission. Her email is lucilla.dearcangelis@unicampania.it.

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>.

EUGENIO LIPPIELLO



is Associate Professor at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy. He received the Ph.D. in Physics from the University of Salerno. His research interests are in the field of non-equilibrium statistical mechanics, including, phase ordering and random magnet dynamics,

granular systems, statistical seismology and seismic hazard of natural phenomena. His email is eugenio.lippiello@unicampania.it.

3D-stacked memory for shared-memory multithreaded workloads

Sourav Bhattacharya Horacio González-Vélez
Cloud Competency Centre, National College of Ireland

Sourav.Bhattacharya@gmail.com, horacio@ncirl.ie

KEYWORDS

3D-stacked memory; memory latency; computer architecture; parallel computing; benchmarking; HPC

ABSTRACT

This paper aims to address the issue of CPU-memory intercommunication latency with the help of 3D stacked memory. We propose a 3D-stacked memory configuration, where a DRAM module is mounted on top of the CPU to reduce latency. We have used a comprehensive simulation environment to assure both fabrication feasibility and energy efficiency of the proposed 3D stacked memory modules. We have evaluated our proposed architecture by running PARSEC 2.1, a benchmark suite for shared-memory multithreaded workloads. The results demonstrate an average of 40% improvement over conventional DDR3/4 memory architectures.

INTRODUCTION

High Performance Computing (HPC) systems typically use custom-built components such as enterprise-grade GPUs, millions of custom CPUs, and above all super-fast interconnection mechanisms for networking and memory. But, despite using the most cutting-edge materials and components available today, it has become clear that one clear limitation relates to the memory-wall challenge, i.e. the imbalance between memory and core/processor performance and bandwidths which has a cascading effect on the overall performance [12], [13]. In fact, memory latency has been identified as a leading cause of overall performance degradation. The fastest memory available today that is used in high performance systems is Error Correcting Code Double Data Rate Synchronous Dynamic Random-Access Memory, also known as *ECC DDR SDRAM*. DDR4 SDRAM, is an abbreviation for Double Data Rate Fourth-Generation Synchronous Dynamic Random Access Memory. It is a type of memory that has a high bandwidth (“double data rate”) interface with a latency under 10 nanoseconds [20]. When this latency is magnified by the number of distant nodes, the overall latency of the supercomputing platform increases. This is the scalability challenge where even the smallest of CPU-RAM latency per node becomes magnified thousands of times and thus results in lower performance [26].

In order to address the latency and scalability challenges in HPC, the solution must be one that works at a granular level. While there are multiple sources through which latency is introduced, we aim to address the most

relevant cause of lag: the CPU-memory intercommunication latency, with the help of 3D stacked memory.

Conventional DDR RAM has four primary timing indicators which are used to indicate the overall speed of the DRAM. They are as follows:

1. CAS Latency (tCL/tCAS): number of cycles taken to access columns of data, after getting column address.
2. RAS to CAS Delay (tRCD): It is the time taken between the activation of the cache line (RAS) and the column (CAS) where the data is stored.
3. Row Precharge Time (tRP): number of cycles taken to terminate the access to a row of data and open access to another row.
4. Row Active Time (tRAS): number of cycles taken to access rows of data, after getting row address.

These four parameters are cumulatively known as *memory timings*. The motivation behind our research is to study 3D stacked memory architectures which can have significantly faster memory timings than the conventional DDR-4 RAM used in HPC. The future for high-speed memory seems to favour 3D-stacked configurations, as demonstrated by the patent fillings from a number of memory manufacturers[18], [4].

Since the experimental fabrication of multiple new memory modules based on 3D stacked memory is not economically feasible for low yields [27], we will use a combination of simulators—namely DESTINY [25], [23] and CACTI-3DD [8]—in order to create a feasible model of 3D stacked memory. These simulators will help us design and architect the underlying architecture and identify the best possible configuration needed to achieve the highest bandwidth and lowest of latency while keeping in mind the temperature, power consumption, power leakage, area efficiency and other relevant parameters. After a suitable design and architecture sample, satisfying the energy efficiency criteria and other parameters is obtained, we will use it to simulate a full system benchmark using Gem5, a simulator that is a modular, discrete event driven platform for computer architecture, comprising of system-level architecture as well as processor micro-architecture [7]. The overall objective is to model a 3D-stacked memory subsystem, running on top of a generic X86 CPU, and then run a performance benchmark normally used in supercomputing environments to evaluate the performance gains the 3D stacked memory architecture provides when compared with a traditional DDR-4 memory based architecture. Figure 1 describes the high level the approach we have taken to model the 3D stacked memory architecture.

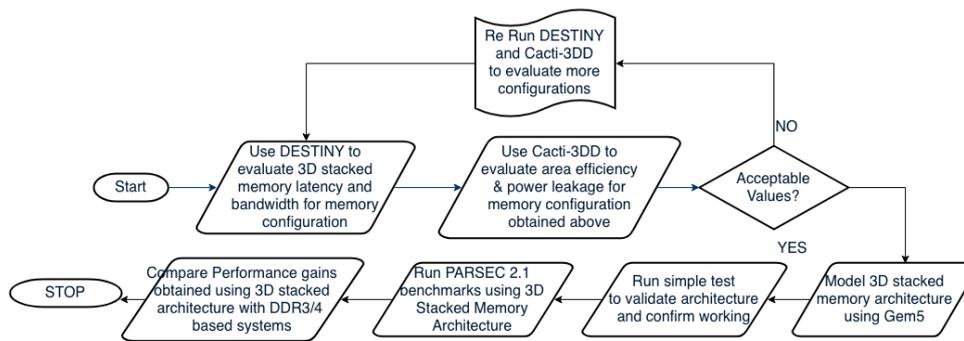


Fig. 1: High Level Approach to Modelling 3D Stacked Memory.

LITERATURE REVIEW

At the turn of this century, researchers focused on making improvements to DRAMs native performance. During any instruction execution, the CPU would have to obtain the next set of instructions from the DRAM. However, as DRAMs locations are off-chip, the apparent attempt was to address delays in accessing data off-chip. By adding components to a chip to increase the memory available in a module, the complexity involved in addressing memory increased significantly [9], [21]. Consequently, the design of a useful interface became more and more challenging. It became evident that communication overhead accounted for some 30% of the DRAM performance, and as a result, moving the DRAM closer to the CPU has become obvious [24].

The next logical step was to try enhancing cache performance, after having failed to solve the latency challenge by improving DRAM performance. The objective was to increase the overall effectiveness of cache memory attached to the processor. Some original ideas in improving cache performance included improving latency tolerance and reduction in cache misses.

Caching can help improve tolerance by doing aggressive prefetching. However, Simultaneous Multithreading or SMT [11] suggested that the processor could be utilised for other work while it was waiting for the next set of instructions. Other approaches included:

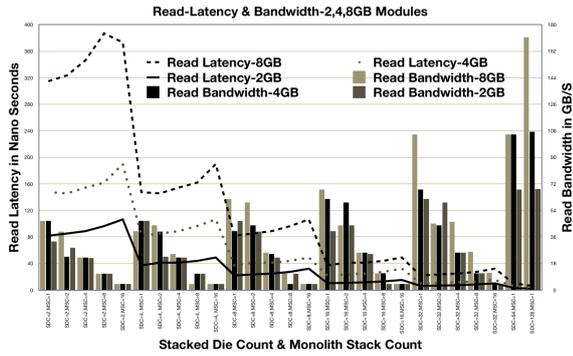
1. Write-buffering: When a write operation to a lower level of cache initiates, the level that generated the write operation no longer has to wait for the process to complete, and can move to the next action. This action can take place assuming there is enough buffer to enable write buffers. This action ensured that a write operation to the DRAM would not result in the latency of DRAM reference.
2. Compression of memory: A compressed memory hierarchy model proposed selectively compressing L2 cache and memory blocks if they could be reduced to half their original size [21]. The caveat with this model was that the overhead for compression and decompression must be less than the amount of time saved by the compression.
3. Critical-word first: Proposed by [17], in the event of a cache miss, the location in the memory that contains the missed word is fetched first. This allows for immediate resolutions to a cache miss. Once the missed word is retrieved and loaded into memory, the rest of the block is

fetched and loaded. This allows for faster execution as the critical word is loaded at the beginning.

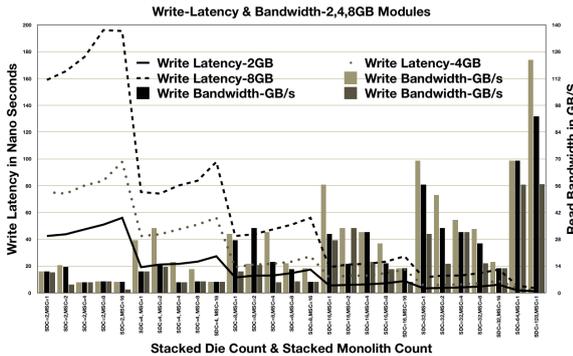
These proposals to improve latency tolerance and improve cache performance worked well when the performance gap between CPU-memory was not extensive. However, some strategies, such as the critical-word first approach did not help much in case the missed word was the first word of the block [1]. In that case, the entire block loading happened the same way it would be loaded typically, which was through contiguous allocation. Similarly, compressing memory did not avoid latency in case of single memory locations, which was the most crucial and significant factor that needed elimination. The only approach that made substantial gains was the write-buffering approach, which was useful if the speed gap between the CPU-memory was huge [10]. The caveat was the size of the buffer, that would keep increasing as the gap in speed between CPU-memory increased.

Similarly, trying to optimise the write performance had minimal use, as most instruction fetch operations are reads. So these approaches also did nothing to reduce the impact of latency. And again, tolerating latency and improving the cache performance becomes extremely difficult when the CPU-DRAM speed differential increases exponentially. After failing to improve the cache performance, researchers looked at reducing the cache misses. A cache miss is when the instruction to be executed is not in the L1 or L2 cache memory and must be retrieved from the main memory [10]. This involves communicating with the main memory, thus invoking the latency issue again. As is evident, by reducing a cache miss, one can bypass the latency tolerance conundrum altogether.

However, improvements in any form of associativity and methods such as memory compression will have much less influence in reducing latency as the cache size keeps growing. Also, managing caches through software will most likely provide the best benefit. Now the biggest challenge in managing cache on a processor using software is the cost involved in losing instructions when using a replacement strategy that is less effective. And as we can see, improving the cache performance is the fastest way to lower the CPU-memory gap in speed. Also, enhancing associativity in cache memory will also provide benefits when leveraged with multithreaded processors, as multithreading and multitasking can hide DRAM latency.



(a) Read-Latency & Bandwidth Comparison.



(b) Write-Latency & Bandwidth Comparison.

Fig. 2: Comparison of StackDie count vs Read-Latency & Bandwidth.

Contribution

To highlight the key elements mentioned above in an approach that will solve the memory wall problem, particularly at an exascale computing level, we are going to detail in the next section our approach to adding 3D stacked memory to a CPU. This architecture will improve the bandwidth, lower the latency while supporting parallelism, associative caching and scalable to the degree that may arguably benefit exascale computing clusters in supercomputers.

METHODOLOGY

In this section, we will describe in detail our approach to create a feasible, energy efficient, working prototype of 3D stacked memory using Destiny, Cacti-3DD, and Gem5.

Using DESTINY to evaluate 3D stacked memory latency and bandwidth

Created as a collaborative design space exploration tool, DESTINY can be used to model 2D and 3D SRAM, eDRAM (enhanced DRAM), STT-RAM (Spin-Transfer Torque Magnetic RAM), and ReRAM (Resistive RAM) [23]. It is a comprehensive modelling tool which can model 22nm to 180nm technology nodes. It can model both conventional and emerging memories, and the results of the modelling are validated against several commercial prototypes. In our case, six parameters have been employed to create a representative 3D-stacked memory architecture: StackedDieCount, PartitionGranularity, LocalTSVProjection, GlobalTSVProjection, TSVRe-

TABLE I: Output Parameters and their valid ranges for the memory size

Output Parameters	Valid Range
Timing - Read Latency	< 13 nano seconds
Timing - Write Latency	< 13 nano seconds
Timing - Refresh Latency	< 13 nano seconds
Bandwidth - Read Bandwidth	> 40 GB per second
Bandwidth - Write Bandwidth	> 30 GB per second

dundancy, and MonolithicStackCount, which are aligned with best practices [22]. We have run a parameter sweep ranging from 128 MB to 8192 MB, and each memory configuration running on varying data bus width, ranging from 32 bits to 1024 bits. For each run, the following output parameters were piped to an output file, with the acceptable values as per the valid range specified in Table I:

After compiling the results of the test runs and simulations, we observed that by varying the parameters: StackedDieCount and MonolithicStackCount, relevant 3D stacked memory configurations are visible. We then run the results through a bar plot to visualise the results and look for the optimum size memory capacity with low latency a high bandwidth capability, which will display the most appropriate quantity of the 3D stacked memory wafer.

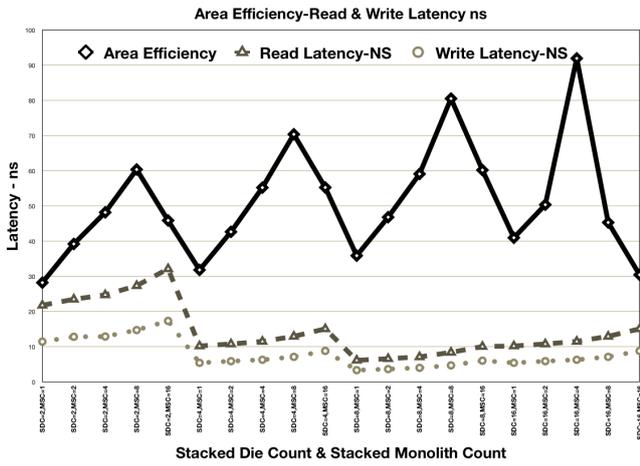
Based on Figures 2a and 2b, the 2GB configuration has been chosen as the prime candidate for the next set of tests with Cacti-3DD, which would help us to model the energy efficiency and power consumption aspects of the selected 3D stacked memory configuration.

Using CACTI-3DD to evaluate Area Efficiency, Power Consumption and Energy Leakage of 3D stacked memory

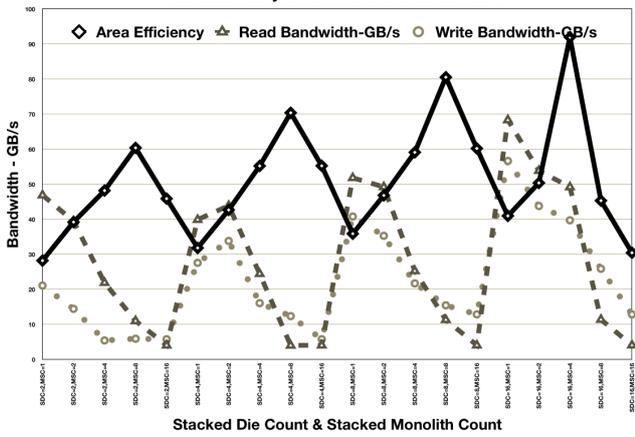
We have then used the 3D-stacked memory configuration obtained from DESTINY to evaluate the area efficiency, power consumption, and thermal efficiency of the 3D stacked memory module. Using a 2GB eDRAM wafer, we had StackedDieCount(SDC) of 1-16 in multiples of 2 and MonolithicStackCount(MSC): 1-16 in multiples of 2. MSC was increased in multiples of 2, from 1,2,4,8,16,32, and for each MSC count, SDC count was kept constant. For each configuration, Cacti-3DD was run to evaluate the area efficiency of the proposed configuration, along with power consumption, power leakage and thermal efficiency. In our case, area efficiency of < 100% and Power-Leakage < 1 Watt.

For the chosen 2GB memory module, the area efficiency and power leakage parameters have been evaluated in order to find the most energy efficient design, with the lowest power leakage. We see a sharp drop in area efficiency after the SDC & MSC with counts up to 16, as displayed in Figures 3a and 3b respectively. We then overlay the area efficiency plots over the power leakage plots, in order to find the SDC & MSC configuration that has the highest bandwidth and lowest latency.

When we evaluate the power leakage parameter and compare it with the area efficiency parameter, with increasing SDC & MSC, we observe that the area efficiency



(a) Read/Write-Latency Comparison.
Area Efficiency-Read & Write Latency ns



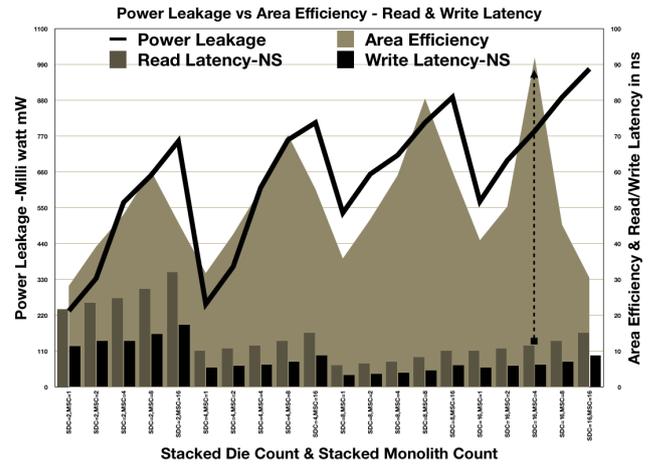
(b) Read/Write-Bandwidth Comparison.
Area Efficiency-Read & Write Bandwidth GB/s

Fig. 3: Efficiency and power leakage evaluation

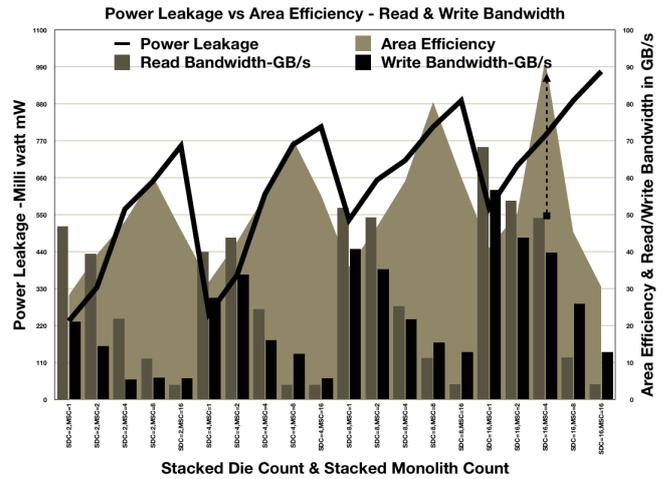
is high at 91% with power leakage under 1 Watt, when the SDC & MSC count is at 16 and 4, as is evident in Figures 4a and 4b. And while this configuration isn't very low in terms of energy leakage, it does showcase the lowest power leakage combined with the lowest latency and highest bandwidth. Hence, we choose this layout for the actual design specification section. In the next section, we will showcase the actual design specification of the 3D stacked memory, based on the configuration parameters detailed above, in the Gem5 simulator. After a successful 3D stacked memory configuration is obtained in Gem5, we will then run a simple hello-world test using Gem 5 to validate the architecture and the simulation of the chip.

Based on the figures 3a, 3b, 4a and 4b, the 2GB candidate with the following specifications using Cacti-3DD and DESTINY was selected for the next stage of design specification using gem5:

```
Bank Organization: 256 x 256 x 8
- Row Activation : 1 / 256 x 1
- Column Activation: 1 / 256 x 1
Mat Organization: 1 x 2
- Row Activation : 1 / 1
- Column Activation: 1 / 2
- Subarray Size : 32 Rows x 512 Cols
Mux Level:
- Senseamp Mux : 1
- Output Level-1 Mux: 1
- Output Level-2 Mux: 2
Local Wire:
- Wire Type : Local Aggressive
- Repeater Type: No Repeaters
```



(a) Power Efficiency-Power Leakage-Latency
Power Leakage vs Area Efficiency - Read & Write Latency



(b) Power Efficiency-Power Leakage-Bandwidth
Power Leakage vs Area Efficiency - Read & Write Bandwidth

Fig. 4: Power Efficiency-Power Parameterisation for Leakage-Latency and Leakage-Bandwidth

```
- Low Swing : No
Global Wire:
- Wire Type : Global Aggressive
- Repeater Type: No Repeaters
- Low Swing : No
Buffer Design Style: Latency-Optimized
Area:
- Total Area = 3.93101mm x 39.1354mm
- Mat Area = 15.3555um x 152.872um
- Subarray Area = 15.3555um x 73.9001um
- TSV Area = 1.96um^2
- Area Efficiency = 91.9679%
Timing:
- Read Latency = 11.5334ns
- Write Latency = 6.31317ns
- Refresh Latency = 3.40672us
- Read Bandwidth = 49.2234GB/s
- Write Bandwidth = 39.741GB/s
Power:
- Read Dynamic Energy = 294.026pJ
- Write Dynamic Energy = 293.86pJ
- Refresh Dynamic Energy = 7.16835uJ
- Leakage Power = 786.093 mW
```

DESIGN SPECIFICATIONS

A. Creating a basic CPU with Cache & Memory Controller in Gem5

When trying to architect the basic 3D-stacked Wide I/O DRAM memories, we are faced with three massive changes to be modelled:

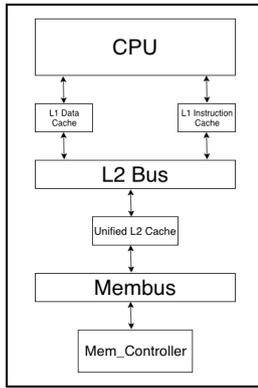


Fig. 5: Simple CPU Layout in Gem5.

1. How to enable 3D stacking of DRAM dies with the help of Through Silicon Via (TSV) interconnects.
2. How to support a minimum of four independent memory channels.
3. How to extend I/O interfaces to x128 bits per channel

Now, when compared to conventional DRAM, a 3D-stacked DRAM architecture offers increased memory bandwidth plus improved energy efficiency, due to the largely increased I/O interface width and significantly reduced I/O power consumption. The reduction in power consumption is achieved by stacking DRAM dies using low capacitance TSVs, compared to the traditional horizontal organisation of DRAM chips on one single plane.

We begin by creating a simple CPU in gem5, with standard CPU core connected to a system-wide memory bus. Please note that setup, installation and configuration of gem5 are not covered in this section. With that, we will connect a simple DDR3 memory channel, also connected to the bus. As gem5 is a modular design simulator, most of its components can be simulated as Sim-Objects such as CPUs, memory controllers, cache memories, buses etc. So we import all the SimObjects from the m5 library, instantiate the system we are going to simulate, set the clock on the system, specify a voltage domain. Next, we need to simulate the memory timings mode for the memory simulation. We also create a simple CPU timing based CPU which executes one single instruction in one clock cycle and then create a memory bus that is connected system-wide. Connecting the cache ports to the CPU, such as the instruction cache port and the data cache port, is the next step. A system cannot function without an interrupt controller, so we create the I/O controller and connect it to the memory bus. So we create a special port in the system to allow the system to read and write memory.

The next step is to create a memory controller and connect it to the memory bus. We use a simple DDR3 controller and connect it to the membus. This creates a system with no caches, temporarily. The next step is to add caches to the CPU model. We create caches separately and import them into the main CPU model file. We create L1 cache, instruction and data, and give it parameters such as associativity, tag_latency, data_latency, response_latency, miss status handling registers etc. Next, we instantiate each cache type, such as L1 data and L1 instruction, and add a size value to each, 16kB for L1 In-

struction and 64kB for L1 Data. Similarly, we create another L2 cache with similar parameters and size 256kB. Now, we need to instantiate the caches and connect them to the interconnect. Once this is done, the process that the CPU needs to execute needs to be set up. Gem5 operates in two modes, syscall emulation mode (SE mode) and full system mode (FS mode).

For now, we will run our system in the SE mode by providing it with a pre-compiled executable. We use a simple "hello world" C program, after we create a process and set the process command to the command we want to run, we set the CPU to use the process as its workload and create the functional execution context in the CPU.

The CPU created by the steps mentioned above is detailed in Figure 5. This CPU is our reference CPU which will be later used to run the PARSEC benchmarks, by leveraging the 3D stacked memory architecture. The objective is to create a CPU model Gem5 understands, and then give it a memory subsystem it can use to load and store data, just like a regular CPU uses DRAM to load and store data.

IMPLEMENTATION

With inputs from [3], [2], [19], [16], we set out to create the 3D stacked memory architecture in Gem5. The overall architecture that was created is displayed below in Figures 6 and 7. This memory architecture was created in a separate file, and was imported into the CPU created in the previous section. It uses the following components unique to the 3D stacked memory architecture: Vault Controllers, Serial Links, Internal Memory Crossbars.

The 3D stacked configuration is arranged in layers of DRAM wafers, each layered on top of the other, and connected to each other with the help of TSVs or Through Silicon Vias. A vault is a vertical interconnect across the four layers, each layer containing 128 MB of DRAM, thus creating a vault size of 512 MB. This can be increased or decreased by adding or removing more layers of DRAM stacked on top of each other. The logic layer and the 3D stacked memory crossbar sits under the base layer of DRAM, and provides routing and access to the vaults. The crossbar helps vaults connect to each other. The system designed here contains four DRAM layers and one logic layer die. Within each cube, the memory is organised and stacked vertically. Each vault has a vault controller that manages memory operations such as timing, refresh, command sequencing. Each vault can be configured with a maximum theoretical bandwidth of 10GB/s, thereby giving the 3D stacked architecture with 8GB of memory a total bandwidth of 160GB/s, which is possible using 2GB wafers with 40GB/s bandwidth, as we have explained previously. After a vault is accessed, we have configured a delay which prevents the vault from being accessed again, just like in regular DRAM. Each crossbar is assigned 4 vaults considered to be local. The other 12 vaults are considered remote, and can be accessed with the help of the crossbar switch. The 3D stacked memory must be refreshed periodically, and it is handled internally by the vault and logic controller.

At the bottom of the 3D stacked memory, we have the

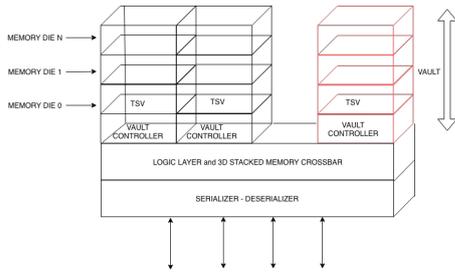


Fig. 6: 3D Stacked Memory Design Architecture

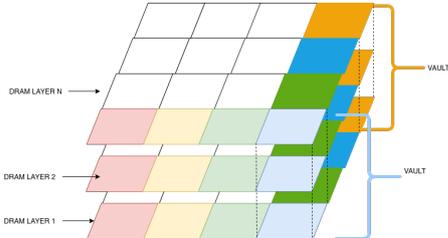


Fig. 7: 3D Stacked Memory - Vaults & Layers

serializer and deserializers, and the high speed links that are comprised of 16 different lanes that can be used to transmit in both transmit and receive directions. Therefore, a theoretical maximum bandwidth of 320 GB/s can be attained with the help of this 3D stacked architecture. Each lane can be configured to run in a full width or half width configuration. From a programmable perspective, the lanes can run in a 16X16, 16X8, 8X16, or 8X8 lane configuration. Figure 8 indicates the relationship between the local vaults, quadrants, crossbars and the remote vaults. The 3D stacked memory architecture created with the parameters and configurations previously described, was compiled and run in Gem5.

EVALUATION

In order to test and evaluate our 3D stacked memory based CPU architecture, we need to run our custom CPU on a clean system that does not have any additional software on it. Table II reports the performance of the custom architecture running PARSEC[5], [6]. We have compared the run times and execution times for the benchmarks using our custom architecture against a standard DDR3/DDR4 memory based system at DDR3/4-

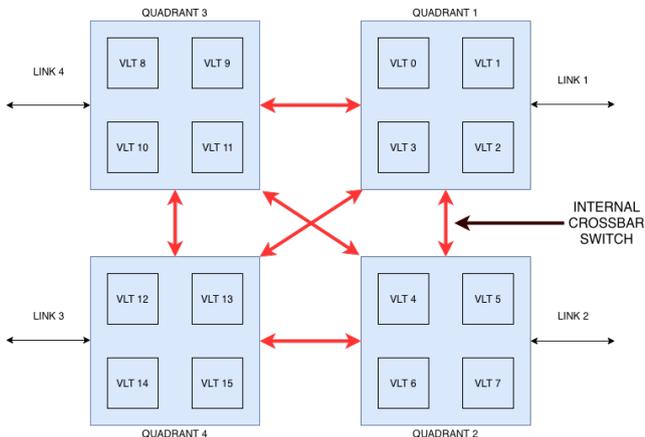


Fig. 8: Vaults, Quadrants and Crossbars

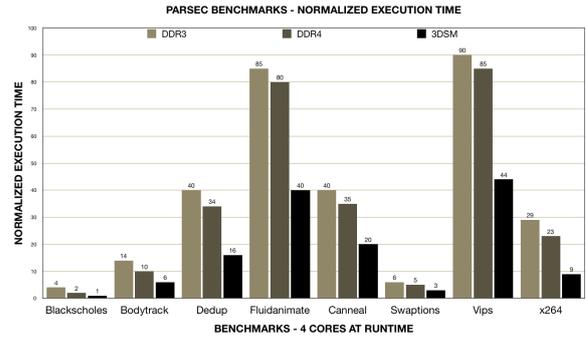


Fig. 9: PARSEC - Multi Core Normalized Execution Times, DDR3/DDR4/3D Stacked Memory.

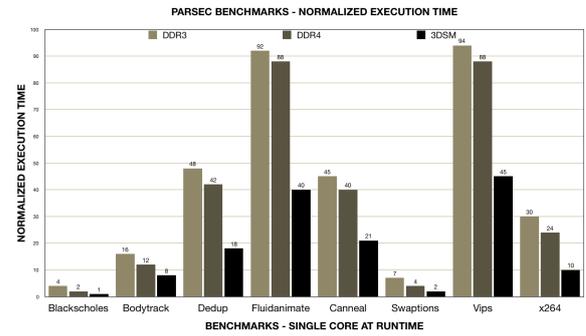


Fig. 10: PARSEC - Normalized Execution Times, DDR3/DDR4/3D Stacked Memory.

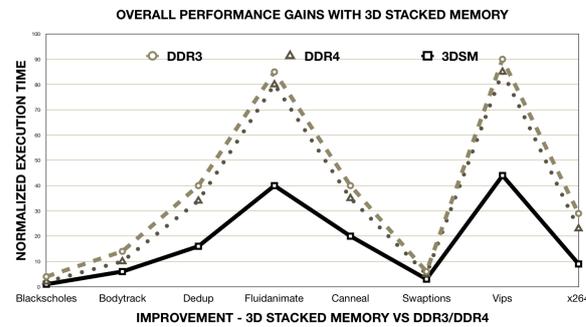


Fig. 11: Performance Gains & Improvement over DDR3/4 memory

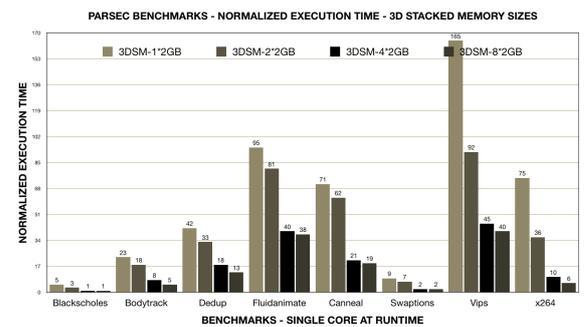


Fig. 12: PARSEC Benchmark Execution times compared by size - 3D Stacked Memory Sizes

TABLE II: Performance Gains - Single and Multi Core PARSEC Tests

Benchmark	Improvement-Single Core	Improvement-Multi
blackscholes	67%	67%
bodytrack	43%	50%
dedup	60%	56%
fluidanimate	55%	51%
canneal	51%	46%
swaptions	63%	45%
vips	50%	49%
x264	63%	65%

3200 with 15-15-15-35 (2T) timings, running the same X86 CPU running at 3.0 GHz.

Figure 10 shows the normalised execution time for 8 of the 13 benchmarks available in PARSEC. The runtimes were compiled for each memory system, using a reference x86 CPU at 3 GHz. Memory capacity was set to 8 GB for each memory type. As is evident from the results, there's not much of a difference between DDR3 and DDR4 in terms of execution speed. However, when we look at our 3D stacked memory architecture, there is an average of 40%-60% reduction in execution time for each benchmark. The benefit is especially evident in memory intensive benchmarks such as Fluidanimate and Vips, both of which are data-parallel models with medium working set data. This was the result of a single core execution, where the reference CPU was launched with just one core running at 3 GHz.

The next set of tests have been executed with multiple cores being launched at runtime. As some of the benchmarks leverage multi threading and multiple cores as well, observing the performance in a multi CPU and multi threaded environment would be extremely relevant. By using the -n switch to specify number of CPUs, we were able to simulate a multi CPU environment. The performance difference, while not hugely different from the single CPU benchmark result, still indicates that some benchmarks are inherently more CPU dependent than memory dependent. In Figure 9, we see the results of using 4 cores assigned to each CPU at run time. As before, all results were normalised according to the execution time of the target architecture with main memory implemented with DDR3. As is evident from the results of two benchmarks displayed in Figure 10 and 9, the 3D stacked memory configuration displays a significant improvement in performance in industrial benchmarks, while delivering improved read & write bandwidth, lower latency than traditional DDR3 and DDR4 memory, while satisfying the area efficiency, power consumption and temperature parameters. By comparing the normalised execution times, we observe the following gains in performance over conventional DDR-3/DDR-4 based memory systems.

For each of the benchmarks evaluated in this paper, performance gains visible, as displayed in Table II and Figure 11, by providing the reference 3.0 GHz CPU with a 3D stacked memory architecture varies from a minimum of 43% in a single-core environment to a maximum of 67% in a multi-core environment. This is a defini-

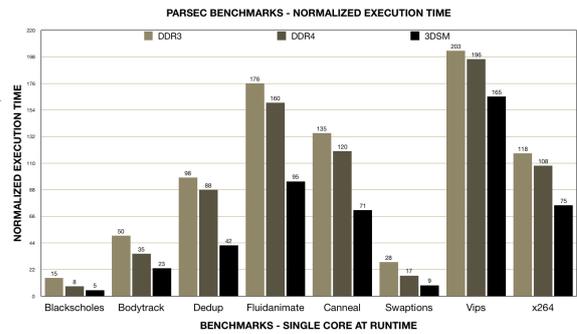


Fig. 13: PARSEC Benchmark Execution times - 2GB - 3D Stacked Memory

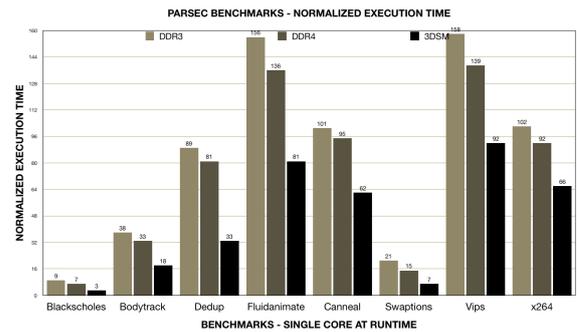


Fig. 14: PARSEC Benchmark Execution times - 4GB - 3D Stacked Memory

itive indication that supercomputing and indeed the memory wall challenge will benefit from 3D stacked memory. By observing significant reduction in the execution times of supercomputing standard benchmarks, we can confidently say that the overall performance gains achievable with 3D stacked memory in the configuration detailed in this paper should be a step in the right direction towards achieving exascale levels of computing.

We have also run the same set of benchmarks by modifying the 3D stacked memory size. By altering the number of stacked monoliths, we have evaluated the performance of 3D stacked memory for sizes ranging from 2GB, with just one 3D stacked memory wafer, to 16GB, comprising of 8 3D stacked memory wafers, each of size 2GB. The normalised execution times are displayed in Figures 12, 13, 14, and 15. We see conclusive evidence, especially in memory size intensive benchmarks such as vips & x264 encode, where higher the memory size, the lesser is the execution time.

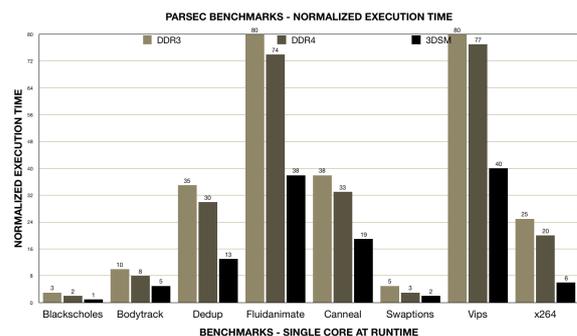


Fig. 15: PARSEC Benchmark Execution times - 16GB - 3D Stacked Memory

CONCLUSIONS AND FUTURE WORK

As we approach exascale levels of computing, we have realised that not just compute, but every single aspect of computing needs to scale massively in order to deliver the expected performance boost. This includes storage, memory, interconnects, space, power consumption and cooling. Not to mention the availability of structured parallel programming frameworks [14], [15] and administration as well. This paper has taken on one of the many challenges we as a race face in achieving and breaching exascale levels of computing. This paper has taken on the challenge of implementing a full scale 3D stacked memory architecture, creating a workable X86 architecture that is compatible with existing CPUs and benchmarks capable of evaluating hardware at supercomputing levels. The results look promising to say the least, with significant improvements visible in memory-intensive benchmarks such as Fluidanimate and Vips, and the system also looks capable of scaling and performing under multi-core environments as well.

This exercise intends to showcase that existing challenges in computing require a different, non-conventional approach such as 3D stacking. The future work on this topic would be to incorporate a machine learning algorithm to evaluate the results of multiple memory sizes, multiple cores, on the PARSEC benchmarks and run the benchmarks at a proper computationally-demanding environment in order to see how much gains are possible in terms of FLOPS. Not just supercomputers, but cloud computing will also benefit from the 3D stacked memory architecture, as many cloud service providers today provide custom instances tailored to running memory intensive workloads. CSPs such as AWS provide X1 instances that are custom built and designed for large-scale and in-memory applications in the cloud, which will benefit tremendously from leveraging a 3D stacked memory architecture.

By providing additional bandwidth, lowered latency, increased and efficient power consumption metrics for systems leveraging 3D stacked memory, cloud service providers will be able to provide high-performance instance capable of running memory intensive workloads such as running in-memory databases such as SAP HANA, big data processing engines like Apache Spark or Presto, and HPC applications. The potential benefits obtainable from such instances will go a long way in providing cheap, high-performance compute platforms to end users.

REFERENCES

- [1] K. Ahmed, J. Liu, A. Badawy, and S. Eidenbenz. A brief history of HPC simulation and future challenges. In *2017 WSC*, pages 419–430, Las Vegas, Dec. 2017. IEEE.
- [2] J. Ahn, S. Yoo, and K. Choi. Low-power hybrid memory cubes with link power management and two-level prefetching. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(2):453–464, 2016.
- [3] E. Azarkhish et al. High performance AXI-4.0 based interconnect for extensible smart memory cubes. In *2015 DATE*, pages 1317–1322, Grenoble, Mar. 2015. IEEE.
- [4] Z. Z. Bandic and other. Multilayer 3D memory based on network-on-chip interconnection. US Patent US 10,243,881 B2, Western Digital Technologies, Irvine, CA, Mar. 2019.
- [5] C. Bienia, S. Kumar, J. P. Singh, and K. Li. The PARSEC bench-

mark suite: Characterization and architectural implications. In *PACT '08*, pages 72–81, Toronto, Oct. 2008. ACM.

- [6] C. Bienia and K. Li. PARSEC 2.0: A new benchmark suite for chip-multiprocessors. In *MoBS 2009*, pages 1–9, Austin, June 2009.
- [7] N. Binkert et al. The Gem5 simulator. *SIGARCH Comput. Archit. News*, 39(2):1–7, Aug. 2011.
- [8] K. Chen et al. CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory. In *2012 DATE*, pages 33–38, Dresden, Mar. 2012. IEEE.
- [9] V. Cuppu, B. Jacob, B. Davis, and T. Mudge. High-performance drams in workstation environments. *IEEE Transactions on Computers*, 50(11):1133–1153, 2001.
- [10] R. Das. Blurring the lines between memory and computation. *IEEE Micro*, 37(6):13–15, 2017.
- [11] S. J. Eggers et al. Simultaneous multithreading: A platform for next-generation processors. *IEEE Micro*, 17(5):12–19, 1997.
- [12] A. Geist and D. A. Reed. A survey of high-performance computing scaling challenges. *Int. J. High Perform. Comput. Appl.*, 31(1):104–113, 2017.
- [13] S. Ghose, T. Li, N. Hajinazar, D. S. Cali, and O. Mutlu. Demystifying complex workload-dram interactions: An experimental study. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(3), Dec. 2019.
- [14] M. Goli and H. González-Vélez. Formalised composition and interaction for heterogeneous structured parallelism. *Int. J. Parallel Program.*, 46(1):120–151, 2018.
- [15] H. González-Vélez and M. Leyton. A survey of algorithmic skeleton frameworks: high-level structured parallel programming enablers. *Softw. Pract. Exp.*, 40(12):1135–1160, 2010.
- [16] R. Hadidi et al. Demystifying the characteristics of 3D-stacked memories: A case study for Hybrid Memory Cube. In *2017 IISWC*, pages 66–75, Seattle, Oct. 2017.
- [17] J. Handy. *The Cache Memory Book*. Academic Press Professional, Inc., San Diego, CA, USA, 1993.
- [18] J. Hopkins and other. 3D memory. US Patent US 10,170,639 B2, Micron Technology, Boise, ID, Jan. 2019.
- [19] G. Kim et al. Memory-centric system interconnect design with hybrid memory cubes. In *PACT '13*, pages 145–156, Edinburgh, 2013. IEEE.
- [20] D. Lee et al. Design-induced latency variation in modern DRAM chips: Characterization, analysis, and latency reduction mechanisms. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):26:1–26:36, June 2017.
- [21] W.-F. Lin. Reducing DRAM latencies with an integrated memory hierarchy design. In *HPCA '01*, pages 301–, Monterrey, Jan. 2001. IEEE.
- [22] S. Mittal. A survey of architectural techniques for managing process variation. *ACM Comput. Surv.*, 48(4):54:1–29, Feb. 2016.
- [23] S. Mittal, R. Wang, and J. Vetter. DESTINY: a comprehensive tool with 3D and multi-level cell memory modeling capability. *J. Low Power Electron. Appl.*, 7(3):23, 2017.
- [24] H. A. D. Nguyen et al. A classification of memory-centric computing. *J. Emerg. Technol. Comput. Syst.*, 16(2):13:1–26, Jan. 2020.
- [25] M. Poremba et al. DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches. In *2015 DATE*, pages 1543–1546, Grenoble, Mar. 2015. IEEE.
- [26] M. Qureshi. With new memories come new challenges. *IEEE Micro*, 39(1):52–53, 2019.
- [27] C. Weis, M. Jung, and N. Wehn. 3D stacked DRAM memories. In P. Franzon, E. Marinissen, and M. Bakir, editors, *Handbook of 3D Integration: Design, Test, and Thermal Management*, volume 4, chapter 8, pages 149–185. Wiley, Weinheim, 2019.

AWS EC2 Spot Instances for Mission Critical Services

Jerry Danysz, Víctor del Rosal, Horacio González-Vélez
Cloud Competency Centre, National College of Ireland
E: jerry.danysz@dancom.eu, {Victor.DelRosal,horacio}@ncirl.ie

KEYWORDS

AWS; spot instance; utility computing; elastic provisioning; SLA; random forest regression; cloud computing

ABSTRACT

For over a decade now, Amazon Web Services (AWS) has offered its spare capacity at a discounted price in the form of EC2 spot instances. This discount comes at the price of variable pricing and sudden instance termination. In this paper, we present a machine-learning solution to one of the challenges when using AWS Spot Instances, namely the termination of the instance on short notice. Our system, Spot Instance Management System (SimS), can effectively manage spot instances and keep up the availability at the desired level using 100-tree Random Forest Regression model. By using a risk assessment mechanism and proactive actions, SimS assures a three-nines SLA using AWS spot instances with lower running costs on workloads for a major European financial institution.

INTRODUCTION

AWS EC2 spot instances have offered a significant discount—up to 90% according to AWS—compared to their dedicated and on-demand/reserved models for over a decade. One of the base assumptions behind spot instances is that price is dynamic and can change anytime based on available capacity and current demand for the type of instance. Consequently, the attractive pricing comes with the trade-offs related to the availability of the spare capacity on a given region at a given period in time.

The use of spot instances requires customers to set a maximum instance price (per hour) they wish to pay and also to understand the risk, e.g. if the price of a given instance changes above the maximum price, the instance could be suddenly terminated and all data lost if not saved elsewhere. Until 2015, the instance termination was executed without prior notification to the customer. Since then, a different type of termination behaviours is offered, ranging from default termination through stopping the instance

to hibernation of the instance. Nonetheless, the usage of spot instances, even if cost-effective, is normally circumscribed to applications that are either non-critical or designed for interruptions.

This paper presents the Spot Instance Management System (SimS) whose main goal is to counteract the sudden/unexpected termination of cloud services running on top of EC2 spot instances. SimS employs a parallelised Random Forest Regression [5], [6] model for continuous response with 100 trees to predict price fluctuations. In this case, rather the regression model allows systems to swiftly detect when price would change and calculate the risk of the instance termination, and then consider the appropriate actions to maintain a 99.9% (a.k.a three nines) Service Level Agreement (SLA).

The actions could be live migration to another availability zone or redeployment of spot instance with a higher maximum bid. It is worth to notice that all actions are done proactively, while the risky instance is still running, therefore minimising the potential downtime of any service using EC2 spot instances.

RELATED WORK

Spot Price Prediction (SPP) has been a topic of research since the introduction of AWS spot instances in 2009. Yehuda et al. [1] predict the spot price by deconstruction and reverse engineering of a hypothetical spot instance algorithm that, despite common assumptions on spot prices, does not necessarily lead to supply- and demand-driven fluctuations but alternatively where prices are randomly generated from dynamic hidden reserve pricing. To confirm their hypothesis, they analyse the spot market and divided it into three pricing epochs with each epoch change at the significant change of SPP pricing models. They acknowledge that there is a market element in the price, but prices are still driven from the hidden reserved price.

In contrast, Singh and Dutta [13] seem not to fully concur with Yehuda et al.'s conclusion given that their model for dynamic price prediction is accounting for global market trends and local seasonality. The dynamic price prediction model is presenting two types of predictions:

short-term (hourly based) and long-term (a week ahead) based on analysis of 9 months of historical data for the top ten most used spot instances. They present the prediction result with an average 9.4% prediction error in short-term prediction and 20% in long-term (five days and more) price prediction.

Accurate price prediction for cloud instances typically relies on assertive workload quantification, which is related to the application type, e.g. HPC [11], [12] or to extrinsic factors such as seasonality or social-demographic factors [14].

Consequently, AWS spot price prediction has been previously modelled using a moving simulation model to create an artificial neural network-based algorithm for price prediction [16]. It employs historical data available for only medium size instances in a period of 7 months to train the MLP model, resulting in 4% prediction error in short-term prediction (hourly prediction) on average for medium size spot instances. This leads to the conclusion that neural network models are well suited for the prediction of price changes of spot instances. Zhao et al. [17] follow a different prediction approach by using a time-based series forecasting method, ARIMA Model (Auto-Regressive Integrated Moving Average) that is lighter compared to machine learning techniques like neural networking. Other approaches [13] have added a seasonal component to their ARIMA model effectively changing it to the SARIMA model. SARIMA has been used to analyse five months of historical data and create a prediction that is close to the average price for a period of 48 hours.

Random forest regression using historic AWS traces has been recently reported in the literature [8]. Similar approaches have previously used Support Vector Poly Kernel Regression (SMOReg), Gaussian Process and Linear Regression [3], and multiple discrete-time modelling [7]. All these models have been trained for the month with 12 months of historical data for the three most used types of instances. Their predictions have been typically generated for short (next hour), medium (half day) and Long (next day) periods. Per the conclusion, the neural network-based algorithms are performing better than others for medium (half day predictions) where SMOReg is better suited for predictions with highly variable months, and random forest regressions seem to deal better with workload variability.

From the above analysis, we can see that there are different heuristic approaches to prediction of spot instance pricing, ranging from reverse engineering and understanding what principles are behind the price level, through the classical statistical approach to the most modern use of arti-

cial intelligence and machine learning techniques. For this research, we want to prove that ML Models, specifically Random Forest Regression, in combination with business-driven automation can achieve a 99.9% SLA whilst accurately predicting price and potential price variation. Our approach has been successfully evaluated to support mission-critical cloud workloads from a major European financial institution.

*

AWS Spot bidding strategies overview

Previous research has attempted to find the best strategy to find a golden bid to assure spot instance availability. Andrzejak et al. [2], question how bidding may be conducted with strict target dates or SLAs, focusing their research on bidding strategies with that goal. Li et al. [10], classify common bidding approaches into three types:

- White box approach where bidding strategies are taking into account interactions between different market participants and effectively bidding can influence a spot price.
- Grey box approach has more individual bidding strategies wherein, contrary to the white-box approach, market interactions are not taken into account, but strategies are focusing mostly on workload, cost and availability of the resources.
- Black box approach, consisting of the most common strategies which derives bidding from historical spot pricing data and do not focus so much on workload, cost and availability, nor on interactions between market participants.

The five most classic strategies in the black box approach have been discussed by Li et al. [10] and by Voorsluys and Buyya [15]:

1. The minimum price, where the bid is based on historical minimum spot price
2. Mean, the bid price is set as the mean of all values of the historical spot price
3. High, the bid price on the maximum price observed in historical data
4. Current, the bid price is set as the value of current spot price
5. On-demand is the bid price equal to the on-demand price of the instance.

The above five strategies have also been incorporated to the solution for reliable provisioning of spot instances by Voorsluys and Buyya [15], combining all five strategies with fault tolerance techniques like migration to assure the most reliable solution for the limitations of spot instances. A survey on spot pricing by Kumar et al. [9] presents four bidding strategies: bidding on near to reserve price; bidding on above the average price calculated from the historical data; bidding close to the on-demand price; and, bidding over the on-demand price. Each of the aforementioned

	LOW	MEDIUM	HIGH
Availability			✓
Integrity			✓
Confidentiality			✓

TABLE I: Business Application Critically Matrix

techniques has its own benefits particularly for the final consumer, but also its costs in terms of the actual deployment and the business interest of the cloud provider.

In contrast to other attempts to use machine learning for AWS spot price prediction [4], the authors researched the possibility of maintaining the agreed service level of 99.9% that is common for a business-critical application while savings costs by running those applications on EC2 Instances. The aim was to test how EC2 spot instances can be used to reliably host mission-critical applications. Underpinned by a parallel Random Forest Regression model, we have employed a real application scenario borrowed from a major European Financial Institution to validate our findings.

METHODOLOGY

- Availability: impact if information availability is affected.
- Integrity: impact if information integrity is affected.
- Confidentiality: information confidentiality level.

The assumption is a scenario where the IT Service Provider is responsible for providing business-critical batch processing and ERP application for a European Financial Institution with agreed Availability of the service on the level of 99.9%. The IT Service Provider and the European Financial Institution have agreed upon a project to run the batch processing application using only AWS Spot Instances for the cost-effectiveness. The criticality matrix shown in Table I determines the application criticality level of Availability, Integrity, and Confidentiality.

As seen above, the application is highly critical to the core business of the European Financial Institution. If the availability of the system and, therefore, the application is compromised, the institution's ability to operate properly is degraded, potentially leading to significant profit loss and reputational damage. Compromise of information integrity may lead to substantial profit losses as well as posing a risk to confidentiality.

Since the system holds sizeable amounts of confidential information subject to General Data

Protection Regulation (GDPR), in case of a confidentiality breach, the Institution could face legal actions based on GDPR provisions as well as widespread loss of customer trust, negatively impacting the ability to conduct business.

Based on the criticality matrix, the following service level requirements are presented:

- Application Criticality: High
- AWS Region: eu-central-1
- Cumulative Downtime including maintenance 3.65 days per year
- Mean Time to Respond: 2 minutes
- Mean Time to Resolve: 15 minutes

Based on the above, in the Service Level Agreement (SLA), the service level has been agreed at three nines (99.9%). The service Level has been measured via the monitoring system *site24x7*¹, which calculates system availability using HTTP response codes and general host response.

We have deployed a Python data analysis module based on random forest regression model using 100 trees, parallelised using 4 instances at a time. To save computing time and costs associated with it, we have decided to limit data analysis to only the EU-CENTRAL-1 region and three selected types of instances `c5.xlarge`, `t3.micro` and `t3.medium`.

The Spot Instance Management System (SIMS) has been developed composed of four main modules:

- *The Data Collection Module:* responsible for downloading and aggregating historical data for EC2 Spot prices.
- *The Data Pump Module:* responsible for moving collected data from S3 Landing zone Bucket to S3 Staging Zone bucket after data collection, where later this data is used by Data Analysis Module for training the machine learning algorithm.
- *The Data Analysis Module:* responsible for analysing historical data gathered by the data collection module and then is responsible for applying the machine learning model based on random forest regression. Execution is started via Amazon Lambda function that is responsible for starting the AWS Fargate Task(Figure 1).
- *The Risk Assessment and Automation module:* responsible for risk analysis of the instance interruption in each of availability zones in the configured region and next for taking the appropriate actions concerning the level of the risk(Figure 2).

The prime idea behind the system is to have automated mechanism that with use of machine learning, can predict the price of the spot instance in the next hour and act accordingly by migrating the affected spot instance to the next availability zone. In case when all availability zones would

¹<https://www.site24x7.com>

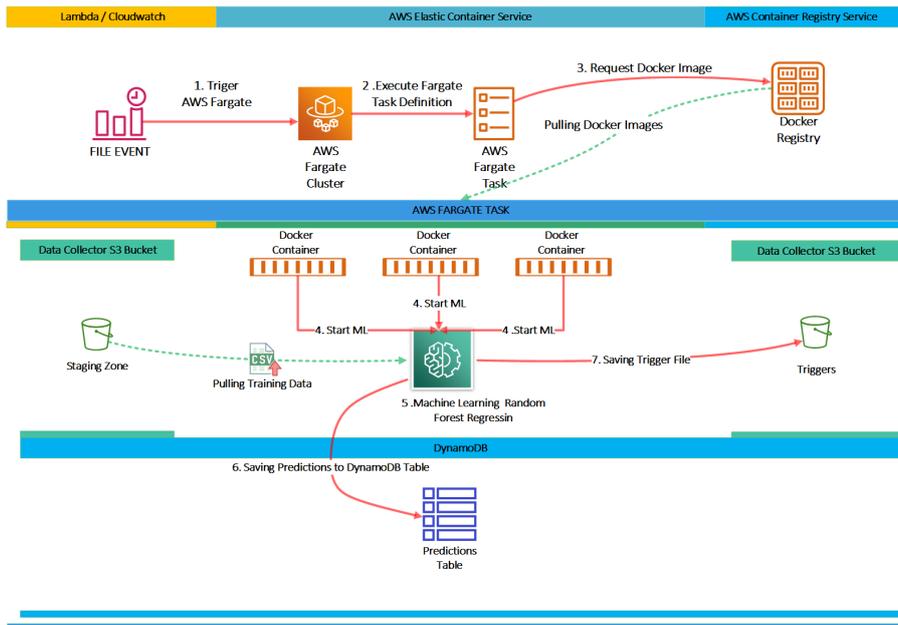


Fig. 1: The Data Analysis Module

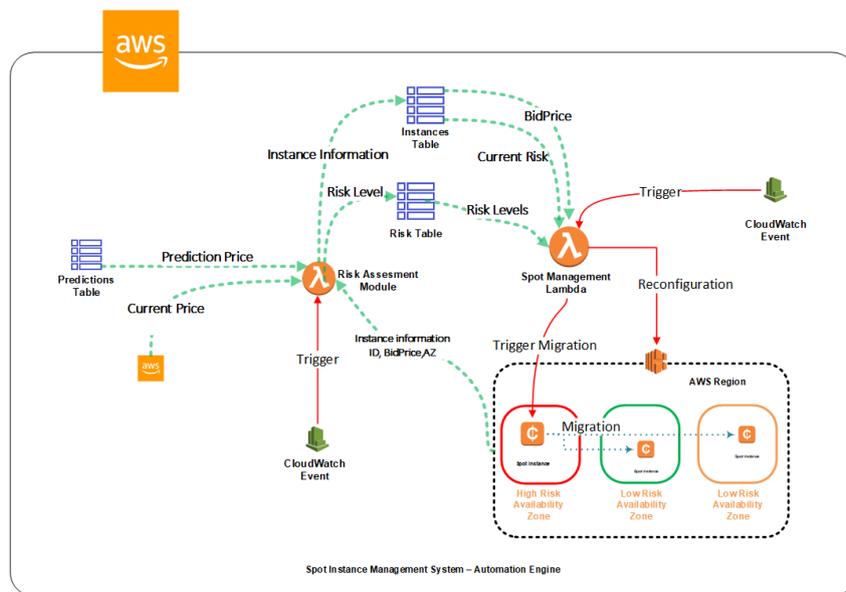


Fig. 2: The Risk Automation Module

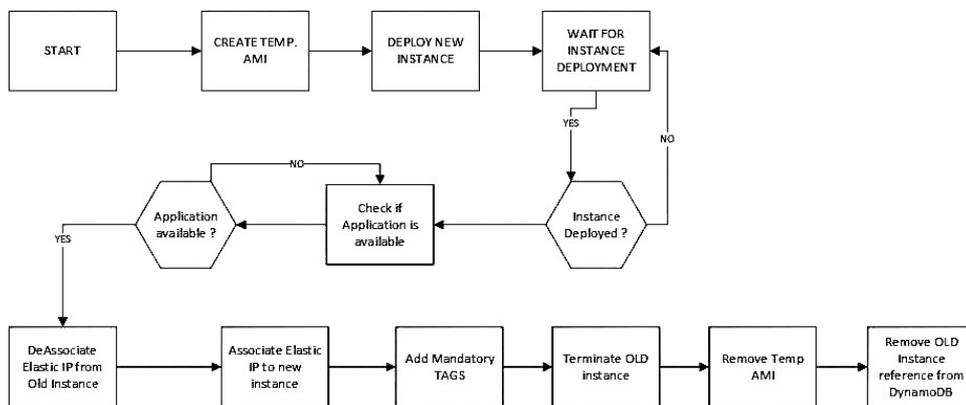


Fig. 3: Instance Migration Process

be labelled as High Risk, the system will redeploy the spot instance with a new bid price increased by 20%.

In the worst-case scenario, where there would be no spare capacity to spin the machine, the system would run the on-demand machine only when capacity is unavailable.

The risk engine is the part of the risk and automation module responsible for the calculation of the risk of instance interruption in each of the availability zones. The risk is calculated based on the maximum bid price, current price, predicted price and the threshold for each of the risks level Low Lr , Medium Mr , and High Hr .

The risk is calculated for each of the availability zones in a given region with the following formulae using Current price (Cp), Maximum bid price (Mb), and Prediction price(Pp) :

$$Lr = Cp < \left(\frac{55}{100} x Mb \right) \text{ AND } (Pp < Mb)$$

$$Mr = Cp > \left(\frac{55}{100} x Mb \right) \text{ AND } Cp < \left(\frac{8}{10} x Mb \right) \text{ AND } (Pp < Mb)$$

$$Hr = Cp > \left(\frac{8}{10} x Mb \right) \text{ AND } (Pp < Mb)$$

In case prediction price (Pp) is larger than the maximum bid price, risk is calculated as High.

If current risk is Medium and there is at least one availability zone with low-risk assessment than migration operation is starting to that zone. If there is no low zone, there is no action taken. If current risk is high and low or medium zones are available, then the migration process will move the instance to the lowest risk zone. If the risk is high in all availability zones, the machine is re-deployed by the automation engine to the same availability zone but with a 20% higher maximum bid price. The migration process is shown in Figure 3.

EVALUATION

For simulation purposes, a Python script has been developed to execute the following steps:

- Change the predicted price for the next full hour to one of the values selected randomly on every execution (on-demand price, predicted price, maximum bid price, current price, mean of all above).
- Execute Risk Recalculation.

After risk is recalculated, we would rely on the automation module to evaluate and execute necessary actions against the SimS Managed instance. For the instance that is not managed (the

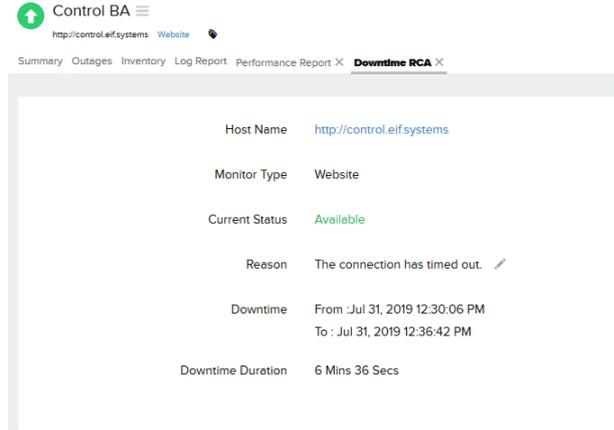


Fig. 4: Root Cause Analysis Report from Monitoring System

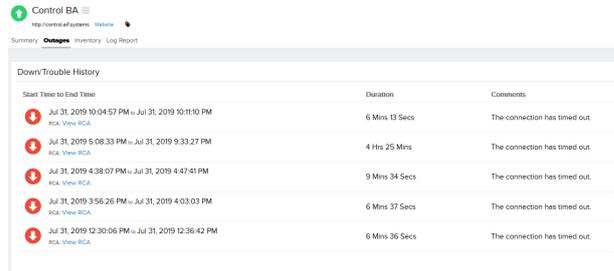


Fig. 5: Outage reports of control system

control instance), the script will execute the following steps: evaluate if the maximum bid price is lower than the predicted price and If the predicted price is higher, terminate the instance after 2 minutes. The authors have also included a simulation of an engineer acting on a given incident created by the monitoring system. A manual intervention required to bring the system back on-line has been calculated to take approximately four minutes and forty five seconds.

The data collection bucket contains landing zone, staging store and triggers, where trigger files at the end of the execution of data collection and data pump are stored.

To gather availability data, it was necessary to find a monitoring system that would be independent of the solution and would provide SLA-type reports and SLA configuration. The authors opted for the Site24x7 SaaS offering and configured it for website monitoring, checking connectivity to HTTP ports of defined targets, response times, DNS response time and general availability by ping, in one minute intervals.

During the migration execution, the following outputs are generated:

- New Temporary Amazon Image (AMI) created based on the current instance
- New Instance deployed to target availability zone based on the created image

Time (UTC +00:00)	Message
2019-08-04	
No older events found at the moment. Retry.	
01:16:06	Gathering Current SIMS Managed Spot Information
01:16:06	Searching DynamoDB for the Entry with ID: i-06afce5df599a9c9cb
01:16:06	Searching DynamoDB for the Entry with ID: t3.medium-Linux/UNIX
01:16:06	Searching DynamoDB for the Entry with ID: t3.medium-Linux/UNIX
01:16:06	Searching DynamoDB for the Entry with ID: t3.medium-Linux/UNIX
01:16:06	Risk Level in eu-central-1a is Low
01:16:06	Risk Level in eu-central-1b is Low
01:16:06	No Action is needed
No newer events found at the moment. Retry.	

Fig. 6: Cloud Watch Log representing No Actions

Time (UTC +00:00)	Message
2019-08-03	
16:52:05	Searching DynamoDB for the Entry with ID: t3.medium-Linux/UNIX
16:52:05	Searching DynamoDB for the Entry with ID: t3.medium-Linux/UNIX
16:52:05	Checking if there is LOW Risk AZ Available
16:52:05	Adding 50% to Maximum Bid Price
16:52:05	Searching DynamoDB for the Entry with ID: i-094ea2c8903d23bc3
16:52:05	Redeployment of the instance with new BidPrice
16:52:05	Creating Image
16:52:05	Preparation: Creating IMAGE
16:52:05	Preparation: Image Created
16:52:05	Deployment: Getting current Bid price for the instance id i-094ea2c8903d23bc3
16:52:05	Searching DynamoDB for the Entry with ID: i-094ea2c8903d23bc3
16:52:05	Deployment: Preparing move of the instance to new availability zone eu-central-1b
16:52:05	Deployment: Spot Request Created. Waiting for fulfillment
16:52:05	Deployment: Instance Deployed to new availability Zone eu-central-1b
16:52:05	Deployment: New Instance Deployed in target AZ with id i-06afce5df599a9c9cb
16:52:05	Deployment: Waiting for Instance to Complete the Boot
16:52:05	Configuration: Switching Traffic from old to new instance
16:52:05	Configuration: Reassigning IP Address 35.157.195.247 from old instance to new
16:52:05	Configuration: Setting SIMS Tags on instance i-06afce5df599a9c9cb
16:52:05	Name: BA-SIMS
16:52:05	domain: eif.systems
16:52:05	SIMS_Managed: True
16:52:05	FQDN: ba.eif.systems
16:52:05	Cleanup: Terminating old Instance i-094ea2c8903d23bc3
16:52:05	Cleanup: Removing image ami-0c35a61c510d42afb
16:52:05	Cleanup: image ami-0c35a61c510d42afb removed
16:52:05	Cleanup: Removing old instanceID i-094ea2c8903d23bc3 from DynamoDB reference Table

Fig. 7: Redeployment of Instance with new Bid Price

- Old Instance Data removal from DynamoDB Instances Table.

As part of the interruption of Control System, we have the following:

- Scenario: Classic Termination of EC2 Spot Instance
- Actors: Simulator
- Desired Outcome: System Terminated and restored

The simulation script to terminate EC2 instances operates with 2 minute delays, simulating the Amazon Notification grace period. It also later simulates the system engineer's input by restoring the service.

The control system during the experiment has been terminated a number of times causing reported outage (Figure 5) and unavailability of the application.

SimS Automated Actions low Risk

- Scenario: risk Level low low low
- Actors: SimS System
- Desired Outcome: no migration initiated

Expected behaviour, if all availability zones are low risk, then no action is taken (Figure6). As presented above, the automation module evaluated the risk and did not perform any actions.

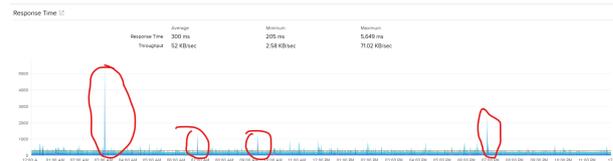


Fig. 8: Response time monitoring of the System

Down/Trouble History	Duration	Comments
Start Time to End Time		No outages for "Yesterday"

Fig. 9: No outage reported by the system. Monitoring system did not report system outage, the only indication of migration is a short spike in response time.

SimS Automated Actions High Risk

- Scenario: Risk Level high high high
- Actors: SimS System
- Desired Outcome: instance redeployed with higher bid price, no system downtime reported by monitoring system

During the execution, the Automation Module will evaluate risk in all availability zone and based on the result will move the instance to another availability zone or redeploy to current with the higher bid price.

In this scenario, The Sims System detected that all availability zones in the region are high risk. In this situation, the system is designed to redeploy the instance with a higher bid price to mitigate the high risk of interruption and therefore to maintain desired availability level by setting up (migrating new instance) and reassigning the elastic IP from Risky instance to newly deployed one, therefore allowing traffic to reach the system without any issue.

5-Day SLA Monitoring

- Scenario: 5-Day SLA Monitoring
- Desired Outcome: SLA Level of 99.9%

A simulation script has been scheduled and running in four hour intervals affecting the classic EC2 Spot instance as well as risk analysis data for the SimS System. Due to random price selection, the exact time and date of the next interruption were not known.

Results show that usage of a proactive system managing spot instances can prove to be valuable if our main goals are low costs and availability of the system using the spot instances.

We can see that if automation is designed to act while a risky instance is still running, automated switch-over is almost seamless but at the price

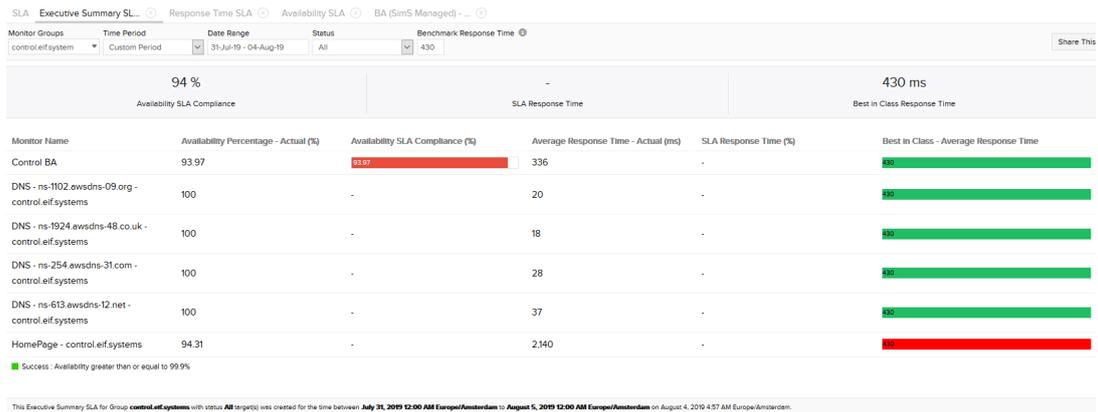


Fig. 10: 5-Day SLA Report for Control System



Fig. 11: 5-Day availability report for Control System



Fig. 13: 5-Day availability report for SimS Managed System

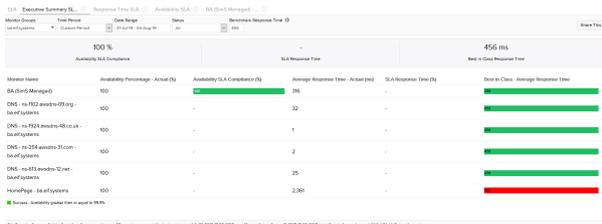


Fig. 12: 5-Day SLA Report for SiMS Managed System

of a drop in performance during the migration of the instance.

For applications that are very response time-sensitive this could be still the issue, as well as for the application where data is written continuously as during the migration the checkpoint (image) of the machine is created and any data written in the time between image creation and switch over of the traffic to the newly deployed machine would be lost.

The question in here would be the trade-off, in case of classic spot instance the customer can face

unexpected termination and if they do not have any solution that would periodically save the data from that instance while running they could face complete data loss in comparing to few seconds of loss in case of SimS Managed System, therefore automated system as one presented in this research can significantly expand possibilities of the application of spot instances as well as can help with reduction of operational costs.

The Spot Instance Management system for its risk analysis is using Random Forest regression-based machine learning model, while this model during our research proved sufficient, the model itself was not a part of in-depth testing and therefore could not be as accurate as it could be hoped. Necessary comparison testing showed us that price predicted are the same as current prices or difference in price is minimal.

Our prediction machine learning model requires a lot of computing power, to calculate price predictions for 150 types of EC2 spot instances for every availability zone in EU region would require over 26 hours running in 4CPU docker containers provided by AWS. Even with those lim-

itations, it has been proved that smart, proactive system that can move around spot instances based on the predicted price and therefore risk calculated from it, can be effective in achieving not only 99.9% availability but even 100%

The authors' research question poses how can we assure SLA Level of 99.9% for service running on EC2 Spot instance, as mentioned above and shown in the evidence, the SimS System is able to effectively manage spot instances and keep up the availability on the desired level and achieve required three nines SLA thanks to its Risk assessment mechanism and proactive actions before any terminate can happen.

With automatic bid rise in case of high risk in all availability zone at the current stage, this could lead to higher than desired bid price and therefore not always attain cost-effectiveness.

CONCLUSIONS

The research question poses how a service level of 99.9% can be assured. In the paper, the authors show that, in principle, Spot instance Management system (SimS) can effectively manage spot instances to keep an availability level of 99.9%

For a system like SimS to be effective and accurate, there is a requirement for a reliable machine learning module to predict spot price in the next hour. In hourly five-day tests, we have simulated a number of interruptions of classic EC2 spot instances, and we combined this with the change of the risk level in availability zones to force the SimS system to act and migrate affected machine if are located in a high risk availability zone. Migration is using image creation (checkpointing).

Currently, due to the limitation in computing power and the aim to run the system as serverless as possible, the support for more than just a subset of instances is limited by the computational power of AWS Fargate docker containers.

Future researchers could take this solution a step further by selecting more sophisticated machine learning models that would take into account not only the historical data but also seasonality, allowing support for a more significant number of instances and availability zones.

In our research, we have presented the theoretical application of the SimS system in the financial sector to run important CRM application. As an overall conclusion, the authors posit that with the minor adjustments mentioned in this chapter, the solution could be of commercial value in a variety of sectors where service availability, as well as low cost, play a pivotal role.

REFERENCES

- [1] O. Agmon Ben-Yehuda, M. Ben-Yehuda, A. Schuster, and D. Tsafir. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation*, 1(3):16, 2013.
- [2] A. Andrzejak, D. Kondo, and S. Yi. Decision model for cloud computing under SLA constraints. In *MASCOTS '10*, pages 257–266, Miami, Aug. 2010.
- [3] S. Arévalos, F. López-Pires, and B. Baran. A comparative evaluation of algorithms for auction-based cloud pricing prediction. In *IC2E*, pages 99–108, Berlin, Apr. 2016. IEEE.
- [4] M. Baughman, C. Haas, R. Wolski, I. Foster, and K. Chard. Predicting Amazon Spot Prices with LSTM Networks. In *ScienceCloud'18*, Tempe, June 2018. ACM.
- [5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [6] A. Cutler, D. R. Cutler, and J. R. Stevens. Random forests. In C. Zhang and Y. Ma, editors, *Ensemble Machine Learning*, pages 157–175. Springer, Boston, 2012.
- [7] R. Keller, L. Häfner, T. Sachs, and G. Fridgen. Scheduling flexible demand in cloud computing spot markets: A real options approach. *Business & Information Systems Engineering*, 62:25–39, 2020.
- [8] V. Khandelwal, A. K. Chaturvedi, and C. P. Gupta. Amazon EC2 spot price prediction using regression random forests. *IEEE Transactions on Cloud Computing*, 8(1):59–72, 2020.
- [9] D. Kumar, G. Baranwal, Z. Raza, and D. P. Vidyarthi. A survey on spot pricing in cloud computing. *Journal of Network and Systems Management*, 26:809–856, 2018.
- [10] Z. Li, M. Kihl, and A. Robertsson. On a feedback control-based mechanism of bidding for cloud spot service. In *CloudCom '15*, pages 290–297, Vancouver, Nov 2015. IEEE.
- [11] A. Marathe et al. Exploiting redundancy and application scalability for cost-effective, time-constrained execution of HPC applications on Amazon EC2. *IEEE Transactions on Parallel & Distributed Systems*, 27(9):2574–2588, 2016.
- [12] D. Petcu et al. Next generation HPC clouds: A view for large-scale scientific and data-intensive applications. In *Euro-Par 2014*, volume 8806 of *Lecture Notes in Computer Science*, pages 26–37, Porto, Aug. 2014. Springer.
- [13] V. K. Singh and K. Dutta. Dynamic price prediction for amazon spot instances. In *2015 48th Hawaii International Conference on System Sciences (HICSS)*, pages 1513–1520. IEEE, 2015.
- [14] P. Smith, H. González-Vélez, and S. Caton. Social auto-scaling. In *PDP 2018*, pages 186–195, Cambridge, Mar. 2018. IEEE Computer Society.
- [15] W. Voorsluys and R. Buyya. Reliable provisioning of spot instances for compute-intensive applications. In *AINA 2012*, pages 542–549, Fukuoka, Mar. 2012. IEEE.
- [16] R. M. Wallace et al. Applications of neural-based spot market prediction for cloud computing. In *ID-AACS '13*, volume 2, pages 710–716, Berlin, Sept. 2013. IEEE.
- [17] H. Zhao, M. Pan, X. Liu, X. Li, and Y. Fang. Exploring fine-grained resource rental planning in cloud computing. *IEEE Transactions on Cloud Computing*, 3(3):304–317, 2015.

A simulation study on a WSN for emergency management

Lelio Campanile

Mauro Iacono

Fiammetta Marulli

Michele Mastroianni

Dipartimento di Matematica e Fisica
Università degli Studi della Campania

”L. Vanvitelli”

viale Lincoln 5

81100, Caserta, Italy

KEYWORDS

Performance evaluation; Wireless Sensor Networks; emergency management; simulation; ns-3; Internet of Things.

ABSTRACT

Wireless Sensors Networks (WSN) are one of the ways to provide the communication infrastructure for advanced applications based on the Internet of Things (IoT) paradigm. IoT supports high level applications over WSN to provide services in a number of fields. WSN are also suitable to support critical applications, as the supporting technologies are consolidated and standard network services can be used on top of the specific layers. Furthermore, generic distributed or network-enabled software can be run over the nodes of a WSN.

In this paper we evaluate and compare performances of IEEE 802.11g and 802.11n, two implementations of the popular Wi-Fi technology, to support the deployment and utilization of an energy management support system, used to monitor the field by a team of firefighters during a mission. Evaluation on an example scenario is done by using ns-3, an open network simulator characterized by its realistic details, to understand the actual limitations of the two standards besides theoretical limits.

I. INTRODUCTION

IoT and WSN are an established approach to implement viable and flexible monitoring infrastructures. The evolution of both hardware market and software support enables a variety of applications, spanning from domotics, Industry 4.0 related solutions and smart cities. While the commodity market offers a large choice of products, including low cost ones, the use of these technologies to support critical applications should be carefully planned and related solutions require proper care and choice of components, implementation and integration.

The technology stack that implements the IoT

paradigm over WSN is now mature and is, at the state, complex and layered. What happens at the system level is not anymore immediately predictable and dominated by determinism as it was in the first, proprietary or embedded products. In fact, there is a large availability of solutions, so that the performances of the layers chosen in a given architecture cannot be predicted on the basis of the declared performances of components. In order to design critical applications and verify the scenarios in which they will operate, possibly in support of safety and protection of human operators or to lower risk in case of intervention in dangerous sites, a simulation based approach that exploits very detailed models of the technological layers should be preferred as a support and reference, particularly when it is not possible, for the variability of operational conditions and setup or their unpredictability, to proceed with validation on the field.

We focus on the domain of emergency management. In particular, the work described in this paper is a part of a larger research activity that aims at studying an Edge computing based support system for emergency response. In the overall approach, a key role is played by IoT technologies in support of the field action of firefighters inside large buildings. Coverage understanding, as well as technological choices, are of paramount importance in the design process: consequently, as we did not find supporting studies after a thorough analysis of the existing literature, here we present the results of a parametric analysis, that has been conducted in a simple scenario that mimics a large warehouse, a typical case for the application domain, that also has the advantage of having a simple geometry, from which general conclusions can be drawn about technological limitations.

In this paper we verified, by means of simulation, the actual possibilities of IEEE 802.11g and 802.11n in a constrained WSN scenario. These two technologies have been used as part of an emergency management support that aims to assist firefighters while entering and moving in a building in which there is a fire. While the second one offers in theory higher performances,

specially on available bandwidth and management, the first is already available in rugged hardware and has more mature commercial implementations, due to its consolidated presence on the market and the feedback resulting from its wide number of installations in different conditions and for different uses. The analysis and the comparison is thus motivated by the need for choosing in a savvy way components according to cost issues, and for solving the make or buy dilemma in critical situations for disposable WSN nodes. The scenario is designed to avoid additional workloads and traffic on a node with respect to the one generated by or directed to the node, to obtain an evaluation that is neutral with respect to routing effects. The emergency management system implements an Edge computing based application over heterogeneous IoT WSN nodes with different characteristics, including Augmented Reality personal support for firefighters that operate on the field and camera based protection solutions for the environments in which they operate: consequently, we explored the potential of these two technologies with synthetic workloads in this perspective.

This paper is organized as follows: Section II presents related works; Section III presents the operational scenario in which the system is to be deployed and a glance on all its components, to describe the complete framework; Section IV introduces the simulation choices and the characteristics of the simulation; Section V presents the results of the simulation campaigns and a short analysis of the outcomes; Section VI closes the paper with final considerations and future works.

II. BACKGROUND AND RELATED WORK

The growth of IoT technology causes the increasingly pressing need to keep computing close to the user or to the application scenario, while keeping the advantages of Cloud computing available and in the loop. This sums up the potential offered by Cloud computing, Mobile computing and IoT, with the purpose of matching both global requirements in terms of cost management, performances and flexibility of resources, and local requirements such as privacy enforcement and resiliency to network problems. With this in mind, Fog computing paradigm standardizes the Edge approach to computing.

With regard to Fog computing, interesting introductory resources are [24], [26] and [14], that propose complementary approaches to the fundamentals of the topic.

The paradigm and the substanding architecture present a richness of open research challenges, that encompass basic aspects like communication protocols, system organization and architectures [16][28][21], and technological solutions [17], advanced aspects like performance evaluation and verification [10], design methodologies, system and software management [12][11][27], data reduction [4], security [25][29], load balancing [20] or policy-related aspects such as legal implications and risk [13][23]. A useful review paper is [15].

A convenient approach to study and evaluate different network topologies without the need of setting up a physical implementation is network simulation. Due to the richness of aspects and parameters that characterize computer networks, selecting the appropriate network simulator to target is a crucial task for researchers. In order to deep into simulators, an extended description and a comparison table between the most relevant network simulators may be found in [6].

In order to perform large-scale network simulation, one of the most widely used tool is Network Simulator 3 (ns-3). ns-3 is stated as a versatile and complete simulator by many authors, and in some benchmarks related to the study of wireless network it proved to be the fastest simulator in terms of computation time [18]. ns-3 is an open source simulator, released in 2006 [1][22], and may be considered as a replacement of an older tool, called ns-2. The simulation environment is released for most of the modern operating systems, and is written in C++ with an optional Python scripting API. It allows researchers and practitioners to study network protocols (mainly Internet-related) and large-scale networked systems in a controlled environment. ns-3 provides community supported modules for a wide variety of network protocols and components, it supports both simulation and emulation that allows including real network portions, it is designed to support large-scale simulations and it is easily extensible and programmable.

ns-3 generates PCAP traces of simulated models, so researchers can easily study or debug the output with standard tools such as Tcpdump [2] or Wireshark [3]. Additionally, there is also a number of external tools provided by the ns-3 community, in this work we have used extensively the Flow Monitor module [7] for the performance monitoring and the ns-3 Simulation Execution Manager (SEM)[19] to perform multiple simulations in a structured and repeatable manner. A systematic literature review on ns-3 is [6].

III. SCENARIO

The reference scenario concerns a system conceived to support firefighting squads during field operations (from [9]). Firefighters are equipped with a number of sensors to monitor their health conditions, an augmented reality device that enriches their personal view with details on the scene they are observing, and a camera. Besides personal equipment, firefighters deploy on the field additional sensors while moving into the incident location. Such sensors are designed to provide various kind of sensing features: some complex sensors with significant computing power can synthesize elementary sensed data into abstract information and integrate sensed data from other sensors. Sensors and personal equipment are powered by batteries and compose a WSN that connects all devices to an Edge server that runs a field command and control application. For a more complete description of the system and of the personal equipment, the reader can refer to [9].

The whole system can be described in terms of three

classes of nodes:

- *Personal Support* (PS): equipment that each fireman wears, including various sensors (e.g. vital parameters, audio, thermal, chemical), an AR visor including a camera, and a local computing device that manages all sensing;
- *Simple Sensor* (SS): there are ordinary WSN nodes with some local computing capability, mainly used for managing interactions with the rest of the system and preprocessing data;
- *Intelligent Sensor* (IS): there are nodes equipped with multiple sensors and can perform significant local computing on sensed data and execute generic tasks, with the possibility of offloading from other nodes and towards other nodes or the Edge server.

The PS nodes may also generate AR additional graphics, and process sensed information to produce comprehensive abstract local status information and interacts with the rest of the system.

In the system designed, the SS nodes can be put directly under the control of PS or IS nodes in order to augment their capabilities, but in this paper we will simplify the problem by only considering the case in which all nodes are directly under the control of the Edge server, in order to evaluate network performances.

The SS nodes, in normal operating conditions, generate a light and regular network traffic towards the Edge server, and IS nodes send larger and less regular network workloads. The PS nodes, besides generating large and non regular traffic, also deals with an additional traffic due to video information, and receive traffic from the Edge server. Furthermore, IS and PS nodes, when available energy on their board is lower than a given threshold, start delegating computing tasks to the Edge server, in order to extend their own active lifetime. For this reason, more intense data traffic is generated on the WSN after an initial reconfiguration and status synchronization data traffic. A similar situation happens (in the opposite traffic direction) whenever IS or PS nodes need to reload software images and/or reconfiguration parameters, which are loaded from the Edge server.

In the next section some simulation-based analyses are presented about the proposed solution, and in the section V some preliminary results are shown.

IV. MODELING AND SIMULATION

Before implementing the real network, wide simulation campaigns are needed in order to assess the technology to be used (WiFi, LoWPAN, etc.). The platform chosen for simulation is ns-3, because of its versatility and simulation performance. As a first attempt to simulate the network, standard WiFi technologies are simulated, such as IEEE 802.11g and 802.11n.

The choice of WiFi technologies is due to their diffusion and performance in term of throughput and data rate; on the other hand, this kind of devices are more power consuming than others (e.g., LoWPAN). Moreover, the outdated 802.11g technology is taken into consideration due to the larger availability of industrial

TABLE I: Values of simulation parameters

Parameters	Values
WiFi Protocol	[802.11g, 802.11n]
N. of nodes	[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
Datarate	[3Mbps, 5Mbps, 7Mbps]

rugged devices than 802.11n.

The used topology is a simple grid with the same horizontal and vertical distance between rows and columns. The Edge server is placed in the upper left corner and the second deployed mode is the Access Point (AP), while the other nodes are placed in rows, with F nodes per row. The distance between rows is ΔX , and the distance between columns is ΔY (see Fig. 2). All nodes directly communicate with the Edge server (see Fig. 3). Obviously, the diagonal of grid is the maximum distance from the grid server to the node (worst case). The version of ns-3 used is 3.30.1, compiled and the simulation ran on a macbook pro 15" equipped with an Intel core i7 2.8GHz with 16Gb of RAM and macOS Mojave 10.14.5.

The ns-3 modules used to set up the simulation are showed in Table II.

The simulation campaigns were aimed at obtaining different metrics for evaluating the performance of the entire network. The metrics that have been taken into consideration are the total throughput that the server can manage, or rather the maximum data flow it can receive, the throughput that each node can manage singularly and finally a derived metric, the number of nodes that are able to guarantee a throughput greater than a threshold value.

In order to obtain the best and most complete simulation results, we decided to run the simulations by varying different parameters:

- *WiFi protocol*: the protocols set in the simulations. We used 802.11 family protocol at level 1 and 2 of the ISO/OSI protocol stack.
- *number of nodes*: the number of simultaneous nodes in each simulation. We used IPv6 protocol at level 3 of the ISO/OSI protocol stack.
- *Datarate*: the continuous dataflow that the node transmits to the Edge Server. We used the UDP protocol at the level 4 of the ISO/OSI protocol stack.

Figure 1 summarizes the ISO/OSI protocol stack setup used in simulations. We execute the simulation with all possible combinations of value, shown in table I, for the above parameters. We then had 72 different simulation scenarios run for the subsequent analysis of the results and evaluation of performance.

V. RESULTS

Keeping in mind that some of the simulated nodes (particularly PS nodes) are part of the gear of firemen in a crisis area, the dispatched simulation campaigns are aimed to detect expected performance values. First of all, it is needed to know how many PS nodes may be used in the crisis area (i.e. how many firemen can work

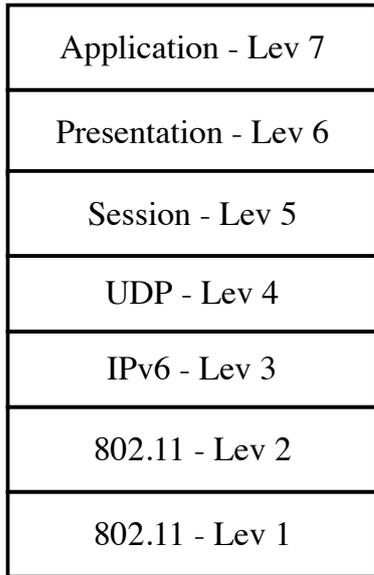


Fig. 1. The ISO/OSI stack used in simulations

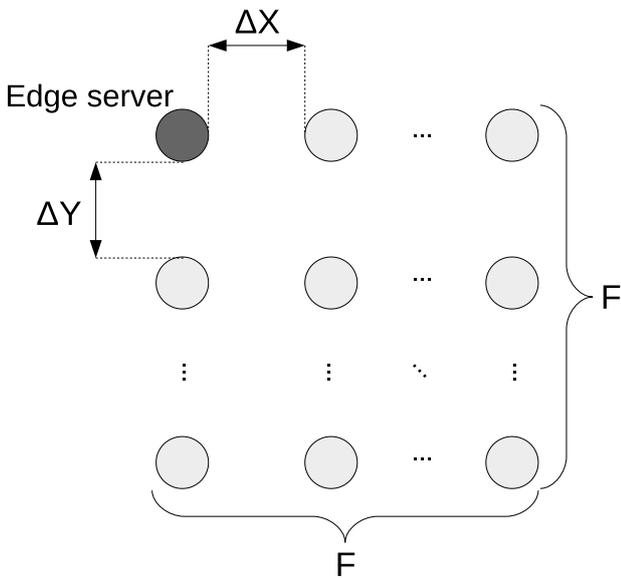


Fig. 2. Spatial configuration of the network and related parameters

in crisis area) without causing performance decay. In order to do this, tests have been conducted varying the numbers of involved nodes. The nodes feed the Edge server with data streams of different data rates (3, 5 and 7 Mbps). For a correct understanding of the results we want to emphasize that in the simulations all the nodes involved start transmitting at maximum speed at the same time. The next figures show throughput values for both IEEE 802.11g (Figure 4) and 802.11n (Figure 5) devices.

Looking at the results, it is clear that the throughput given by IEEE 802.11g is widely inadequate. In fact, Figure 4 shows that, even with a 3 Mbps data stream, the throughput is very low also with a low number of

TABLE II: Used ns-3 modules for the simulation setup

Modules	Description
Internet	general IP, TCP and UDP protocols implementation
IPv6	IPv6 protocol implementation
Mobility	a model that helps to allocate devices
Spectrum	it aims at providing support for modeling the frequency-dependent aspects of communications [5]
Propagation Loss	modeling of propagation loss and propagation delay
Wi-Fi	implementation of ns-3 models for wi-fi (IEEE 802.11)
Flow Monitor	the module uses probes, installed in network nodes, to track the packets exchanged by the nodes to measure the performance of network protocols [8]
Application	modeling different applications at ISO/OSI level 7

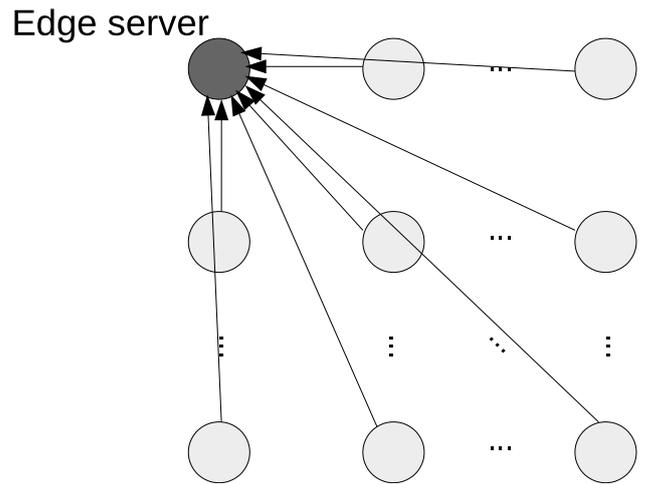


Fig. 3. Structure of the connections and information flow considered for the experiments

nodes. Figure 5, instead, shows that IEEE 802.11n is really promising, allowing sufficient performance, up to 8 PS nodes with a data stream of 3 Mbps.

Established that 802.11g is inadequate for implementation of the designed system, the focus shifts on 802.11n. The analysis continues with another simulation campaign, in which the number of nodes that are "up" (i.e. nodes that are able to make use of data rate/2) is evaluated. The following Figure 6 shows that, with a data rate of 3 Mbps, 8 nodes are fully available; with a data rate of 7 Mbps, at least 5 nodes may be fully available.

The next graph, in Figure 7, is also more interesting, and shows the mean performance of PS nodes versus total number of nodes. In this Figure it is easy to see

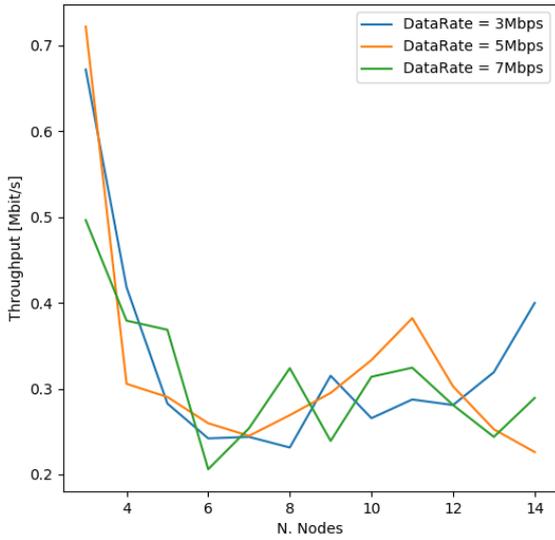


Fig. 4. Throughput of the network when using IEEE 802.11g

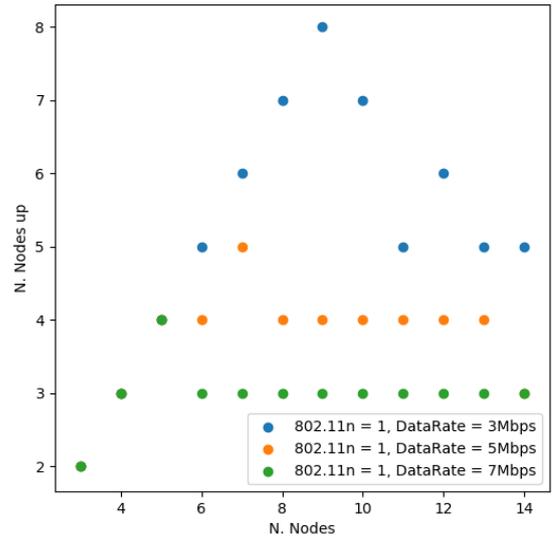


Fig. 6. Nodes "up" versus total number of nodes (IEEE 802.11n)

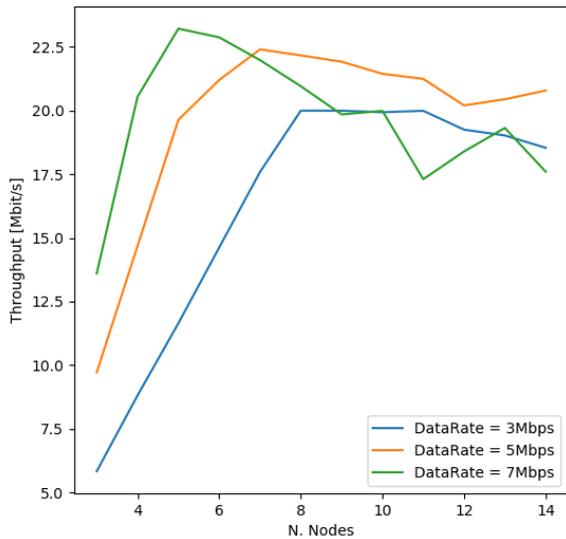


Fig. 5. Throughput of the network when using IEEE 802.11n

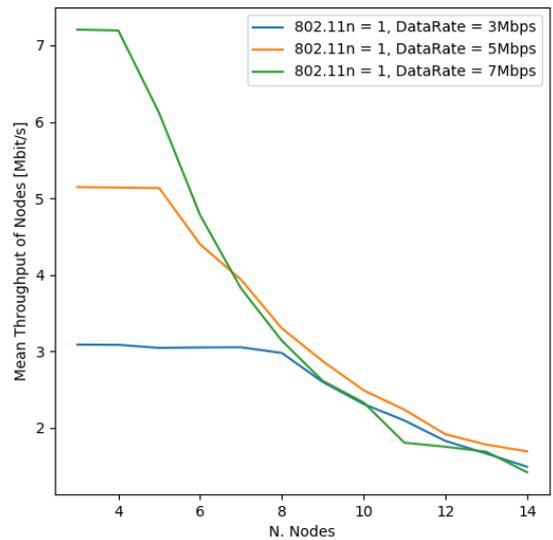


Fig. 7. Mean throughput of nodes versus total number of nodes (IEEE 802.11n)

that, with a data stream of 3 Mbps, up to 8 PS nodes may be used at the same time.

The same test has been performed also on a network based on IEEE 802.11g devices. The results, shown in Figure 8, confirm that 802.11g is not able to perform well enough for the case study; in fact, throughput decays quickly, even when using a little number of PS nodes.

VI. CONCLUSIONS

In this paper, the network simulator ns-3 is used to study the performance of an Edge system conceived to support an advanced emergency management sys-

tem. This system, based on three different types of sensors (personal, base and intelligent sensors) is simulated considering the possible implementation based on IEEE 802.11g or 802.11n network devices. A large test campaign has been performed, and the results show that 802.11g is inadequate for implementation, whilst 802.11n shows promising results. In this case, it is possible to use up to eight PS nodes with a data rate of 3 Mbps, which may be acceptable for the purpose of the system.

Future work includes extending simulation using different network technologies, such as LoWPAN, and

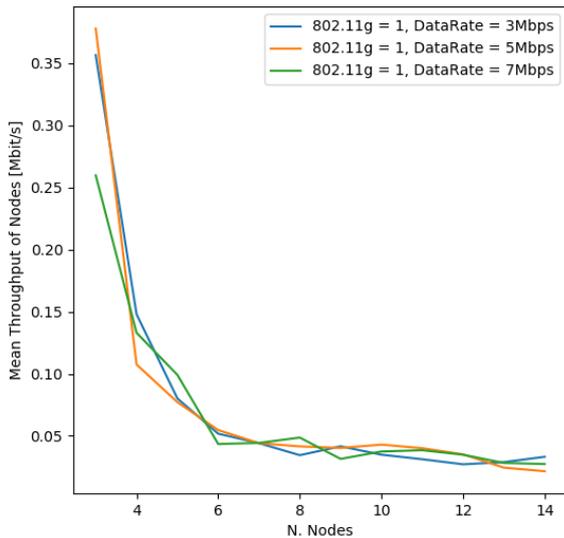


Fig. 8. Mean throughput of nodes versus total number of nodes (IEEE 802.11g)

deepening into the model to study the effect of different policies on the nodes behaviour. Furthermore, energy management in the sensor network will be considered, as well as more complex routing policies.

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 by project “Attrazione e Mobilità dei Ricercatori” Italian PON Programme (PON_AIM 2018 num. AIM1878214-2).

REFERENCES

[1] NS3. <https://www.nsnam.org/>. Accessed: 2019-06-03.

[2] Tcpdump. <https://www.tcpdump.org/>. Accessed: 2019-06-03.

[3] Wireshark. <https://www.wireshark.org/>. Accessed: 2019-06-03.

[4] M. Aazam and E.-N. Huh. Fog computing and smart gateway based communication for Cloud of Things. pages 464–470, 2014.

[5] N. Baldo and M. Miozzo. Spectrum-aware channel and phy layer modeling for ns3. In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS '09*, pages 2:1–2:8, ICST, Brussels, Belgium, Belgium, 2009. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[6] L. Campanile, M. Gribaudo, M. Iacono, F. Marulli, and M. Mastroianni. Computer network simulation with ns-3: A systematic literature review. *Electronics*, 9(2):272, Feb 2020.

[7] G. Carneiro, P. Fortuna, and M. Ricardo. Flowmonitor - a network monitoring framework for the network simulator 3 (ns-3). *ACM*, 5 2010.

[8] G. Carneiro, P. Fortuna, and M. Ricardo. Flowmonitor - a network monitoring framework for the network simulator 3 (ns-3). *ACM*, 5 2010.

[9] E. Cavalieri d’Oro, S. Colombo, M. Gribaudo, M. Iacono, D. Manca, and P. Piazzolla. Modeling and evaluating a complex Edge computing based systems: An emergency man-

agement support system case study. *Internet of Things*, 6:100054, 2019.

[10] M. Chiang and T. Zhang. Fog and IoT: An overview of research opportunities. *IEEE Internet of Things Journal*, 3(6):854–864, Dec 2016.

[11] A. V. Dastjerdi and R. Buyya. Fog computing: Helping the Internet of Things realize its potential. *Computer*, 49(8):112–116, Aug 2016.

[12] M. Desertot, C. Escoffier, and D. Donsez. Towards an autonomous approach for edge computing: Research articles. *Concurr. Comput. : Pract. Exper.*, 19(14):1901–1916, Sept. 2007.

[13] C. Esposito, A. Castiglione, F. Pop, and K. K. R. Choo. Challenges of connecting Edge and Cloud computing: A security and forensic perspective. *IEEE Cloud Computing*, 4(2):13–17, March 2017.

[14] M. Ficco, C. Esposito, Y. Xiang, and F. Palmieri. Pseudodynamic testing of realistic Edge-Fog Cloud ecosystems. *IEEE Communications Magazine*, 55(11):98–104, Nov 2017.

[15] M. Hajibaba and S. Gorgin. A review on modern distributed computing paradigms: Cloud computing, Jungle computing and Fog computing. *Journal of Computing and Information Technology*, 22(2):69–84, 2014.

[16] Z. Hao, E. Novak, S. Yi, and Q. Li. Challenges and software architecture for Fog computing. *IEEE Internet Computing*, 21(2):44–53, Mar. 2017.

[17] M. B. A. Karim, B. I. Ismail, W. M. Tat, E. M. Goortani, S. Setapa, J. Y. Luke, and H. Ong. Extending Cloud resources to the Edge: Possible scenarios, challenges, and experiments. In *2016 International Conference on Cloud Computing Research and Innovations (ICCCRI)*, pages 78–85, May 2016.

[18] A. R. Khan, S. M. Bilal, and M. Othman. A performance comparison of open source network simulators for wireless networks. In *2012 IEEE International Conference on Control System, Computing and Engineering*, pages 34–38, Nov 2012.

[19] D. Magrin, D. Zhou, and M. Zorzi. A simulation execution manager for ns-3: Encouraging reproducibility and simplifying statistical analysis of ns-3 simulations. In *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWIM '19*, page 121–125, New York, NY, USA, 2019. Association for Computing Machinery.

[20] J. Oueis, E. Strinati, and S. Barbarossa. The Fog balancing: Load distribution for small cell Cloud computing. volume 2015, 2015.

[21] H. D. Park, O.-G. Min, and Y.-J. Lee. Scalable architecture for an automated surveillance system using Edge computing. *J. Supercomput.*, 73(3):926–939, Mar. 2017.

[22] G. F. Riley and T. R. Henderson. *The ns-3 Network Simulator*, pages 15–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[23] R. Roman, J. Lopez, and M. Mambo. Mobile Edge computing, Fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, 2016.

[24] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, Oct 2016.

[25] J. Shropshire. Extending the Cloud with Fog: Security challenges & opportunities. 2014.

[26] L. Vaquero and L. Rodero-Merino. Finding your way in the fog: Towards a comprehensive definition of fog computing. *Computer Communication Review*, 44(5):27–32, 2014.

[27] M. Villari, M. Fazio, S. Dustdar, O. Rana, and R. Ranjan. Osmotic computing: A new paradigm for Edge/Cloud integration. *IEEE Cloud Computing*, 3(6):76–83, Nov 2016.

[28] S. Yi, Z. Hao, Z. Qin, and Q. Li. Fog computing: Platform and applications. In *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pages 73–78, Nov 2015.

[29] S. Yi, Z. Qin, and Q. Li. Security and privacy issues of fog computing: A survey. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9204:685–695, 2015.

AUTHOR BIOGRAPHIES



LELIO CAMPANILE is a PhD. student 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. His email is lelio.campanile@unicampania.it.



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. His research activity is mainly centered on the field of performance modeling of complex computer-based systems, with a special attention for multiformalism modeling techniques. His email is mauro.iacono@unicampania.it. More information about his activities is available at his website <http://www.mauroiacono.com>.



FIAMMETTA MARULLI is an Assistant Professor in Computing Systems 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. Her email is fiammetta.marulli@unicampania.it.



MICHELE MASTROIANNI is currently the Data Protection Officer of Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy, and is also a research associate at Dipartimento di Matematica e Fisica of the same University, 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, and has been Network Manager at the same University, project leader and expert for many local, regional and national technical projects. He also teaches

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

PROBABILITY MODEL OF CONCEPTS RECOVERY IN SMALL SAMPLE LEARNING

Alexander A. Grusho, Nick A. Grusho,
Michael I. Zabezhailo and Elena E. Timonina
Federal Research Center
"Computer Science and Control"
of the Russian Academy of Sciences
Vavilova 44-2, 119333, Moscow, Russia
Email: grusho@yandex.ru, info@itake.ru,
zabezhailo@yandex.ru, eltimon@yandex.ru

Vladislav V. Kulchenkov
Lomonosov Moscow State University
Faculty of Computational
Mathematics and Cybernetics
Leninskiye Gory 1-52, 119991, Moscow, Russia
Email: vlad.kulchenkov@gmail.com

KEYWORDS

Small sample learning, random graph model for series of small samples, concept learning.

ABSTRACT

Many information security monitoring systems and controlling of IoT systems receive information in the form of short messages, which can be considered as small samples. Concepts are considered as classes of small samples that allow you to determine the correctness of monitoring systems. The paper is devoted to the problem of recovering concepts on observations of series of small samples.

Probabilistic model of appearance of series of small samples is introduced. To define concepts, the probabilistic dependency is used within series of small samples. The case of series of length 2 of small samples is considered. This assumption allowed the construction of a random graph and provided its probability-statistical analysis. Asymptotic approximations of probability distributions in the series scheme are used to identify ranges of parameter values that better define the structure of concepts. The set of parameter values is defined, at which the structure of concepts is uniquely determined with probability which tends to 1.

INTRODUCTION

Many information security monitoring systems (Grusho et al., 2015, 2019a) and controlling of IoT systems receive information in the form of short messages, which can be considered as small samples.

Unlike classical mathematical statistics, the meaning of a "small sample" does not involve obtaining information by statistical processing of data contained in one small sample. The information required for data analysis is intended to be obtained by statistical processing of a set of small samples.

However, many problems arise with small samples analysis (Etz and Arroyo, 2015). One major problem is that the distribution on the set of small samples may not be homogeneous even under conditions of independence of obtaining small samples.

Let's consider the example of problems with small samples. Assume that the part of small samples (k of Communications of the ECMS, Volume 34, Issue 1, Proceedings, ©ECMS Mike Steglich, Christian Mueller, Gaby Neumann, Mathias Walther (Editors)
ISBN: 978-3-937436-68-5/978-3-937436-69-2(CD) ISSN 2522-2414

small samples) is obtained according to the distribution P_0 , and the remaining m of small samples are obtained according to the distribution P_1 , $P_0 \neq P_1$. If the problem of recovering samples with distribution P_1 is considered, and the probability of "false alarm" (a sample from distribution P_0 is by mistake taken as the sample from distribution P_1) has to be made as little as possible, then if $m/k \rightarrow 0$ and $k \rightarrow \infty$ the probability to classify incorrectly the sample from distribution P_1 , as a rule, is limited from below to a constant (Grusho et al., 2019a; Axelsson, 1999).

Due to the inertia of information systems, monitoring system messages are often received in series, reflecting the close states of a computer system or a network.

The purpose of monitoring systems is to identify anomalies in the operation of monitored objects. Technologies for detecting anomalies based on the construction of models of normal behavior are known (Tukey, 1977). However, a set of incoming messages does not always have simple structure (Grusho et al., 2016). It is not always possible to apply regression methods (Tukey, 1977). For example, when the network changes its behavior, many parameters of the network functioning change. If the device shows several modes of operation, it is necessary to base their description on analysis of incoming small samples using machine learning procedures.

Recently, a lot of machine learning methods have been developed (see, for example, (Jordan and Mitchell, 2015; Bramley, 2017)). Machine learning techniques based on small samples have also been studied in detail (Shu et al., 2019). One of the main scenarios in such learning is based on Concept Learning. In (Shu et al., 2019) concepts are called classes of small samples, which allow to determine the correctness of monitoring systems. The purpose of Concept Learning is to recognize concepts by a small number of small samples based on previously observed concepts. The second purpose of this approach is to recover a set of concepts. In the future, the terminology of small sample learning theory will be used, where concepts refer to classes of samples to which membership needs to be determined for newly arrived small samples. One more problem is the concept construction (Shu et al., 2019).

The problem of recovering of concepts is recognized

as one of the main challenges in the theory of machine learning (small sample learning). As noted above, because of the inertia of information systems, messages are often received in series, reflecting the close states of the computer system or the network.

Next, we assume that the data comes from the series of small samples. Each series is often associated with a single concept. At the same time, the number of concepts is unknown, but it is finite. Each concept will be described by a set of samples. The paper (Grusho et al., 2019b) deals with the case where the series are unambiguously related to the same concept. In this paper, the mathematical model of building a series of small samples is described as a random graph. This model reflects the possibility of series from different concepts. Analysis of this model helped to prove that asymptotically with probability which tends to 1, the set of series uniquely recovers concepts and their number.

MATHEMATICAL MODEL

Let the finite alphabet A be defined. Monitoring system messages are sent to the analysis system by short formalized messages reflecting the states of different sensors. For simplicity, we consider that messages have the same length r . However not all messages are admissible, i.e. there is a set of admissible messages $X \subseteq A^r$, $|X| = n$. Each message can be considered as a small sample. The analysis system can operate in several automatically switched modes. This corresponds to the existence on set X of several the priori existing concepts that we will denote through M_1, \dots, M_k . The number of concepts k and concepts themselves are not defined.

The task of machine learning is to recover the concepts themselves on the set of accumulated small samples. The paper assumes that small samples with a high probability are received from any one concept on series of the length l . For simplicity, let us assume that $l = 2$.

Let time T be represented by the set of natural numbers. Series of small samples of the size $l = 2$ appear sequentially and independently of each other. If a series of small samples (x_1, x_2) consists of small samples belonging to the same concept, then the probability of this event equals to $\alpha > 0$. If x_1 belongs to one concept and x_2 belongs to another concept, then the probability of such series equals to $1 - \alpha > 0$. In other words, let the control sequence of 1 and 0 be obtained by the Bernoulli scheme with the probability of appearance of 1 equaling to α and the probability of appearance of 0 equaling to $1 - \alpha$. If the element of control sequence equals to 1 at the given time, a concept from one of the sets M_1, \dots, M_k is equiprobably selected, and a pair of small samples (x_1, x_2) is independently and equiprobably selected from this concept. If the element of control sequence equals to 0 at the given time, two different concepts are randomly and equiprobably selected, from each of which independently and equiprobably selects a small sample x_1 and a small sample x_2 .

Suppose the series sequence is infinite. Then due to independent appearance of series from Borel-Cantelli's lemma follows that each pair (x_1, x_2) , $x_1, x_2 \in X$, will

be met in an infinite set of times.

Let's consider the random graph $G^{(n)}$ with nodes on the set X and with edges corresponding to appearance of each pair (x_1, x_2) , $x_1, x_2 \in X$, at least once. Then it follows from the above that the graph $G^{(n)}$ is complete for any n . It follows that graph $G^{(n)}$ carries no information about the structure of concepts and their number.

Let $n \rightarrow \infty$. Note that in this case the sequence $G^{(n)}$ also carries no information about the structure of concepts and their number.

Let's consider now a great number of random graphs $\{G^{(n)}\}$ which are defined on X , $|X| = n$, by means of a chain of length T of randomly chosen series of small samples from the set X . The random graph $G_T^{(n)}$ may not be a complete graph. Therefore, in graphs $G_T^{(n)}$ some regularities may be revealed related to the structure of concepts and their number. Obviously, at small T , these regularities manifest weakly, and at large T , they disappear altogether. However, there is area of T , where these regularities can be identified. Using the set of graphs $\{G_T^{(n)}\}$, the algorithm for recovering the set of concepts of M_1, \dots, M_k and estimating their number k will be constructed.

PROPERTIES OF GRAPHS $G_T^{(n)}$

Let's assume that the number of concepts $k < \infty$ is fixed, and it is unknown. Each small sample from X belongs to some concept M_i , the numbers of small samples in concepts M_1, \dots, M_k satisfy the following conditions:

$$\begin{aligned} |M_i| &= n\varepsilon_i, \quad 0 < \varepsilon_i < 1, \\ \sum_{i=1}^k \varepsilon_i &= 1, \quad i = 1, \dots, k \end{aligned} \tag{1}$$

Let a chain of series of small samples of length T be constructed. On this data the graph $G_T^{(n)}$ is defined. This graph is the source for the concept recover algorithm.

Lemma 1. Let $n \rightarrow \infty$, $\alpha = 1 + o(1)$,

$$T = \frac{(n \ln n)^2}{\alpha} (1 + o(1)),$$

then in the random graph $G_T^{(n)}$ the subgraphs formed by nodes M_i , $i = 1, \dots, k$, are complete graphs with the probability which tends to 1.

Proof. Consider the classic task of allocations of particles into boxes (Kolchin et al., 1978), where boxes are pairs of the kind (x, x') , where x, x' belong to the same concept, and different particles are placed in these boxes according to the probability scheme built above at the moments of time when there are units in the control sequence. For each concept M_i , the number of such pairs is equal to

$$\binom{|M_i|}{2}.$$

Let in the control sequence of the length T and random choosing of the concept M_i gets N_i of units. If N is the number of all units in the sequence of the length T , then by the definition of the defined probability measure

$N_i = \frac{N}{k}(1 + o(1))$ with probability $1 + o(1)$, where the convergence of the probability to 1 is denoted by $1+o(1)$. At the same time

$$P(N = \alpha T(1 + o(1))) = 1 + o(1).$$

From here for all $i = 1, \dots, k$

$$P(N_i = \frac{\alpha T}{k}(1 + o(1))) = 1 + o(1).$$

The mathematical expectation of the number of empty boxes μ_0 (i.e. the number of edges missing in the graph $G_T^{(n)}$ on the set of nodes M_i) in equiprobable scheme of allocations for each $i = 1, \dots, k$, equals to (Kolchin et al., 1978):

$$E_i(\mu_0) = \binom{|M_i|}{2} \left(1 - \frac{1}{\binom{|M_i|}{2}} \right)^{N_i}. \quad (2)$$

It follows that simultaneously for all concepts the mathematical expectation of the number of empty boxes for all concepts is equal to $o(1)$.

By Markov's inequality (Shiryayev, 1984), the probability that at least one node will not fall into the corresponding complete graph is equal to $o(1)$.

Thus, all concepts in $G_T^{(n)}$ produce complete graphs with probability which tends to 1. Lemma 1 is proved.

In complete graphs obtained in the random graph $G_T^{(n)}$ on subsets of nodes $M_i, i = 1, \dots, k$, the value of considered parameter T does not take into account the following possibility.

Let node x belong to the set M_i . However, it is possible that edges resulting from zeros in the control sequence will also allow the node x to be attached to the complete graph arising on another concept. Such a situation will be called a learning error, as it introduces ambiguity in the description of the membership of the node x to one of concepts.

The following lemma shows that the probability of any error tends to zero when

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1))$$

and

$$1 - \alpha = O\left(\frac{1}{\ln n}\right).$$

Lemma 2. Let $n \rightarrow \infty, 1 - \alpha = O\left(\frac{1}{\ln n}\right)$,

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1)).$$

Then the probability of any error, i.e. that there exists a node belonging to any two concepts, tends to zero.

Proof. Let node $x \in M_i$. Then x generates an error with the set $M_j, j \neq i$, if fixed $|M_j|$ edges corresponding to some zeros of the control sequence connect x to all nodes of the set M_j . The probability of this event for fixed zeros in the control sequence is equal to

$$\left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}.$$

The mathematical expectation of the number of such events on the set of zeros of the control sequence is equal to

$$\binom{T - N}{|M_j|} \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}. \quad (3)$$

Then the mathematical expectation of the number of errors generated on the set M_i by edges connecting M_i with M_j is equal to

$$|M_i| \cdot \binom{T - N}{|M_j|} \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}. \quad (4)$$

Let's use the inequality (Riordan, 1958)

$$\binom{m}{r} \leq \frac{m^r}{r^r(m-r)^{m-r}}.$$

Having substituted the received estimates and values $|M_i|, |M_j|$, we receive for some $\varepsilon > 0$ that the next estimate takes place for formula (4) in conditions of Lemma 2:

$$ne^{Cn} \left(\frac{\ln n}{n} \right)^{\varepsilon n} = o(1), \quad (5)$$

where $C > 0, C = const$. This estimation is true for every pair $(i, j), i \neq j$, of concepts. Lemma 2 is proved.

ALGORITHM FOR CONCEPTS RECOVERY

Let's denote via \mathcal{B} the algorithm for concept recovery. Let \mathcal{A} be the algorithm for allocating the maximum complete subgraph in a graph. The result of \mathcal{A} in $G_T^{(n)}$ we will denote via $G_T^{(n)}(1)$. Delete then graph $G_T^{(n)}(1)$ from the graph $G_T^{(n)}$ (remove nodes of the graph $G_T^{(n)}(1)$ and the associated edges). In the remaining graph, using algorithm \mathcal{A} , we will allocate the maximum complete subgraph $G_T^{(n)}(2)$, etc. Thus, complete subgraphs $G_T^{(n)}(1), \dots, G_T^{(n)}(\hat{k})$ define sets of nodes $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$. These are the first \hat{k} steps of the algorithm \mathcal{B} .

After \mathcal{A} has finished its work algorithm \mathcal{B} has to correct the sets $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$. The procedure is as follows. If the node x does not belong to any of sets $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$, then we calculate its connectivity with each of these sets. Let \hat{M}_i be the set having maximum connectivity to the node x . Attach this node to \hat{M}_i . Repeat this procedure with each node that does not belong to any of sets $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$. If ambiguity occurs, any of the allowed sets is arbitrarily selected. If x is an isolated node, then any of sets $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$ is arbitrarily selected.

After all the nodes from the set X are distributed across sets $\hat{M}_1, \dots, \hat{M}_{\hat{k}}$, let's find the area of values of parameter T at which $k = \hat{k}$ and with probability which tends to 1, $M_i = \hat{M}_i, i = 1, \dots, k$.

Lemma 3. With probability which tends to 1, for each $i = 1, \dots, \hat{k}$ there exists a number $j = 1, \dots, k$ such that $\hat{M}_i = M_j$.

Proof. Consider the set \hat{M}_1 and assume that the proposition of lemma 3 for \hat{M}_1 is not performed. This means that the set \hat{M}_1 consists of several subsets of

sets $M_i, i = 1, \dots, k$. Since the graph $G_T^{(n)}(1)$ is the maximum complete subgraph, and the subgraphs of the graph $G_T^{(n)}$ generated on sets $M_i, i = 1, \dots, k$, are complete graphs, the number of nodes in the graph $G_T^{(n)}(1)$ is greater than or equal to the number of nodes in each set $M_i, i = 1, \dots, k$. That is, there is $\delta > 0$ such that $|\hat{M}_1| \geq n\delta$. Then there exists the number $j = 1, \dots, k$ such that $|\hat{M}_1 \cap M_j| \geq n\varepsilon, \varepsilon > 0$. In fact, if the intersection \hat{M}_1 with each set $M_i, i = 1, \dots, k$, is equal to $o(n)$, it contradicts the condition $|\hat{M}_1| \geq n\delta$ because $k < \infty$.

In addition, there is a node $x \in \hat{M}_1 \cap M_i$ for some $i \neq j$. For a pair $(x, \hat{M}_1 \cap M_i)$, an estimate similar to (3) is fair, namely

$$\left(\frac{T - N}{|M_j \cap \hat{M}_1|} \right) \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j \cap \hat{M}_1|}, \quad (6)$$

where $i \neq j$. It follows that for (6) the estimation (5) is fair. It means that there are no nodes in the set \hat{M}_1 that do not belong to M_j with probability which tends to 1.

According to the algorithm \mathcal{B} , the graph $G_T^{(n)}(1)$ is removed from the graph $G_T^{(n)}$ together with the set M_j . Since the following graph is the maximum complete graph on the set of nodes $X \setminus M_j$, all reasonings fair to \hat{M}_1 are true for \hat{M}_2 . Then there is a number $i = 1, \dots, k$ such that $\hat{M}_2 = M_i$ with probability which tends to 1.

Obviously, at the k -th step, the maximum complete subgraph is defined on the remaining set M_s and all elements of the set X are exhausted. It follows that $k = \hat{k}$ and after the corresponding renumbering, equality $M_i = \hat{M}_i, i = 1, \dots, k$ is fair. Lemma 3 is proved.

Theorem. Let $n \rightarrow \infty, 1 - \alpha = \frac{C}{\ln n}$, where $C = \text{const}, C > 0$,

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1)).$$

Then

$$P(k = \hat{k}, M_i = \hat{M}_i, i = 1, \dots, k) \rightarrow 1.$$

Proof. The proposition of the theorem follows from lemmas 1-3 and the description of the algorithm \mathcal{B} of graphs formation $G_T^{(n)}(1), \dots, G_T^{(n)}(\hat{k})$.

It follows from the theorem that at large n there is an area of values of T in which concepts are correctly recovered with great probability. At the same time, at very small and very large values of T , the structure of concepts cannot be determined in principle.

CONCLUSION

1. In the paper it is proved that in the process of using machine learning to recover concepts in the original small sample data, bad results are obtained not only when the set of accumulated samples is small, but also when the set of accumulated samples becomes very large. It is shown that there is a relation between the parameters of the number of possible small samples and the accumulated data, in which

the structure of the set of concepts is uniquely determined with probability which tends to 1.

It follows that it is not true that in all tasks increasing of the set of accumulated data improves the quality of learning. Therefore, it is necessary to highlight the parameter value areas when the quality of training is the best.

2. It is convenient to use asymptotic approximations of probability distribution in the series scheme to identify parameter value areas that define the structure of concepts better.
3. The paper does not consider estimates of the complexity of algorithms for the concepts recovery. Currently, there are many algorithms for allocating complete graphs and comparing the complexities of these algorithms is not part of the task of this paper. However, it should be noted that if it is possible to construct a linear order in each of recovered concepts, the task of classifying a newly incoming small sample is solved quite quickly by known algorithms.
4. The paper does not consider the possibility of building the best estimation of the probability of the correct recovering of the concepts. As noted in paragraph 1 of the Conclusion, the problem of proving the existence of an area of parameters in which the estimation of the probability of the correct recovery of concepts tends to 1 has been solved. However, preliminary studies have shown that limitations (1) on the power of concepts can be weakened.

Acknowledgements

This work was partially supported by the Russian Foundation for Basic Research (grant No. 18-29-03081).

REFERENCES

- Grusho, A., N. Grusho, E. Timonina, and S. Shorgin. 2015. "Possibilities of secure architecture creation for dynamically changing information systems". *Systems and Means of Informatics* 25, No. 3, 78–93.
- Grusho, A., N. Grusho, and E. Timonina. 2019. "The bans in finite probability spaces and the problem of small samples". In *Distributed computer and communication networks*. V.M. Vishnevskiy, K.E. Samouylov, and D.V. Kozyrev (Eds.), Lecture Notes in Computer Science, vol 11965. Springer, Cham, 578–590.
- Etz, K. E., J. A. Arroyo. 2015. "Small Sample Research: Considerations Beyond Statistical Power". *Prev Sci*, 1033–1036.
- Axelsson, S. 1999. "The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection". In *Proc. of the 6th Conference on Computer and Communications Security*, 1–7.
- Tukey, J.W. 1977. *Exploratory data analysis*. Addison Wesley. 711 p.

- Grusho, A., N. Grusho, and E. Timonina. 2016. "Detection of anomalies in non-numerical data". In *Proceedings of 8th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops*. IEEE, Piscataway, N.J., 273–276.
- Jordan, M. I., and T. M. Mitchell. 2015. "Machine learning: Trends, perspectives, and prospects". *Science* 349, Iss. 6245, 255–260.
- Bramley, J.W. 1977. *Constructing the world: Active causal learning in cognition*. London: University College London. PhD Thesis. 361 p.
- Shu, J., X. Zongben, and M. Deyu. 2019. "Small sample learning in big data era". Available at: <https://arxiv.org/abs/1808.04572>.
- Grusho, A. A., M. I. Zabezhailo, N. A. Grusho, and E. E. Timonina. 2019. "Concepts forming on the basis of small samples". *Informatics and applications* 13, Iss. 4, 81–84.
- Kolchin, V. F., B. A. Sevast'yanov, V. P. Chistyakov. 1978. *Random allocations, Scripta Series in Mathematics*. V. H. Winston and Sons, Washington, DC, xi+262 p.
- Shiryayev, A. N. 1984. *Probability*. Addison Wesley. 711 p. Translated from the Russian by R. P. Boas. Graduate Texts in Mathematics, 95. Springer-Verlag, New York, xi+577 p.
- Riordan, John. 1958. *An introduction to combinatorial analysis*. John Wiley and Sons, New York.

AUTHOR BIOGRAPHIES

ALEXANDER A. GRUSHO, Professor (1993), Doctor of Science in physics and mathematics (1990). He is principal scientist at Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences and Professor of Moscow State University.

Research interests: probability theory and mathematical statistics, information security, discrete mathematics, computer sciences.

His email is grusho@yandex.ru.

NICK A. GRUSHO has graduated from the Moscow Technical University. He is Candidate of Science (PhD) in physics and mathematics. At present he works as senior scientist at Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS).

Research interests: probability theory and mathematical statistics, information security, simulation theory and practice, computer sciences.

His email is info@itake.ru.

VLADISLAV V. KULCHENKOV has graduated from the Moscow State University, Faculty of Computational Mathematics and Cybernetics.

Research interests: probability theory and mathematical statistics, machine learning, optimization theory, data mining, financial risk.

His e-mail address is: vlad.kulchenkov@gmail.com.

ELENA E. TIMONINA has graduated from the Moscow Institute of Electronics and Mathematics and obtained the Candidate degree (PhD) in physics and mathematics (1974). She is Doctor in Technical Science (2005), Professor (2007). Now she works as leading scientist in Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS).

Research interests: probability theory and mathematical statistics, information security, cryptography, computer sciences.

Her email is eltimon@yandex.ru.

MICHAEL I. ZABEZHAILO has graduated from the Institute of Physics and Technology and gained the Candidate degree (PhD) in theoretical computer science (1983). He is Doctor of Science in physics and mathematics (2016). Now he works as Head of laboratory in Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

Research interests: mathematical foundations of artificial intelligence, reasoning modeling, information security, theoretical computer sciences.

His email is: zabezhailo@yandex.ru.

A SIMPLE DISPATCHING POLICY FOR MINIMIZING MEAN RESPONSE TIME IN NON-OBSERVABLE QUEUES WITH SRPT POLICY OPERATING IN PARALLEL

Mikhail Konovalov and Rostislav Razumchik
 Institute of Informatics Problems of the FRC CSC RAS
 Vavilova, 44-2, 119333, Moscow, Russia
 Email: mkonovalov@ipiran.ru, rrazumchik@ipiran.ru

KEYWORDS

shortest remaining processing time, server farm, non-observable, load balancing, customer assignment, dispatching, mean response time

ABSTRACT

Consider a non-observable system with a single dispatcher and $N \geq 2$ single server queues operating in parallel and independently. Each queue uses shortest remaining processing time discipline for scheduling the jobs. Jobs from a single flow arrive one by one to the dispatcher, which must immediately route them to one of the queues. Each decision is irrevocable. The dispatcher does not have any online information about the system and can base its decisions only on the job size distribution, job's inter-arrival time distribution, server's speeds, time instants of previously arrived jobs and previous routing decisions. Under these conditions, one is interested in the routing policies, which minimize the job's long-run mean response time. New simple single-parameter policy is proposed which is applicable in case of i.i.d. arrivals and i.i.d. service times and which, according to the numerical experiments, always outperforms the optimal probabilistic policy and may outperform the deterministic policy (under medium and low load).

INTRODUCTION

In this paper consideration is given to the problem of optimal scheduling in parallel non-observable single-server queues each with SRPT (shortest remaining processing time) scheduling discipline, fed by a single flow of jobs. The structure of the system is the following. There is one dispatcher, which does not have a queue for storing the jobs and thus routes arriving jobs immediately to one of the queues. Each queue has infinite capacity and a single server, which serves the jobs one by one according to the SRPT discipline (see Schrage and Miller (1966)). The non-observability means that the dispatcher has only static information about the system: cumulative distribution function (CDF) of job's inter-arrival times, CDF of job's size and servers' speeds. Any online information about the system's state (like queue sizes, remaining work etc.) is not available to the dispatcher.

Under the assumptions made above, one is interested in routing policies¹, which minimize job's long-run mean response time.

For the latest literature on non-observable queues and, in particular, dispatching systems one can refer to Anselmi (2017); Brun (2016); Grosf et al. (2019); Hyytiä et al. (2012); Hassin and Snitkovsky (2017); Konovalov and Razumchik (2018); Lingenbrink and Krishnamurthy (2017) and references therein. Performance analysis of queues with SRPT discipline has been the subject of extensive research in the past and still remains to be (see Grosf et al. (2018, 2019); Kruk (2019); Scully et al. (2019) and references therein).

To our knowledge the the problem of optimal routing in unobservable dispatching systems with SRPT scheduling in queues has not been studied before. Under non-observability only two class of routing policies are available to the dispatcher²: probabilistic and deterministic.

According to a probabilistic (also known as Bernoulli, random) policy a job is routed to the queue n , $1 \leq n \leq N$, with the probability p_n independently of the previous decisions. If the arrival flow is Poisson with rate λ , then the arrival process to each queue remains Poissonian and the queue n is the M/GI/1-SRPT queue with the arrival rate $p_n \lambda$. If the server's speed of the queue n is $v^{(n)}$, the job size distribution is $B(x)$ (with the mean $\mathbf{E}X = \int_0^\infty x dB(x)$), then, whenever the queue n is stable, the stationary mean sojourn time $\mathbf{E}[T_n]^{\text{SRPT}}$ is equal to

$$\mathbf{E}[T_n]^{\text{SRPT}} = \int_0^\infty \frac{\int_0^x u (1 - B(uv^{(n)})) du}{x^2 (1 - \lambda p_n \int_0^x u dB(uv^{(n)}))} dx. \quad (1)$$

The optimal probabilistic routing policy³, further referred to as RND-opt, is the probability distribution (p_1, p_2, \dots, p_N) that minimizes the job's mean response time

$$\sum_{n=1}^N p_n \mathbf{E}[T_n]^{\text{SRPT}} \quad (2)$$

under the constraint $0 \leq p_n \lambda \mathbf{E}(X/v^{(n)}) < 1$ for each n . This problem does not allow an analytical solution but

¹Throughout the paper the terms routing policy, dispatching policy and algorithm are used as synonyms.

²This is in sharp contrast with the partially and fully observable dispatching systems, for which a wider range of routing policies exists.

³This is a well-known traffic/resource allocation problem (see, for example, Ibaraki and Katoh (1988)).

usually can be solved numerically at a satisfactory level⁴. If the arrival process is a general renewal process with the mean $1/\lambda$ and SCV (squared coefficient of variation) equal to C_A^2 , then it is known that the arrival process to each of the N queues is also a renewal process with the mean $1/(\lambda p_n)$ and SCV equal to $1 + (C_A^2 - 1)p_n$. Each queue is thus a GI/GI/1-SRPT queue and (1)–(2) are not valid any more. We are unaware of any analytic or a simulation procedure for computing the optimal N -tuple (p_1, p_2, \dots, p_N) . Yet simulation seems to be the only way to obtain a reasonable solution. The quality of the solution heavily depends on the algorithm (adaptive or meta-heuristic) used to perform the exhaustive search over the domain.

The other feasible class of policies in the considered setting are known as deterministic policies (see Hordijk and van der Laan (2004)). According to a deterministic policy jobs are dispatched according to the prescribed order i.e. according to the known infinite sequence a_1, \dots, a_n, \dots , where a_i is the queue number, whereto the i -th arriving job is routed. Round-Robin policy (RR policy), which rotates the jobs between the queues in the cyclic order, is the special case of such a policy (see Combé and Boxma (1994)). In the heterogeneous system i.e. when the servers' speed are not equal, the RR policy may lead to infinite mean response times and thus one has to pick up a different deterministic policy⁵. Finding the optimal deterministic sequence for N single server queues in parallel, is a difficult problem for which no general procedure exists so far. Yet very good results can be achieved by special deterministic sequences – billiard sequences, which can be constructed using greedy algorithms. One of such algorithms further referred to as SG (Special Greedy) is given in (Hordijk and van der Laan, 2004, p.184). According to the SG policy the i -th arriving job is routed to the queue a_i :

$$a_i = \operatorname{argmin}_{1 \leq n \leq N} \left(\frac{x_n + \kappa^n(i-1)}{p_n} \right), \quad (3)$$

where $\kappa^n(i)$ is equal to the the number of jobs (among the first i jobs) sent to the queue n so far, p_n are the optimal probabilities of the RND policy (i.e. the solution of (2)) and x_n are properly chosen non-negative rational numbers⁶.

The drawback of the SG algorithm as well as of the RND policy is the need to calculate or estimate $(N-1)$ parameters. But as numerical experiments and analytic

⁴Though one may face some technical problems, if $B(x)$ has very large variance.

⁵If the system is homogeneous i.e. $v^{(1)} = v^{(2)} = \dots = v^{(N)}$ then the intuition suggests that the optimal dispatching policy with respect to the stationary mean response time is the one which maximizes the inter-arrival times to each of the N queues. Thus in the homogeneous case the RR policy is optimal policy irrespective of the inter-arrival time distribution and the job size distribution. To our knowledge the proof of this statement is known only in special cases (see, for example, Ephremides et al. (1980); Liu and Towsley (1994)).

⁶For example, if the servers' speeds are all different, then one can put (as in Anselmi and Gaujal (2011)) $x_n = 1$ if n is the fastest server i.e. if $v^{(n)} = \max_{1 \leq j \leq N} v^{(j)}$ and $x_n = 0$ otherwise.

analysis show (see, for example, Anselmi (2017)), deterministic routing is in general more effective than probabilistic routing with respect to the job's long-run mean response time.

The RND and SG policies seem to be the only policies available in the literature, which are applicable in the considered setting. In this paper a new simple routing policy is being proposed, which can outperform both the RND-opt and the SG policy. This new routing policy, further referred to as the AA policy (Arrival Aware), is based on the following intuitive idea: if the dispatcher can memorize its previous routing decisions and also the time instants at which those decisions were made, then this information must help in the problem of reducing the job's long-run mean response time.

The rest of the paper is organized as follows. In the next section the detailed description of system is given. The third section contains the overview of the AA policy, and is followed by the section with the numerical comparison of the AA policy with the RND-opt and SG algorithms. In the concluding section the discussion of the obtained results is presented.

MODEL DESCRIPTION AND ASSUMPTIONS

The system consists of $N \geq 2$ single server infinite capacity queues, operating in parallel. The queues are numbered from 1 to N . The server's speed of queue n is denoted by $v^{(n)}$, $1 \leq n \leq N$. The service discipline employed in each queue is SRPT. Jockeying between queues is not allowed. Inter-arrival times between jobs, which arrive one by one, and their sizes are i.i.d. with the known CDF $A(x)$ and $B(x)$ respectively. Upon receiving a job the dispatcher must immediately route it to one of the queues.

Fix an arbitrary integer $i \geq 1$. Let $0 \leq t_1 < \dots < t_i$ denote the arrival instants of the first i jobs and let y_1, y_2, \dots, y_{i-1} be the first $i-1$ routing decisions i.e. y_j is the server whereto the job arrived at instant t_j was routed. Each y_j takes a value from the set $\{1, 2, \dots, N\}$. For the i -th job arrived at time instant t_i , the dispatcher in order to make a routing decision may use only the following information:

- the values of t_1, t_2, \dots, t_i ,
- the values of y_1, y_2, \dots, y_{i-1} ,
- the inter-arrival time distribution $A(x)$ and the jobs size distribution $B(x)$,
- the values $v^{(1)}, v^{(2)}, \dots, v^{(N)}$ of the servers' speeds.

Online information (like the arriving job size, current queues' sizes etc.) is not available. The objective of the dispatcher is to route jobs in such a way, which minimizes job's long-run mean response time.

OVERVIEW OF THE NEW POLICY

Assume that the system starts working at $t_0 = 0$ and the remaining workload in queue n (including server) at t_0 is equal to $b_n \geq 0$. Denote by $0 < t_1 < \dots < t_i < \dots$ the jobs' arrival instants and by $y_1, y_2, \dots, y_i, \dots$ the routing decisions. Let $\omega_i^{(n)}$ be the sojourn time in the queue n (i.e. the sum of waiting time and service time) of the i -th job arrived at time instant t_i and routed to queue n . The good routing decision y_i would be such, which prescribes to send the i -th job to the queue for which the value $\omega_i^{(n)}$ is minimum⁷. Unfortunately under the SRPT service discipline it seems to be impossible to compute $\omega_i^{(n)}$ and thus the routing decision y_i based on $\omega_i^{(n)}$ cannot be computed as well. The new routing policy, which is being proposed (AA policy), suggests to keep the same rule for choosing y_i but to replace $\omega_i^{(n)}$ with the other value (see $u_i^{(n)}$ below), which can be computed.

Fix the positive real $c > 0$. Let us associate with the i -th job arriving at the dispatcher, N numbers, say $u_i^{(1)}, \dots, u_i^{(N)}$, which are defined recursively as follows:

$$\begin{aligned} \tilde{u}_i^{(n)} &= \max\left(0, u_{i-1}^{(n)} - (t_i - t_{i-1})\right), \quad 1 \leq n \leq N, \quad i \geq 1, \\ u_i^{(n)} &= \begin{cases} \tilde{u}_i^{(\tilde{y}_i)} + \frac{c}{v^{(\tilde{y}_i)}}, & \text{if } n = \tilde{y}_i, \\ \tilde{u}_i^{(n)}, & \text{otherwise,} \end{cases} \end{aligned}$$

where⁸

$$\begin{aligned} \tilde{y}_i &= \operatorname{argmin}_{1 \leq n \leq N} \left(\tilde{u}_i^{(n)} + \frac{c}{v^{(n)}} \right), \\ u_0^{(1)} &= b_1, \dots, u_0^{(N)} = b_N. \end{aligned}$$

Now everything is ready to define the AA policy⁹: route the i -th job to the queue $y_i = \tilde{y}_i$. The pseudo code for the policy is given below (see Algorithm 1). Unlike the RND and SG policies, the AA policy depends only on the single parameter c , which must be somehow estimated¹⁰.

⁷This idea led to some fruitful results for unobservable FIFO queues, see Konovalov and Razumchik (2018).

⁸When $\operatorname{argmin}()$ is being evaluated, ties are broken in the favour of the fastest server and randomly between the fastest servers.

⁹The idea behind the algorithm is the following. Assume that in addition to the considered (primary) system there are $M \geq 1$ analogous (secondary) systems running in parallel, but which are *fully observable*. The m -th secondary system has a single dispatcher, N servers with speeds $v^{(1)}, v^{(2)}, \dots, v^{(N)}$, and the job size distribution of the arriving jobs is equal to $B^{(m)}(x)$. Secondary systems do not have independent job arrivals. Instead jobs' arrivals to all M secondary systems are synchronized with the jobs' arrivals to the original system. Upon arrival of the i -th job to the original system, the i -th job arrives to each secondary system. The dispatcher in the m -th secondary system, independently of other dispatchers, based on the job size distribution $B^{(m)}(x)$ and available remaining workloads in the queues, chooses for the i -th job the queue, which minimizes its virtual sojourn time in the m -th system, say \tilde{y}_i^m . Note that each \tilde{y}_i^m takes a value in the set $\{1, \dots, N\}$. Let y_i^* be the most frequent value among $\tilde{y}_i^1, \dots, \tilde{y}_i^M$. If y_i^* is not unique, then the tie is broken in the favour of the fastest server, and randomly among the fastest servers. Once the y_i^* is chosen all \tilde{y}_i^m are put equal to y_i^* i.e. $\tilde{y}_i^m = y_i^*$, $1 \leq m \leq M$, and the dispatcher in the original system routes the i -th job to server y_i^* . The AA policy is the special case of the described scheme, when $M = 1$ and $B^{(1)}(x) = 0$ for $x \leq c$ and $B^{(1)}(x) = 1$ for $x > c$.

¹⁰In all the numerical examples presented, the values of c were estimated on the trial-and-error basis using Monte-Carlo simulation.

We put further discussion of the AA policy off to the last section and proceed below with some numerical examples, which demonstrate its performance.

Algorithm 1 High-level description of the implementation procedure for the AA policy

```

function NEXTDECISION( $N, v^{(1)}, \dots, v^{(N)}, u_{i-1}^{(1)}, \dots, u_{i-1}^{(N)}, t_i, t_{i-1}, c$ )
  for  $n = 1 \rightarrow N$  do
     $u_i^{(n)} = \max\left(0, u_{i-1}^{(n)} - (t_i - t_{i-1})\right)$ 
  end for
   $y_i = \operatorname{argmin}_{1 \leq n \leq N} \left( u_i^{(n)} + c/v^{(n)} \right)$ 
   $u_i^{(y_i)} = u_i^{(y_i)} + c/v^{(y_i)}$ 
  return  $y_i, u_i^{(1)}, \dots, u_i^{(N)}$ 
end function

```

^a The function $\text{NEXTDECISION}(N, v^{(1)}, \dots, v^{(N)}, u_{i-1}^{(1)}, \dots, u_{i-1}^{(N)}, t_i, t_{i-1})$ returns for the i -th arriving job the routing decision y_i based on the i -th job arrival instant t_i and the arrival instant t_{i-1} of the $(i-1)$ -th job, servers' speeds $v^{(1)}, \dots, v^{(N)}$, auxiliary values $u_{i-1}^{(1)}, \dots, u_{i-1}^{(N)}$ and c .

^b The values $u_0^{(1)}, \dots, u_0^{(N)}$ are the initial remaining workloads in the queues (including server). For example, if the whole system is initially empty $u_0^{(1)} = \dots = u_0^{(N)} = 0$.

^c The positive real value c is the parameter of the algorithm, which must be set manually.

NUMERICAL EXPERIMENT

Numerical results presented below show the values of the job's long-run mean response¹¹ time under the three dispatching policies: RND-opt, SG and AA. The probabilities (p_1, p_2, \dots, p_N) of the RND-opt policy are the solution of (2); the SG policy uses (3) and (p_1, p_2, \dots, p_N) ; the AA policy uses Algorithm 1 and the values of the parameter c from Tables 2 and 4.

Two systems are considered: the system with two servers (see Table 1) and 8 servers (see Table 3). In each case servers have different speeds, the incoming flow of jobs is Poisson and the mean job size EX is equal to 1. The considered job size distributions are: exponential, uniform with $\text{SCV} = 0,083$ and bimodal with $\text{SCV} = 4$.

The numerical results evidence that if the system's load is not heavy, the AA policy (which is based both on the inter-arrival times and the decision history) always outperforms the RND-opt policy and can be also better than the SG policy. The relative gain (with respect to the RND-opt policy) is higher for the low variable job size distributions and shrinks with the increase of the job size variability and the system's size.

Although under heavy load the AA policy is always better than the RND-opt policy, it seems not to be case when compared with the SG policy. Yet, based on our numerical experiments, it is hard¹² to make a conclusion here.

¹¹All the values of the job's long-run mean response (except for the RND-opt policy) were obtained by simulation (within the simulation framework described in Konovalov and Razumchik (2014); Konovalov (2014)).

¹²High variance of the response time under heavy load complicates the matter.

Table 1: Job’s long-run mean response time in the system with 2 servers ($N = 2$). The server’s speeds are $v^{(1)} = 1/3$ and $v^{(2)} = 2/3$. The job arrival flow is Poisson with rate λ . Mean job size $\mathbf{EX} = 1$. The offered load to the whole system is $\rho = \lambda \mathbf{EX} / \sum_{i=1}^2 v^{(i)} = \lambda$.

		$\rho = 0, 1$	$\rho = 0, 2$	$\rho = 0, 3$	$\rho = 0, 4$	$\rho = 0, 5$	$\rho = 0, 6$	$\rho = 0, 7$
Exp(1)	RND-opt	1,627	1,796	2,033	2,328	2,643	3,035	3,595
	SG	1,627	1,796	2,034	2,274	2,487	2,765	3,190
	AA	1,627	1,785	1,970	2,170	2,413	2,723	3,168
U[0,5; 1,5]	RND-opt	1,630	1,808	2,072	2,384	2,731	3,137	3,887
	SG	1,630	1,808	2,071	2,290	2,426	2,695	3,078
	AA	1,623	1,763	1,916	2,088	2,296	2,570	2,984
Bimodal	RND-opt	1,627	1,795	2,031	2,325	2,641	3,037	3,614
	SG	1,627	1,795	2,031	2,292	2,556	2,888	3,383
	AA	1,627	1,790	2,002	2,232	2,503	2,851	3,362

^a Exp(1) – exponentially distributed job size with mean 1.

^b U[0,5; 1,5] – uniformly distributed on [0,5; 1,5] job size.

^c Binomial – job size distribution with two values 0, 5 and 9 with the probabilities 16/17 and 1/17 correspondingly.

Table 2: The values of the parameter c of the AA policy in Table 1.

		$\rho = 0, 1$	$\rho = 0, 2$	$\rho = 0, 3$	$\rho = 0, 4$	$\rho = 0, 5$	$\rho = 0, 6$	$\rho = 0, 7$
Exp(1)		0,705	0,9	1,24	1,24	1,24	1,24	1,2
U[0,5; 1,5]		1,15	1,2	1,15	1,19	1,24	1,21	1,18
Bimodal		0,1	0,6	1,12	1,17	1,25	1,23	1,21

Table 3: Job’s long-run mean response time in the system with 8 servers ($N = 8$). The server’s speeds are $v^{(1)} = 1/36$, $v^{(2)} = 2/36$, $v^{(3)} = 3/36$, $v^{(4)} = 4/36$, $v^{(5)} = 5/36$, $v^{(6)} = 6/36$, $v^{(7)} = 7/36$ and $v^{(8)} = 8/36$. The job arrival flow is Poisson with rate λ . Mean job size $\mathbf{EX} = 1$. The offered load to the whole system is $\rho = \lambda \mathbf{EX} / \sum_{i=1}^8 v^{(i)} = \lambda$.

		$\rho = 0, 1$	$\rho = 0, 2$	$\rho = 0, 3$	$\rho = 0, 4$	$\rho = 0, 5$	$\rho = 0, 6$	$\rho = 0, 7$
Exp(1)	RND-opt	5,45	6,24	7,08	8,05	9,24	10,84	13,14
	SG	5,19	5,67	6,19	6,81	7,59	8,63	10,06
	AA	5,02	5,51	6,07	6,72	7,54	8,63	10,21
U[0,5; 1,5]	RND-opt	5,48	6,29	7,19	8,25	9,60	11,47	14,36
	SG	5,13	5,48	5,86	6,29	6,76	7,47	8,35
	AA	4,79	5,09	5,42	5,84	6,37	7,07	8,06
Bimodal	RND-opt	5,46	6,24	7,07	8,04	9,26	10,86	13,26
	SG	5,32	5,89	6,57	7,35	8,32	9,66	11,50
	AA	5,24	5,86	6,53	7,32	8,32	9,70	11,70

Table 4: The value of the parameter c of the AA policy in Table 3.

		$\rho = 0, 1$	$\rho = 0, 2$	$\rho = 0, 3$	$\rho = 0, 4$	$\rho = 0, 5$	$\rho = 0, 6$	$\rho = 0, 7$
Exp(1)		1,75	1,7	1,45	1,45	1,44	1,336	1,255
U[0,5; 1,5]		1,21	1,23	1,25	1,27	1,27	1,25	1,21
Bimodal		1,5	1,65	1,7	1,59	1,52	1,42	1,4

SUMMARY

According to the numerical experiments, the proposed routing policy outperforms the RND-opt policy across all values of the system’s load. With respect to the deterministic policy SG (see (3)), it gives lower long-run mean response¹³ time under low and medium load. The relative gain of the AA policy with respect to the SG policy is not high (not greater than 5% in the studied cases), and for job size distributions with low variance it is higher than

for those with high variance.

In general, the gain of the AA policy depends on the properties of the job size distribution, on the system’s structure and the value of the policy parameter c . As can be seen from the Tables 2 and 4 the value of c seems to depend on the number of servers, their speeds and the system’s load. So far we could not find a formula for c . More understanding is needed here.

The performance improvement promised in the Tables 1 and 3, comes almost for free. The new policy can be implemented in the dispatcher at very limited costs

¹³And lower variance in most cases.

(see Algorithm 1). Unlike the SG and RND-opt policy, which depend on $N - 1$ parameters, the AA policy depends only on a single parameter, which seems to make the problem of its estimation easier.

It is also worth noticing that the described behaviour of the AA policy (with respect to RND and SG policies) remains the same for i.i.d. inter-arrival times as well as for other service disciplines in the queues (for example, FIFO).

REFERENCES

- Anselmi, J. 2017. Asymptotically optimal open-loop load balancing. *Queueing Systems*. Vol. 87. No. 3-4. Pp. 245–267.
- Anselmi, J. and B. Gaujal. (2011) The price of forgetting in parallel and non-observable queues. *Perform. Eval.* Vol. 68. No. 12. Pp. 1291–1311.
- Brun, O. 2016. Performance of non-cooperative routing over parallel non-observable queues. *Probability in the Engineering and Informational Sciences*. Vol. 30. No. 3. Pp. 455–469.
- Combé, M.B., Boxma, O.J. 1994. Optimization of static traffic allocation policies. *Theor. Comput. Sci.* Vol. 125. No. 1. Pp. 17–43.
- Ephremides, A., P. Varaiya, and J. Walrand. 1980. A simple dynamic routing problem. *IEEE Transactions on Automatic Control*. Vol. 25. No. 4. Pp. 690–693.
- Grosf, I., Z. Scully, and M. Harchol-Balter. 2019. Load balancing guardrails: keeping your heavy traffic on the road to low response times. *SIGMETRICS Perform. Eval. Rev.* 47, 1 (December 2019), 910.
- Grosf, I., Z. Scully, and M. Harchol-Balter. 2018. SRPT for multiserver systems. *Perform. Eval.* Vol. 127-128. Pp. 154–175.
- Hassin, R. and R.I. Snitkovsky. 2017. Strategic customer behavior in a queueing system with a loss subsystem. *Queueing Systems*. Vol. 86. No. 3-4. Pp. 361–387.
- Hordijk, A. and D.A. van der Laan. 2004. Periodic routing to parallel queues and billiard sequences. *Math. Method. Oper. Res.*, 2004. Vol. 59. No. 2. Pp. 173–192.
- Hyttiä, E., S. Aalto, and A. Penttinen. 2012. Minimizing slowdown in heterogeneous size-aware dispatching systems. *SIGMETRICS Perform. Eval. Rev.* Vol. 40. No. 1. Pp. 29–40.
- Ibaraki, T.I. and N. Katoh. 1988. *Resource Allocation Problems*. Cambridge: MIT Press.
- Kruk, Ł. (2019) Diffusion limits for SRPT and LRPT queues via EDF approximations. In: Phung-Duc T., Kasahara S., Wittevrongel S. (eds) *Queueing Theory and Network Applications*. QTNA 2019. Lecture Notes in Computer Science. Vol. 11688. Pp. 263–275.
- Konovalov, M.G. and R.V. Razumchik. 2018 Improving routing decisions in parallel non-observable queues. *Computing*. Vol. 100. No. 10. Pp. 1059–1079.
- Konovalov, M. and R. Razumchik. 2014. Simulation Of Task Distribution In Parallel Processing Systems. *Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems*. Pp. 657–663.
- Konovalov, M.G. 2014. Building a simulation model for solving scheduling problems of computing resources. *Systems and Means of Informatics*. Vol. 24. No. 4. Pp. 45–62. (in Russian)
- Lingenbrink, D. and K. Iyer. 2017. Optimal Signaling Mechanisms in Unobservable Queues with Strategic Customers. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, New York, NY, USA. Pp. 347–347.
- Liu, Z. and D. Towsley. 1994. Optimality of the Round-Robin Routing Policy. *Journal of Applied Probability*. Vol. 31. No. 2. Pp. 466–475.
- Schrage, L.E. and L.W. Miller. 1966. The queue M/G/1 with the shortest remaining processing time discipline. *Oper. Res.* Vol. 14. No. 4. Pp. 670–684.
- Scully, Z., M. Harchol-Balter, and A. Scheller-Wolf. 2019. Simple near-optimal scheduling for the M/G/1. *SIGMETRICS Perform. Eval. Rev.* Vol. 47. No. 2. Pp. 24–26.

AUTHOR BIOGRAPHIES

MIKHAIL KONOVALOV is a Doctor of Sciences in Technics and holds position of the principal scientist 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 research activities are focused on adaptive control of random sequences, modelling and simulation of complex systems. His email address is mkonvalov@ipiran.ru.

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

METHOD FOR BOUNDING THE RATE OF CONVERGENCE FOR ONE CLASS OF FINITE-CAPACITY MARKOVIAN TIME-DEPENDENT QUEUES WITH BATCH ARRIVALS WHEN EMPTY

Anastasiya Kryukova,
Viktoriya Oshushkova
Vologda State University, Russia,
Email: krukovanastya25@mail.ru,
oshushonok@yandex.ru

Alexander Zeifman
Vologda State University, Russia,
Institute of Informatics Problems
of the FRC CSC RAS, Russia,
Vologda Research Center RAS
Email: a_zeifman@mail.ru

Rostislav Razumchik
Institute of Informatics Problems
of the FRC CSC RAS, Russia,
Email: rrazumchik@ipiran.ru

KEYWORDS

Birth and death process, inhomogeneous process, batch arrivals, bounds, ergodicity, queueing systems.

ABSTRACT

Consideration is given to one class of inhomogeneous birth and death processes with finite state space and additional transitions from the origin, which may be used to study the queue-length process in finite-capacity Markovian time-dependent queues with possible batch arrivals when empty. The latter means that customers may arrive in batches only during the periods when the system is idle. All possible transition intensities are allowed to be state-dependent non-random functions of time. Method based on Lyapunov functions, which allows one to obtain ergodicity bounds, is presented. Short numerical example is given.

INTRODUCTION

In this short note we revisit the problem of bounding the rate of convergence to the limiting regime (whenever it exists) of the queue-length process in finite-capacity queues of type $M_n(t)/M_n(t)/1/(S-1)$ (and some other queues, for example, $M_n(t)/M_n(t)/S/0$) with possible batch arrivals when empty (see section 2). The queue-length process in such a queue, further denoted by $X(t)$, can be described by one subclass of continuous-time Markov chains – inhomogeneous birth and death processes with additional transitions from and to origin¹. Under the assumption that the state space is countable, this subclass was studied in Zeifman et al. (2016, 2017); Zeifman, Korotysheva et al. (2017). It was shown (particularly in (Zeifman et al., 2017, Eq. (15))) that under the presence of disasters and some other conditions on batch arrival intensities, it is possible to obtain the ergodicity bounds² using the method based on the logarithmic

¹In order to keep the connection with the past research, we note that in some papers, which consider related Markov chains, transitions “to the origin” are called “mass exodus” and “from the origin” – “resurrection” Li and Zhang (2017) and “mass arrivals” Chen and Renshaw (1997); Zhang and Li (2015).

²And also uniform in time error bounds of truncation that allow the (approximate) calculation of the limiting performance characteristics,

norm of linear operators and special transformations of the intensity matrix. In order to obtain the ergodicity bounds for the considered finite-capacity queue with possible batch arrivals, one could try to the same approach. But in the case of finite state space (and the state space of $X(t)$ is finite) and in the absence of disasters, the ergodicity bounds cannot³ be obtained using the logarithmic norm method. Thus a different approach is needed. In section 3 it is shown, that the method based on Lyapunov functions can be used⁴ to find the upper bound on the rate of convergence. When the transition intensities are periodic the method yields the constant decay parameter (spectral gap). Section 4 concludes the paper with a short numerical example. Since for Markov chains with a finite state space apparently no general method for the construction of Lyapunov functions can be suggested, the results of section 3 may be of independent interest.

BIRTH AND DEATH PROCESS

Let $X(t)$ be an inhomogeneous continuous-time Markov chain with the state space $\mathcal{X} = \{0, 1, 2, \dots, S\}$. A transition, whenever it occurs from state 0, can be to any state $i > 0$, $i \in \mathcal{X}$, and has intensity $q_{0i}(t)$. A transition from state $i > 0$ can be only to neighbouring states, i.e. either to state $(i-1)$ with intensity $\mu_i(t)$ or to state $(i+1)$ with intensity $\lambda_i(t)$. All transition intensities $q_{0i}(t)$, $\mu_i(t)$ and $\lambda_i(t)$ are allowed to be non-random functions of time and, if so, are required to be continuous in t for $t \geq 0$.

Denote by $p_{ij}(s, t) = \Pr\{X(t) = j | X(s) = i\}$, $i, j \geq 0$, $0 \leq s, t \leq t$ transition probabilities of $X(t)$ and by $p_i(t) = \Pr\{X(t) = i\}$ probability that Markov chain $X(t)$ is in state i at time t . Let $\mathbf{p}(t) = (p_0(t), p_1(t), \dots, p_S(t))^T$ be probability distribution vector at time t .

whenever the limiting regime exists, see (Zeifman et al., 2017, Theorem 4.1). It is also worth noticing that the obtained results allow one to perform the analysis of the “inhomogeneous generalization” of the queues considered in Chen and Renshaw (1997, 2004); Li and Zhang (2017); Pakes (1997); Zhang and Li (2015).

³Even though the logarithmic norm does exist, it does not allow one to obtain any meaningful ergodicity bounds.

⁴This method was also successfully applied to a different process in Zeifman et al. (2020).

Given any proper initial condition $\mathbf{p}(0)$, the probabilistic dynamics of $X(t)$ is described by the forward Kolmogorov system of differential equations

$$\frac{d\mathbf{p}(t)}{dt} = A(t)\mathbf{p}(t), \quad (1)$$

where $A(t)$ denotes the transposed intensity matrix, i.e.

$$A(t) = \begin{pmatrix} a_{00}(t) & \mu_1(t) & 0 & 0 & \dots & 0 & 0 \\ a_{10}(t) & a_{11}(t) & \mu_2(t) & 0 & \dots & 0 & 0 \\ a_{20}(t) & \lambda_1(t) & a_{22}(t) & \mu_3(t) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ a_{S0}(t) & 0 & 0 & 0 & \dots & \lambda_{S-1}(t) & -\mu_S(t) \end{pmatrix}.$$

Note that all column sums of $A(t)$ are equal to zero for $t \geq 0$ and thus $A(t)$ is essentially non-negative i.e. all its off-diagonal elements are non-negative for any $t \geq 0$.

ERGODICITY BOUNDS

Throughout the paper by $\|\cdot\|$ we denote the Euclidean norm, i. e., $\|\mathbf{p}(t)\| = \sqrt{\sum_{i \in X} p_i(t)^2}$.

Recall that a Markov chain $X(t)$ is called weakly ergodic, if $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \rightarrow 0$ as $t \rightarrow \infty$ for any initial conditions $\mathbf{p}^*(0)$ and $\mathbf{p}^{**}(0)$, where $\mathbf{p}^*(t)$ and $\mathbf{p}^{**}(t)$ are the corresponding solutions of (1). The rate at which this difference tends to zero is called the rate of convergence.

As it was mentioned in the Introduction the method based on the logarithmic norm (Zeifman et al. (2018)) does not allow one to obtain ergodicity bounds for the considered $X(t)$. In the rest of the section we show that it is possible to find the upper bound on the rate of convergence using Lyapunov functions.

Using the normalization condition it can be verified that the first equation of (1) i.e. $p_0'(t) = a_{00}(t)p_0(t) + \mu_1(t)p_1(t)$ is identical to

$$\frac{dp_0(t)}{dt} = (a_{00}(t) - \mu_1(t))p_0(t) + \mu_1(t) - \mu_1(t) \sum_{i=2}^S p_i(t).$$

Thus the system (1) can be rewritten as

$$\frac{d\mathbf{p}(t)}{dt} = A^*(t)\mathbf{p}(t) + \mathbf{f}(t), \quad (2)$$

where $\mathbf{f}(t) = (\mu_1(t), 0, \dots, 0)^T$.

Let $\mathbf{p}^*(t)$ and $\mathbf{p}^{**}(t)$ be the solutions of (2) corresponding to different initial conditions $\mathbf{p}^*(0)$ and $\mathbf{p}^{**}(0)$. Then for the vector $\mathbf{z}(t) = \mathbf{p}^*(t) - \mathbf{p}^{**}(t) = (z_1(t), z_2(t), \dots, z_{S+1}(t))^T$ we have

$$\frac{d\mathbf{z}(t)}{dt} = A^*(t)\mathbf{z}(t). \quad (3)$$

Fix $S + 1$ positive numbers, say d_1, \dots, d_{S+1} and put $w_i(t) = d_i z_{i-1}(t)$, $1 \leq i \leq S + 1$. Multiply the previous equation from the right by D^{-1} and from the left by D , where $D = \text{diag}(d_1, d_2, \dots, d_{S+1})$. Then (3) in terms of the vector $\mathbf{w}(t) = (w_1(t), w_2(t), \dots, w_{S+1}(t))^T$ has the form:

$$\frac{d\mathbf{w}(t)}{dt} = A^{**}(t)\mathbf{w}(t). \quad (4)$$

where $A^{**}(t) = DA^*(t)D^{-1}$. It is important to notice that the coordinates of $\mathbf{w}(t)$ can be of arbitrary signs i.e. they have no probabilistic meaning.

Let $\mathbf{w}(t)$ be the solution of (4). By differentiating $V(t) = \sum_{k=1}^{S+1} w_k^2(t)$, we obtain

$$\begin{aligned} \frac{dV(t)}{dt} &= \sum_{k=1}^{S+1} 2w_k(t) \frac{dw_k(t)}{dt} = \\ &= -2 \sum_{i=1}^{S+1} \sum_{j=1}^{S+1} (-a_{i-1,j-1}^{**}(t)) w_i(t) w_j(t). \end{aligned} \quad (5)$$

From (5) it follows that if one finds a set of positive numbers $\{d_i, 1 \leq i \leq S + 1\}$ and a function $\beta^*(t)$ satisfying

$$\frac{dV(t)}{dt} \leq -2\beta^*(t)V(t), \quad (6)$$

for any $\mathbf{w}(t)$ being the solution of (4), then

$$\|\mathbf{w}(t)\| \leq e^{-\int_0^t \beta^*(\tau) d\tau} \|\mathbf{w}(0)\|,$$

for any initial condition $\mathbf{w}(0)$.

Assume that $q_{0i}(t)$, $\mu_i(t)$ and $\lambda_i(t)$ do not depend on t . Thus the matrix $A^{**}(t) = A^{**} = (a^{**})_{i,j=0}^S$ is equal to

$$A^{**} = \begin{pmatrix} a_{00} - \mu_1 & 0 & -\mu_1 \frac{d_1}{d_3} & -\mu_1 \frac{d_1}{d_4} & \dots & -\mu_1 \frac{d_1}{d_{S+1}} \\ a_{10} \frac{d_2}{d_1} & a_{11} & \mu_2 \frac{d_2}{d_3} & 0 & \dots & 0 \\ a_{20} \frac{d_3}{d_1} & \lambda_1 \frac{d_3}{d_2} & a_{22} & \mu_3 \frac{d_3}{d_4} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{(S-1)0} \frac{d_S}{d_1} & 0 & 0 & 0 & \dots & \mu_S \frac{d_S}{d_{S+1}} \\ a_{S0} \frac{d_{S+1}}{d_1} & 0 & 0 & 0 & \dots & -\mu_S \end{pmatrix}.$$

By plugging the explicit values of the entries of the matrix A^{**} in (5), we get

$$\begin{aligned} \frac{1}{2} \frac{dV(t)}{dt} &= (a_{00} - \mu_1) w_1^2(t) + \sum_{i=1}^{S-1} a_{ii} w_{i+1}^2(t) + \\ &+ a_{10} \frac{d_2}{d_1} w_1(t) w_2(t) - \mu_S w_{S+1}^2(t) + \\ &+ \sum_{i=2}^S \left(a_{i0} \frac{d_{i+1}}{d_1} - \mu_1 \frac{d_1}{d_{i+1}} \right) w_1(t) w_{i+1}(t) + \\ &+ \sum_{i=2}^S \left(\lambda_{i-1} \frac{d_{i+1}}{d_i} + \mu_i \frac{d_i}{d_{i+1}} \right) w_i(t) w_{i+1}(t). \end{aligned}$$

Put $d_i = d_1 \sqrt{\frac{\mu_1}{a_{i0}}}$ for $1 \leq i \leq S + 1$. Thus each term with the coefficient $(a_{i0} \frac{d_{i+1}}{d_1} - \mu_1 \frac{d_1}{d_{i+1}})$ is equal to zero and the previous equality can be rewritten in the form

$$\begin{aligned} \frac{1}{2} \frac{dV(t)}{dt} &= (a_{00} - \mu_1) w_1^2(t) + \sum_{i=1}^{S-1} a_{ii} w_{i+1}^2(t) + \\ &+ a_{10} \frac{d_2}{d_1} w_1(t) w_2(t) - \mu_S w_{S+1}^2(t) + \\ &+ \left(\lambda_1 \sqrt{\frac{\mu_1}{a_{20}}} \frac{d_1}{d_2} + \mu_2 \sqrt{\frac{a_{20}}{\mu_1}} \frac{d_2}{d_1} \right) w_2(t) w_3(t) + \\ &+ \sum_{i=3}^S \left(\lambda_{i-1} \sqrt{\frac{a_{i-1,0}}{a_{i0}}} + \mu_i \sqrt{\frac{a_{i0}}{a_{i-1,0}}} \right) w_i(t) w_{i+1}(t). \end{aligned}$$

Finally, by applying the same reasoning as in (Zeifman et al., 2020, Theorem 4), one can show that there exists a positive number β^* and a set of numbers $\{\alpha_i, 1 \leq i \leq S + 1\}$ such that

$$\begin{aligned} \frac{dV(t)}{dt} = & -2\beta^* \sum_{k=1}^{S+1} w_k^2(t) - \\ & -2 \sum_{k=1}^S (\alpha_k w_k(t) - \alpha_{k+1} w_{k+1}(t))^2. \end{aligned} \quad (7)$$

Thus the following bound on the rate of convergence holds:

$$\|\mathbf{w}(t)\| \leq e^{-\beta^* t} \|\mathbf{w}(0)\|, \quad (8)$$

where β^* is the decay parameter (spectral gap) of $X(t)$. Unfortunately the closed-form expression for β^* cannot be obtained. For a given matrix A^{**} it is computed algorithmically-wise (see (Zeifman et al., 2020, Section 4)). Note that since the state space of $X(t)$ is finite, (8) can be rewritten in the form:

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \leq M e^{-\beta^* t}, \quad (9)$$

for some positive M and thus β^* is the decay parameter of the original Markov chain $X(t)$.

Now assume that all the transition intensities $q_{0i}(t)$, $\mu_i(t)$ and $\lambda_i(t)$ are non-random functions of time t . For simplicity we also assume that the transition rates are periodic with period 1 and thus $X(t)$ may have a periodic limiting regime. Let $a_{i0}(t) = 2 + \cos(2\pi t)$, $\lambda_i(t) = \lambda(t) = 2 + \sin(2\pi t)$, $\mu_i(t) = i(2 + \cos(2\pi t)) = i\mu(t)$ for $1 \leq i \leq S$.

Then the non-zero entries of the matrix $A(t)$ are:

$$\begin{aligned} a_{i0}(t) &= 2 + \cos(2\pi t) = \mu(t), \\ a_{i-1,i}(t) &= i\mu(t), \\ a_{i+1,i}(t) &= \lambda(t), \\ a_{00}(t) &= -S\mu(t), \\ a_{i,i}(t) &= -(\lambda(t) + i\mu(t)). \end{aligned}$$

Put $d_i = 1$ for all $1 \leq i \leq S + 1$. By computing the entries of the matrix $A^{**}(t)$ and plugging them into (5), we get

$$\begin{aligned} -\frac{1}{2} \frac{dV(t)}{dt} = & (S + 1)\mu(t)w_1^2(t) - \mu(t)w_1(t)w_2(t) + \\ & + \sum_{i=2}^S (\lambda(t) + (i - 1)\mu(t)) w_i^2(t) + S\mu(t)w_{S+1}^2(t) - \\ & - \sum_{i=2}^S (\lambda(t) + i\mu(t)) w_i(t)w_{i+1}(t). \end{aligned}$$

The right part of the previous relation can be bounded from below i.e. it can be shown that

$$\begin{aligned} -\frac{1}{2} \frac{dV(t)}{dt} \geq & S\mu(t)w_1^2(t) + \\ & + \sum_{i=2}^S k_i(t) (w_i(t) - w_{i+1}(t))^2 + \\ & + \mu(t) \left[\frac{3}{2} (w_1^2(t) + w_2^2(t)) + \left(\frac{w_1(t)}{\sqrt{2}} - \frac{w_2(t)}{\sqrt{2}} \right)^2 \right]. \end{aligned}$$

where $k_i(t) = \min \{\lambda(t), i\mu(t)\}$.

From here, using the same arguments as in the proof of (Zeifman et al., 2020, Theorem 4) and using the fact that the intensities are periodic functions, one can establish the existence of a positive constant β^* such that the bounds (8) and (9) on the rate of convergence hold. Just like in the homogeneous case the closed-form expression for β^* cannot be obtained and for a given matrix A^{**} the value of β^* is computed algorithmically-wise.

NUMERICAL EXAMPLE

Consider the $M_t/M_t/1/(S - 1)$ queue with FIFO service and batch arrivals when empty. Let $X(t)$ be the queue-length process. If at time t there is at least one customer in the system then new customers arrive according to inhomogeneous Poisson process with intensity $\lambda(t)$. But if at time t the system is empty ordinary customers arrive in bulk (or groups) in accordance with a inhomogeneous Poisson process of intensity $S^{-1}\lambda(t)$ for a group of size n , $n = 1, 2, \dots, S$. Whenever server becomes free customer from the queue (if there is any) enters server and its service time has exponential distribution with parameter $\mu(t)$.

Let the maximum number of customers in the system be equal to $S = 100$. Denote by $E(t, k) = E(X(t)|X(0) = k)$ the conditional expected number of customers in the queue (including server) at instant t , provided that initially (at instant $t = 0$) k customers were present. The probability of the empty queue $p_0(t)$ and the values of $E(t, k)$, computed using the ergodicity bounds obtained above, are shown in Fig. 1–4.

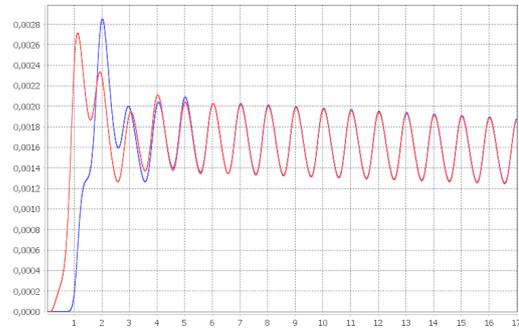


Figure 1: Probability of empty queue $p_0(t)$ for $t \in [0, 17]$ with initial conditions $X(0) = 35$ (blue) and $X(0) = 10$ (red).

ACKNOWLEDGEMENTS

This research was supported by Russian Science Foundation under grant 19-11-00020.

REFERENCES

Chen, A., E. Renshaw. 1997. The $M/M/1$ queue with mass exodus and mass arrivals when empty. *Journal of Applied Probability*, **34**(1), 192–207.

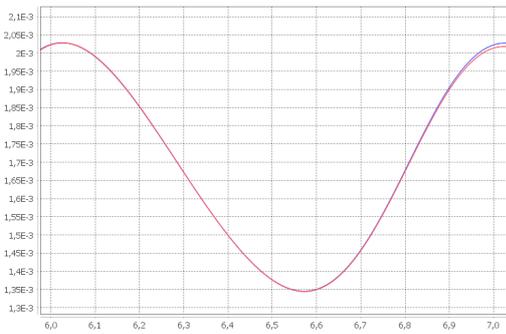


Figure 2: Probability of empty queue $p_0(t)$ for $t \in [6, 7]$ with initial conditions $X(0) = 35$ (blue) and $X(0) = 10$ (red).

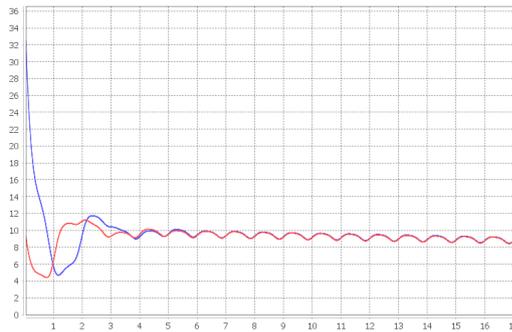


Figure 3: Expected number of customers in the system $E(t, k)$ for $t \in [0, 17]$ with initial conditions $X(0) = 35$ (blue) and $X(0) = 10$ (red).

Chen, A., E. Renshaw. 2004. Markovian bulk-arriving queues with state-dependent control at idle time. *Advances in Applied Probability*, **36** (2), 499–524.

Li, J., L. Zhang. 2017. $M^X/M/c$ Queue with catastrophes and state-dependent control at idle time. *Frontiers of Mathematics in China*, **12**(6), 1427–1439.

Pakes, A.G. 1997. Killing and resurrection of Markov processes. *Stochastic Models*, **13** (2), 255–269.

Zeifman, A., Y. Satin, A. Korotysheva, V. Korolev, V. Bening. 2016. On a class of Markovian queuing systems described by inhomogeneous birth-and-death processes with additional transitions. *Doklady Mathematics*, **94**, 502–505.

Zeifman, A., A. Korotysheva, Y. Satin, R. Razumchik, V. Korolev, S. Shorgin. 2017. Ergodicity and truncation bounds for inhomogeneous birth and death processes with additional transitions from and to origin. *Stochastic Models*, **33**, 598–616.

Zeifman, A., A. Korotysheva, Y. Satin, K. Kiseleva, V. Korolev, S. Shorgin. 2017. Bounds for Markovian queues with possible catastrophes. In *Proceedings of 31st European Conference on Modelling and Simulation ECMS 2017*, 628–634.

Zeifman, A., R. Razumchik, Y. Satin, K. Kiseleva, A. Korotysheva, V. Korolev. 2018. Bounds on the rate of convergence for one class of inhomogeneous Markovian queuing models with possible batch arrivals and services. *International Journal of Applied Mathematics and Computer Science*. **28**, 141–154.

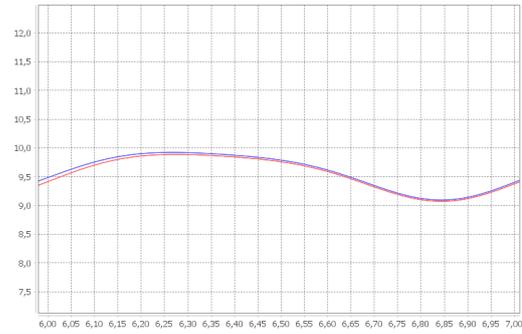


Figure 4: Expected number of customers in the system $E(t, k)$ for $t \in [6, 7]$ with initial conditions $X(0) = 35$ (blue) and $X(0) = 10$ (red).

Zeifman, A., Y. Satin, A. Kryukova, R. Razumchik, K. Kiseleva, G. Shilova. 2020. On the Three Methods for Bounding the Rate of Convergence for some Continuous-time Markov Chains. *International Journal of Applied Mathematics and Computer Science*. **30**, arXiv preprint arXiv:1911.04086.

Zhang, L., J. Li. 2015. The M/M/c queue with mass exodus and mass arrivals when empty. *Journal of Applied Probability*, **52**, 990–1002.

AUTHOR BIOGRAPHIES

ANASTASIYA KRYUKOVA is a senior lecturer at Vologda State University. Her current research activities focus on queueing theory. Her email is krukovanastya25@mail.ru

VIKTORIYA OSHUSHKOVA is a graduate student, Department of Applied Mathematics, Vologda State University. Her email is oshushonok@yandex.ru

ALEXANDER ZEIFMAN is a Doctor of Science in physics and mathematics, professor and the Head of Department of Applied Mathematics at Vologda State University. He also holds the senior scientist position at the Institute of Informatics Problems of the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences and the principal scientist position at Vologda Science Center of the Russian Academy of Sciences. His current research activities focus on inhomogeneous Markov chains and queueing theory. His email is a_zeifman@mail.ru.

ROSTISLAV RAZUMCHIK received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as a leading research fellow at the Institute of Informatics Problems of the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences. His current research activities are focused on queueing theory and its applications for performance evaluation of stochastic systems. His email is rrazumchik@ipiran.ru.

AUTHOR INDEX

- 91 Agoston, Kolos C.
177 Ali, Farid
228 Aranha, Claus
256 Bagi, Katalin
73, 78 Bagirova, Anna
308 Banares-Alcantara, Rene
56 Bandinelli, Romeo
368 Bhattacharya, Sourav
56 Bindi, Bianca
129 Biro, Peter
168 Boeling, Jari
65 Boros, Eszter
11 Bozhkova, Valentina P.
111 Brzeczek, Tomasz
243 Buffoli, Andrea
384 Campanile, Lelio
243 Ceresoli, Federico
287, 334 Claus, Thorsten
235 Collonval, Frederic
228 Cunha, Sergio
228 da Silva Torres, Ricardo
376 Danysz, Jerry
361 De Arcangelis, Lucilla
214 de Oliveira, Felipe F.
376 del Rosal, Victor
153 Denkova, Zapryana
146, 153 Denkova-Kostova, Rositsa
41 Dietrich, Alexander
250 Doczi, Martin O.
328 Doka, Tamas
104 Dudas, Fanni
177 El-Kilany, Ayman
25, 32 El-Mihoub, Tarek
11 Ershov, Egor I.
48 Fabry, Jan
56 Fani, Virginia
207 Fonseca, Icaro A.
301 Fottner, Johannes
207, 214 Gaspar, Henrique M.
221
116 Gelanyi, Ildiko
368, 376 Gonzalez-Velez, Horacio
146, 153 Goranov, Bogdan
97 Gribaudo, Marco
393 Grusho, Alexander A.
393 Grusho, Nick A.
129 Gyetvai, Marton
294 Haag, Stefan
168 Haghbayan, Mohammad-H.
19 Hameed, Ibrahim A.
190 Hassani, Marwan
235 Hatledal, Lars I.
287, 334 Herrmann, Frank
328 Horak, Peter
361, 384 Iacono, Mauro
139 Immonen, Eero
139 Immonen, Paula
146 Ivanova, Kristina
354 Jakobik, Agnieszka
84 Juhasz, Peter
104 Juhasz, Peter
275 Kaczorek, Tadeusz
19 Karlsen, Anniken Th.
183 Kelleher, John D.
161 Kleppe, Paul S.
129 Klimentova, Xenia
48 Klimova, Anna M.
398 Konovalov, Mikhail
146, 153 Kostov, Georgi
197 Krauthann, Raphael A.
197 Kruse, Tobias
403 Kryukova, Anastasiya
84 Kuerthy, Gabor
393 Kulchenkov, Vladislav V.
48 Kuncova, Martina
228 Lavinias, Yuri
301 Lienert, Thomas
361 Lippiello, Eugenio
177 Mahmoud, Ayat
97 Manini, Daniele
384 Marulli, Fiammetta
384 Mastroianni, Michele
266 Mate, Tamas
177 Mazen, Sherif
322 Meier, Klaus-Juergen
221 Monteiro, Thiago G.
228 Moura, Felipe

197	Mueller, Hinnerk J.	250	Szoedy, Robert
139	Murashko, Kirill	65	Sztano, Gabor
122	Muratov-Szabo, Kira	345	Tacchella, Alberto
19	Nasar, Wajeeha	345	Tacchella, Armando
116	Nemeth, Andras O.	168	Tahir, Anam
104	Nemeth-Durko, Emilia	354	Tchorzewski, Jacek
11	Nikolaev, Dmitry P.	153	Teneva, Desislava
214	Nishimoto, Kazuo	334	Terbrack, Hajo
25, 32	Nolle, Lars	25, 32	Tholen, Christoph
41	Nowitzki, Mario	393	Timonina, Elena E.
256, 261	Orosz, Akos	5	Tolujevs, Jurijs
403	Oshushkova, Viktoriya	287	Trost, Marco
322	Pankratz, Vincent	41	van de Sand, Ron
129	Pedroso, Joao P.	122	Varadi, Kata
129	Pettersson, William	116	Varga, Erzsebet T.
228	Pinciroli Vago, Nicolo O.	91	Vaskoevi, Agnes
97	Pironti, Marco	129	Viana, Ana
97	Pisano, Paola	84	Vidovics-Dancs, Agnes
168	Plosila, Juha	301	Wenzler, Florian
214	Prata Vieira, Daniel	315	Wozniak, Adrian P.
122	Prepuk, Andrea	393	Zabekhailo, Michael I.
308	Qi, Wenchan	403	Zeifman, Alexander
25	Ralle, Oliver	221, 235	Zhang, Houxiang
139	Ranta, Samuli	250, 261	Zwierczyk, Peter T.
197	Rausch, Peter	266	
398, 403	Razumchik, Rostislav		
41	Reiff-Stephan, Joerg		
161	Rekdalsbakken, Webjoern		
228	Rodrigues, Daniele		
25	Rofallski, Robin		
183	Rogers, Eoin		
183	Ross, Robert J.		
5	Saifutdinov, Farid		
97	Scuotto, Veronica		
322	Selmair, Maximilian		
11	Shepelev, Denis A.		
146, 153	Shopska, Vesela		
73, 78	Shubat, Oksana		
294	Simon, Carlo		
221	Skourup, Charlotte		
139	Sovela, Janne		
190	Spennath, Yorick		
280	Steglich, Mike		
197	Stumpf, Michael		
84	Szaz, Janos		
122	Szodorai, Melinda		