Copyright **© ECMS2018**

Publisher: **European Council for Modelling and Simulation**

ISBN: 978-0-9932440-6-3 (print) **ISBN:978-0-9932440-7-0 (CD)**

ISSN 2522-2414 (print) **ISSN 2522-2430 (CD)**
ISSN 2522-2422 (online)

Cover pictures: **front side: Foto/Montage © Jens Werner**
back side: Foto © Axel Biewer

Printed by: **Digitaldruck Pirrot GmbH**
66125 Sbr.-Dudweiler,
Germany

PROCEEDINGS

32nd European Conference on Modelling and Simulation ECMS 2018

May 22nd – May 25th, 2018
Wilhelmshaven, Germany

Edited by:

Lars Nolle

Alexandra Burger

Christoph Tholen

Jens Werner

Jens Wellhausen

Organized by:

ECMS - European Council for Modelling and Simulation

Hosted by:

Jade University of Applied Sciences, Wilhelmshaven

Sponsored by:

Jade University of Applied Sciences, Wilhelmshaven

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 2018 ORGANIZATION

Conference Chair

Lars Nolle

Jade University of Applied Sciences
Germany

Conference Co-Chairs

Alexandra Burger

Jade University of Applied Sciences
Germany

Christoph Tholen

Jade University of Applied Sciences
Germany

Programme Chair

Jens Werner

Jade University of Applied Sciences
Germany

Programme Co-Chair

Jens Wellhausen

Jade University of Applied Sciences
Germany

President of European Council for Modelling and Simulation

Khalid Al-Begain

University of South Wales, United Kingdom

Vice-President of European Council for Modelling and Simulation

Lars Nolle

Jade University of Applied Sciences, Wilhelmshaven, Germany

Managing Editor

Martina-Maria Seidel

St. Ingbert, Germany

COUNCIL BOARD 2018

Khalid Al-Begain University of South Wales United Kingdom	President of European Council for Modelling and Simulation and Past President of European Council for Modelling and Simulation (2006-2010)
Lars Nolle Jade University of Applied Sciences, Wilhelmshaven Germany	Vice-President of European Council for Modelling and Simulation elected in 2016
Evtim Peytchev Nottingham Trent University United Kingdom	Past President of European Council for Modelling and Simulation (2012-2014) Treasurer (2015-today)
Andrzej Bargiela United Kingdom	Past President of European Council for Modelling and Simulation (2002-2006, 2010-2012)
Eugène Kerckhoffs The Netherlands	Founder, Past President and Historian of ECMS
Christoph Tholen Jade University of Applied Sciences, Wilhelmshaven Germany	Conference Co-Chair of ECMS 2018
Jens Werner Jade University of Applied Sciences, Wilhelmshaven Germany	Programme Chair of ECMS 2018
Kata Váradi Corvinus University of Budapest, Corvinus Business School Hungary	Conference Co-Chair of ECMS 2017
Peter T. Zwierczyk Budapest University of Technology and Economics Hungary	Programme Co-Chair of ECMS 2017

EDITORIAL BOARD 2018

INTERNATIONAL PROGRAMME COMMITTEE

Finance and Economics and Social Science

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

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

Ágnes Vidovics-Dancs
Corvinus University of Budapest, Corvinus Business School , Hungary

Simulation in Industry, Business, Transport and Services

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

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

Romeo Bandinelli
University of Florence, Italy

Simulation of Intelligent Systems

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

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

Modelling, Simulation and Control of Technological Processes

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

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

Simulation and Optimization

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

Co-Chairs:
Thorsten Claus
Technical University Dresden, Germany

Michael Manitz
University of Duisburg-Essen, Germany

Simulation of Fluid-Mechanically Effective Microstructures and Combustion Processes

Track Chair: **Karsten Oehlert**
Jade University of Applied Sciences, Germany

Co-Chairs:
Peter Charles
FH Bielefeld University of Applied Sciences, Germany

Konrad M. Hartung
Jade University of Applied Sciences, Germany

Multiphysical Finite Element Simulation

Track Chair: **Tamara Bechtold**
Jade University of Applied Sciences, Germany

Co-Chairs:
Chengdong Yuan
Jade University of Applied Sciences, Germany

Siyang Hu
Jade University of Applied Sciences, Germany

Finite – Discrete - Element Simulation

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

Co-Chair: **János P. Rádics**
Budapest University of Technology and Economics, Hungary

High Performance Modelling and Simulation

Track Chair: **Mauro Iacono**
Seconda Università degli Studi di Napoli, Italy

Co-Chairs:
Agnieszka Jakobik
Cracow University of Technology, Poland

Rostislav V. Razumchik
Institute of Informatics Problems FRC CSC RAS,
Russian Federation

Honorary Track Chair:
Joanna Kolodziej
Cracow University of Technology, Poland

IPC Members in Alphabetical Order

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

Anna Bagirova, Ural Federal University, Russian Federation

Enrico Barbierato, Politecnico di Milano, Italy

Dénes Bencsik, Robert Bosch Kft., Hungary

Simona Bernardi, Centro Universitario de la Defensa, Zaragoza, Spain

Jan Bielański, AGH University of Science and Technology, Poland

Aleksander Byrski, AGH University of Science and Technology, Poland

Damián F. Cerero, Universidad de Sevilla, Spain

Kevin Chalmers, Edinburgh Napier University, United Kingdom

Petr Chalupa, Tomas Bata University in Zlin, Czech Republic

Adriana E. Chis, National College of Ireland Dublin, Ireland

Marina Chukalina, Russian Academy of Science, Russian Federation

Catherine Cleophas, RWTH Aachen, Germany

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

Ciprian Dobre, University Politehnica of Bucharest, Romania

Sergei Dudin, Belarusian State University, Belarus

Virginia Fani, University of Florenz, Italy

Thomas Farrenkopf, Techn. Hochschule Mittelhessen Friedberg, Germany

Robert Forstner, SimPlan AG in Regensburg, Germany

Michael Fuchs, OptWare GmbH in Regensburg, Germany

Ildikó Gelányi, Corvinus University of Budapest, Hungary

Andreas Goldmann, ITV Leibniz Universität Hannover, Germany

Daniel Grzonka, Cracow University of Technology, Poland

Michael Guckert, Techn. Hochschule Mittelhessen Friedberg, Germany

Gábor Hámori, Corvinus University of Budapest, Hungary

Florian Herbst, ITV Leibniz Universität Hannover, Germany

Daniel Hernandez, Sosa University of Las Palmas de Gran Canaria, Spain

Dennis Hohlfeld, University of Rostock, Germany

Markus Höltermann, ITV Leibniz Universität Hannover, Germany

Daniel Honc, University of Pardubice, Czech Republic

Teruaki Ito, The University of Tokushima, Japan

Jácint Juhász, Babes-Bolyai University, Romania

Petra Kalfmann, Corvinus University of Budapest, Hungary

Bogumil Kamiński, Warsaw School of Economics, Poland

István Keppler, Szent István University, Hungary

György Kerényi, Budapest University of Technology, Hungary

Petia Koprinkova, Bulgarian Academy of Sciences, Bulgaria

Victor Korolev, Moscow State University, Russian Federation

Ina Kortemeier, University Duisburg-Essen, Germany

Martin Kotyrba, University of Ostrava, Czech Republic

Julia Kruk, Belarusian National Technical University, Belarus

Mateusz Krzysztoń, Warsaw University of Technology and NASK, Poland

Marek Kubalcik, Tomas Bata University in Zlin, Czech Republic

Frederick Lange, Maschinenfabrik Reinhausen Regensburg, Germany

Alexander H. Levis, George Mason University Washington, USA

Boris Lohmann, Technical University in Munich, Germany

Andrea Marin, Università Ca'Foscari di Venezia, Italy

Agnieszka Mars, Jagiellonian University Cracow, Poland

Radomil Matousek, Brno University of Technology, Czech Republic

Radek Matusu, Tomas Bata University in Zlin, Czech Republic

Nicolas Meseth, University of Applied Sciences in Osnabrueck, Germany

Frank Morelli, University of Applied Sciences in Pforzheim, Germany

Christian Müller, University of Applied Sciences in Wildau, Germany

Maximilian Munninger, University Duisburg-Essen, Germany

Valeriy Naumov, Service Innovation Research Institute, Finland

Catalin Negru, University Politehnica of Bucharest, Romania

Ewa Niewiadomska-Szynkiewicz, Warsaw Univ of Techn and NASK, Poland

Lars Nolle, Jade University of Applied Sciences Wilhelmshaven, Germany

Jakub Novak, Tomas Bata University in Zlin, Czech Republic

István Oldal, Szent István University, Hungary

Mihály Ormos, Corvinus University of Budapest, Hungary

Francesco Palmieri, Università degli Studi di Salerno, Italy

Libor Pekar, Tomas Bata University in Zlin, Czech Republic

Dóra Petróczy, Corvinus University of Budapest, Hungary

Michal Pluhacek, Tomas Bata University in Zlin, Czech Republic

Florin Pop, University Politehnica of Bucharest, Romania

Simone Righi, Hungarian Academy of Sciences, Hungary

Boris Rohal-Ilkiv, Slovak University of Technology in Bratislava, Slovakia

Michael Römer, Martin-Luther-Universität Halle-Wittenberg, Germany

Juliette Rouchier, GREQAM-CNRS, France

Konstantin Samouylov, Peoples' Friendship University, Russian Federation

Leonid Sevastyanov, Peoples' Friendship University, Russian Federation

Oleg Shestakov, Moscow State University, Russian Federation

Sergey Ya. Shorgin, Instit of Inform. Problems FRC CSC RAS, Russian Federation

Markus Siegle, Universität der Bundeswehr München, Germany

Stelios Sotiriadis, University of Toronto, Canada

Grażyna Suchacka, Opole University, Poland

Katarzyna Sum, Warsaw School of Economics, Poland

Grzegorz Szafrański, Institute of Finance, Lodz University, Poland

János Száz, Corvinus University of Budapest, Hungary

Magdalena Szmajduch, Cracow University of Technology, Poland

Piotr Szuster, Cracow University of Technology, Poland

Pawel Szynkiewicz, Polish Academy of Science Warsaw, Poland

Armando Tacchella, Università degli Studi di Genova, Italy

Kristóf Tamás, Corvinus University of Budapest, Hungary

Enrico Teich, Technical University of Dresden, Germany

Marco Trost, Technical University of Dresden, Germany

Christopher Tubb, University of South Wales, United Kingdom

Tobias Uhlig, Universität der Bundeswehr München, Germany

Nikolai Ushakov, Norwegian University of Science and Technology, Norway

Ward van Breda, VU University of Amsterdam, The Netherlands

Enrico Vicario, Università degli Studi di Firenze, Italy

Andrea Vinci, CNR – National Research Council of Italy, Italy

Narayanan Viswanath, Government Engineering College Thrissur, India

Jaroslav Vitku, GoodAI, Czech Republic

Thorsten Vitzthum, University of Applied Sciences in Regensburg, Germany

Eva Volna, University of Ostrava, Czech Republic

Andrzej Wilczyński, Cracow University of Technology, Poland

Victor Zakharov, Inst of Inform Problems FRC CSC RAS, Russian Federation

Alexander Zeifman, Vologda State University, Russian Federation

Bihary Zsolt, Corvinus University of Budapest, Hungary

PREFACE

The European Conference on Modelling and Simulation (ECMS), which is the 32nd annual conference of the European Council for Modelling and Simulation, is an independent forum for academics and practitioners dedicated to research, development, and applications of modelling and simulation, not only from Europe but from all over the world. This year, the conference is hosted by Jade University of Applied Sciences in Germany.

Jade University of Applied Sciences, which was founded in 2009, is a state-owned university in the northwest of Lower Saxony. It has three campuses in the towns of Wilhelmshaven, Oldenburg, and Elsfleth. The university is made up of six departments and offers 37 undergraduate and 13 postgraduate degree courses. It is home to more than 7,600 students and 500 staff. Its main emphasis is on maritime, engineering, and business studies, in which modelling and simulation are of great importance.

The 2018 edition of ECMS once again brings together experts and practitioners from the field of modelling and simulation to present and to discuss their ideas, research, and challenges. This year's conference includes three new tracks, namely Simulation of Fluid-mechanically Effective Microstructures and Combustion Processes, Multi-physical Finite Element Simulation and Finite Discrete Element Methods. It attracted high quality submissions from 22 countries on five different continents. We are proud to present our distinguished keynote speakers to this conference, Professor Thomas Bäck from the University of Leiden in the Netherlands and Professor Frederic Stahl from the University of Reading in the United Kingdom. Professor Bäck is talking about algorithms for simulation-based optimisation problems whilst Professor Stahl is talking about building adaptive data mining models on streaming data in real-time.

We hope this book will serve as a reference to researchers and practitioners in the field, as well as an inspiration to those interested in the area of modelling and simulation. The chairpersons of this conference would like to thank all authors for their contributions and the track chairpersons for organising the reviewing process. Our special thanks go to the referees for their time and their efforts in reviewing all submitted papers. With their expertise and with valuable comments in most cases, they helped to maintain a high scientific quality of the conference. Furthermore, our thanks are due to Martina-Maria Seidel for running the ECMS office and to Professor Khalid Al-Begain, the president of the European Council on Modelling and Simulation. Last but not least, we would like to express our gratitude to the various members of the Department of Engineering Sciences of Jade University of Applied Sciences for their support.

We are looking forward to welcoming you in Wilhelmshaven

Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen

Wilhelmshaven, May 2018

TABLE OF CONTENTS

Plenary Talks - Abstracts

Finance and Economics and Social Science

Simulation in Industry, Business, Transport and Services

Simulation of Intelligent Systems

Modelling, Simulation and Control of Technological Processes

Simulation and Optimization

Simulation of Fluid-Mechanically Effective Microstructures and Combustion Processes

Multiphysical Finite Element Simulation

Finite – Discrete - Element Simulation

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems - Special Session -

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

ECMS 2018

SCIENTIFIC PROGRAM

Plenary Talk

Algorithms for Simulation-Based Optimization Problems

Thomas Bäck

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
The Netherlands
t.h.w.baeck@liacs.leidenuniv.nl

ABSTRACT

Many industries use simulation tools for virtual product design, and there is a growing trend towards using simulation in combination with optimization algorithms for tuning simulation input parameters. The requirements for optimization under such circumstances are often very strong, involving many design variables and constraints and a strict limitation of the number of function evaluations to a surprisingly small number (often around one thousand or less).

Tuning optimization algorithms for such challenges has led to very good results obtained by variants of evolution strategies and of Bayesian optimization algorithms. Evolutionary algorithms are nowadays standard solvers for such applications. In the presentation, sample cases from industry are presented, and their challenges are discussed in more detail. Results of an experimental comparison of contemporary evolution strategies [1] on the black box optimization benchmark (BBOB) test function set for a small number of function evaluations are discussed, and further enhancements of contemporary evolution strategies are outlined. In addition, we will also briefly discuss the concept of Bayesian global optimization [2] and its connection with research in evolutionary strategies, motivated by a generalized infill criterion [3]. The corresponding algorithms typically combine so-called metamodels (i.e., data-driven nonlinear regression models using Gaussian process or random forests for regression) with state-of-the-art optimization algorithms for the identification of new sampling points, based on the above mentioned infill criterion. Essentially, the latter is also again a multimodal objective function defined over the metamodel, which in turn requires an optimizer to solve the optimization problem defined by the infill criterion.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Our practical examples are motivated by industrial applications. A typical challenge is to find innovative solutions to a design optimization task. Based on a suitable optimization algorithm, which often also needs to deal with multiple, conflicting objectives, an application of this concept to an industrial design optimization task is discussed in the presentation.
Discussing these applications and the variants of evolution strategies applied, the capabilities of these algorithms for optimization cases with a small number of function evaluations are illustrated.

References

[1] T. Bäck, C. Foussette, P. Krause: Contemporary Evolution Strategies, Springer, Berlin, 2013.
[2] D. R. Jones, M. Schonlau, and W. J. Welch: Efficient global optimization of expensive black-box functions, *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.
[3] H. Wang, B. van Stein, M. Emmerich, T. Bäck: A new acquisition function for Bayesian optimization based on the moment-generating function. *IEEE International Conference on Systems, Man, and Cybernetics, Banff, Canada, 2017*, pp. 507-512.

Biography

Thomas Bäck is full professor of computer science at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands, since 2004.
He received his PhD in Computer Science (under supervision of Hans-Paul Schwefel) from Dortmund University, Germany, in 1994, and then worked for the Informatik Centrum Dortmund (ICD) as department leader of the Center for Applied Systems Analysis. From 2000 - 2009, Thomas was President of NuTech Solutions GmbH and CTO of NuTech Solutions, Inc. In his work with industrial partners in the industry 4.0 domain, he provides data mining, optimization software and services to companies such as, e.g., BMW, Daimler, Ford, Honda, Tata Steel, and KLM. Thomas Bäck has more than 300 publications and authored a book on evolutionary algorithms, (Evolutionary Algorithms: Theory and Practice).

He is co-editor of the Handbook of Evolutionary Computation and the Handbook of Natural Computing, and co-author of the book Contemporary Evolution Strategies (Springer, 2013). He is editorial board member of a number of journals and has served as program chair for major conferences in evolutionary computation. He received the best dissertation award from the Gesellschaft für Informatik (GI) in 1995 and is an elected fellow of the International Society for Genetic and Evolutionary Computation for his contributions to the field. In 2015, he received the IEEE Evolutionary Computation Pioneer Award for his contributions in synthesizing evolutionary computation.

Building Adaptive Data Mining Models on Streaming Data in Real-Time, an Outlook on Challenges, Approaches and Ongoing Research

Frederic Theodor Stahl

University of Reading
Department of Computer Science
f.t.stahl@reading.ac.uk

ABSTRACT

Advances in hardware and software, in the past two decades have enabled the capturing, recording and processing of potentially large and infinite streaming data. As a consequence the field of research in Data Stream Mining has emerged building Data Mining models, workflows and algorithms enabling the efficient and effective analysis of such streaming data at a large scale. Application areas of Data Stream Mining techniques include real-time telecommunication data, telemetric data from large industrial plants, credit card transactions, social media data, Smart Cities, IoT, etc. Some applications allow the data to be processed modelled and analysed in batches by traditional Data Mining approaches. However, others require the model building and analytics to take place in real-time as soon as new data becomes available i.e. to accommodate infinite streams and fast changing concepts in the data. This talk discusses challenges, barriers, opportunities and recent research on Micro-Cluster based Data Stream Mining models to overcome these barriers.

Biography

Frederic Stahl is Associate Professor in Data Science at the University of Reading and has been working in the field of Data Mining and Knowledge Discovery from Data (KDD) for the last 12 years. In particular he has been working in the research domain

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

of Big Data Analytics. His research interests here are in (i) developing scalable parallel algorithms for building Data Mining models on large volumes of data; (ii) developing algorithms for building self-adaptive Data Mining models for real-time streaming data and (iii) applications in Big Data Analytics. He currently leads a small group of 5 PhD students working on many aspects of Data Mining and Data Stream Mining. In previous appointments Frederic worked as a Lecturer at Bournemouth University and as Senior Research Associate at the University of Portsmouth. He obtained his PhD in 2010 from the University of Portsmouth with the title “Parallel Rule Induction” and his Engineering Diploma in Bioninformatics in 2006 from the University of Applied Science Weihenstephan (Germany). He has published over 50 articles in peer-reviewed conferences and journals. He is heavily involved in the BCS SGAI, the Specialist Group on Artificial Intelligence of the British Computer Society. Here he serves as an elected committee member, is the main organiser of the UK Symposium on Knowledge Discovery and Data Mining, co-organiser of the societie’s annual International Conference on Artificial Intelligence and Guest Editor of the conference’s Special Issue journal (Expert Systems, Wiley).

Finance and Economics and Social Science

DEVELOPING AND CALIBRATING AN ABM OF THE PROPERTY LISTING TASK

Enrique Canessa
Facultad Ingeniería y Ciencias, CINCO
Universidad Adolfo Ibáñez
Av. P. Hurtado 750, Viña del Mar, Chile
E-mail: ecanessa@uai.cl

Sergio E. Chaigneau
Escuela Psicología, CINCO
Universidad Adolfo Ibáñez
Av. D. Las Torres 2640, Santiago, Chile
E-mail: sergio.chaigneau@uai.cl

Carlos Barra
Independent Researcher
Las Pimpinelas 880, Concon, Chile
E-mail: c.barra@vtr.net

KEYWORDS
Property listing task, Conceptual Agreement Theory, Agent-based modeling, intersubjective variability.

ABSTRACT

We present an Agent Based Model (ABM), which shows how subjects may perform the Property Listing Task (PLT), which is widely used across psychology. In the PLT, subjects list properties that describe concepts in their minds. This ABM views the PLT as a communicative process, by which agents list properties for a concept, so that they can achieve a given level of agreement among them, i.e. produce properties that are more or less the same among agents. Using Conceptual Agreement Theory (CAT), we model that agreement in the ABM and are able to derive functional forms that the ABM's outputs should follow. The results show that agents produce agreement curves that indeed match the stated functional forms and also reasonably follow the corresponding curves obtained from empirical data. Thus, the ABM can now be used to better understand the PLT and also be further developed to model how subjects may stop listing properties for a concept according to some criteria based on the achieved level of agreement.

INTRODUCTION

The Property Listing Task (PLT) is widely used across psychology to learn about concepts that occur in natural language (e.g., CANARY, CAR, ALLIGATOR). In the PLT, subjects are asked to list properties that are true of a given concept (e.g., to the request of listing properties true of the concept CANARY, *has feathers, is a pet* and *sings* might be produced). When responses are coded (e.g., *has feathers* and *is feathered* are counted as tokens of the same property) and lists are accumulated across individuals, frequency distributions of conceptual properties are obtained. It is widely assumed that these distributions reflect concepts' underlying semantic structure (Cree and McRae 2003), which in turn is thought to reflect the statistical structure of the categories under study (e.g., McNorgan, Kotack, Meehan and McRae 2007).

Although the PLT is widely used (e.g., in cognitive psychology, social psychology, cognitive neuroscience, neuropsychology, consumer psychology), little is known about what people are doing when they perform the task (Santos, Chaigneau, Simmons and Barsalou 2011; Wu and Barsalou 2009). Perhaps, this is because the task is deceptively simple; just a report of those properties that are associated with a given concept in an individual's mind. A fact that can be easily overlooked in the PLT (and one that speaks against its purported simplicity) is that conceptual content varies across individuals. Note that unless one wants to argue that people have identical concepts in their minds, variability is a necessary feature of concepts. As will become clear throughout our current work, and in contrast to other approaches, we see this variability as central to understanding the PLT. In other approaches, the solution to the problem of variability in PLT has been to treat inter-subject variability as measurement error which needs to be discarded. Consequently, a common practice in PLT is to delete those properties produced by less than a given proportion of the sample (e.g., in McRae et al. 2005, properties produced by less than 5 out of 30 participants were dropped from the analyses).

Notably, in contrast to all previous research on this topic, in the current work we deliberately focus on conceptual variability. As will become clear in what follows, in our analysis the PLT is not a neutral report of conceptual content. Rather, it is a communication task in which people try to produce a list of properties with which others would agree. To achieve this agreement, and given that people probably know that others may view a given concept differently (e.g., due to learning, context, or point of view), they need to carefully tailor their property lists. Thus, our main goal in the present study is to model the PLT as a communication task using Agent-Based Modeling (ABM) and see whether we can relate variability in conceptual content to the PLT. Additionally, we want to assess whether the ABM's outputs are similar to those obtained from the corresponding empirical data. Having such an ABM, will then allow us to better understand the PLT.

CONCEPTUAL AGREEMENT THEORY, CONCEPTUAL VARIABILITY AND THE PLT

To achieve the stated goals, we use our Conceptual Agreement Theory (CAT; Canessa and Chaigneau 2016; Chaigneau, Canessa and Gaete 2012), and use the probability of true agreement ($p(a1)$) and the probability of illusory agreement ($p(a2)$). These probabilities represent how likely it is to find that others agree with

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

one's list in the PLT. To understand them, consider the following. In the PLT, after raw data have been collected and properties have been coded, all that researchers have are different lists of properties produced by individual subjects. Each of these lists can be thought of as a sample of conceptual content taken from the total population of properties produced by participants in the study. These samples vary in length and in content (respectively, the number of properties and the specific properties they contain). True agreement probability (*p(a1)*) is defined as the probability that any given property contained in one randomly chosen list produced by a subject in a PLT study in response to a given focal concept, will be also contained in another randomly chosen list produced in response to the same focal concept (i.e., the probability that agreement will be found regarding specific properties that are associated to the concept). This is a measure of the agreement that an individual can expect to find for his list, and it will achieve a value of 0 (zero) only when all subjects in a PLT study produce lists with completely different contents in response to a given concept. In contraposition, it will achieve a value of 1.0 only when all subjects in a PLT study produce the same list. Similarly, illusory agreement probability (*p(a2)*), is defined as the probability that a property contained in one random list produced by a subject in a PLT study in response to a given alternative concept, will be also contained in a randomly chosen list produced in response to the focal concept (i.e., the probability that agreement will be found regarding general properties that are associated to the concept). As will become clear further ahead, this measure will achieve a value of 0 (zero) only when there are no properties common to both concepts. Note that, when applied to concepts that are ordinates within the same superordinate (e.g., CANARY and PARROT are ordinates relative to BIRDS), *p(a2)* should achieve a value greater than zero because there should exist common properties for both concepts (i.e., otherwise, there would be no way for them to share the superordinate).

As shown in previous work, the mathematical formulation of CAT easily allows computing an estimate of *p(a1)* and *p(a2)* by using Equations (1) and (2) (Canessa and Chaigneau 2016):

$$p(a1) = \frac{s_1}{k_1} \tag{1}$$

$$p(a2) = \frac{s_1 u}{k_1 k_2} = p(a1)\,\frac{u}{k_2} \tag{2}$$

In Equations (1) and (2), s_1 is the average number of properties listed for a concept by subjects, k_1 is the total number of properties that describe a focal concept *C1* (i.e., the total number of properties after coding), k_2 is the same for an alternative concept *C2*, and u is the number of properties that simultaneously describe both concepts (i.e., the number of properties that belong to both concepts *C1* and *C2*). In the PLT, as people list properties, s_1 (or s_2 for the alternative concept) increases and so does k_1 (k_2). Importantly, the longer the individual lists of properties produced by subjects are, the higher the probability that those lists will contain low frequency properties that are not repeats in other lists. Take for example the list *has feathers*, *is a pet*, *tweets*, produced in response to CANARY. The longer this list becomes, the higher the probability that it will contain a property like *I have one at home*, which is probably seldom found in other lists. This implies a nonlinear relation between s_1 and k_1. To understand why, consider the following. When s_1 is small, a given increase in the mean number of properties produced yields a relatively small number of low frequency properties, and therefore, a modest increase in k_1. In contrast, when s_1 is larger, the same given increase in the mean number of produced properties yields a relatively larger number of low frequency properties, and thus a larger increase in k_1. Thus, we can state that the increase in k_1 due to an increase in s_1 depends on the value of s_1, or that the rate of increase of k_1 relative to the increase in s_1 is related to the value of s_1. This can be mathematically expressed by the following differential equation:

$$\frac{dk_1}{ds_1} = c\, s_1 \tag{3}$$

This simple equation can be solved by separating variables and integrating both sides:

$$\int dk_1 = \int c\, s_1 ds_1 \tag{4}$$

$$k_1 = c\,\frac{s_1^2}{2} + a \tag{5}$$

, where a is an arbitrary constant, and also without losing generality, we can write:

$$k_1 = a + b\, s_1^2 \tag{6}$$

Given that Equation (6) reflects only part of the relation between s_1 and k_1, i.e., other processes may also intervene in that relation, we can use Equation (6) as the guiding functional form in a regression analysis to see whether it empirically holds. Indeed, regression equations like Equation (6) explain on average almost 50% of the variance in k_1, and parameters a and b are statistically significant across many concepts (Chaigneau, Canessa, Barra and Lagos 2017). Thus, if our ABM models the PLT, the corresponding ABM´s outputs should also show that relationship. Moreover, replacing Equation (6) in Equations (1) and (2) yields expressions for *p(a1)* and *p(a2)*, which should also be obtained in PLT's ABM:

$$p(a1) = \frac{s_1}{k_1} = \frac{s_1}{a + b\, s_1^2} \tag{7}$$

$$p(a2) = \frac{s_1 u}{k_1 k_2} = \frac{s_1 u}{(a + b\, s_1^2)^2} \tag{8}$$

Recall that earlier we stated that the PLT is a communicative task in which individuals attempt to tailor property lists with which others will agree. The reader

possesses now the necessary information to understand why. In our view of the PLT, depending on their communicational goal (i.e., being specific, being general, pursuing a combination of both, or being precise), people would probably select different properties to tailor their lists. Using *p(a1)* and *p(a2)*, we can express these goals as: (1) trying to be specific implies maximizing *p(a1)*; (2) trying to be general implies maximizing *p(a2)*; (3) trying to combine specificity and generality implies simultaneously maximizing *p(a1)* and *p(a2)*; (4) trying to be precise implies maximizing *p(a1)* and simultaneously minimizing *p(a2)*.

Regardless of which of the above-stated goals better represents what people attempt to do when performing the PLT, our ABM should reasonably model those possibilities, i.e., model *p(a1)* vs. s_1 per Equation (7) (corresponding to the first alternative: maximize *p(a1)*); *p(a2)* vs. s_1 per Equation (8) (corresponding to the second alternative: maximize *p(a2)*); *p(a1)* + *p(a2)* vs. s_1 (corresponding to the third alternative: simultaneously maximize *p(a1)* and *p(a2)*); and *p(a1)* – *p(a2)* vs. s_1 (corresponding to the fourth alternative: simultaneously maximize *p(a1)* and minimize *p(a2)*).

DESCRIPTION OF PLT'S ABM

We designed a simple ABM that implements the PLT, abiding by the *KISS* (keep it simple stupid) principle (Axelrod 1997), with the goal of modeling people's observed behavior when performing it. Note that we necessarily make some assumptions regarding how concepts are represented in their minds, which we will specify shortly. Note that because of space constraints, we don't strictly follow the ODD (Overview, Design concepts, Details) protocol (Grimm et al. 2006) to present the ABM, but we comply with including most of the material suggested in it.

First, the ABM creates *N* agents and assigns to each of them two frequency distributions of properties that describe the focal (*C1*) and alternative concept (*C2*). These frequency distributions were obtained from an actual PLT study (Devereux et al. 2014) and the ABM code has a routine that can input and process those two frequency distributions. These distributions correspond to our first representational assumption: property distributions obtained through the PLT, reflect people's average mental representation of the concepts at hand. To represent the properties, numbers are used, which identify each property. Each of those numbers has a given frequency. Note that *C1* and *C2* may have properties with the same number, which represent the common properties, and the total number of those common properties is *u*.

Second, per our previous discussion on conceptual variability, the ABM assigns to each agent's *C1* and *C2* concepts a given number of different properties (*NDP*) (properties that are not shared by any other agent). That number is sampled for each agent from U(0, *MAXNDP*), where *MAXNDP* can be set and is the maximum number of different properties that an agent can have. These *NDP*s properties correspond to our second representational assumption: as discussed above, conceptual content varies across individuals. The ABM code automatically creates those different properties, so that their respective identification numbers are different from the ones assigned to properties shared by all agents and different from the exclusive properties assigned among agents. To illustrate this whole process, imagine two agents (A1 and A2), and that for A1 *NDP* = 3 and for A2 *NDP* = 1. Also assume that *C1* = {5100,2300,4500,1400} and *C2* = {3000,5100,1400} with *FC1* = {10,8,9,10} and *FC2* = {15,16,14}, where *FC1* and *FC2* are the frequency of each property in *C1* and *C2*. Note that in this case $u = 2$ and that those frequency distributions are an input to the ABM. For this situation, A1 and A2 will have a set of properties equal to *C1* and *C2* with respective frequencies *FC1* and *FC2*. Also, A1 will have in its set of properties that represent *C1* and *C2*, three exclusive properties, for example {1, 2, and 3} for *C1* and {10, 20, 30} for *C2*. Similarly, A2 will have one exclusive property in its set of *C1*, for example {100}, and one exclusive property in set *C2*, for example {130}. These two sets are the potential properties for *C1* and *C2* in each agents' mind, which each agent might list. After the initialization stage, the ABM performs the following actions, which comprise a simulation step:

a) From all the *N* agents, randomly select without replacement one agent.
b) This agent samples with replacement and using roulette-wheel selection, a property for *C1* (from its set of properties that represent *C1*) and another for *C2* (from its set of properties that represent *C2*), and adds them to separate lists that represent the properties listed for *C1* and *C2*.
c) Repeat actions a) and b) until all agents had done them.

It is worth noting that in step b), roulette-wheel selection, means that a property's probability of being selected and listed is equal to the ratio of its frequency to the sum of all the frequencies of the properties contained in the given set (*C1* or *C2*) for a specific agent, which include the corresponding exclusive properties. For example, for A1 and concept *C1* and property 5100, the probability of A1 selecting it is: 10 / (10 + 8 + 9 + 10 + 1 + 1 + 1) = 10 / 40 = 0.4

In each simulation step, after all agents have performed actions a) and b), the ABM calculates k_1, k_2 and *u*. That is done in exactly the same way as it is completed in the PLT coding stage. The ABM collects the lists of listed properties for *C1* and *C2* from all agents, merges them and counts the non-repeating properties for *C1* (*C2*), obtaining k_1 (k_2). For calculating *u*, the ABM counts the total number of non-repeating properties that are contained in both merged *C1* and *C2* lists. Additionally, to calculate s_1 (s_2, although we do not use this output for now), each agent reports to the ABM the number of non-repeating properties that are contained in its list of listed properties for *C1*(*C2*), and the ABM calculates the mean

of those numbers. Note that per Equations (1) and (2), having k_1, k_2, s_1, and u is all we need to calculate $p(a1)$ and $p(a2)$. An example of the calculation will make the whole process clear.

For simplicity, consider the same two agents A1 and A2 as before. Suppose that A1 has listed for $C1$ = {5100, 2, 5100, 2300} and for $C2$ = {1400, 1400, 10, 5100}. Similarly, A2 listed for $C1$ = {2300, 5100, 5100, 5100} and for $C2$ = {3000, 3000, 1400, 130}. Then, for the given simulation step, the merged list of properties for $C1$ = {5100,2,5100,2300,2300,5100,5100,5100} and for $C2$ = {1400,1400,10,5100,3000,3000,1400,130}. Hence, k_1 = 3 (3 non-repeating properties = {5100,2,2300}), k_2 = 5 (5 non-repeating properties = {1400, 10, 5100, 3000, 130}), u = 1 (1 non-repeating property belongs to the merged list of properties for $C1$ and $C2$ = {5100}). Also, A1 has listed 3 non-repeating properties for $C1$ and 3 for $C2$, whereas A2 has listed 2 for $C1$ and 3 for $C2$. Thus, s_1 = (3 + 2) / 2 = 2.5 and s_2 = (3 + 3) / 2 = 3.0. Applying Equations (1) and (2) for $C1$: $p(a1)$ = 2.5 / 3 = 0.833 and $p(a2)$ = 0.833 x 1 / 5 = 0.167. All those values are the outputs of the ABM, which will be used in analyzing the results of experiments. Finally, for the interested reader, the ABM is coded using Netlogo version 5.3.1 (Wilensky 1999) and is available upon request from the authors, along with data that allow to replicate the experiments.

EXPERIMENTS AND RESULTS

Our general question is whether the ABM might produce curves of k_1 vs. s_1 that follow the relation stated by Equation (6) and also whether the curves of $p(a1)$ and $p(a2)$ vs. s_1 (calculated by the ABM as already shown in the previous section), mimic the ones implied by Equations (7) and (8), and also the corresponding values of $p(a1) + p(a2)$ and $p(a1) - p(a2)$ (per the discussion in the section Conceptual Agreement Theory, conceptual variability and the PLT). Additionally, those curves should reasonably match the corresponding ones obtained from empirical data.

To that purpose, we used data obtained from the Centre for Speech, Language and the Brain concept property norms (from here and on, the CSLB norms; Devereux et al. 2014). These norms contain empirical conceptual property frequency distributions obtained from a sample of subjects performing the PLT for a large number of concepts. These concepts can be organized in several superordinate sets. From the CSLB norms, we selected 5 superordinates: Musical Instruments (M.I.), Fruits (Fr.), Clothing Items (C.I.), Birds (Br.) and Weapons (W.). From each superordinate, we selected a focal ($C1$) and alternative concept ($C2$). Table 1 shows those superordinates along with the corresponding concepts and other information that will be used. Those superordinates and corresponding concepts were selected so that we could get different values for u (number of common properties between $C1$ and $C2$), and for the a and b coefficients obtained from regression analysis using Equation (6) as the relational form. Note that in the regression equations, R^2 ranged from 0.23 to 0.59 (mean 0.43), so that we can generalize our findings to concepts that reasonably or more than reasonably follow Equation (6). We must note that to obtain the points necessary to calculate the coefficients of the parabolas for the CSLB data, we used the following procedure. The a and b parameters are not computed for a single concept, but for all concepts in the same superordinate (e.g., Musical Instruments) by regressing their corresponding k_1 values over their s_1 values. In contrast, our analysis of the ABM's outputs was done on each pair of concepts (thus, on only two k_1 and s_1 values; not enough to compute a single parabola). Hence, to obtain more data points at the concepts' level, we obtained the a and b coefficients for the corresponding superordinate and then used the ABM to obtain multiple points for the two concepts, which are part of the superordinate. To do so, we set up the ABM with the frequency distribution of each concept and adjusted *MAXNDP* (the only free input parameter to the ABM), so that we could approximately obtain the a and b coefficients of the superordinate. Then, we run the ABM to obtain (s_1, k_1) pairs for the concepts, i.e. we run the ABM and looked at the $p(a1) + p(a2)$ vs. s_1 curve and visually determined the corresponding values of s_1 and k_1 where that curve reached its maximum. We did that 20 times and obtained 20 different (s_1, k_1) pairs. Finally, using those pairs, we calculated a regression equation for obtaining the a and b coefficients for the pair of concepts, shown in Table 1.

Table 1: Data for superordinates and other information obtained/calculated using the CSLB norms

Superordinate	Concepts	Values calculated for a, b and u
Musical Instruments	$C1$ = accordion $C2$ = bagpipes	a = 26.7 b = 0.110 R^2 = 0.44 u = 10
Fruits	$C1$ = apple $C2$ = apricot	a = 26.1 b = 0.070 R^2 = 0.59 u = 11
Clothing Items	$C1$ = blouse $C2$ = cloak	a = 23.8 b = 0.096 R^2 = 0.37 u = 7
Birds	$C1$ = canary $C2$ = parakeet	a = 29.2 b = 0.114 R^2 = 0.54 u = 16
Weapons	$C1$ = axe $C2$ = machete	a = 24.7 b = 0.210 R^2 = 0.23 u = 10

Note: a and b see Equation (6) were obtained from regressing k_1 on s_1, all coeffs. are statistically significant at least at the 0.05 level

The experiments done with the ABM consisted in keeping fixed the number of agents N = 30 (because that

was the number of subjects used in the CSLB norms), and entering the frequency distribution of properties for *C1* and *C2* obtained from the CSLB norms for each superordinate, one at a time. Then, for each superordinate, we adjusted by trial-and-error, *MAXNDP* (the maximum number of different properties that an agent can have), so that we could obtain a reasonable match between the outputs of the ABM and the corresponding values calculated from the CSLB data. Note that in our experiments, the only free input parameter to the ABM is *MAXNDP*. Recall that this value represents conceptual variability across individuals, and arguably cannot be measured, which justifies it being a free parameter. Given that the ABM has random processes, the final analyses were done on the mean outputs of 20 replications, for each superordinate. The stopping condition for each run was set at 400 time-steps. We selected that condition because visual analyses indicated that the values of the time series for $p(a1)$, $p(a2)$, $p(a1) + p(a2)$ and $p(a1) - p(a2)$ were all past the values of s_1 that optimize those expressions, per the discussion in the section Conceptual agreement theory, conceptual variability and the PLT.

To present the results, we first show the graphs of k_1 vs. s_1 (for concept *C1*) for each superordinate in Figure 1 for data obtained from the ABM.

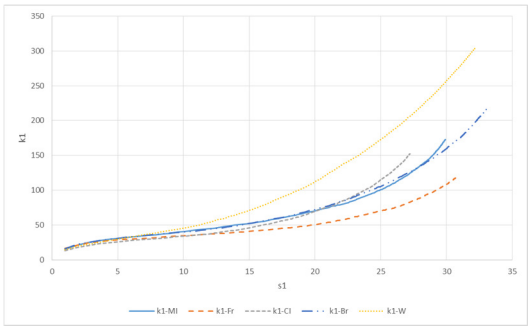


Figure 1: Avg. k_1 vs. s_1 for superordinates (data obtained from the ABM)

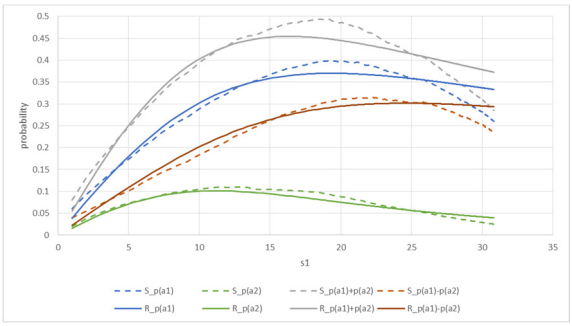
A visual analysis of Figure 1 indicates that indeed the curves of k_1 vs. s_1 for all superordinates exhibit a shape consistent with Equation (6) for positive a and b coefficients (a parabola whose axis is parallel to the y-axis and opens in the positive y-axis direction). To quantitatively corroborate that, we regressed k_1 on s_1 for a relevant range of s_1 (a range corresponding to the mean s_1 of *C1* for each superordinate ± 3 std. deviations). Table 2 shows the results of the regressions. As can be seen, the a and b coefficients obtained from the ABM data are very close to the ones of the CSLB data. Additionally, note that the R^2 values for the regression equations using the ABM data are very high, which means that the parabolas fit very well that data, i.e., the ABM data exhibit a shape very similar to the predicted parabolas. We got these suspiciously large R^2 values, because we are modelling data obtained from the same ABM, as explained before. In summary, we can say that the curves of k_1 vs. s_1 obtained from ABM's corresponding outputs fit the ones of the CSLB data for all five superordinates. Note also that these fits are achieved with only one free input parameter (i.e., the *MAXNDP*).

Table 2: Regression results of k_1 on s_1 for superordinates

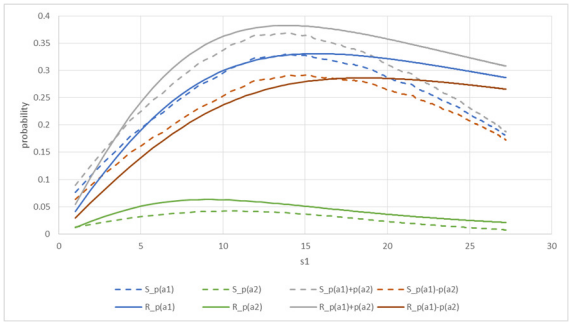
Superordinate and *MAXNDP* for best fit	Reg. coeffs. for ABM data	Reg. coeffs. for CSLB data	Range of s_1 and R^2 for reg. eqs. for ABM data
Musical Instruments *MAXNDP* = 10	a = 26.5 b = 0.150	a = 26.7 b = 0.110	s_1 min = 2 s_1 max = 11 R^2 = 0.96
Fruits *MAXNDP* = 7	a = 26.0 b = 0.072	a = 26.1 b = 0.070	s_1 min = 1 s_1 max = 17 R^2 = 0.89
Clothing Items *MAXNDP* = 10	a = 23.82 b = 0.100	a = 23.8 b = 0.096	s_1 min = 3 s_1 max = 16 R^2 = 0.99
Birds *MAXNDP* = 15	a = 29.31 b = 0.103	a = 29.2 b = 0.114	s_1 min = 3 s_1 max = 18 R^2 = 0.99
Weapons *MAXNDP* = 23	a = 24.6 b = 0.211	a = 24.7 b = 0.210	s_1 min = 4 s_1 max = 18 R^2 = 0.99

Note: a and b coeffs. are statistically significant at least at the 0.05 level

The next analyses are aimed at corroborating whether the curves corresponding to Equations (7) and (8) for the real CSLB data are similar to the ones obtained from the ABM. To do that, we calculated and graphed the $p(a1)$, $p(a2)$, $p(a1) + p(a2)$ and $p(a1) - p(a2)$ vs. s_1 curves for the CSLB data using Equations (7) and (8) and the corresponding a and b coefficients and u shown in Table 1, for concept *C1*, and thus varying s_1. Then, we calculated the same curves, but using the $p(a1)$ and $p(a2)$ obtained from the ABM's outputs. Recall that the ABM computes $p(a1)$ and $p(a2)$ per Equations (1) and (2) and using the k_1, k_2, s_1, and u calculated by the procedure explained at the end of section Description of PLT's ABM. Thus, we are comparing different curves obtained from different data and using different procedures. Given that we do not have enough space to show all the five graphs for each of the superordinates, Figure 2 presents those curves for only two superordinates: Fruits and Clothing Items. Fruits exhibits the best match based on the Mean Absolute Percentage Error (MAPE) for the $p(a1)$, $p(a2)$, $p(a1) + p(a2)$ and $p(a1) - p(a2)$ vs. s_1 curves, and Clothing Items the worst one (see Table 3, second column). Figure 2 indicates that the better fit of Fruits over Clothing Items stems from the right tail of the curves. For that region, the difference between the simulated curves (dashed lines) and the real ones (solid lines) is smaller for Fruits than for Clothing Items. However, the match around the values of s_1 which maximize the curves is better for Clothing Items than for Fruits. Remember that per the discussion in the section Conceptual agreement theory, we want to model the maximization of those curves.



(a)



(b)

Figure 2: *p(a1)*, *p(a2)*, *p(a1)* + *p(a2)* and *p(a1)* – *p(a2)* vs. s_1 curves for real CSLB and ABM data: (a) Fruits, (b) Clothing Items (simulated = dashed lines, real = solid lines)

Thus, using only the MAPE for the entire range of s_1 might be misleading for our purposes. Hence, to better asses the fit between the simulated and real curves we calculated several goodness of fit figures. Along with MAPE, we also computed the Root Mean Squared Error (RMSE). Additionally, those two indices were calculated for the entire range of s_1 and also for a selected range of s_1. We established that range for each superordinate by taking into account the minimum s_1 and the value of s_1 that is located past the maximum value of *p(a1)*, *p(a2)*, *p(a1)* + *p(a2)* and *p(a1)* – *p(a2)* in those respective ABM's simulated curves (dashed lines). For example, looking at Figure 2 (a), the range is approximately from $s_1 = 1$ to $s_1 = 20$ for Fruits and Figure 2 (b) shows that for Clothing Items that range is about $s_1 = 1$ to $s_1 = 16$. We acknowledge that we used an approximate visual procedure to set those ranges, but we think it is appropriate for our purposes (see a further discussion of this point later on). Table 3 presents the indices along with the corresponding ranges (actually only the upper s_1, since the lower $s_{1\,is}$ always 1). From Table 3 we can see that the match of the real curves to the ones obtained from the ABM are rather good for Musical Instruments, Fruits and Weapons, all of them with a MAPE for all points below 20%. However, recall that one of our goals for building the ABM is to model the behavior of subjects when listing properties in the PLT. As already discussed, we hypothesize that people performing the PLT may try to list a number of properties (s_1) that may maximize *p(a1)*, *p(a2)*, *p(a1)* + *p(a2)*, or *p(a1)* – *p(a2)*.

Table 3: Goodness of fit indices for assessing the match between real and simulated *p(a1)*, *p(a2)*, *p(a1)* + *p(a2)* and *p(a1)* – *p(a2)* vs. s_1 curves

Super.	MAPE all points	MAPE range s_1	RMSE all points	RMSE range s_1
M.I.		$s_1\,max = 25$		$s_1\,max = 25$
	16.1%	3.59%	0.0442	0.0086
	23.0%	16.9%	0.0059	0.0072
	16.5%	4.28%	0.0490	0.0128
	16.3%	5.42%	0.0396	0.0094
	$\bar{x} = 18.0$	$\bar{x} = 7.55$	$\bar{x} = 0.0347$	$\bar{x} = 0.0095$
Fr.		$s_1\,max = 20$		$s_1\,max = 20$
	10.8%	5.97%	0.0427	0.0173
	20.5%	11.8%	0.0105	0.0106
	11.9%	6.44%	0.0528	0.0264
	9.59%	6.46%	0.0329	0.0113
	$\bar{x} = 13.2$	$\bar{x} = 7.67$	$\bar{x} = 0.0347$	$\bar{x} = 0.0164$
C.I.		$s_1\,max = 16$		$s_1\,max = 16$
	23.8%	4.67%	0.0768	0.0089
	50.4%	30.7%	0.0138	0.0174
	26.0%	7.38%	0.0890	0.0205
	22.2%	11.9%	0.0652	0.0184
	$\bar{x} = 30.6$	$\bar{x} = 13.7$	$\bar{x} = 0.0612$	$\bar{x} = 0.0162$
Br.		$s_1\,max = 23$		$s_1\,max = 23$
	16.4%	5.25%	0.0410	0.0115
	33.4%	6.07%	0.0096	0.0040
	18.1%	4.42%	0.0503	0.0125
	14.7%	7.95%	0.0321	0.0118
	$\bar{x} = 20.6$	$\bar{x} = 5.92$	$\bar{x} = 0.0333$	$\bar{x} = 0.0100$
W.		$s_1\,max = 20$		$s_1\,max = 20$
	11.9%	3.33%	0.0188	0.0065
	28.9%	14.2%	0.0032	0.0056
	12.7%	3.16%	0.0209	0.0080
	11.7%	6.67%	0.0169	0.0091
	$\bar{x} = 16.3$	$\bar{x} = 6.84$	$\bar{x} = 0.0150$	$\bar{x} = 0.0073$

Hence, the agents in our further development of the present ABM should include rules to stop listing properties accordingly. Thus, the most relevant range of s_1 for the outputs of the ABM is from the beginning of the listing process ($s_1 = 1$) to the value of s_1 that maximizes the given ABM's outputs (i.e., the simulated curves, see dashed lines in Figure 2). Given that we do not exactly know at which simulated value of s_1 agents will actually stop listing properties, and to be on the safe side, we think that a good stopping limit will be the value of s_1 past the one which maximizes the given simulated output. As Table 3 shows for those ranges of s_1 (see columns 3 and 5), the fit is much better, with a MAPE that is half or less the value of the MAPE for all the points. The same conclusions can be reached when analyzing the RMSEs. For RMSEs, the differences between RMSEs for all points and the ones for the relevant s_1 ranges are generally even bigger than for MAPEs. The ratios of RMSEs for all points to the ones for the relevant s_1 ranges go from 2.1 to 3.8, whereas the ratios for the corresponding MAPEs span from 1.7 to 3.5.

DISCUSSION/CONCLUSIONS

As we stated in this paper, the PLT is widely used across psychology, but little is known about what people are doing when they perform the task. By viewing the task as a communicative process and applying CAT, we were

able to mathematically model part of it, and then developed an ABM that models the PLT from a subject's point of view. The fit obtained between real and ABM´s data is a first step in trying to model and better understand the PLT. Of course, these are preliminary results, given that we must assess whether we will obtain the same good fit to other PLT data. Additionally, given that the fit between real and ABM's data do not guarantee per se the validity of the model (Macy and Willer 2002), we must still work on validation. However, the effort in developing the ABM by finding relevant theory and practice and then merging it in a coherent model, allowed us to assess that the model may be plausible and to gain valuable insights regarding the PLT. Also, the appropriate fit of the real data to the outputs of the ABM encourages us to continue developing and using the ABM to unravel the subject's mental processes behind the PLT. We think that the development of the ABM can take two different but complementary ways. The ABM suggests that the core mechanisms embedded in the rules that govern agents are part of a probabilistic sampling process. Thus, we could refine that part of the ABM to gain further comprehension of that process and then use it to better model that part of the PLT, perhaps with a probabilistic mathematical model. Also, per our vision that subjects performing the PLT are maximizing the relation between *p(a1)*, *p(a2)*, *p(a1)* + *p(a2)*, or *p(a1)* – *p(a2)* (i.e., maximize *p(a1)* and minimize *p(a2)*) and the number of properties they list (s_l), we could develop an extension to the present ABM, so that agents have rules that allow them to stop listing properties when they reach the maximum of those curves. Of course, many issues must be solved to do that, like how agents will calculate *p(a1)* and *p(a2)* in their minds and without actually explicitly knowing the properties listed by other agents. Hence, agents will need to perform something like a PLT in their minds, using information from the property frequency distributions of concepts they have in their minds and guessing when they might have achieved the sought property listing stopping condition. Although that refinement of the ABM seems plausible, but also difficult, it will help us to better understand the PLT. The work we have done until today shows that it is possible.

ACKNOWLEDGMENT

This work was supported by CONICYT, FONDECYT (Fondo Nacional de Ciencia y Tecnologia of the Chilean Government) grant Nr. 1150074 to the first two authors.

REFERENCES

Axelrod, R. 1997. "Advancing the art of simulation in the social sciences". In *Simulating Social Phenomena 1997*, R. Conte, R. Hegselmann, and P. Terna (Eds.). Springer, Berlin, 21-40.

Canessa, E. and S. Chaigneau. 2016. "When are concepts comparable across minds?" *Quality & Quantity* 59, No. 3, 1367-1384.

Chaigneau, S.; E. Canessa; C. Barra, C.; and R. Lagos. 2017. "The role of variability in the property listing task". *Behavior Research Methods*, DOI 10.3758/s13428-017-0920-8.

Chaigneau, S.E.; E. Canessa; and J. Gaete. 2012. "Conceptual agreement theory". *New Ideas in Psychology* 30, No. 2, 179-189.

Cree, G. S. and K. McRae. 2003. "Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns)". *Journal of Experimental Psychology: General* 132, 163-201.

Devereux, B. J.; L.K. Tyler; J. Geertzen; and B. Randall. 2014. "The Centre for Speech, Language and the Brain (CSLB) concept property norms". *Behavior Research Methods* 46, No. 4, 1119–1127.

Grimm, V. et al. 2006. "A standard protocol for describing individual-based and agent-based models". *Ecological Modelling* 198, 115-126.

Macy, M.; R. Willer. 2002. "From factors to actors: Computational Sociology and Agent-Based Modeling". *Annual review of sociology* 28, 143-166.

McNorgan, C.; R.A. Kotack; D.C. Meehan; and K. McRae. 2007. "Feature–feature causal relations and statistical co-occurrences in object concepts". *Memory & Cognition* 35, No. 3, 418–431.

McRae, K.; G.S. Cree; M.S. Seidenberg; and C. McNorgan. 2005. "Semantic feature production norms for a large set of living and nonliving things". *Behavior Research Methods* 37, No. 4, 547-559.

Santos, A.; Chaigneau, S. E.; Simmons, W. K.; and Barsalou, L. W. 2011. Property generation reflects word association and situated simulation. *Language and Cognition* 3, No. 1, 83-119.

Wilensky, U. 1999. NetLogo. http://ccl.northwestern.edu/netlogo/, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wu, L.L. and Barsalou, L.W. 2009. "Perceptual simulation in conceptual combination: Evidence from property generation." *Acta Psychologica* 132, 173-189.

AUTHOR BIOGRAPHIES

ENRIQUE CANESSA is associate professor at U. Adolfo Ibáñez, Chile. He holds a PhD in MIS/CIS (2002), a Certificate of Graduate Studies in Complex Systems (2001) and an MBA (1991) from the University of Michigan, USA. His research interests include the study of sociology and cognitive psychology using ABM.

SERGIO E. CHAIGNEAU is full professor at U. Adolfo Ibáñez, Chile and holds a PhD in Cognitive and Developmental Psychology (2002) from Emory University, USA, and a M.A. (1995) in Experimental Psychology from the University of Northern Iowa, USA. His research interests include the study of causal categorization and inter-subjective agreement.

CARLOS BARRA holds a postgraduate degree in Integrated Political Science (2005) from the Chilean Naval Academy, an MBA (1999) from IEDE, Chile, and a M.Sc. in Computer Engineering (1996) from UTFSM, Chile. His research interests include the study of complex systems.

ECONOMETRIC MODELLING OF TIME SERIES RELATIONSHIP BETWEEN FERTILITY AND INCOME FOR THE RUSSIAN POPULATION: METHODOLOGICAL ISSUES

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

KEYWORDS
Regression models, time series, modelling, fertility, income

ABSTRACT

Finding determinants of demographic processes is a highly topical issue in countries with negative demographic trends. Our research was aimed at studying the relationships between fertility and income indicators in Russia. The period under review was 2000 to 2016. To explore the correlation between the time series, we used the methodology of estimating trend deviation. We applied analytical smoothing to model trends, estimating regression models. To assess the strength of relationship between the time series, we analysed correlation between regressions' residuals. The results of our analysis showed no relationship between people's incomes and fertility rates. The research we carried out into time series dynamics did not confirm the results of other studies based on static data. Accordingly, this raises questions about the methodology for analyzing the relationship between dynamic processes with a high volatility of input data. Evidently to receive reliable and stable results, multi-dimensional analysis methods should be integrated into the study of relationships between dynamic time series, including preliminary multi-dimensional data classification. This will enable carrying out analysis on homogenous territorial or temporal segments, which would be more methodologically sound.

INTRODUCTION

Finding determinants of demographic processes is a highly topical issue in countries with negative demographic trends, including today's Russia. Birthrates in Russia are significantly below the simple reproduction rate. In Soviet times, the Total Fertility Rate (TFR) reached this level, starting to decline in the 1980s. In 2000, the TFR was at its lowest at 1.195, just 56.9% of the replacement fertility rate. Despite the positive dynamic of the TFR in Russia from 2001, the pace of change is very slow and the potential for growth has been exhausted. The adverse demographic situation in Russia is exacerbated by low life expectancy and high mortality rates.

Given these factors, studying demographic processes to seek out determinants and develop robust forecasts is highly topical. Fertility is of utmost importance, since it determines natural population growth. However, this demographic process provokes the most vigorous scientific and political debate in Russia – questions related to fertility determinants and ways to regulate it remain unresolved.

The issues of the nature of the relationship between fertility indicators and population income levels have been the subject of scientific research by economists and demographers since the 18th century. The nature of the relationship between fertility and people's wealth has been evaluated with respect to different research subject groups.

The first group compares fertility indicators for territories (countries, regions and so on) with different levels of incomes. For example, figure 1 shows a scatter plot of the relationship between GDP levels and TFR in different countries for 2015 (Fertility rate 2017; GDP per capita 2017). The values on the plot suggest an inverse relationship between these variables.

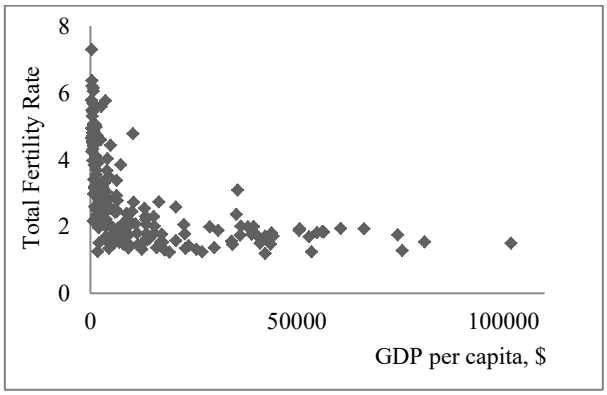


Figure 1: Correlation between GDP per capita and TFR in different countries (Fertility rate 2017; GDP per capita 2017)

The second group evaluates families/women with different levels of income within the same region. For example, in their study, Maleva and Sinyavksaya identified that for Russian women, growth in income

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

was accompanied by a decline in the average number of children per woman (Maleva and Sinyavskaya 2007). This is shown in Figure 2, presented in decile groups of women, clustered by income.

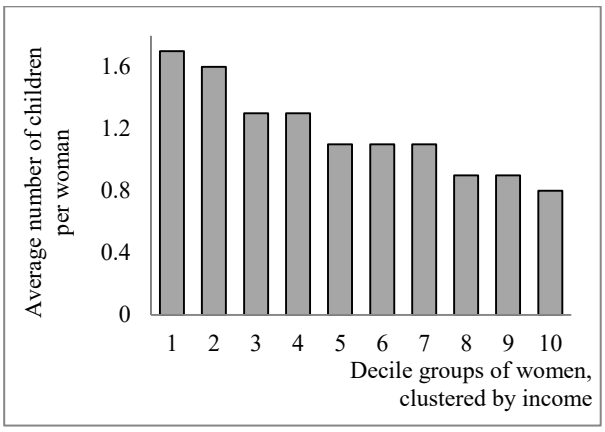


Figure 2: Average number of children per woman in groups of women, clustered by income (Maleva and Sinyavskaya 2007)

Finally, the third group comprises specific territories (countries, regions) whose populations display changing incomes and fertility rates. Recall Adam Smith's words that "poverty… seems even to be favourable to generation" (Smith 1827:33). However, contemporary researchers are cautious about declaring an adverse relationship between income levels and birth rates (Maleva and Sinyavskaya 2007). This is connected to the fact that, for example, countries with similar economic dynamics may show varying rates of inverse fertility dynamics. Moreover, low fertility is observed in countries with different levels of income across the population.

Stopping short of making sweeping statements about declining fertility due to growing incomes, researchers highlight two reasons for the presumed relationship.

Firstly, as people's incomes grow, potential parents start to assign more value to the quality, rather than quantity, of children. This entails greater investment (money, work, time) in the quality of children's human capital (Lee and Mason 2010; Becker et al. 1990). Understanding that their resources are finite, parents are forced to choose between quantity and quality of children and, in the context of a developed economy, choose to improve the latter.

The second reason is related to restrictions that children place upon their parents, first and foremost on mothers. They are connected to, inter alia, the so-called "motherhood wage penalty" – a drop in earnings that follows childbirth. When people's incomes are high, this decline becomes more pronounced (Van Bavel 2010; Begall and Mills 2012). Additional adverse factors include fewer opportunities for professional self-actualisation and impediments to working mothers' career development. For countries with developed economies and high levels of income, these factors undermine reproductive intentions and consequently lead to a decline in birth rates.

Despite the fact that such reasons may drive declining fertility in high-income countries, incentives that seek to mitigate them are often monetary. In particular, there is an existing stereotype in contemporary Russia, that sufficient material resourcing boosts fertility. This belief led to the introduction of the so-called maternity capital initiative in 2007, which seeks to boost income levels and guarantees for women who decide to have a second child.

As such, despite extensive evidence of a lack of a positive relationship between income and fertility, in Russia, economic factors of life continue to be seen as a key determinant of fertility. This sees the implementation of corresponding demographic policy measures, at a high cost to the state.

On the other hand, we believe that such a clearly mistaken view could hardly have been placed at the heart of Russia's demographic policy. One could thus suppose that the real determinant of fertility is not economic conditions per se, as the way they are perceived by people, for instance, through expectations of future economic and social stability. From this standpoint, maternity capital, as a way to improve fertility, may be viewed not as a means of genuine improvement of material welfare, but attention on the part of the authorities to family matters and an orientation towards family values, which in turn raises people's confidence in the future and creates certain life guarantees.

In light of this, our research was aimed at studying the relationships between fertility and income indicators in Russia, as well as people's subjective assessments of their welfare.

DATA AND METHODS

For our study we used data on annual dynamics for time series of socio-demographic indicators. The data was sourced from World Bank and Russian Federal State Statistics Service resources. The period under review was 2000 to 2016, which was chosen for two reasons: firstly, the availability of corresponding open-source data and secondly, the comparability of indicators as regards the span of the time series.

The analysis covered the following variables:

- *Total Fertility Rate (TFR).* The average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given fertility rate at each age. TFR is a more direct measure of the level of fertility than the crude birth rate, since it refers to births per woman (Total Fertility Rate 2017).

- *GDP per capita (constant 2010 US$)*. Gross domestic product divided by midyear population. Data are in constant 2010 U.S. dollars (GDP per capita 2017).

- *Consumer confidence index – general, as well as for women and men separately*. A generalised index, which is calculated on the basis of five individual indices: past and expected changes to personal material welfare; past and expected changes to the economic

situation in Russia; favourability of conditions for capital purchases.

These indices are calculated on the basis of data obtained in the course of representative polls (Consumer confidence index 2017).

- *Coefficient of income differentiation.* Characterizes the degree of social stratification and is defined as a ratio between the average levels of money income of 10 percent of the population with the highest income and 10 percent of the population with the lowest income (Coefficient Gini 2017).

- *Coefficient Gini – index of income inequality.* Characterizes the level of deviation of the actual volume of distribution of income of population from the line of their even distribution. The value of coefficient may vary from 0 to 1, and the higher the value of the indicator, the less even is the distribution of income in the society (Coefficient Gini 2017).

We chose these variables as we sought to include both objective and subjective indicators of people's incomes. Objective variables included statistical indicators of levels and variability of income; subjective ones covered sociological indicators connected to people's assessments of the economic situation as a whole and their material welfare in particular. As such, we tested three hypotheses in our research:

1. Birth rates are correlated with people's incomes (in this hypothesis GPD per capita was used as an objective indicator of people's incomes);

2. Birth rates are correlated with people's subjective perceptions as regards their income and material welfare (in this hypothesis the consumer confidence index was used as the subjective measure of welfare levels);

3. Birth rates are correlated with the level of income inequality across the population (in this hypothesis, we used the coefficient of income differentiation and the index of income inequality as fertility determinants).

To explore the correlation between the time series, we used two approaches. The first one is the methodology of estimating trend deviation. We applied analytical smoothing to model trends, estimating regression models for the stated time series. We used Ordinary Least Squares as the method for estimating the unknown parameters in regression models. To assess the strength of correlations between the time series, we used the Pearson correlation (based on a study of correlations between residuals in regression models). The second approach entailed assessing a multiple regression model, where time was one of the explanatory variables. Including the time variable into the model allowed assessing the specificity of the influence of income on the level of fertility, while excluding trend effects.

To test the hypothesis about the possible influence of income on fertility growth we assessed the models and tested for a correlation between the studied variables using synchronized data, and also data about income levels with one and two-year lags. The idea was that fertility changes after a change in income, but not straightaway, with some lag (for example, 1-2 years).

RESULTS

Studying the fertility time series showed that over the course of the considered period, the growth of this indicator was well-approximated by a linear trend. The main results of modelling the total fertility rate trend are shown in tables 1-2 (model 1). As the presented data show, the quality of the approximation for the model is very high with a determination coefficient in excess of 95%. All parameters of the equation are statistically significant and show that between 2000 and 2016 the total fertility rate in Russia grew 0.039 points on average.

Studying GDP per capita dynamics showed that this dynamic is best approximated by a quadratic function. Yet the largest part of the initial data (2000-2013) fits the ascending arc of the parabola, while the three most recent years (2014-2016) showed a slight decrease in Russia's GDP per capita. However we do not consider it appropriate to speak of an established adverse trend toward a decline in people's incomes. The main results of modeling the trend for GDP per capita are shown in tables 1-2 (model 2).

Table 1: Model Summary
(dependent variable: model 1 – TFR; model 2 – GDP per capita)

Model	R Square	Adjusted R Square	Std. Error of the Estimate	F	Sig.
1	0.959	0.957	0.042	355.143	0.000
2	0.966	0.961	352.287	197.296	0.000

Table 2: Coefficients
(dependent variable: model 1 – TFR; model 2 – GDP per capita)

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	Constant	1.094	0.023	47.145	0.000
	Years	0.039	0.002	18.845	0.000
2	Constant	5259.955	289.748	18.154	0.000
	Years	819.624	74.103	11.061	0.000
	Years^2	-27.474	4.001	-6.866	0.000

Given the growth in both the TFR and income levels over a long period of time (2000 to 2013), we tested the hypothesis of an existing correlation between these two factors in the course of subsequent analysis. During this period, their dynamics were well approximated by a

linear trend; hence we used a multiple linear regression model. TFR was the dependent variable in the model, while GDP per capita and time were predictors. The main parameters of the model are shown in tables 3-4 (model 1). As the presented data show, no relationship between TFR and GDP per capita was established – the parameters of the equation were not statistically significant.

Table 3: Model Summary
(dependent variable: TFR;
predictors: model 1 – GDP per capita;
model 2 – GDP per capita with a 1-year lag;
model 3 – GDP per capita with a 2-years lag)

Model	R Square	Adjusted R Square	Std. Error of the Estimate	F	Sig.
1	0.950	0.941	0.041	104.333	0.000
2	0.939	0.928	0.046	84.323	0.000
3	0.937	0.925	0.047	81.463	

Table 4: Coefficients
(dependent variable: TFR;
predictors: model 1 – GDP per capita;
model 2 – GDP per capita with a 1-year lag;
model 3 – GDP per capita with a 2-years lag)

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	Constant	1.338	0.146	9.159	0.000
	GDP per capita	-4.160E-005	0.000	-1.703	0.117
	Years	.057	0.011	5.372	0.000
2	Constant	1.183	0.152	7.798	0.000
	GDP per capita	-1.732E-005	0.000	-.607	0.556
	Years	.047	0.013	3.592	0.004
3	Constant	1.104	0.144	7.650	0.000
	GDP per capita	-2.428E-006	0.000	-0.080	0.938
	Years	.040	0.015	2.783	0.018

Similar analysis with account of a possible lag effect also did not confirm any correlation (models 2 and 3 in tables 3-4). An analysis based on a study of the correlation of residuals in regression models also did not uncover any relationship between TFR and GPD per capita in Russia. Pearson correlation coefficients were not high, nor statistically significant (table 5).

Table 5: Correlations between TFR and GDP per capita (model 1 – GDP per capita; model 2 – GDP per capita with a 1-year lag; model 3 – GDP per capita with a 2-years lag)

Model	Indicator	Value
1	Pearson Correlation	-0.457
	Sig. (2-tailed)	0.101
	N	14
2	Pearson Correlation	-0.180
	Sig. (2-tailed)	0.538
	N	14
3	Pearson Correlation	-0.024
	Sig. (2-tailed)	0.935
	N	14

To test the hypothesis about the possibility that people's subjective perceptions about income and material welfare influence fertility we studied time series for consumer confidence indices. We analysed three indices – general, male and female. The analysis we carried out showed that between 2000 and 2016 these indices did not show a single development trend. Until 2007, consumer confidence grew rather stably, however this gave way to a period of high volatility with a trend toward sharp decline (figure 3). Accordingly, given the uninterrupted growth in TFR throughout this period, the hypothesis about a correlation between fertility and subjective ideas about income levels and material welfare was not confirmed.

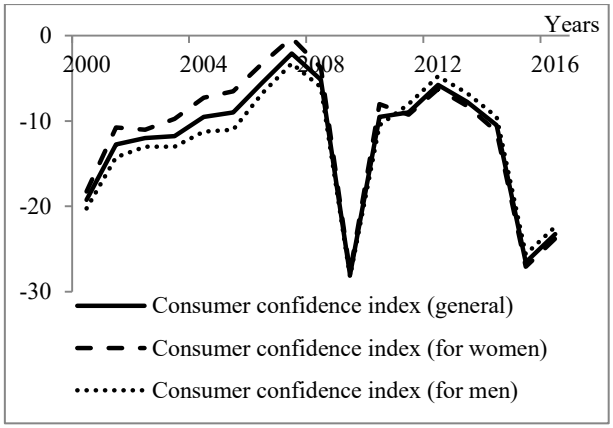


Figure 3: Dynamic of indices of consumer confidence for the Russian population

To test the hypothesis about income inequality possibly affecting fertility we studied time series of the coefficient of income differentiation and the index of income inequality. The completed analysis showed that between 2000 and 2016 both indicators did not have a single growth trend (figure 4). Thus before 2007 the level of inequality in people's income distribution grew,

followed by a short period of stabilization and then decline. As such, given the uninterrupted growth in TFR throughout the same period, the hypothesis that fertility and income inequality are correlated has not been confirmed.

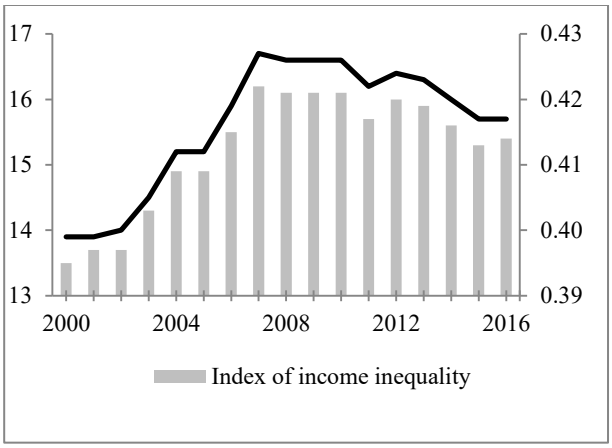


Figure 4: Dynamics of indicators for income differentiation of the Russian population

DISCUSSION

The results of our analysis showed no relationship between people's incomes and fertility rates. We note that this is a departure from the results of many other studies (Maleva and Sinyavskaya 2007; Jones et al. 2011). We believe there could be a number of reasons behind this.

1. There is considerable differentiation between Russian regions as regards fertility rates and people's incomes. Table 6 presents data from 2016 to illustrate this variability.

Table 6: Differentiation of incomes and fertility by Russian regions

Indicators	Min	Max	Ratio of max to min
TFR	1.32	3.35	2.54
Coefficient of income differentiation	4.7	7.6	1.62
GDP per capita (rubles)	116007	2047998	17.65

We note that such differentiation is observed over the entire period we analysed. This highlights an important methodological problem as regards the appropriateness of using average values for analysis. Indeed, given the volatility of input data for income and fertility that characterize the overall situation in the country, they can hardly be considered valid indicators. This in turn may be why similar types of research (fertility and income, fertility and GPD per capita), which uses time series of nationwide indicators as variables, often contradict one another.

One possible way out could be analysis conducted for a single period but for different regions, or for groups of people with different incomes. It is worth noting that the results obtained in different studies of this kind are also rather contradictory. For instance, the aforementioned Maleva and Sinyavskaya noted a negative correlation between income and the number of children for Russian women (Maleva and Sinyavskaya 2007). At the same time, Konig identified that in Hungary high earnings strengthened women's reproductive intentions (whereas in Germany, for example, the correlation was much weaker) (Konig 2012). Moreover, analysis carried out using data from a single period does not lend itself to talking about reliability of results and their stability over time. In the end, the situation for a particular year may be determined by some one-off phenomenon and thus cannot be seen as confirmation of an established cause-and-effect relationship. The results of monitoring research, repeated from time to time using a single methodology, could provide such confirmation. However, research of this type is not currently undertaken in Russia.

It may be that given a high differentiation of indicators, it would be more appropriate to study relationships between different socio-economic and demographic factors across similar types of regions, rather than for the country as a whole. These regions could be identified on the basis of, for example, cluster analysis (using Ward's method). Regions may be clustered by income level or TFR, with subsequent time series analysis inside each typified group. We have partly used such an approach toward time series analysis in our previous research (Shubat et al. 2016; Shubat et al. 2017). The results we obtained, which showed a lack of relationship between fertility and income, define the expedience of differentiating groups beforehand. Each cluster may have its own answer as to whether income is a determinant of fertility for that group of regions.

2. There may be different determinants when birth order is taken into account. For example, research by Wood et al. across seven European countries identified specific economic and institutional factors that particularly influence the birth of second children (Wood et al. 2016); a study by Anna Matysiak and Ivet Szalma showed that the same demographic policy measures may have different impact upon the birth of the first or the second child (Matysiak and Szalma 2014) and so on.

We believe that variability in determinants for children of different birth order is even more likely in Russia. We consider that there is a connection with demographic policy measures aimed at stimulating children of a particular birth order. For example, in 2007, Russia introduced financial incentives for second children – maternity capital (Federal law 256-FZ 2006); in 2011 and 2012 regions introduced payments for third children, whereas from 2018 there is a significant monthly payment aimed at stimulating the birth of first children (Bill 333958-8 2017).

Analysing the relationship between people's incomes and children of different birth order may help to identify a correlation between income and the birth of children of specific order. The same could be fairly said for groups of the population cut by income levels – such correlations may exist in a specific group/ groups by income to be levelled out by other determinants. Unfortunately, this assumption is difficult to verify, as Russian statistics on birth order are not collected.

3. Subjective, rather than objective factors may have a greater influence on fertility. It is known that reproductive behaviour is carried out against a background of certain stereotypes as regards parenthood and childbirth, which influence structural elements and determine social norms on the number of children (Newson 2005), the approach to parenting (Voroshilova. 2016), ideas about the advantages and disadvantages of having children for future parents (Bagirova and Shubat 2014) and so on.

Subjective factors may include ideas about family and children that exist in different religious doctrines. Given Russia's multi-confessional landscape, this too may play a significant role. Orthodox Christianity, the most widespread religion in the country, does not declare having and raising children as the goal of marriage. Orthodoxy does not say that parents have to give children a certain upbringing and education, create for them cultural, material and social living conditions that would allow them to flourish in later life. Meanwhile, having children is encouraged in Islam and in the Muslim tradition the role of the mother is a woman's most important, sacred and invested responsibility.

4. A high "cost of parenting", which is traditionally seen as the reason for reduced fertility in developed countries, manifests itself quite peculiarly in Russia. This may be due to the fact that unlike developed countries with a positive correlation between income and education (for example, this was shown by Rodriguez-Pose and Tselios (Rodriguez-Pose and Tselios 2009)), such a correlation began to manifest weakly only in post-Soviet times. As such, the price of parenting for developed countries, «paid» by mothers with a high income and a higher education, is an issue first of all for mothers with a high level of education, which is not always backed by high income. In this case the so-called subjective cost of maternity (unlike the "objective" "so-called mommy tax, the lost lifetime income a woman can expect by becoming a mother" (Crittenden 2010)). This subjectively perceived cost of motherhood is made up of different components – fewer opportunities for professional self-actualisation, less time for leisure, personal development, external communication and so on. As such, a high "cost of motherhood" in Russia may be a reason for low fertility not for women with high income, but for educated women.

CONCLUSIONS

The research we carried out into time series dynamics did not confirm the results of other studies based on static data on the relationship between fertility and people's incomes. Accordingly, this raises questions about the methodology for analyzing the relationship between dynamic processes with a high volatility of input data. Evidently to receive reliable and stable results, multi-dimensional analysis methods should be integrated into the study of relationships between dynamic time series, including preliminary multi-dimensional data classification. This will enable carrying out analysis on homogenous territorial or temporal segments, which would be more methodologically sound.

We note that the demographic policy measures that are being currently implemented in Russia are developed based on an assumed relationship between people's incomes and fertility. At the same time, there is no unambiguous confirmation of a cause-and-effect relationship. Moreover, there may be a different, inverse, relationship between these indicators: a large number of children may drive greater professional engagement due to the need to provide for a family. The role of latent factors also cannot be ruled out – for example, the provision of housing, orientation toward family values and so on, which may vary across different groups of the population with different income levels and be connected to fertility rates more than income levels.

We believe there is opportunity to extend our research by adapting the methodology for modelling and carrying out multi-dimensional analysis to study demographic processes in Russia. This is necessary to find substantiated answers to questions about real determinants of fertility in Russia. The search for such determinants is particularly topical today, when experts consider the potential for population growth and its reproductive potential to be exhausted.

ACKNOWLEDGMENTS

The article is processed as one of the outputs of the research project "Fertility and parenting in Russian regions: models, invigoration strategies, forecasts", supported by the President of Russian Federation, project no. NSh-3429.2018.6.

REFERENCES

Bagirova A. and O. Shubat. 2014. "Parenthood image and its development in conception of parents work". *Sotsiologicheskie Issledovaniya*, Vol 4, 103-110.

Becker, G.S., Murphy, K.M. and R. Tamura. 1990. "Human capital, fertility and economic growth". *Journal of Political Economy*, Vol 98, Issue 5, S12-S37.

Begall, K. and M.C. Mills. 2012. "The influence of educational field, occupation, and occupational sex segregation on fertility in the Netherlands". *European Sociological Review*. Vol 9, Isuue 4, 720-742. doi: 10/1093/esr/jcs051

Bill 333958-8 “On monthly payments to families with children”. 2017. URL: http://sozd.parlament.gov.ru/bill/333958-7 (access date 04.01.2018)

Coefficient Gini and Coefficient of income differentiation data. Federal State Statistic Service. 2017. URL: http://www.gks.ru/wps/wcm/connect/ rosstat_main/rosstat/ru/statistics/population/level/# (access date 13.11.2017).

Consumer confidence index data. Single inter-departmental information and statistical system (SIDIS). Rosstat, Moscow. 2017. URL: https://fedstat.ru/indicator/33651 (access date 13.11.2017).

Crittenden, A. 2010. “The Price of Motherhood: Why the Most Important Job in the World is Still the Least Valued”. New York, Picador.

Federal law 256-FZ from 29 December 2006 “On additional measures of state support for families with children”.

Fertility rate, total. World Bank data. 2017. URL: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN/ (access date 12.01.2018).

GDP per capita (current US$). World Bank data. 2017. URL: https://data.worldbank.org/indicator/NY.GDP. PCAP.CD?end=2016&start=2014 (access date 12.01.2018).

Jones, L.E., Schoonbroodt, A. and M. Tertilt. 2011. “Fertility theories. Can they explain the negative fertility-income relationship?”. In *Demography & the economy 2011*, J.B. Shoven (Ed.). Chicago: University of Chicago Press, 43-100.

König, S. 2012. *Higher Order Births in Germany and Hungary: Comparing Fertility Intentions in a National Context*. Mannheim: Mannheimer Zentrumfür Europäische Sozialforschung.

Lee, R. and A. Mason. 2010. “Fertility, human capital and economic growth over the demographic transition”. *European Journal of Population*, Vol 26, Issue 2, 159-182.

Maleva, T. and O. Sinyavskaya. 2007. “Socio-economic factors of fertility in Russia: Empirical measurement and social policy challenges”. In T. Maleva and O. Sinyavskaya (Eds.), *Parents and Children, Men and Women in Family and Society*. URL: http://www.socpol.ru/publications/pdf/PiDMiG1_end.indd .pdf (access date 08.01.2018).

Matysiak, A. and I. Szalma. 2014. “Effects of parental leave policies on second birth risks and women's employment entry”. *Population,* Vol 69, Isuue 4, 659-698. doi: 10.3917/popu.1404.0659

Newson, L., Postmes, T., Lea, S. and P. Webley. 2005. “Why are modern families small? Toward an evolutionary and cultural explanation for the demographic transition”. *Personality and social psychology review,* Vol 9, Issue 4, 360-375. doi: 10.1207/s15327957pspr0904_5

Rodriguez-Pose, A. and V. Tselios. 2009. “Education and income inequality in the regions of the European Union”. *Journal of regional science,* Vol 49, Issue 3, 411-437. doi: 10.1111/j.1467-9787.2008.00602.x

Shubat O., Bagirova A., Abilova M. and A. Ivlev. 2016. “The Use of Cluster Analysis for Demographic Policy Development: Evidence From Russia”. In *Proceedings of the 30th European Conference on Modelling and Simulation (*Regensburg, Germany, May 31st-June 03rd, 2016), 159-165.

Shubat O., Bagirova A. and I. Shmarova. 2017. “The Use Of Cluster Analysis To Assess The Demographic Potential Of Russian Regions”. In *Proceedings of the 31th European Conference on Modelling and Simulation (*Budapest, Hungary, May 23rd-May 26th, 2017), 53-59.

Smith, A. An Inquiry Into the Nature and Causes of the Wealth of Nations. Printed at the University Press for T. Nelson and P. Brown, 1827. Edinburgh

Total Fertility Rate data. Single inter-departmental information and statistical system (SIDIS). Rosstat, Moscow. 2017. URL: https://fedstat.ru/indicator/31517 (access date 15.11.2017).

Voroshilova, A. 2016. “Content analysis of parenting image on a social network”. In *The 10th International Days of Statistics and Economics: Conference Proceedings* (Prague, Czech Republic, September 8-10, 2016), 2007-2014.

Van Bavel, J. 2010. “Choice of study discipline and the postponement of motherhood in Europe: The impact of expected earnings, gender composition and family attitudes”. *Demography*, Vol 47, 439-458.

Wood, J.; Neels, K.; and J. Vergauwen. 2016 “Economic and Institutional Context and Second Births in Seven European Countries”. *Population research and policy review*. Vol 35, Issue 3, 305-325. doi: 10.1007/s11113-016-9389-x.

AUTHOR BIOGRAPHIES

OKSANA SHUBAT is an Associate Professor of Economics at Ural Federal University (Russia). She has received her PhD in Accounting and Statistics in 2009. Her research interests include demographic processes, demographic dynamics and its impact on human resources development and the development of human capital (especially at the household-level). Her email address is: o.m.shubat@urfu.ru and her Web-page can be found at http://urfu.ru/ru/about/personal-pages/O.M.Shubat/

ANNA BAGIROVA is a professor of economics and sociology at Ural Federal University (Russia). Her research interests include demographical processes and their determinants. She also explores issues of labour economics and sociology of labour. She is a doctoral supervisor and a member of International Sociological Association. Her email address is: a.p.bagirova@urfu.ru and her Web-page can be found at http://urfu.ru/ru/about/personal-pages/a.p.bagirova/

DYNAMICS OF VOLATILITY SPILLOVER BETWEEN STOCK AND FOREIGN EXCHANGE MARKET: EMPIRICAL EVIDENCE FROM CENTRAL AND EASTERN EUROPEAN COUNTRIES

Ngo Thai Hung
Department of Finance
Corvinus University of Budapest
1093 Budapest, Hungary
E-mail: ngothai.hung@gmail.com

KEYWORDS

Volatility spillover, Central and Eastern European Countries, ARCH, GARCH, EGARCH, Exchange rate, Stock market.

ABSTRACT

We use an Exponential Generalised Autoregressive Conditional Heteroskedasticity (EGARCH) model to investigate the asymmetric volatility spillover effects between the stock market and foreign exchange market in Hungary, Poland, the Czech Republic, Romania and Croatia for the pre- and post- financial crisis periods. The whole of the study period covers from 1st April 2000 to 29th September 2017. The results reveal bidirectional volatility spillover between stock and foreign exchange market of Hungary in all periods and Poland in the post-crisis period, unidirectional volatility spillover in Croatia in the pre-crisis period and from the stock market to exchange market in the Czech Republic during two periods. In the post-crisis period, the two financial markets show the non-presence of the volatility spillover between them in Croatia. Furthermore, empirical results from our analysis provide valuable insights to investors, multinational companies and economic policymakers regarding financial decision making.

INTRODUCTION

The interlinkage between the stock and foreign exchange markets has attracted the attention from academic researchers and practitioners alike for a long time and offers meaningful insight into them. It is widely acknowledged that rapid growth in international financial markets has become substantially more integrated in recent years. The drastic increasing trend in financial assets has been followed by increasing in high demand and supply of foreign currencies. The interdependency has been generated by both the high demand of currencies and equity flows, leading to some degree of interdependency in both securities and currencies. According to (Kanas 2000), the huge increases in interdependency has also increased the volatility spillovers between stock and foreign exchange markets, positive and significant spillovers of volatility transmission may increase the nonsystematic residual international portfolio risk faced by global investors, this might reduce the gains from international portfolio diversification. In reality, volatility analysis helps us to understand the information mechanism in both financial markets including price and volatility spillovers across markets, shock propagation across foreign exchange and stock markets and hedging strategy problems (Chaker Aloui 2007).

In recent finance literature, there is a lot of interest in the financial press on the relationship between returns in the stock and foreign exchange markets because the financial roles of both variables play in prominent portfolio decisions and economic development. Theory suggests two broad channels which link return in stock and exchange markets. The first approach known as the flow-oriented models of exchange rates (Dornbusch and Fischer 1980), this approach posits that there exists a positive linkage between exchange rates and stock prices, and specifically centering on the current account and trade balance. The second approach known as the stock-oriented models of exchange rates (Frankel 1983), these models suggest that the exchange rate is determined by the demand and supply of financial assets such as equities and bonds. More recently, the information of volatility spillovers between the two financial markets has been studied by many researchers for different countries. However, this paper distinguished itself from the previous studies in the following ground. A lot of studies exist in the developed and emerging markets, but for developing markets such as Hungary, Poland, Czech Republic, Romania and Croatia, there are only two investigations of (Lucia Morales 2008) and (Fedorava and Saleem 2010) regarding volatility spillover effect between stock and foreign exchange markets. (Lucia Morales 2008) undertook his study on pre-and post-europe periods in Hungary, Poland, Czech Republic and Slovakia, while (Fedorava and Saleem 2010) focused on the questions of volatility spillover effects of Poland, Hungary, Russia, and the Czech Republic. However, in this research, we studied in the Hungarian, Polish, the Czech Republic, Romanian and Croatian financial markets on subprime financial crisis period, and the methodology adopted is not the same with that of our purposed study.

In this paper, we employ the EGARCH model to address some critical research questions. First, the persistence and asymmetric effects in the conditional volatility of daily returns on stock and exchange indices in Central and Eastern European countries in the pre-and post-crisis periods. Second, whether there is a relationship between the two financial markets or not in these

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

countries. Besides, making comparisons between different countries and various time periods has been performed.

The organization of the paper is as follow. The next section gives the relevant literature review. Section 3 discusses the EGARCH model. Section 4 presents and discusses the estimation results of EGARCH model. Section 5 summarizes the study and concludes with some implications.

LITERATURE REVIEW

There is a rich empirical literature which exists on the investigation of the volatility transmission mechanism of the dynamic linkage between the exchange rates and the stock market. Many of these studies are based on the Generalized ARCH(GARCH) framework to examine volatility spillovers between two financial markets in different countries. It is clear that in the context of the literature of the volatility spillover can be divided into three key points: first, a bidirectional volatility spillover between two markets; second, an unidirectional flow of volatility from stock market to exchange market and vice versa; third, non-persistence of the volatility spillover between two financial markets.

The first study of analyzing volatility spillovers was conducted by (Kanas 2000), he used daily data for the period from January 1st, 1986 to February 28th, 1998 and investigated six industrialised countries, namely the US, the UK, Japan, Germany, France and Canada by employing the bivariate EGARCH model for conditional variances. He found evidence of spillovers from stock returns to exchange rate returns for all countries except Germany, the non-persistence of spillovers from exchange to the stock market.

Yang and Doong (2004) applied the bivariate EGARCH model on weekly data from May 1st, 1979 to January 1st,1999 to examine the nature of the mean and volatility spillover between stock and foreign exchange markets for the G-7 countries. Their empirical evidence supports the asymmetric volatility spillover effect from the stock market to the foreign exchange market in France, Italy, Japan and the US.

Chaker Aloui (2007) explored the nature of the mean, volatility and causality transmission mechanism between stock and foreign exchange markets for the US and some major European markets (France, Germany, Belgium, Spain, Italy). The dataset consisted of daily closing exchange rates and stock indexes for these countries. The asymmetric volatility transmission was illustrated by EGARCH model. He found the asymmetric and long-range persistence volatility spillover effect and evidence of causality in mean and variance in the two markets for both pre and post-europe periods. Additionally, the author confirmed that the stock returns had a more significant effect on the foreign exchange rate for the two subsamples.

Volatility spillovers between stock returns and foreign exchange rates in four Central and Eastern European countries (Hungary, Czech Republic, Poland and Slovakia) was studied by (Lucia Morales 2008). The author applied daily data for the 1999-2006 period that was divided into two sub-periods, pre- Europe and post-Europe. The analyses were carried out on EGARCH model in which apparently confirmed that there was non-existence of significant volatility spillover from stock to foreign exchange markets in these countries. However, the overall results were the lack of significant spillovers from exchange rates to stock returns and volatility in both markets tended to decrease after the countries joined the European Union.

Fedorava and Saleem (2010) investigated the dynamic volatility spillover between stock and currency markets in emerging Central and Eastern European markets of Poland, Hungary, Russia, and the Czech Republic. By estimating a bivariate GARCH-BEKK model using weekly returns. The findings showed strong evidence of direct linkages between the equity markets in term of both returns and volatility and currency markets. The unidirectional volatility spillovers from currency to stock markets had been highlighted for all countries except the Czech Republic in this research.

Natalia and Helena (2014) used a multivariate asymmetric GARCH model to examine volatility spillovers between stock and currency markets in Asian economies, consisted of 2893 observations daily indices, in the period 2003 to 2014. Their results presented evidence of bidirectional volatility spillovers between both markets, independently of the individual country's level of development.

Mozumder et al. (2015) examines the volatility spillover between stock prices and exchange rates in three developed and three emerging countries including Ireland, Netherlands, Spain, Brazil, South Africa and Turkey, across the recent pre-financial-crisis, crisis and post-crisis periods, using weekly data and employing a bivariate EGARCH model. The study concluded that there are the asymmetric volatility spillover effects between both markets in both developed and emerging economies during the financial crisis. Namely, the findings indicated that there is a unidirectional volatility spillover effect running from stock returns to exchange rate returns in developed countries. Volatility spillover direction between the two markets is opposite in the emerging countries, but there is a bidirectional volatility spillover both financial markets in Brazil.

Dynamics of volatility spillover between the stock market and foreign exchange market in Asian countries (China, Hong Kong, India, Japan, Pakistan and Sri Lanka) was empirically investigated by (Jebran and Iqbal 2016) by using EGARCH model. This study considered daily data from January 4th, 1999 to January 1st, 2014. Their research pointed out bidirectional asymmetric volatility spillover between the stock market and foreign exchange market of Pakistan, China, Hong Kong and Sri Lanka. For India, the findings shown unidirectional transmission of volatility from stock to exchange markets. Nevertheless, the analysis also

confirmed no evidence of volatility spillover in both markets in case of Japan.

It is clear from the above review of relevant literature that there are mixed results on volatility spillover effects from various periods as well as different countries. This study aims at contributing to the existing literature by filling the gap of exploring knowledge about the volatility transmission mechanism of the dynamic linkage between the exchange rates and the stock market in the selected countries by adopting an empirical approach based on a multivariate EGARCH model. Also from EGARCH model, examining the relationship between stock and exchange rate movements has been estimated, the questions of the previous researches have centered only on the first moments of the joint stock and exchange rate distributions. Another contribution of our study is considering daily data for the pre-crisis period of 9 years and the post-crisis period of 10 years to research in the long run because daily data capture more information than weekly and monthly data and so as to ascertain the extent to which the recent financial crisis affected the link in question. Furthermore, empirical results from our analysis are of great interest to investors, multinational companies and economic policymakers regarding financial decision making.

METHODOLOGY

Data

The data set consists of daily closing stock and exchange rate prices for five Central and Eastern European countries, we took daily data covering the period from 1st April 2000 to 29th September 2017. The entire investigation period is subdivided into two sub-periods: Pre-crisis period: 1st April 2000 to 29th August 2008. Post-crisis period: 1st September 2008 to 29th September 2017. The whole period in present study divides into pre and post financial crisis period on the basis of certain justifications. The reason for collecting daily data is to capture more the precise information content of changes in stock prices and exchange rates than doing with weekly or monthly data (Jebran and Iqbal 2016), and better able to capture the dynamics between variables (Agrawal et al. 2010). The sample five European countries includes Hungary, Poland, Czech Republic, Romania and Croatia and their stock indexes are: Budapest Stock Exchange BUX, Warsaw Stock Exchange WIG, Prague Stock Exchange PX, Bucharest Stock Exchange BET and Zagreb Stock Exchange respectively. The national currencies of these countries are Hungarian Forint HUF, Polish Zloty PLN, Czech Koruna CZK, Romanian Leu RON and Croatian Kuna HRK respectively.The exchange rate series are from the European countries are stated in US dollars per local currency (note that value of the dollar). Because stock markets operate for five trading days from Monday to Friday and foreign exchange markets operate six trading days (excluding weekends and holidays), this research makes a common data series by adjusting the dates of both the stock and exchange rate indices. The data for our empirical investigation is obtained from Bloomberg, accounted by the Department of Finance, Corvinus University of Budapest. The daily return data series are calculated as $R_{i,t} = 100 \times \ln(P_{i,t}/P_{i,t-1})$, where $P_{i,t}$ is the price level of market i (i = 1 for the stock market and i = 2 for the foreign exchange rate) at time t. The plots of stock prices and exchange rate series for five countries in sample illustrate the volatility that occurs in bursts. (not reported to conserve space).

Model Specification

Unit root test

The stationary of the series is considered by commonly checking through Phillips and Perron PP (1988) and Dickey and Fuller ADF (1979) methods to ensure that the results of the analysis are not spurious. These tests have been implemented to confirm that the data were stationary at level.

EGARCH model

The empirical study for examining whether the volatility of stock returns affects and is affected by the volatility of exchange rate returns is aimed to be captured by employing Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) model developed by (Nelson 1991). The EGARCH specification is applied to test whether the volatility spillover effects are asymmetric. The simple GARCH model enforces a symmetric effect of volatility (positive shocks) and is not able to capture asymmetric shocks (negative shocks) because of the conditional variance being a function of lagged residuals and not their signs. There is no such restriction in EGARCH model on the parameters, the EGARCH model is able to capture both symmetric and asymmetric shocks. Therefore, numerous empirical studies based on the EGARCH framework to specify volatility spillovers between different financial assets in different countries. For instance, except for foregoing mentioned scholars, (Mishra et al. 2007; Choi et al. 2010; Adiasi et al. 2008; Qayyum and Kamal 2006; Okpara and Odionye 2012; Beer and Hebein 2011). In this study, we applied EGARCH(1,1) model to examine the transmission mechanism of volatility separately for each selected country.

EGARCH model for volatility spillover from foreign exchange market to stock market

$$R_t = \alpha_0 + \alpha_1 R_{t-1} + \alpha_2 R_{t-1(ER)} + \varepsilon_t \quad (1)$$

$$h_{t(SP)} = \beta_0 + \beta_1 h_{t-1} + \beta_2 \left| \frac{\varepsilon_t - 1}{\sqrt{h_{t-1}}} \right| + \phi \frac{\varepsilon_t - 1}{\sqrt{h_{t-1}}} + \delta_{(resid(ER))} \quad (2)$$

The equation (1) and (2) represent the EGARCH(1,1), which is applied for examining volatility spillover from foreign exchange market to stock market in each country.

EGARCH model for volatility spillover from stock market to foreign exchange market

$$K_t = a_0 + a_1 K_{t-1} + a_2 K_{t-1(SP)} + \varepsilon_t \quad (3)$$

$$h_{t(ER)} = \gamma_0 + \gamma_1 h_{t-1} + \gamma_2 \left| \frac{\varepsilon_t - 1}{\sqrt{h_{t-1}}} \right| + \varphi \frac{\varepsilon_t - 1}{\sqrt{h_{t-1}}} + \psi_{(resid(SP))} \quad (4)$$

Table 1: Description of Parameters Equations (1)-(4)

Explanation	S	E
The conditional mean equation	(1)	(3)
The conditional variance equation	(2)	(4)
Return	R_t	K_t
Intercept	α_0	a_0
Measuring the effects of previous day's return on today's return	α_1	a_1
Measuring the effects of exchange rates changes on stock returns	α_2	
Measuring the effects of stock returns on exchange rate returns		a_2
Error term	ε_t	ε_t
Log of conditional variance	$h_{t(SP)}$	$h_{t(ER)}$
Volatility constant	β_0	γ_0
Function of volatility (consistency)	β_1	γ_1
Volatility reaction to change in news	β_2	γ_2
Measuring asymmetric effect of volatility	ϕ	φ
Volatility spillover	δ	ψ

Note: S is stock return, E is exchange rate return

The procedure for measuring volatility spillover of this study is implemented in the following stages. An initial step we provide descriptive statistics for stock and exchange rate returns to summarize the statistical characteristics of our sample. We then carry out the stationary test including ADF and PP test on each of the concerned variables. Next step, identifying and estimating an autoregressive and moving average (ARMA) model for the mean equation, using the residuals of the mean equation to test for ARCH effect (the significant value of chi-square depicts ARCH effect in the underlying variable). EGARCH model shall be employed on data in which ARCH effect exists. After making sure that there exists ARCH effect, we have specified and estimated the volatility spillover between stock market and foreign exchange rate market. Finally, residual diagnostics have been performed.

RESULTS

Descriptive statistics for stock and foreign exchange returns as well as unit root are reported in Table 2 and Table 3. The analyses reveal that sample means of stock return are positive and significantly different from zero for five countries except for the Czech Republic and Croatia in the post-crisis period. The sample variances range from 0.69% for Croatia to 1.62%. for Hungary. Similarly, Hungarian and Polish exchange markets have the highest daily average return over the study period. Skewness and Kurtosis coefficients indicate that return series are far from the normal distribution, this is formally confirmed by The Jarque-Bera test statistics. Finally, all exchange rate changes and stock returns series are found to be stationary at level (e.i I(0)) at the 1% significance level according to the PP & ADF statistics.

Table 4 represents the results of the purpose of ARCH effect for the underlying variables (stock prices and exchange rate) over the study periods. The ARCH effect illustrates the presence of autocorrelation and heteroskedasticity issues in data. The result shows that there is the strong evidence of the existence of ARCH effect in all concerned series. EGARCH (1,1) can be employed on data having ARCH effect in data.

Examining volatility spillover between stock and exchange market by using EGARCH (1,1) model is the final step. We have studied on each market information spillover separately for each country. First, we have conducted analyses by examining volatility spillover from the foreign exchange market to the stock market, after that we have continued to examine volatility spillover from the stock market to the exchange market. For selecting the appropriate lag length of each model, the basis of Akaike's information criterion has been selected.

Table 5 presents the EGARCH estimations for both the mean and conditional variance equations. The mean equation results show that the changes in the exchange rate have a significant negative impact on stock returns of Hungry, Poland over the study period and the Czech Republic, Romania, Croatia in the post-crisis period, nevertheless, insignificant for the Czech Republic, Romania and Croatia in the pre-crisis period. The significant negative impact of foreign exchange market on stock market reveals that changes in exchange market could reduce stock returns of these countries, this decreases the profitability and stock prices of the firms. The negative effect of exchange market would create fluctuation in trade balance and the competitiveness of the country. As a result, it would decrease real income and economic growth (Jebran and Iqbal 2016). The results of the negative impact of the foreign exchange rate market on the stock return are similar to those of (Aloui 2007; Yang and Doong 2011; Jebran and Iqbal 2016). The findings support the theoretical prediction of stock-oriented models in which reported that there is a negative relationship between foreign exchange rate and stock price. In case of the Czech Republic, Romania and Croatia, there is an insignificant effect of exchange rate on stock returns in the pre-crisis period may postulate the effective hedging strategies against the currency risk in these countries. The results of our empirical analysis indicate that there is also an insignificant linkage between stock market changes and foreign exchange rate dynamics for all countries over study period except Romania in the post-crisis period. The mean equation reveals that stock market fluctuations have a significant negative effect on exchange rate returns in Romania recent years, it supports the theoretical prediction of

portfolio balance model which points that exchange rates respond to the demand for and supply of stock market. Consequently, an increase in domestic stock prices will lead investors to sell foreign assets in the market for the purpose of purchasing domestic assets. This result in line with a study like (Jebran and Iqbal 2016). Additionally, as regards to other countries, the weak or no impact of stock prices on exchange rate supports the theoretical prediction of monetary approach which presumes that there is no linkage between exchange rate and stock prices. These results are consistent with (Ngo Thai Hung 2017).

Turning to the second moment interdependencies, the variance equation results indicate coefficient δ which measures volatility spillover form exchange market to stock market and indicates whether this spillover is asymmetric, is statistically significant for Hungary, Poland in all the periods, Romania and Croatia in the pre-crisis period. For the pre-crisis period, the coefficient is positive in case of Hungary, Poland and Croatia, while negative in case of Romania. For the post-crisis period, the coefficient is positive for Hungary and Poland, while negative in case of Croatia. The positive coefficient illustrates that foreign exchange market volatility is increasing the volatility of the stock market, on the other hand, negative coefficient shows that foreign exchange market volatility is decreasing the volatility of the stock market. In case of Hungary, Poland and the Czech Republic, the findings are consistent with (Lucia Morales 2008; Chaker Aloui 2007; Natalia and Helena 2014). ψ measures volatility spillover from stock prices to exchange rates. The coefficient is statistically significant for Hungary and the Czech Republic in all periods, however, insignificant in case of Poland and Romania in the pre-crisis period and Croatia in the post-crisis period. The coefficient is negative in all cases. This negative coefficient describes that stock market volatility is decreasing the volatility of foreign exchange market. It is important to note that we find no volatility spillover from exchange market to stock market in case of Czech Republic in both periods and Romania and Croatia in the post-crisis period, and no volatility spillover from the stock market to the exchange market in case of Poland and Romania in the pre-crisis period and Croatia the post-crisis period. These results are in line with (Fedorava and Saleem 2010) and inconsistent with (Lucia Morales 2008).

Briefly, the results are mixed when we compare the volatility spillover with different countries and during two periods because changes in volatility spillover between stock returns and foreign exchange market have changed over time, in particular, have increased in post-crisis period, it is consistent with the notion that financial market integration has increased after crisis period. The results show that there is a bidirectional volatility spillover between stock and foreign exchange market in Hungary in all period and Poland in the post-crisis period that represents the information inefficiency of these stock markets, while unidirectional volatility spillover in Croatia in pre-crisis period and from the stock market to the exchange market in the Czech Republic during two periods. However, no volatility spillover for Croatia in the post-crisis period, this implies that effective strategies against the stock market and exchange rate fluctuations. The absence of volatility spillover from exchange market to stock market in case of the Czech Republic (two periods), Romania and Croatia (post-crisis period) could indicate the effective hedging strategies against currency risk. Finally, the asymmetric spillovers from stock returns to exchange rates have all the positive signs, they are interpreted as follows: good news has a greater impact on volatility than unexpected bad news. On the other hand, the asymmetric spillovers from exchange rates to stock returns have all the negative signs implies that negative shocks generate more volatility than positive shocks of the same magnitude.

In order to evaluate the robustness of the estimation results, we examined the ARCH effect on the residuals of each model to determine whether the ARCH effect still exists in model. The null hypothesis is that there is ARCH effect. It can be seen in Table 5, the results of ARCH test illustrate that we find strong evidence that there is no ARCH effect for all series considered except only for Romania when we estimate the volatility spillover from the stock market to the foreign exchange market in the pre-crisis period. It is similar to the study of (Melik Kamisly et al. 2015) and also a limitation of this investigation. Hence, modeling the EGARCH model can successfully capture the price volatility interaction between stock and exchange markets.

CONCLUSION

In this study, we have investigated the empirical dynamics of volatility spillover effects between the stock market and foreign exchange market in Central and Eastern European countries i.e… Hungary, Poland, the Czech Republic, Romania and Croatia across the pre-crisis and post-crisis periods using EGARCH model. Our empirical evidence shows that there is a bidirectional volatility spillover between stock and foreign exchange markets in Hungary in all periods and Poland in the post-crisis period. The results also reveal unidirectional volatility spillover in Croatia in the pre-crisis period and from the stock market to the exchange market in the Czech Republic during two periods. In the post-crisis period, the two financial markets show the non-presence of the volatility spillover between them in Croatia. The spillovers are asymmetric in nature in all financial markets. Volatility spillover from stock returns to exchange rates have decreased after the crisis period. The volatility persistence indicates that there was volatility persistence in all series in all periods, in general, the persistence of exchange market volatility was found greater than stock market volatility.

Our findings have several important economic and finance implications for economic policymakers and investors. First, international portfolio managers and

hedgers may be better able to understand how the two financial markets interrelated overtime, they might provide them benefit in forecasting the behavior of one market by capturing the other market information. Second, the information concerning the nature of volatility transmission across stock and exchange markets in the country would be important for policymakers and decision makers from an economic stability perspective as financial markets integration through exchange rates imply financial sector integration. Third, for investors, this should be particularly important when they want to make an efficient portfolio, can apply these results in reducing their risk, increase their returns and make decisions in the selected markets.

REFERENCES

Agrawal G., Srivastav A.K. and Srivastava A. 2010. "A Study of Exchange Rates Movement and Stock Market Volatility". *International Journal of Business and Management* 5, No. 12 (Dec), 62-73.

Beer. F and Hebein. F. 2011. "An Assessment of The Stock Market and Exchange Rate Dynamics in Industrialized and Emerging Markets". *International Business Economic Research Journal* 7, No.8, 59–70.

Chaker Aloui. 2007. "Price and Volatility Spillovers Between Exchange Rates and Stock indexes for the pre-and post-Euro period". *Quantitative Finance* 7, No. 6(Dec), 669-685.

Choi DF, Fang V, Fu TY. 2010. "Volatility Spillovers Between New Zealand Stock Market Returns and Exchange Rate Changes Before and After the 1997 Asian Financial Crisis". *Asian Journal of Finance & Accounting* 1, No. 2, 106–117.

Dornbusch. R and Fischer S.1980. "Exchange Rates and The Current Account". *American Economic Review* 70, No. 5, 960-971.

Dickey, David A., and Wayne A. Fuller. 1979. "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American statistical association* 74 No. 366a , 427-431.

Frankel J. A. 1983. "Monetary and Portfolio-Balance Models of Exchange Rate Determination' in Bhandari J S and Putnam B H (eds) Economic Interdependence and Flexible Exchange Rates", *Cambridge: MIT Press*.

Fedorova E. and Saleem K. 2009. "Volatility Spillovers Between Stock and Currency Markets: Evidence from Emerging Eastern Europe". Czech Journal of Economics and Finance 60, No. 6, 519-533.

Jerban K. and Iqbal A. 2016. "Dynamics of Volatility Spillover Between Stock Market and Exchange Market: Evidence from Asian Countries". *Financial Innovation* 2, No. 3, 1-20.

Kanas A. 2000. "Volatility Spillovers Between Stock Return and Exchange Rate Changes: International Evidence". *Journal of Business Finance and Accounting* 27, No. 3, 447-467.

Lucia Morales. 2008. "Volatility Spillovers Between Stock Returns and Foreign Exchange Rates: Evidence from Four Eastern European Countries". *In the conference proceedings of the Financial management Association (FMA) European Conference*, 4-6 June 2008, Prague, the Czech Republic.

Mishra A. K, Swain N. and Malhotra D. K. 2007. "Volatility Spillover Between Stock and Foreign Exchange Markets: Indian Evidence". *International Journal of Business* 12, No. 3, 343-359.

Mozumder N., Vita G., Kyaw K. S. and Larkin C. 2015. "Volatility Spillover Between Stock Prices and Exchange Rates: New Evidence Across the Recent Financial Crisis Period". *Economic Issues* 20, No. 1, 43-64.

Melik Kamisli, Serap Kamisli and Mustafa Ozer. 2015. "Are Volatility Transmissions Between Stock Market Returns of Central and Eastern European Countries constant or dynamic? Evidence from MGARCH Models". *In the conference proceedings of* 10th MIBES Conference, 15-17 October 2015, Larisa, Greece, 190-203.

Natalia V. and Helena C. 2014. "Volatility Transmission Between the Stock and Currency Markets in Emerging Asia: the Impact of the Global Financial Crisis". *Research Institute of Applied Economics* 31, 1-26.

Nelson D. B. 1991. "Conditional Heteroskedasticity in Asset Returns: A New Approach". *Econometrica* 59, No. 2, 347-370.

Ngo Thai Hung. 2017. "An Empirical Test on Linkage between Foreign Exchange Market And Stock Market: Evidence From Hungary, Czech Republic, Poland And Romania". *European Scientific Journal* 13, No. 31, 25-38.

Okpara G.C. and Odionye J.C. 2012. "The Direction of Volatility Spillover between Stock Prices and Exchange Rate: Evidence From Nigeria. *Elixir Finance* 42, 6410–6414.

Qayyum A. and Kamal A. 2006. "Volatility Spillover Between the Stock Market and the Foreign Market in Pakistan". *Pakistan Institute of Development Economics Working Papers* 7.

Phillips, Peter CB, and Pierre Perron. 1988. "Testing for a unit root in time series regression." *Biometrika* 75, No. 2 , 335-346.

Yang S. and Doong S. 2004. "Price and Volatility Spillovers Between Stock Prices and Exchange Rates: Empirical Evidence from The G-7 Countries". *International Journal of Business and Economics*, 3 No. 2, 139-153.

Table 2: Descriptive statistics of daily return of stock indices

Countries	Hungary	Poland	Czech	Romania	Croatia
Panel A. Pre- crisis period					
Mean (%)	0.04	0.03	0.05	0.11	0.07
SD (%)	1.39	1.30	1.26	1.78	1.39
Skewness	-0.0883	-0.2559	-0.1862	-0.1390	0.5367
Kurtosis	4.3107	5.1726	5.4575	21.9218	16.726
Jarque-Bera	157.88**	449.76**	559.39**	313320**	16342.68**
PP test	-44.47*	-44.92*	-44.96*	-43.33*	-44.93*
ADF test	-44.49*	-44.79*	-44.96*	-43.36*	-44.92*
N	2166	2166	2173	2099	2069
Panel B. Post- crisis period					
Mean (%)	0.02	0.02	-0.01	0.01	-0.02
SD (%)	1.62	1.22	1.4927	1.52	1.18
Skewness	-0.1033	-0.5254	-0.6210	-0.7250	0.0734
Kurtosis	11.4429	7.9297	20.3306	15.5740	27.2175
Jarque-Bera	6746.18**	2407.276	28642.34**	15240**	55425.20
PP test	-45.07*	-34.53*	-39.97*	-44.89*	-44.53*
ADF test	-35.42*	-42.73*	-44.53*	-44.93*	-25.32*
N	2270	2274	2277	2283	2268

Table 3: Descriptive statistics of daily changes in exchange rate return

Countries	Hungary	Poland	Czech	Romania	Croatia
Panel A. Pre- crisis period					
Mean (%)	-0.0196	-0.026	-0.033	0.0129	-0.02
SD (%)	0.78	0.6937	0.7049	0.6516	0.69
Skewness	0.4141	0.2921	0.0008	0.9338	-0.0863
Kurtosis	6.1554	5.3131	4.2452	20.066	5.0085
Jarque-Bera	960.48**	513.65**	140.39**	25778.4**	350.36*
PP test	-46.89*	-43.80*	-47.20*	-48.94*	-50.02*
ADF test	-46.89*	-43.80*	-47.19*	-48.80*	-49.99*
N	2166	2166	2173	2099	2069
Panel B. Post- crisis period					
Mean (%)	0.02	0.02	0.01	0.02	0.01
SD (%)	1.0559	1.04	0.85	0.77	0.69
Skewness	0.1971	0.2039	-0.1349	0.2228	-0.0347
Kurtosis	6.0054	6.6307	8.3099	6.6083	5.6903
Jarque-Bera	869.04**	1264.76**	2681.99**	1257.42**	684.43**
PP test	-47.65*	-47.85*	-48.39*	-45.12*	-47.68*
ADF test	-47.61*	-47.71*	-48.38*	-34.59*	-47.68*
N	2270	2274	2277	2283	2268

*Note (Tabel 2&3): SD represents standard deviation. N: Observations. ** denotes the level of significance at 5%, * indicates p < 1%. Critical value at 1%, 5% and 10% are -3.43, -2.86 and -2.56 respectively.*

Table 4. ARCH test

Countries	Hungary	Poland	Czech	Romania	Croatia
Panel A: Pre-crisis period (Stock indices)					
Constant	1.254*	1.021*	0.892*	1.525*	1.137*
AR(1)	0.061*	0.042**	0.071*	0.390*	0.271*
ARCH test	80.09*	110.57*	141.74*	362.65*	240.30*
Panel B: Post-crisis period (Stock indices)					
Constant	1.137*	0.579*	0.659*	0.942*	0.544*
AR(1)	0.290*	0.020	0.228*	0.306*	0.297*
ARCH test	363.55*	309.09*	559.10*	411.67*	550.14*
Panel A: Pre-crisis period (Exchange rates)					
Constant	0.515*	0.331*	0.402*	0.2085*	0.335*
AR(1)	0.097*	0.137*	0.012	0.567*	0.108*
ARCH test	25.01*	72.17*	21.37*	576.77*	59.96*
Panel B: Post-crisis period (Exchange rates)					
Constant	0.498*	0.408*	0.291*	0.256*	0.276*
AR(1)	0.042**	0.158*	0.084*	0.038	0.173*
ARCH test	251.82*	403.89*	325.83*	291.60*	151.88*

Note: ARCH test is the arch effect test

Table 5: Volatility spillover between stock and foreign exchange market

Countries	Hungary	Poland	Czech	Romania	Croatia
Volatility spillover from foreign exchange market to stock market					
Panel A: Pre-crisis period					
α_0	0.041	0.053**	0.074*	0.075*	0.099*
α_1	0.033	0.061*	0.049**	0.143*	0.037
α_2	-0.083**	-0.116*	-0.022	0.021	-0.014
β_0	-0.082*	-0.080*	-0.133*	-0.221*	-0.167*
β_1	0.943*	0.979*	0.939*	0.916*	0.924*
β_2	0.147*	0.112*	0.192*	0.406*	0.285*
ϕ	-0.052*	-0.028*	-0.095*	-0.033*	-0.008
δ	0.055*	0.014*	-0.0006	0.134*	-0.092*
ARCH LM(1)	0.019(0.89)	3.21(0.07)	4.48(0.03)	0.39(0.53)	2.14(0.14)
Panel B: Post-crisis period					
α_0	0.031	0.033	0.005	0.027	0.015
α_1	-0.016	0.052**	0.024	0.067*	0.055*
α_2	-0.11*	-0.102*	-0.111*	-0.049**	-0.066*
β_0	-0.110*	-0.094*	-0.159*	-0.285*	-0.156*
β_1	0.986*	0.986*	0.980*	0.957*	0.987*
β_2	0.152*	0.123*	0.209*	0.393*	0.200*
ϕ	-0.063*	-0.045*	-0.067*	-0.062*	-0.036*
δ	0.016**	0.027*	0.004	0.011	-0.001
ARCH LM(1)	0.63(0.42)	0.76(0.37)	0.61(0.43)	2.38(0.12)	0.06(0.79)
Volatility spillover from the stock market to foreign exchange market					
Panel A: Pre-crisis period					
a_0	-0.017	-0.033**	-0.031**	0.078*	-0.036*
a_1	0.010	0.072*	-0.013	0.0004	-0.072*
a_2	-0.018	-0.004	-0.016	-0.002	0.0002
γ_0	-0.130*	-0.216*	-1.025**	-0.100*	-0.057*
γ_1	0.864*	0.901*	-0.367**	0.993*	0.995*
γ_2	0.080*	0.181*	0.073**	0.123*	0.069*
φ	0.078*	0.074*	0.073*	-0.038*	-0.001
ψ	-0.046*	-0.010	-0.051**	-0.004	-0.015*
ARCH LM(1)	0.48(0.48)	0.60(0.43)	0.01(0.90)	80.7(0.00)	0.02(0.86)
Panel B: Post-crisis period					
a_0	0.011	0.016	0.017	0.011	0.014
a_1	-0.026	-0.04**	-0.005	-0.008	-0.032
a_2	-0.001	-0.025	0.015	-0.023**	-0.018
γ_0	-0.046*	-0.062	-0.077*	-0.050*	-0.047
γ_1	0.996*	0.994*	0.992*	0.993*	0.996*
γ_2	0.058*	0.078*	0.091*	0.057*	0.047*
φ	0.025*	0.028*	0.019*	0.0211*	0.029*
ψ	-0.016*	-0.022*	-0.021*	-0.012*	-0.007
ARCH LM(1)	0.08(0.76)	1.25(0.26)	0.34(0.55)	0.14(0.70)	4.49(0.03)

*Note: ** denotes the level of significance at 5%, * indicates p < 1%. Numbers in parentheses are probability.*

FUZZY LOGIC MODELLING OF THE RUSSIAN DEMOGRAPHIC SPACE

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

KEYWORDS
Fuzzy logic, clustering, demographic potential, Russian regions

ABSTRACT

Demographic processes are extremely difficult to manage and require the different methods of their research. Our studies aimed at modelling the demographic space of Russian regions by using fuzzy clustering. Our analysis is based on the indicators of the regions' demographic potential. We used our own original methodology combining the statistical procedures of fuzzy clustering and expert survey data. We considered indicators characterizing the reproduction potential and variables characterizing the potential quality of the future population. As a result of fuzzy clustering, five clusters were formed. Our experts evaluated the reproduction potential and the quality of the future population for each cluster. The data for each region were used to calculate their reproduction potential and the quality of their future population. In comparison to hard clustering, fuzzy clustering enhances the flexibility of evaluation: our assesments of each region do not depend exclusively on the potential of the nearest cluster, as we also take into consideration the region's possible similarities with other neighbouring clusters with different potential. Such modelling allows us to identify those Russian regions that could be considered as 'growth points' in the implementation of demographic policy.

INTRODUCTION

Demographic processes are extremely difficult to manage for a number of reasons. Firstly, they are influenced by a range of external and internal political, economic, cultural, and religious factors. Secondly, demographic processes are inertial: it will take time even for the most efficient demographic policy to make a difference. In addition, the goals of the country's demographic policy might change over the course of time, which will result in a further slowdown of demographic processes. Thirdly, demographic processes and situations in some regions can differ considerably. Given these factors, studying demographic processes to seek out determinants and develop robust forecasts is highly topical. Fertility is of utmost importance, since it determines natural population growth. However, this demographic process provokes the most vigorous scientific and political debate in Russia – questions related to fertility determinants and ways to regulate it remain unresolved.

It should be noted that the latter situation can be found only in certain countries: for instance, some researchers point out differences in the demographic development between northern and southern Italy (Pastuszka 2017), provinces of China (Wang et al. 2017), and Russian regions (Shubat et al. 2017). In Russian regions, demographic indicator values vary greatly (see table 1).

Table 1: Minimax values of demographic indicators in Russian regions in 2016 (Russian Regions 2017)

Indicators	Minimum		Maximum		Ratio of maximum to minimum values
	Value, ‰	Region	Value, ‰	Region	
Crude birth rate	9.2	Leningrad	23.2	Tuva	2.5
Crude death rate	3.3	Ingushetia	17.9	Pskov	5.4
Crude marriage rate	3.8	Ingushetia	8.6	Kam-chatka	2.3
Crude divorce rate	0.9	Chechnya, Ingushetia	6.4	Ma-gadan	7.1
Migration rate	-103.2	Chukotka	140.7	Mos-cow	-

The uncertainty of the demographic situation in countries with highly imbalanced regional development can be reduced if we apply adequate mathematical and statistical tools for demographic analysis and modelling. One such tool is clustering. Traditional cluster analysis, however, does not always provide results reliable enough to develop a demographic policy. One of the reasons for this is that clustering is often based on the researcher's intuition, which also means subjectivity.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

To enhance objectivity in cluster analysis when modelling demographic processes, we can use fuzzy logic. The fuzzy set theory appeared in the mid-1960s, when Lotfi A. Zadeh proposed a new approach to describe objects and systems which are difficult to formalize (Zadeh 1965). Zadeh showed that for more realistic modelling, it is necessary to go beyond the traditionally accepted standard quantitative methods and introduce so-called linguistic variables into the analysis. The values of these variables can be words or sentences, which enables us to take into account the uncertainty or 'fuzziness' of human knowledge. One of Zadeh's key ideas was to extend classical set theory. In the case of fuzzy sets, the membership function can vary between 0 and 1 rather than just take the value of 0 or 1.

Fuzzy set methods make it possible for one and the same object to belong to several or even all clusters at the same time but with different membership degrees. In many cases, the results of fuzzy clustering seem to be more natural and visual than those of hard clustering. Thus, the application of fuzzy clustering enables us to formalize the uncertainty inherent in demographic processes.

It should be noted that fuzzy set theory is used comparatively rarely in demography. Generally, fuzzy logic methods are used for population forecasts (Sasu 2010; Abbasov and Mamedova 2003). In the Russian Science Citation Index, which is a national database that contains over 12 million scientific publications by Russian authors, only five such publications have been registered since 2005. Seletkov and Martsenyuk have used fuzzy sets to analyze population dynamics (Seletkov and Martsenyuk 2016); Shokin and Fedorov have investigated the relationship between dynamic macroeconomic and demographic indicators in Russian regions (Shokin and Fedorov 2012); and Kulikova and Nikishina have forecasted the population of a specific region (Kulikova and Nikishina 2015).

This research is aimed at modelling the demographic space of Russian regions by using fuzzy clustering. Our analysis is based on the indicators of the regions' demographic potential. Thus, such modelling allows us to identify those Russian regions that could be considered as 'growth points' in the implementation of demographic policy.

The paper is structured as follows: first, we justify the need to use methods with unclear logic to analyse the demographic space of Russian regions. Then we describe the methodology we developed to study demographic potential of Russian regions. Afterwards we outline the results of applying it to assess the potential of population reproduction and the quality potential of the future population of Russian regions. We also explain possible reasons for the imbalance between these two types of demographic potentials and describe possible points of debate in our analysis.

DATA AND METHODS

In this research, we used our own original methodology combining the statistical procedures of fuzzy clustering and expert survey data. The main stages in the application of this methodology are as follows:

1. We created a database on the demographic potential of Russian regions. It should be noted that demographic potential is a comparatively new concept in social and economic studies. There are different approaches to evaluating demographic potential: Goraj et al. measure regions' demographic potential with the help of quantitative population indicators (Goraj et al. 2016), while Dobrokhleb and Zvereva believe that qualitative indicators (in particular life expectancy) can also serve as indicators of demographic potential (Dobrokhleb and Zvereva 2016). We believe that to measure the demographic potential of regions we should include both quantitative and qualitative indicators.

Therefore, this research considers two sets of indicators: the first set deals with a region's reproduction potential and reflects the quantitative aspects of its demographic potential. These variables include the following:

- the rate of natural increase, which is the crude birth rate minus the crude death rate;
- the proportion of children (aged between 0 and 15) in the population;
- the rate of reproduction potential realization – the quotient of the total number of child births by the total number of conceptions (calculated as the sum of the number of child births and abortions).

The second set describes the qualitative aspect of demographic potential (the potential quality of the future population). These variables include the following:

- the proportion of healthy children in the overall population of school-age children;
- the disability rate in the child population (proportion of disabled children aged between 0 and 15);
- the juvenile crime rate (proportion of juvenile delinquents aged 14-17);
- the involvement of children in supplementary educational programs (proportion of children aged 5-18 provided supplementary educational services);
- the proportion of state (municipal) educational institutions meeting modern standards in the total number of such institutions.

The data were provided by the Federal State Statistics Service. Due to certain peculiarities of the national system for the collection of statistical data, the available information did not refer to one year but characterized the situation in the given regions in 2015 and 2016. The resulting database included the characteristics of the demographic potential of 79 Russian regions (six regions were excluded because the necessary data was lacking).

2. To evaluate the demographic potential of these regions, fuzzy clustering was applied on the basis of the fuzzy *c*-means (FCM) algorithm (Bezdek 1981). At the preliminary stage, the initial data were normalized.

It is convenient to consider the result of the clustering of M elements into c clusters by using the following characteristic function:

$$U = [\mu_{ki}], \mu_{ki} \in \{0,1\}, k \in \overline{1,M},\ i \in \overline{1,c}, \quad (1)$$

where the kth row of matrix U shows the membership degree of the kth object in the ith cluster (0 designates that the object does not belong to the cluster while 1 means that it does). The membership function in the fuzzy set can take any values within the interval [0, 1].

The algorithm of fuzzy c-means has the following parameters: c is the number of clusters while m is the exponential weight $m \in (1, \infty)$. Exponential weight m affects the matrix of membership degrees. The higher m is, the fuzzier the final c-means matrix; at $m \to \infty$, the degrees of membership tend to $1/c$, which is a bad solution as all objects belong to all clusters with the same membership degree.

In this analysis, Euclidean distance was used as a measure of distance. This algorithm enabled us to solve the task of criterion minimization:

$$\sum_{i=\overline{1,c}} \sum_{k=\overline{1,M}} (\mu_{ki})^m \|V_i - X_k\|^2, \quad (2)$$

where V_i are cluster centres; X_k are clustering elements; $m \in (1, \infty)$ is the exponential weight; and μ_{ki} is the membership degree of the kth element in the ith cluster. The values of cluster centres V_i are calculated as

$$V_i = \sum_{k=\overline{1,N}} (\mu_{ki})^m X_k / \sum_{k=\overline{1,N}} (\mu_{ki})^m,\ i = \overline{1,c} \quad (3)$$

3. The analysis further included expert evaluation of the resulting cluster centroids. Five specialists in demography separately evaluated the reproduction potential X_i^1 and the potential quality of the future population X_i^2 in each cluster. Then their evaluations were converted into numerical values (a high level corresponded to 5 and a low level to 1).

4. Expert evaluations of the clusters and the membership degrees of the regions in the clusters were used to calculate the reproduction potential P_i^1 and the quality of the future population P_i^2 in each region:

$$P_i^1 = \sum_{j=1}^{c} \mu_{ij} X_i^1,\ P_i^2 = \sum_{j=1}^{c} \mu_{ij} X_i^2,\ i = \overline{1,M} \quad (4)$$

5. The calculated values of the regions' potential were then used to rank these regions. The resulting values R_i^1 and R_i^2 are the ordinal indicators of a region's potential compared to other regions (these indicators show how high or low the given region ranks according to its potential).

6. To develop a differentiated and focused demographic policy framework, it is necessary to consider not only the values of each region's demographic potential, but also how balanced this demographic potential is. Therefore, values R_i^1 and R_i^2 were used to calculate the indicator of balance Q_i:

$$Q_i = |R_i^1 - R_i^2|,\ i = \overline{1,M}. \quad (5)$$

It is clear that the lower value Q_i is, the more balanced the demographic situation in the region is in terms of its reproduction potential and the quality of its future population.

RESULTS

1. As a result of fuzzy clustering, five clusters were formed (exponential weight $m = 1.4$). Table 2 shows the values of cluster centroids.

Table 2: Cluster centroids

Cluster	Indicators characterizing the reproduction potential		
	Rate of natural increase, %	Proportion of children in the population, %	Rate of realization of reproduction potential, %
1	13.4	29.8	0.87
2	1.2	20.7	0.62
3	-3.6	16.3	0.69
4	-1.4	18.2	0.63
5	1.1	18.6	0.74

Cluster	Indicators characterizing the potential quality of the future population				
	Proportion of healthy children, %	Disability rate in the child population, %	Juvenile delinquency rate, %	Involvement of children in the system of supplementary education, %	Proportion of modern educational institutions, %
1	78.2	4.30	0.15	40.9	68.2
2	83.6	2.03	1.62	54.5	79.4
3	80.6	1.98	0.81	70.2	82.7
4	78.4	1.90	1.38	71.7	81.1
5	82.0	2.02	0.67	55.7	82.3

Fuzzy clustering does not provide an exact list of regions for each cluster. Thus, we obtained models of demographic situations in different regions of the country. For instance, the first model (Cluster 1) is characterized by a high level of reproduction potential: this cluster has a high rate of natural increase, a large proportion of children, and a high rate of reproduction

potential realization. The potential quality of the future population in this demographic model is extremely low, as this model combines the lowest proportion of healthy children with the highest disability rate, the lowest involvement of children in supplementary education, and the lowest proportion of modern education institutions.

2. Our experts evaluated the reproduction potential X_i^1 and the quality of the future population $X_i^2 (i = \overline{1, M})$ for each cluster (table 3). It can be observed that there is a lack of balance between the two kinds of potential in the majority of clusters.

Table 3: Expert evaluation of the potential of cluster centres

Cluster	Reproduction potential		Potential quality of the future population	
	Expert evaluation	Numeric value in expert evaluation	Expert evaluation	Numeric value in expert evaluation
1	high	5	low	1
2	medium	3	above medium	4
3	low	1	above medium	4
4	low	1	medium	3
5	above medium	4	above medium	4

3. In the next step, the data for each region were used to calculate their reproduction potential P_i^1 and the quality of their future population P_i^2. The results of these calculations were then used to rank regions R_i^1, R_i^2 and to estimate balance Q_i of the two types of potential. In other words, the regions were ranked according to the two types of potential and how balanced they are.

It should be noted that some regions did not belong exclusively to one cluster (based only on the membership degrees). However, expert evaluations and membership degrees P_i^1, P_i^2 made it possible to get a clear picture of these regions' reproduction potential and the potential quality of their population in the future.

Let us now focus on the case of Vologda region. The values of the membership degrees of this region in the clusters are 0.000113, 0.553551, 0.033691, 0.395077, and 0.017568, respectively. Therefore, this region cannot be described as representing only one demographic model since its membership degrees in the second and fourth clusters are quite high.

According to table 3, experts evaluated the reproduction potential of this region in the second and fourth clusters as 'medium' and 'low', respectively. The quality of the future population is evaluated as 'above medium' and 'medium', respectively. These data alone are insufficient to identify the demographic potential of Vologda region.

Further calculations and ranking provided us with a more detailed picture of this region's demographic potential. The region received the following expert evaluations: $P_1 = 2.16, P_2 = 3.6$. Accordingly, the region ranked $R_1 = 40, R_2 = 56$ in terms of its reproductive potential and the potential quality of its future population. Taking into consideration the overall number of the regions in this study (79), the ranking position of Vologda region demonstrates that this region's reproduction potential corresponds to the medium level, while the potential quality of its future population is below the medium level. As for the balance of these two types of potential, the region ranks twenty-first (that is, in the upper third of the ranking) and can be described as a region with comparatively balanced levels of these two types of potential.

Thus, fuzzy clustering enabled us to make an exact evaluation of the demographic potential of the regions whose characteristics were shared by several clusters. These regions' evaluations depended on the values of the demographic potential of the closest clusters proportional to the degrees of membership in these clusters. Such a result would be impossible to achieve if we applied only clear clustering algorithms, which would mean that the evaluation of the region's potential would be equal to the evaluation of the potential of the nearest cluster, regardless of the possible similarities with other clusters.

DISCUSSION

This research undoubtedly contains a number of debatable points, one of which is the fundamentally ambiguous solution of the clustering task. It is known that there is no single or best criterion for clustering quality. Fuzzy clustering means that the researcher uses subjective criteria to set the number of clusters and the exponential weight, which affects the results. If the number of clusters is too small, however, some groups of regions with unique characteristics will not be revealed. An increase in the number of clusters leads to the creation of clusters with close centroids which do not manifest significant differences in terms of the region's demographic potential.

It should be noted that the value of the exponential weight is determined depending on how the clustering results are going to be used. If the level of *m* is too high, the final results are 'excessively averaged'; if it is too low, all regions get the same evaluation as the cluster they belong to with a high membership degree, even if they are actually shared by several clusters.

In further analysis, we can turn from fuzzy clusters to standard hard clusters using a specific list of regions for every cluster, which will enable us to develop a more focused demographic policy. Such a transition can be done in various ways:

– for example, we can use the α-level set. Let us assume that a region belongs to a cluster if its degree of membership in this cluster is higher

than the given value α (α∈[0,1]). This method, however, has certain peculiarities which should be taken into account. For example, if the value of α is high enough, some regions may correspond to none of the clusters (if all degrees of membership are smaller than α). If the value of α is low enough, some regions may belong to two or more clusters;

– another way is to use the principle of maximum membership degree. In this case, if a region's membership degree to a certain cluster is maximal, we will consider that it belongs to this cluster.

In order to interpret these findings correctly, we should keep in mind the following.

Firstly, there are different reasons for the imbalanced demographic development of Russian regions:

1. Cultural and religious reasons. Russia is a multinational and multifaith country. For example, in regions with a high share of Muslim population the quantitative indicators of demographic potential tend to be higher. These regions are generally characterized by strict adherence to family traditions, with people preferring home-based care for pre-school children and having strong views about child care and children's physical development and health. In such regions, the potential quality of the population is usually lower than the reproduction potential.

2. Economic reasons. Russia is a country with a high level of differentiation between its regional economies. For example, the maximum GRP per capita in Tumen region exceeds the minimum GRP in Ingushetia by 14 times. Moreover, different Russian regions have different urbanization levels: for example, in Moscow region, the share of the rural population is 7.5% while in the Republic of Altai, it is 70.8% (9.4 times larger). In regions with a comparatively high standard of living, a developed economy, and a high level of urbanization, the potential quality of the population tends to be higher than the reproduction potential. In these regions, the population of reproductive age primarily seeks to realize themselves professionally, which means that childbirth is often delayed, resulting in the inevitable decline of birth rates and reproduction potential.

To improve the demographic situation in Russian regions, it is essential to achieve the right balance between the reproduction potential and the potential quality of the population, while simultaneously enhancing both types of potential. These are the objectives that should underpin the demographic policy of the regions and the country in general. The full-fledged realization of the population's quality potential, together with a high reproduction potential, will provide regions with the amount and quality of human capital necessary for their economic and socio-cultural development.

Secondly, the ranking of the regions according to their reproduction potential and potential quality of their population demonstrates that a special set of demographic measures should be developed and implemented for certain groups of regions. For instance, in regions with high reproduction potential and low potential quality of the population, it is necessary to improve the institutions responsible for the development of human capital, that is, institutions working in the sphere of health care, education, and culture. In some regions, the potential quality is high while the quantitative indicators are low, which requires measures which are traditionally used in Russia to stimulate population growth. It might also be necessary to redistribute resources from the regions which are at the top of the reproduction potential ranking and, therefore, do not need these measures.

CONCLUSIONS

The original methodology applied in this study has several advantages. Firstly, it considerably reduces the experts' workload, as they did not have to provide evaluation for all 79 regions of Russia. What the experts had to do was evaluate five clusters as models of the demographic situation. Secondly, this methodology reduces the subjectivity level of expert evaluation because in the final evaluation of the regions' demographic potential we used the objective membership degrees we received through fuzzy clustering. Thirdly, in comparison to hard clustering, fuzzy clustering enhances the flexibility of evaluation: our assesments of each region do not depend exclusively on the potential of the nearest cluster, as we also take into consideration the region's possible similarities with other neighbouring clusters with different potential. Furthermore, ranking decreases the impact that the clustering parameters have on the final result, which makes the study more objective.

ACKNOWLEDGMENTS

The article is processed as one of the outputs of the research project "Fertility and parenting in Russian regions: models, invigoration strategies, forecasts", supported by the President of Russian Federation, project no. NSh-3429.2018.6.

REFERENCES

Abbasov, A.M. and M.H. Mamedova. 2003. "Application of fuzzy time series to population forecasting". In *Proceedings of CORP 2003* (Vienna, Austria, February 25th-February 28th, 2003).

Bezdek, J.C. 1981. "Pattern Recognition with Fuzzy Objective Function". New York, Plenum Press.

Dobrokhleb, V.G. and N.V. Zvereva. 2016. "The Potential of Modern Russian Generations", *Economic and Social Changes-Facts Trends Forecast,* Vol 44, Issue 2, 61-78.

Goraj, S., Gwiazdzinska-Goraj, M. and A. Cellmer. 2016. "Demographic Potential and Living Conditions in Rural Areas of North-Eastern Poland". *MSED-2016: 10th International Days of Statistics and Economics,* 482-493.

Kulikova, V.P. and O.A. Nikishina. 2015. "Statistical Modelling and Forecasting of the Region's Demographic Development", *Journal of Omsk Regional Institute,* Vol 1, Issue 1-1, 44-46.

Pastuszka, S. 2017. “Regional Differentiation of The Demographic Potential in Italy and Poland”. *Comparative Economic Research-Central and Eastern Europe,* Vol 20, Issue 3, 137-159.

Russian Regions. Socio-Economic Indicators. 2017. Statistical Book. Rosstat, Moscow.

Sasu, A. 2010. “An application of fuzzy time series to the Romanian population”, *Bulletin of the Transilvania of Brasov*, Vol 3(52), Series III: Mathematics, Informatics, Physics, 125-132.

Seletkov, I.P. and M.A. Martsenyuk. 2016. “Modelling urban growth using matrix fuzzy cellular automata”. In *Proceedings of the Russian Conference of Young Scientists on Mathematics and Interdisciplinary Research (Perm, Russia, May 16th-May 19th, 2016),* 238-241.

Shokin, I.V. and S.V. Fedorov. 2012. “The forecasting method for the mutual influence between dynamic macroeconomical and demographic indicators in regions of Russian Federation (applying some methods of scenario analysis)”, *European Social Science Journal*, Vol 9-2 (25), 382-389.

Shubat, O., Bagirova, A. and I. Shmarova. 2017. “The Use of Cluster Analysis to Assess the Demographic Potential of Russian Regions”. In *Proceedings of the 31th European Conference on Modelling and Simulation (*Budapest, Hungary, May 23rd-May 26th, 2017), 53-59.

Wang, F., Zhao, L. and Zh. Zhao. 2017. “China's family planning policies and their labor market consequences”. *Journal of Population Economics*, Vol 30, Issue 1, 31-68.

Zadeh, L. 1965. “Fuzzy sets”. *Information and Control*, Vol 8, 338-353.

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

OKSANA SHUBAT is an Associate Professor of Economics at Ural Federal University (Russia). She has received her PhD in Accounting and Statistics in 2009. Her research interests include demographic processes, demographic dynamics and its impact on human resources development and the development of human capital (especially at the household-level). Her email address is: o.m.shubat@urfu.ru and her Web-page can be found at http://urfu.ru/ru/about/personal-pages/O.M.Shubat/

ALEXANDER AKISHEV is a PhD student of Institute of Natural Sciences and Mathematics at Ural Federal University (Russia). His research interests include fuzzy clustering and its applications to various applied fields. His email address is: alexander.akishev@urfu.me

OPTIONS WITH STOCHASTIC STRIKE PRICES

János Száz, CSc
Ágnes Vidovics-Dancs, PhD, CIIA

Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest, Hungary
E-mail: agnes.dancs@uni-corvinus.hu

KEYWORDS
option pricing, Monte Carlo simulation

ABSTRACT

In the option pricing theory, the exercise price is constant by definition. The early generalizations of the Black-Scholes formula aimed to get rid of the constant nature of some parameters (the risk-free interest rate or the volatility). The focus of this paper is on generalizing the basic option pricing techniques in another direction: by allowing the strike price to be a random variable or change across time. For this purpose, we will examine a European call option with binomial random strike price and an American put option with time- and state-varying strike price. Taking the exercise price as a random variable might be considered as a bridge between the Balck-Scholes and the Margrabe-model.

INTRODUCTION

Option pricing is one of the favourite topics in quantitative finance. The basic models such as that of Black and Scholes (1973) and Merton (1973) are well-known and used all over the world. In the decades following after their publication, the original framework was extended and generalized in a few directions. One fruitful direction is examining options with various underlying assets. The original model assumed a stock with no dividend, but the underlying might be a currency or an index with continuous dividend yield as well. The famous Black (1976) model is an extension for pricing options on futures or valuing interest rate caps and floors.

Another popular way of improving the model is relaxing the assumption that the volatility of the underlying is constant. One of the most famous stochastic volatility models is that of Heston (1993).

In this paper, we deal with a direction where the strike price is not constant. We show two cases how this feature might appear: the first case is a European option with a binomial strike price, the second is an American option with time- and state-varying strike price. Throughout the paper, we assume that the reader is familiar with the basic Black-Scholes-Merton and binomial option pricing framework.

EXCHANGE OPTIONS

There are many extensions for the BS-model, from our point of view the most important is that of Margrabe (1978) who provided a closed form solution for the price of exchange options. These products grant the holder the right to exchange one risky asset to another risky asset at maturity. It is assumed that both risky assets follow geometric Brownian motion, that is under the risk-neutral Q-measure:

$$dS_{\mathrm{A}} = rS_{\mathrm{A}}dt + \sigma_{\mathrm{A}}S_{\mathrm{A}}dW_{\mathrm{A}} \text{ and}$$
$$dS_{\mathrm{B}} = rS_{\mathrm{B}}dt + \sigma_{\mathrm{B}}S_{\mathrm{B}}dW_{\mathrm{B}},$$

where, W_{A} and W_B are standard Wiener processes under Q, with correlation ρ.

It can be shown (for a derivation see for example Medvegyev–Vidovics-Dancs–Illés, 2015) that the price of the exchange option with payout function $\max(\mathrm{S_{AT}} - \mathrm{S_{BT}}; 0)$ is the following:

$$S_{A0}N(d_1) - S_{B0}N(d_2)\ , \tag{1}$$

$$d_1 = \frac{\ln\left(\frac{S_{\mathrm{A0}}}{S_{\mathrm{B0}}}\right)+\frac{\sigma^2}{2}T}{\sigma\sqrt{T}} \text{ and } d_2 = \frac{\ln\left(\frac{S_{\mathrm{A0}}}{S_{\mathrm{B0}}}\right)-\frac{\sigma^2}{2}T}{\sigma\sqrt{T}}$$

$$\sigma = \sqrt{\sigma_A^2 + \sigma_B^2 - 2\sigma_A\sigma_B\rho},$$

and N stands for the cumulative distribution function of the standard normal distribution. The formula (1) is often called Margrabe-formula.

It is easy to see that Margrabe's model for pricing the exchange option is a generalization of the Black-Sholes-Merton model. A simple call option is a special exchange option, where S_B is deterministic (σ_B=0) and the value of S_B at maturity is the strike price. On the other hand, a simple put option is a special exchange option as well, where S_A is deterministic (σ_A=0) and the value of S_A at maturity is the strike price. Substituting these into Margrabe's formula leads back to the Black-Scholes formula (Figure 1a and 1b).

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

S	K	σ	r	T	BScall
100	130	20%	4%	2	4,83

SA(0)	σA	SB(0)	σB	T	rho	Margrabe
100	20%	120	0%	2	0%	4,83

Figure 1: Connection between Black-Scholes and Margrabe-formulas

What we would like to do in this paper is to build a bridge between the Black-Scholes-Merton and the Margrabe-model. Hence, we will analyse options the strike price of which is stochastic or time-varying, but will not replace it with another traded asset (i.e. with another geometric Brownian motion) as Margrabe did. With other words, our analysis can be interpreted as another (not "Margrabe-style") extension of the famous Black-Scholes-Merton framework.

BINARY STRIKE PRICES

As a first step, we assume that the strike price K is a random variable with two possible outcomes: K_1 and K_2. Under the risk neutral measure, K_1 will occur with probability q_K, and K_2 with probability $(1-q_K)$:

$$K = \begin{cases} K_1, & q_K \\ K_2, & 1 - q_K \end{cases}$$

We will price this option with Monte Carlo simulation and compare the result with c_1 and c_2, the prices of simple call options with strikes K_1 and K_2. We will use the expression 'stochastic option' for the option with the stochastic strike price. (We cannot use the 'binary option' term since it has a widespread and different meaning in the literature.) The price of the stochastic option will be denoted by c_K.

Monte Carlo simulation

Let us assume that the price of the underlying asset (S) follows geometric Brownian motion, with trend parameter equal to the risk-free interest rate (of course, we are talking about the dynamics under Q-measure):

$$dS = rSdt + \sigma S dW.$$

We simulated the cash flows of the option with binary K. In a sample of 10,000 realizations, we got the frequencies showed in Figure 2. The parameters used are: S_0=100, r=5%, σ=20%, T=3, K_1=60, K_2=75, q_K=50%. On the figure, we used the same realizations of S_T for all the three options.

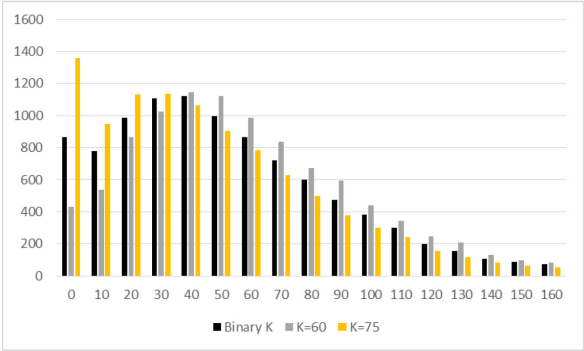


Figure 2: Payout frequencies

As for the price of the stochastic option, it will be obviously between the prices of the simple options with strike prices K_1 and K_2. It is also trivial that the higher the q_K parameter is, the closer the price of the stochastic option is to the price of the option with K_1. We plotted this relationship in Figure 3. Apart from q_K, the parameters are the same as before.

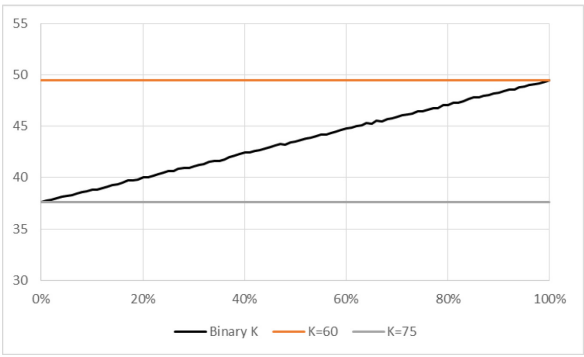


Figure 3: Prices as a function of q_K

The most interesting question is the relationship between the price of the stochastic option and the weighted average of the prices of the deterministic ones. Let us denote the latter one with c_{AVG}:

$$c_{AVG} \triangleq q_K c_1 + (1 - q_K) c_2 \qquad (2)$$

Using Monte Carlo simulation again with 10,000 realizations, we observed that the price of the stochastic option is very close to c_{AVG}, actually it fluctuates around it. We repeated the Monte Carlo simulation 1,000 times and plotted the results on Figure 4, with the same parameters we used so far. On the figure, the constant c_{AVG} is calculated by the Black-Scholes formula.

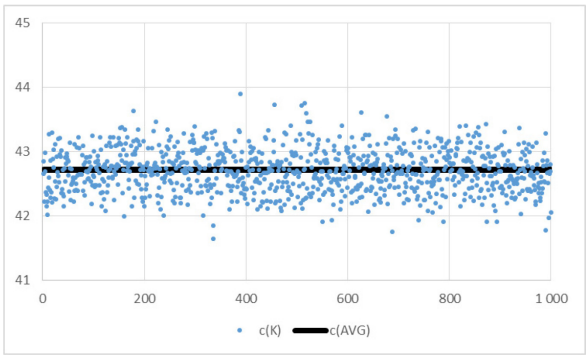


Figure 4: Monte Carlo prices

Binomial tree

To understand the nature of the stochastic option better, we analyse the problem in a one-period binomial model, with r=0. Let S_0 be 100 and the two possible values of S_1 are S_d=60 and S_u=140. In this case, the risk-neutral one-period probability (q) is 50%. As for the values of K_1 and K_2, it is worth distinguishing among six cases, according to their relation to S_d and S_u. Table 1 summarizes these cases and also contains the chosen values of the strike prices in each case.

Table 1: The six cases

Label	Case	K_1	K_2
A	$K_1<K_2<S_d<S_u$	10	30
B	$K_1<S_d<K_2<S_u$	10	100
C	$K_1<S_d<S_u<K_2$	10	170
D	$S_d<K_1<K_2<S_u$	80	120
E	$S_d<K_1<S_u<K_2$	80	170
F	$S_d<S_u<K_1<K_2$	170	200

It is easy to see that Case F is not really interesting. In this case, none of the examined options will be exercised, and hence their value is zero. This is true for the stochastic option as well.

Now we set q_K=40% and calculate c_1, c_2, and c_K with the well-known way: the price of the option is the risk-neutral expected value of its future cash flow. The results are summarized in Table 2, where we also showed c_{AVG}.

Table 2: Binomial prices

Case	K_1	K_2	c_1	c_2	c_K	c_{AVG}
A	10	30	90	70	**78**	**78**
B	10	100	90	20	**48**	**48**
C	10	170	90	0	**36**	**36**
D	80	120	30	10	**18**	**18**
E	80	170	30	0	**12**	**12**
F	170	200	0	0	**0**	**0**

We can observe that in this numerical example $c_K=c_{AVG}$ in all the six cases. Now we show that this is not due to the special values we have chosen for the parameters.

For simplifying the analyses, we introduce the following notation:

$$c_{i,j} = \max(S_j - K_i, 0),$$

where i={1,2} and j={u,d}.

Thus, $c_{i,j}$ is the cash flow of the call option with strike price K_j, if the price of the underlying is S_j.

Under the risk-neutral measure, the cash flows of the options analysed are discrete random variables with the following possible values and probabilities. In the indices, T refers to the fact that we are talking about the cash flows to be paid at the maturity of the option.

$$c_{1,T} = \begin{cases} c_{1,u} & q \\ c_{1,d} & 1-q \end{cases}$$

$$c_{2,T} = \begin{cases} c_{2,u} & q \\ c_{2,d} & 1-q \end{cases}$$

According to the risk-neutral valuation, the prices of the deterministic options with our notations are:

$$c_1 = qc_{1,u} + (1-q)c_{1,d} \quad (3)$$

$$c_2 = qc_{2,u} + (1-q)c_{2,d} \quad (4)$$

Using the definition of (2) we end up with the following equation:

$$c_{AVG} = q_K\left[qc_{1,u} + (1-q)c_{1,d}\right] + \\ +(1-q_K)\left[qc_{2,u} + (1-q)c_{2,d}\right] \quad (5)$$

Now let us determine the distribution of the stochastic option's cash flow.

$$c_{K,T} = \begin{cases} c_{1,u} & q_K q \\ c_{2,u} & (1-q_K)q \\ c_{1,d} & q_K(1-q) \\ c_{2,d} & (1-q_K)(1-q) \end{cases}$$

From this, the price of the stochastic option is the following.

$$c_K = q_K q c_{1,u} + (1-q_K)qc_{2,u} + q_K(1-q)c_{1,d} \\ + (1-q_K)(1-q)c_{2,d} \quad (6)$$

Comparing (5) and (6) shows us that $c_{AVG}=c_K$.

Relaxing the assumption that r=0 would not change our results since both (5) and (6) would be multiplied by the same discount factor.

VARYING STRIKE PRICE

In this section, we analyse an American put option with a time-varying strike price. As a first step, we compare the simple European and American put options in the binomial framework (Table 3, bold numbers indicate the early exercise).

Table 3: American put with fix K*

n	S0	K	IC	u	d	q	DF	dt	DFn
3	**200**	**210**	**1.25**	**2**	0.5	**0.50**	0.8	1	0.512

S(i,j)	0	1	2	3
0	200			
1	100	400		
2	50	200	800	
3	25	100	400	1 600

eu put	0	1	2	3
0	**33.0**			
1	64.8	17.6		
2	118	44	0	
3	**185**	**110**	**0**	**0**

Intrin	0	1	2	3
0	10			
1	**110**	0		
2	**160**	10	0	
3	185	110	0	0

am put	0	1	2	3
0	**51.0**			
1	**110**	18		
2	160	44	0	
3	185	**110**	0	0

PV EV	0	1	2	3
0	**51**			
1	82	**18**		
2	118	44	0	
3				

**Spot price development (S), European put value (eu put), intrinsic value of the put (Intrin), American put value (am put) and the present value of the expected value (PV EV). Time is changing vertically; the states are horizontally (j- number of up moves). IC = 1+ r, DF=1/IC (discount factor), DFn is the DF for n periods (years)*

The value of the right to sell at time T=n=3 for K=200, when the price of the underlying is S=200, is p=33 (European put), while the American put (right to sell the underlying *until* T) has a value of 51. The exercise takes place either in the state (1,0) when S=100 and we sell the underlying for K=210, or at maturity in the state (3,1). In the state (2,1) we could sell the underlying for K=210 with a gain of 10 (S=200), but it is better to hold it, since we can expect a payout of 55, and its discounted value is 44.

Now, let us determine the value of an American put when K is a deterministic function of time (i) and the underlying price (the state, j). Table 4 compares this option with the previous, simple American put. We can see that this rather strange put option has a higher value (**58.2** as compared to **51.0**), reflecting the fact that the strike price is increasing.

Table 4: American put with varying K*

n	S0	K	IC	u	d	q	DF	dt	DFn
3	**200**	**210**	**1.25**	**2**	0.5	**0.50**	0.8	1	0.512

S(i,j)	0	1	2	3
0	200			
1	100	400		
2	50	200	800	
3	25	100	400	1 600

am put	0	1	2	3
0	**51.0**			
1	110	18		
2	160	44	0	
3	185	110	0	0

K	0	1	2	3
0	210			
1	220	240		
2	230	250	270	
3	240	260	280	300

am p K	0	1	2	3
0	**58.2**			
1	120	25.6		
2	180	64	0	
3	215	160	0	0

**Spot price development (S), simple American put value (am put), exercise price as function of time and state (K), American put value with varying K(am p K);*

We can separate the effect of the increasing K due to *time* (Table 5, left side), and due to *price changes* of the underlying (Table 5, right side). The second one leads to a stochastic exercise price, but the source of randomness is the uncertainty in the underlying price itself (perfect correlation).

Table 5: Separated effects of the varying strike price

K	0	1	2	3
0	210			
1	220	220		
2	230	230	230	
3	240	240	240	240

57.0

K	0	1	2	3
0	210			
1	210	230		
2	210	230	250	
3	210	230	250	270

52.3

It is a question whether the 58.2 – 51 = **7.2** option premium increase can or cannot be divided along this separation of exercise increase. The early exercise feature of the American options can lead to several complications, this is why we do not have a closed form formula even for the simplest case.

CONCLUSION

Relaxing the assumption that the strike price is constant might lead to very interesting and novel research in the option pricing field. In this paper, we showed two possible ways how this generalization may be initiated. For the first sight, in case of traditional financial options, it can be rather strange to imagine that K is volatile. However, in the real world applications of the option pricing principal – the martingale approach –, sometimes it is the best way to treat *K* as a deterministic or stochastic function, changing over time and/or across states. For example, Kornai's famous *Soft Budget Constraint* phenomenon could be treated as an American put option with a not well-defined exercise price: the agent that gets into financial trouble does not know when it will be bailed out, and the extension of the bail-out is stochastic (might be zero as well).

REFERENCES

Black, F. 1976. “The pricing of commodity contracts” *Journal of Financial Economics*, Issue 3, 167-179.

Black, F. – Scholes, M. 1973. “The Pricing of Options and Corporate Liabilities” *The Journal of Political Economy*, Vol. 81, Issue 3, 637-654.

Heston, S. 1993. “A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options” *The Review of Financial Studies,* Vol. 6, Issue 2, 327-343.

Margrabe, W. 1978. “The Value of an Option to Exchange One Asset for Another” *Journal of Finance*, Vol. 33, Issue 1, 177-186.

Medvegyev, P – Vidovics-Dancs, Á – Illés, F. 2015. *FX Derivatives* In: Mastering R for Quantitative Finance, Packt Publishing, Birmingham, 115-136.

Merton, R. 1973. “Theory of Rational Option Pricing” *The Bell Journal of Economics and Management Science*, Vol. 4, Issue 1, 141-183.

AUTHOR BIOGRAPHIES

János SZÁZ, CSc is a full professor at the Department of 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 adjunct professor at the Department of Finance at Corvinus University of Budapest. Her main research areas are government debt management in general and especially sovereign crises and defaults. She worked as a junior risk manager in the Hungarian Government Debt Management Agency in 2005-2006. Currently, she is chief risk manager of a Hungarian asset management company. Her e-mail address is: agnes.dancs@uni-corvinus.hu

Competitiveness and finance of supply chains: Considerations on optimisation

Péter Juhász, PhD, CFA
János Száz, DSc
Sándor Misik
Department of Finance
Corvinus University of Budapest
1093, Budapest, Fővám tér 8.
E-mail: peter.juhasz@uni-corvinus.hu

KEYWORDS
Growth, Cost of capital, Added value, Efficiency, Seasonality.

ABSTRACT

The financial aspect of supply chain (SC) management is a somewhat neglected research area, while earlier papers showed that it has a strong link to competitiveness. Our main contribution to literature is to analyse the competitiveness effects of SC financial management decisions under perfect information and cooperation among SC members the absence of which may distort empirical findings. Our simulation-based research shows that even in case of perfect foresight seasonality decreases the profitability and the ability to grow while increasing the capital need. But, we also conclude that cooperation of SC members may reduce this additional capital need while enhancing the profitability and the growth, thus leading to higher competitiveness. This cooperation may be achieved through regulating payment terms or introducing special fees to be paid by the SC members to the dominating company of the SC. Thus, an economic policy aiming at providing cheap capital to firms at a lower level of SC or increasing their added value at the costs of other SC levels may decrease the competitiveness of the SC in whole.

INTRODUCTION

When analysing supply chains (SCs), usually the management of flow of (1) goods and services, (2) rights, (3) information and knowledge (technology), and (4) financial resources are listed as critical issues (Pfohl – Gomm, 2009). This article focuses on this later point.
Based on an extensive survey of Indian firms, More and Basu (2013) highlight that the most critical challenge is the lack of shared vision among SC members (SCMs). The unpredictability of cash flows resulting from delayed financial transactions, poor automatization of financial processes, and weak knowledge of SC finance tools are among the fundamental problems. They call for more collaboration among SCMs to increase the financial stability of the SC. However, what would be the SC like if this cooperation were perfect? In this article, we focus on possible optimisation of the payment process, assuming no information barriers among the SCMs.

LITERATURE REVIEW

Recently various papers focused on SCs, particularly on the competitiveness of them and the new methods to solve finance issues. Still not too many articles examined how competitiveness and financing of SCs are connected.

Competitiveness of supply chains

The term of supply chain management appeared in the literature in the 1980's. Since then, many papers proved that the efficient management of SCs increases the competitiveness of the individual SCMs and so the total of the SC (Marcuta and Marcuta, 2013).
At the same time, measuring SC competitiveness could be very complicated. It is usually measured by the sum of production costs, quality offered and flexibility of the network. Marwah et al. (2014) emphasise that both increased efficiency of SCMs and the improvement of SC activities themselves may lead to improved competitiveness.
Based on case studies on Indian automotive component manufacturers, Joshi et al. (2013) even identified 24 factors of competitiveness within eight groups (cost, flexibility, quality, delivery, buyer-supplier relationship, technology, environmental factors, and customer demand).
Instead of focusing on such influencing (input) factors, we may estimate competitiveness from the output side by measuring the growth of sales, export or employment, in addition to achieved profitability and capital efficiency (business performance) of the SCMs. As for UK Oil and Gas industry Yusuf et al. (2014) found three SC agility factors with high correlation to the business performance: "cooperating to compete" (long-term partnership, reward based on team performance etc.), "mastering change and uncertainty" (rapid decision making, proactive response to changes etc.) and "leveraging the impact of people and information" (information accessibility, team spirit etc.).
Hult et al. (2007) underline that in SCs culture of competitiveness and knowledge development have a positive association with performance. They highlight that during turbulent times the link to knowledge development becomes stronger, while culture of competitiveness seems to lose its effects.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Finance for supply chains

Literature on SC finance usually takes one of two perspectives: papers either focus on products of financial institutions to cope with accounts payable and receivable issues or concentrate on the whole of the SC and the reduction of the working capital need (inventories included) and sometimes also on financing invested assets (Gelsomino et al., 2016).

Still, both of these research directions are far from being complete. Pfohl and Gomm (2009) underline that contrary to flow of goods and information only limited research was done in the field of financing supply chains. Even in those, the cost of capital stayed mostly neglected. Those are the amount of capital needed, the cost of that capital and the flow of cash achieved by employing the given capital that determine the value of the given company. Therefore, it is not only by individual inventory, process, and cash management but also by collaboration and synchronisation among supply chain members and optimisation of funding costs that we may enhance value creation. Based on this logic, we should not only minimise the capital need of the SC, but extraordinary efforts should be made to achieve that the highest amount of capital need emerges at the SCM facing the lowest cost of financing. Of course, when optimising, we also have to consider the duration of that financing need. To be able to do so, Pfohl and Gomm (2009) underline the importance of the information flow among SCMs.

Cavenaghi (2013) highlight that this information is needed not only by SCMs but also by the banks providing financing to them as no matter which member of the SC they finance at the end of the day the financial institutions take the payment risk of the final customer. The management of these new complex and integrated systems call for new risk management tools instead of the standard methods (Chun-Lian, 2016).

Based on case studies, Liebl et al. (2016) emphasise the vast opportunity reverse factoring may offer in supply chain financing. In those cases, buyers seek the help of financial institutions to be able to pay suppliers early to reduce the risk of shocks a bankruptcy at earlier SC levels may generate. This tool is more often used by buyers with a weaker bargaining power as they seem to focus on strengthening of the relationship to key suppliers with a flawless track record.

Based a theoretical optimisation model, Wuttke et al. (2016) showed that introducing SC finance program (where thanks to the main buyer suppliers receive financing on their account receivable at preferred term in exchange for accepting longer payment terms) is a dynamic process where timing is an essential factor. They conclude that the immediate introduction of such a system is not always beneficial for the buyer. It seems that high procurement volume and long initial payment terms both promote the introduction. Extending deadlines under SC financing, which may limit the number of suppliers, is only advantages buyers with lower financing cost, high procurement volume and long initial payment term.

Focusing on the management practices, based on a sample of 110 Malaysian electronics manufacturer, Sundram et al. (2011) identified six dimensions having a significant effect on the SC performance. Like Basu (2013), they found that agreed (1) vision and goals (i.e. a kind of central coordination) are the most critical factor, but also (2) strategic supplier partnership, (3) information sharing, (4) information quality, (5) postponement strategy and (6) risk and reward sharing play a statistically significant role.

Finance of supply chains and competitiveness

The performance of an SCM is strongly linked to that of the SC. Using a Romanian sample, Gyula (2013) showed that the financial, marketing and innovation performance of the SC have a positive and statistically significant impact on the overall organisational performance.

Filbeck et al. (2016) proved for US automotive manufacturers that supply chain disruptions do not only affect negatively the share price of the company being hit but also those of the competitors. This link was particularly strong in bear markets, but not present for Japanese carmakers.

Pino et al. (2010) modelled a SC in a multi-agent system to show that even in case of a flat final demand a vast variation could emerge in the demand lower level SCMs face, called the ‘‘bullwhip effect’’. They conclude that this variation caused by the separate management of the SCMs can be dramatically reduced using MASs methodology. That is why our simulation built on perfect information assumes no such distortions. They underline that removing fluctuations from demand reduces the capital need of SCMs.

MODEL DESCRIPTION

Our model focuses on the financial management issues of an SC. We examine how different financial parameters influence the competitiveness (measured by the ability to grow) of the SC, and how perfect cooperation would transform financing and payment terms to maximise shareholder value (total cash flow achieved).

The SC in our model has three levels: A sells to final customer (market), B is the main supplier of A and the main buyer of C that purchases raw material at a price of 10 per unit. A, B, and C could each be considered as a single company, or a representative merged firm for the given level in the SC. Only A has sales outside of the SC. We assume perfect foresight regarding the demand level. Each level needs one period (month) to produce its final product from the product purchased from its supplier. Thus, in any given period A produces the required quantity in line with the market demand (D_t), but places an order with company B equal to (D_{t+1}). B produces this amount but places an order with C equal to D_{t+2}. So, C manufactures at time t the amount needed at time t+2.

When manufacturing, the firms have to pay immediately for wages (any cost not related to SCMs they purchase from), but they pay for the SC products P days later. Payment terms may not be the same for different SCMs.

As manufacturing needs one period, suppliers need to deliver at the start of the period, thus if P=0 payment to suppliers is due at the beginning of the period.
At the end of the period, all SCMs deliver their products to their buyers but collect the income only R days later. (R may be different for all players.) Due to the set-up of the SC, $P_A=R_B$ and $P_B=R_C$.
Two measures control profitability of the SCMs. Added value (AV) of their product is added to the price of their supplier to calculate their selling price. But a given percentage of AV (Wage%) has to be spent on wages and other costs due immediately.
The simulation starts with setting up manufacturing capacities: we assume that to perform production for each of the SCMs 1 unit/piece invested asset is needed. We have to purchase the machines by the start of the actual manufacturing period, so the payment for the machines takes place a period earlier. If production increases, the additional investment is deducted from the accumulated cash. (Equation 3 and 12.) Initial capacity setup takes place for all companies at the period -2 and that investment is considered as part of the initial capital need.
As a next step, the cash flow of each period is calculated, and the result is added to the opening cash balance. To evade bankruptcy, each of the SCMs has to hold a certain amount of cash at the start of the simulation, representing their working capital (WC) need. This WC (together with the machines) is financed at a cost, though. Cost of capital (CoC, e.g. interest payment or dividend required) may be different for each of the firms. The cash balance is decreased at the end of each period by the starting amount of capital (covering WC and initial machines) times CoC.
This calculation method assumes that firms need to hold a WC enough for to survive the total simulation period right from the start (capital may not enter or leave the company, e.g. there is no dividend payment). There is no loss of capacity due to the usage of the machines and during the simulation period neither the price of the products or the machines changes.
The simulation covers 30 periods, where the first period is the one in which A first sells its products, implying that manufacturing at C starts in period -2. To measure the competitiveness of the SC, we calculate the individual and total amount of start-up capital required to survive simulation period, total additional cash amount generated by the end of the simulation (as a measure of profitability), and maximum growth the SC may survive using a certain amount of capital.
SC is facing a final market demand for the product of A that is calculated based on equation 1 and 2.

$$D_t = D_0 * \prod_{i=1}^{t}(1+g_i) * (1+s) \quad (1)$$

$$s = a * sin(c * (t-1)) \quad (2)$$

D stands for the amount of demand, t indicates time (starting from 1), g shows the growth rate of the period, s for the seasonality trend. Constants a and c describe the form and size of seasonality effect and their value were chosen to be 25 percent and 101 respectively. $D_0*(1+g_1)$ equals to 100 in all cases.
The cash flow of any period is calculated using formula 3 and is added to the initial cash amount.

$$CF_t = Income_t - Wages_t - Mat_t - Inv_t \quad (3)$$

where

$$Income_t = \\ = (x * Q_{t-int(\frac{R}{30})-1} + (1-x)Q_{t-int(\frac{R}{30})}) * SPrice \quad (4)$$

$$x = \frac{R}{30} - int\left(\frac{R}{30}\right) \quad (5)$$

$$Q_{A,t} = D_t \quad Q_{B,t} = D_{t-1} \quad Q_{C,t} = D_{t-2} \quad (6)$$

$$Sprice_A = Sprice_B + AV_A$$
$$Sprice_B = Sprice_C + AV_B$$
$$Sprice_C = Sprice_{Raw} + AV_C \quad (7)$$

$$Wages_t = Q_t * Wages\%_t \quad (8)$$

$$Mat_t = \\ = (y * Q_{t-int(\frac{P}{30})-1} + (1-y)Q_{t-int(\frac{P}{30})}) * PPrice \quad (9)$$

$$y = \frac{P}{30} - int\left(\frac{P}{30}\right) \quad (10)$$

$$Pprice_A = Sprice_B$$
$$Pprice_B = Sprice_C$$
$$Pprice_C = Sprice_{Raw} \quad (11)$$

$$Inv_t = max(0, (Q_{t+1} - Q_t) * 1) \quad (12)$$

The initial cash is determined by iteration that aims to find the minimum amount enough to have all of the end of period cash balances (from the period -2 to 30) above 0.

SIMULATION RESULTS

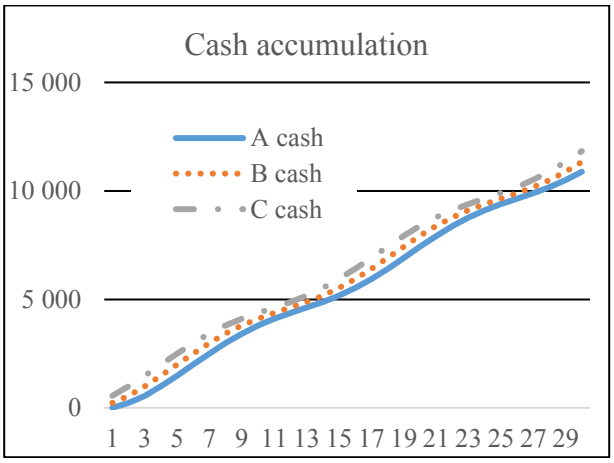
First, to have a reference point, we run the simulation with the parameters in Table 1. We picked 30 days (1 period) as a payment term for all participants. As procurement takes place at the start of the period and the sale happens at the end of it, this means that there is a financing gap of 1 period for all players. Demand was flat at 100 pieces for all the periods.

Table 1: Base scenario

Firm	A	B	C
Payment days	30	30	30
Added value	10	10	10
Wage (% AV)	60%	60%	60%
Cost of capital	1%	1%	1%

Our result shows that the SC needs altogether 4899 units of initial capital to set up, but due to the stable operating cash flow of 400 for all SCMs in each period at the end of period 30, there will be 29633 extra cash accumulated. Both investment need and cash profit are distributed equally among the firms. To be able to grow by 1 percent monthly (12.7 percent yearly), this system needs 30 units (0.6 percent) of additional capital. Thus, accumulated cash rose to 33660.

When adding the seasonality effect to the non-growing market demand, the capital need rises to 5244, while cash accumulated decreased by 2.8 percent to 28807. (This increase is in line with the results of Pino et al. (2010).) When adding 1 percent growth, capital need climbs to 5278, while accumulated cash reaches 32426. This means that in case of growth the seasonality boosted investment by 7 percent while reducing profit by 3.7 percent. Hence, our model supports well the empirical experience that fluctuations in demand may raise capital need, slow growth, and cut back on the profitability of the supply chains (More – Basu, 2013) even in case of perfect foresight.



Figures 1: Base case with seasonality

While the base scenario investigated a SC where members were identical, usually we find huge differences among SCMs. We examined two further cases. (1) SC build on smaller firms and controlled by a big multinational and a (2) distribution channel where the huge producer sells its localised products first to regional and then to local retailers. In the first case, added value content and market power of the firms increases along the SC, while in the second case the opposite is likely.

Table 2 shows the parameters of the scenario describing the manufacturing SC of a large multinational company (e.g. a global car manufacturer). We assumed that both A and B could achieve longer payment terms than their receivable turnover days, but C still has to pay for the raw materials after 30 days. (The market pays to A in 30 days.) Note that the total added value and cost is the same as in the base case.

Table 2: Manufacturing SC of a large multinational company

Firm	A	B	C
Payment days	45	60	30
Added value	15	10	5
Wage (% AV)	60%	60%	60%
Cost of capital	0.5%	1.0%	1.5%

The total start-up capital need of this SC is 6242 (27.4 percent more than the base case) 43 percent of which is needed in company C characterised by the highest cost of capital and lowest profitability (only 40% of its 5 added value remains with the company). When seasonality added, the minimum capital requirement climbs by further 10 percent to 6780. (The growth is similar for all SMCs.) The total of accumulated cash by the end of the last period reaches 27523, 59 percent of which remains with firm A investing only 30 percent of the total capital. Only 5 percent of the return was realised by company C who was the top investor. This finding is again in line with empirical results: the companies at the bottom of the SC complain about weak profitability and high investment need.

Would this SC be more competitive if payment terms remained the same as in the base case? The answer is positive with no doubt. Total capital need when seasonality included is 5226 (24 percent less), while total extra cash accumulated climbs by 3.3 percent to 28422. Capital need is more fairly distributed (A: 39.1%, B: 33.5%, C: 27.5%) just like accumulated cash (A: 52.1%, B: 33.7%, C: 14.1%). The only problem is that this results in A receiving 8.7 percent less of cash, while B faces a decrease of 3.1 percent so that C could get 187.8 percent more. It is clear to see that by coordination both A and B could keep its old profit by receiving compensation from C that would then end up with a 64.4 percent growth.

When adding 1 percent growth, the capital need of the coordinated system (same payment terms) is 23.4 percent less, while cash accumulated is 3.3 percent more. It seems that in case of a strict capital constraint reducing the burden on the SCM with the highest financing cost by offering more advantageous payment terms would be for the benefit of the whole SC and also all the individual SCMs. These findings are in line with Bassu (2013) and Pfohl and Gomm (2009). It seems that it is in the interest of the most powerful SCM not to use its position on extending its payment terms instead to convince the other SCMs to take part in an overall cooperation system.

At the same time, there is another significant conclusion. Many countries support local firms to be a member of multinational SCs expecting a general improvement in the performance of the economy. But, easing on capital constraints by state subsidies lessens the pressure for cooperation and thus reduces the competitiveness of the given firm and SC too.

Another common goal of countries hosting mostly firms joining global SCs at a lower level is to enhance the added value content of the local companies. Let us examine, how relocating some of the high added value functions would modify the competitiveness of the SC. If added value of C would amount to 15, while that of A is cut to 5 in the no-cooperation case (without growth and seasonality) capital need grew by 10 percent, while final cash raised only by 1 percent. When focusing on firm C alone, it will see its capital need to be increased by more than 46 percent (as higher AV implies more wage to pay asking for more WC), while its final cash amount will grow by more than 483 percent.

This result means that by achieving the relocation competitiveness (capital efficiency) of the whole SC decreases while that of C increases radically. Because now a more significant part of the total SC capital need is financed at a country with a higher cost of capital and C improves at the expense of B and A, in the long run, all SCMs will be in a worse position. So, moving more AV to earlier level if SC located in less favourable countries is not realistic if the decision is to be made by A dominating the SC and it is not even advantageous for C in the long run.

Our third scenario describes a retail chain. In this case, C is dominant with the highest AV and best financing position. It is by offering advantageous payment term to its buyers (very often own subsidiaries) that financing is provided to A and B operating with a higher cost of capital due to their smaller size and less advantageous location (e.g. riskier countries). Critical parameters are summed up in Table 3.

Table 3: Retail SC of a large multinational company

Firm	A	B	C
Payment days	60	60	30
Added value	5	10	15
Wage (% AV)	60%	60%	60%
Cost of capital	1,5%	1,0%	0,5%

This SC needs a total capital of 7468 and accumulates a total of 29607 cash. This structure transfers profit from C to A. A invests 18.2 percent of total capital but receives 24.8 percent of the cash, while C invests 51.4 percent and gets 45.9 percent only. (B has a share of almost 30 percent in both cases.) This allocation could be particularly advantageous if all SCMs belong to the same group and A faces a lower corporate tax rate.

If 1 percent growth is added, the capital need grows by 0.9 percent, while total final cash increases by 13.7 percent. Interestingly, capital need at B climbs by 1.2 percent, while that of A and C only by 0.8 percent. Adding seasonality to the base case causes similar distortions. Total initial capital need grows by 9.4 percent, but while this increase is 8.7 percent for A and 8 for C, B suffers a boost of 12.3 percent. This result calls attention to the fact that the growth and the fluctuation of demand may put very different burdens to SCMs even if no structural change occurs within the SC.

To reduce investment need at C, we may try to balance the return distribution back towards that of the investment. A way for this could be C to charge some fee to A (e.g. for the brand, marketing, know-how, licence fee). For to reflect this transfer from A to C, the manufacturing cost expressed in percentage of AV (Wage) should be modified. To evade distortion, we should keep the total of these costs across the SC constant. Given the original AV and Wage values, these expenses amounted to 18 (60%*5+60%*10+60%*15). For example, assuming a compensation per piece of 1.5, we have to modify Wage ratio of A up to 90 percent, and that of C down to 50 percent in our model.

When doing so, SC will need (without growth or seasonality) 2 percent less capital and produce 0.9 percent more total cash. In other words, this step improves the financial competitiveness of the SC. Under the new rules, A loses 62.3 percent of its original final cash balance, while C receives 35.7 additional cash. This restructuring leaves B is entirely unaffected, what is the main difference in this model between charging a fee and modifying payment terms. Therefore, the fees to be paid by the SCMs to the controlling entity are tools to fine tune the system, that is, they offer a method to force cooperation on SCMs. This new structure including fee payment performs better not only in case of growth, but also in case of seasonality, and when controlling for both of them. (Capital need diminished by 2 percent, total final cash increased by 1.1-1.5 percent.) This outcome is in line with Hult et al. (2007) promoting cooperation in turbulent times and Sundram et al. (2011) addressing fair risk and reward sharing as one of the SC success factors. Our finding implies that in case we assume a rational control over the SC by the dominant player national authorities may decrease the competitiveness of the SC if questioning the rightfulness and limiting the amount of such fees (see transfer pricing regulations).

Table 4 summarises our findings in details. Base scenarios refer to the primary assumptions related to the three major cases (identical firms, SC of a large multinational company and retail chain with a dominant actor). Relative changes are calculated to the base scenarios within each case.

Table 4: Summary of scenarios and results

Case	Scenario	Parameters*					Relative changes to Base scenarios			
							Flat demand		+1% growth in demand	
		Payable turnover days	Added value	Wage ratio	Cost of capital	Seaso-nality	Initial capital	Cash accu-mulated	Initial capital	Cash accu-mulated
Case 1: Identical firms	Base scen.	30-30-30	10-10-10	60%-60%-60%	1%-1%-1%	No	-	-	-	-
	Scenario 1					Yes	7%	-3%	6%	-4%
Case 2: Production chain	Base scen.	45-60-30	15-10-5	60%-60%-60%	0.5%-1%-1.5%	No	-	-	-	-
	Scenario 1					Yes	9%	-4%	9%	-5%
	Scenario 2	30-30-30				No	-22%	2%	-22%	2%
	Scenario 3					Yes	-16%	-1%	-16%	-2%
	Scenario 4	45-60-30	5-10-15			No	10%	1%	10%	2%
Case 3: Retail chain	Base scen.	60-60-30	5-10-15	60%-60%-60%	1.5%-1%-0.5%	No	-	-	-	-
	Scenario 1					Yes	9%	-4%	9%	-5%
	Scenario 2			90%-60%-50%		No	-2%	1%	-2%	1%
	Scenario 3					Yes	7%	-3%	7%	-4%

*Listed parameter values refer to Firm A - Firm B - Firm C in the given order.

MAJOR FINDINGS AND CONCLUSION

Our simulations have confirmed that fluctuations in demand cause fall back in growth, profitability and an increase in the capital need even in case of perfect information, so it is not only the uncertainty about the future affecting performance and competitiveness adversely.

We also showed that cooperation among SCMs might allow for reducing the total investment need while boosting the profitability and the ability to grow, in other words, improves the competitiveness. At the same time, we concluded that easing the capital constraint by state subsidies may hurt the competitiveness of the SC dominated by a big company by reducing the motivation for cooperation.

Our results also imply that relocating more of the added value generation of the SC to firms with weak bargaining power (high working capital need) and a high cost of capital decreases the competitiveness of the SC. Thus, for a long-term advantage, economic policy should also focus on improving macro conditions and payment terms beside of raising added value content of the local firms.

We also showed that growth of the SC might ask for very different additional investment from SCMs even if none of the structural variables changes. At the same time, fees paid by SCMs to the controlling company may offer a tool to enforce cooperation among independent firms. Using them wisely may help to optimise the performance of the SC and boost its competitiveness. In such cases, too conservative national transfer pricing systems may weaken the SC competitiveness.

REFERENCES

Cavenaghi, E. 2013. "Supply-chain finance: The new frontier in the world of payments", *Journal Of Payments Strategy & Systems*, 7, 4, 290-293.

Chun-Lian, Z. 2016. "Risk assessment of supply chain finance with intuitionistic fuzzy information", *Journal Of Intelligent & Fuzzy Systems*, 31, 3, 1967-1975.

Filbeck, G., S. Kumar, J. Liu and X. Zhao. 2016. "Supply chain finance and financial contagion from disruptions", *International Journal Of Physical Distribution & Logistics Management*, 46, 4, 414-438.

Gelsomino, L., R. Mangiaracina, A. Perego, and A. Tumino. 2016. "Supply chain finance: a literature review", *International Journal Of Physical Distribution & Logistics Management*, 46, 4, 348-366.

Gyula, L. F. 2013. "Analysis of the Impact of the Supply Chain Performance on the Overall Organisational Performance", *Annals Of The University Of Oradea*, Economic Science Series, 22, 1, 1505-1510.

Hult, G., D. Ketchen, and M. Arrfelt. 2007. "Strategic supply chain management: Improving performance through a culture of competitiveness and knowledge development", *Strategic Management Journal*, 28, 10, 1035-1052.

Joshi, D., B. Nepal, B, A. Rathore, and D. Sharma. 2013. "On supply chain competitiveness of Indian automotive component manufacturing industry", *International Journal Of Production Economics*, 143, 151-161.

Liebl, J., E. Hartmann, and E. Feisel. 2016. "Reverse factoring in the supply chain: objectives, antecedents and implementation barriers", *International Journal Of Physical Distribution And Logistics Management*, 46, 4, 393-413.

Marcuta, L. and A. Marcuta. 2013. "Role of supply chain management in increasing the competitiveness of companies in a global context", *Scientific Papers: Management, Economic Engineering In Agriculture & Rural Development*, 13, 1, 227-229.

Marwah, A., G. Thakar, and R. Gupta. 2014. "A confirmatory study of supply chain performance and competitiveness of Indian manufacturing organisations", *International Journal For Quality Research*, 8, 1, 23-37.

More, D. and P. Basu. 2013. "Challenges of supply chain finance: A detailed study and a hierarchical model based on the experiences of an Indian firm", *Business Process Management Journal*, 19, 4, 624-647.

Pfohl, H. and M. Gomm. 2009. "Supply chain finance: Optimizing financial flows in supply chains", *Logistics Research*, 1, 3-4, 149-161.

Pino, R., I. Fernández, D. Fuente, J. Parreño, and P. Priore. 2010. "Supply chain modelling using a multi-agent system", *Journal Of Advances In Management Research*, 7, 2, 149-162.

Sundram, V. P. K., A. R. Ibrahim, and V. G. R. C. Govindaraju. 2011. "Supply chain management practices in the electronics industry in Malaysia: Consequences for supply chain performance", *Benchmarking: An International Journal*, 18, 6, 834-855.

Wuttke, D., C. Blome, H. Sebastian Heese, and M. Protopappa-Sieke. 2016. "Supply chain finance: Optimal introduction and adoption decisions", *International Journal Of Production Economics*, 178, 72-81.

Yusuf, Y. Y., A. Gunasekaran, A. Musa, M. Dauda, N. M. El-Berishy, S. Cang. 2014. "A relational study of supply chain agility, competitiveness and business performance in the oil and gas industry", *International Journal of Production Economics*, 147, 531-543.

AUTHOR BIOGRAPHIES

PÉTER JUHÁSZ received his master in Economics and PhD in Business Administration from the Corvinus University of Budapest, where he is associate professor of Finance. He also serves as the secretary of CFA Society Hungary. His field of research covers financial modelling, business valuation, corporate finance, and corporate risk management. Besides, he regularly works as a trainer and coach and acts as a consultant for SMEs. His e-mail address is `peter.juhasz@uni-corvinus.hu`.

JÁNOS SZÁZ is a full Professor at the Department of Finance at the Corvinus University of 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`.

SÁNDOR MISIK holds an MA in Finance and is a PhD student at the Department of Finance at the Corvinus University of Budapest. His main research focus is on the implied correlations in the Fx markets. He works as Financial Risk Management Expert at the MOL Group. He received his ACIIA charter in 2010. His e-mail address is `smisik@mol.hu`.

HEALTHCARE DEMAND SIMULATION MODEL

Bożena Mielczarek and Jacek Zabawa
Faculty of Computer Science and Management
Wrocław University of Science and Technology
Wyb. Wyspiańskiego 27
50-370, Wrocław, Poland
E-mail: Bozena.Mielczarek@pwr.edu.pl

KEYWORDS
Simulation, Demand, Hospital, Demography.

ABSTRACT

The aim of this paper is to study the influence of demography on the demand for healthcare services. The research is carried on in Wrocław Region (WR), Poland. We apply the system dynamic method and aging chain approach to simulate the number of individuals belonging to the respective age-gender cohorts. We consider such demographic descriptive parameters as birth and death rates, life expectancy and migration factors. Then, the discrete event simulation model is used to predict the annual demand for emergency hospital care, as registered at the hospitals located in the WR. The historical data on hospital admissions are drawn from National Health Fund regional branch. The input parameters describing the population are calculated based on historical and forecasted rates of primary demographic parameters, retrieved from various databases and official projections published by the Polish Central Statistical Office (CSO). The simulation predicts that between 2011 and 2020 the WR population will grow by 4.5% and the population aged 60+ will increase by 16.2%. Over the same period the number of arriving patients, compared to 2011, will be higher by 1.52%. Furthermore, the noticeable differences will be observed in the number of arrivals between particular hospitals.

INTRODUCTION

The proper estimation of the prospective demand for healthcare services is critical when supporting the decision planning processes at the regional or national level. The valid prediction of population demand for healthcare services can directly contribute to the improvement of people's access to health facilities. Information on the expected patients' arrivals to healthcare units is necessary when attempting to diagnose, correct, and improve the performance of the healthcare system. The credible assessment of future demand determines the optimal allocation of the available resources, which are usually insufficient to meet the population's needs. The initial assumptions on the predicted level of demand have to be made in order to evaluate the economic and/or clinical effectiveness of medical procedures, treatment therapies, and preventive and screening programs.

The forecasts of healthcare demand are essential at different management levels and in relation to different frames of reference. The predictions of demand may be formulated for different groups of healthcare services, such as primary, hospital, or post-hospital care. The estimations may be directed to the particular type of healthcare services (i.e., radiology services) or to the selected healthcare units (i.e., outpatient departments, hospital wards). Furthermore, they might be focused on the particular age-gender cohort (young women, elderly persons) or on the group of patients with the particular diagnosis (i.e., cardiac patients).

Arrival patterns are usually defined based on the historical data or on-site observations. These analyses are, however, focused on the supply aspect of the particular healthcare unit, and they fail when the challenge is to estimate a population's needs at the regional level and/or for the longer time horizon. The overall objectives of the regional health policy planning have a much broader context that requires the inclusion of the random and uncertain factors, and consideration of the constant changes observed in the demographic and health structures of the population. Because of these factors, the undesirable accumulation of demand may be observed despite the satisfactory average level of supply (calculated for the specific period). The incorrectly estimated demand may lead to erroneous assumptions when formulating the health policy, both for the particular supplier and at the regional or national level.

We attempt to develop a conceptual framework of demand simulation models, where the estimation of demand for healthcare services is based not only on the historically registered population needs but also on demographical, geographical, and temporal aspects. In this study, the relations between the volume of demand and the structure of population demography (age, gender, average life expectancy, birth rates), geographical aspects (allocation of health services providers, diversified population density across the region), and temporal aspects (daily, monthly, yearly arrival schedules) will be explored.

A DEMAND PERSPECTIVE OF HEALTHCARE SIMULATION MODELS

The computer simulation approach was selected because it is well suited to tackle problems related to healthcare

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

management. The advantage of a simulation model is its ability to test any modification to fully understand the problem and to estimate the variability involved in the observed process. Furthermore, the healthcare system is a highly complex one and a simulation paradigm seems to be the best choice in this case.

The demand perspective may be taken as the main criterion in the classification of healthcare simulation models. Mielczarek (2014) divides the models into three basic groups: improvements, disease, and strategic, depending on the purpose of estimating the future demand.

The first group (improvements) of healthcare simulation models concentrates on the current work of units that provide the healthcare services. The models try to suggest improvements to the internal organisation of the unit assuming a certain level and structure of demand. The models are used for resource allocation, staff scheduling, admission planning, and managing patient flows in the healthcare units. They are concerned with formulating an overall diagnosis, identifying bottlenecks, and suggesting the changes that could improve the system's performance. The classical problem formulation is usually as follows: how to change the operation of the system to satisfy the output measures, given a certain level and characteristic of the demand?

Models from this group may concern:

- different clinical settings, such as operating theatres, outpatient clinics ambulatory care units, or diagnostic departments (Bowers, Mould 2004), (Persson, Persson 2009), (Testi, Tanfani & Torre 2007), (Rohleder et al. 2011);
- complex centers such as multi-unit hospitals, multi-facility outpatient centers (Cochran, Bharti 2006), (Matta, Patterson 2007);
- treatment processes such as radiation therapies, cataract surgeries, or local stroke care services (Werker et al. 2009), (Comas et al. 2008), (Bayer et al. 2010).

The goal of the models from group 2 (disease) is to study the cost-effectiveness or clinical effectiveness of medical procedures, medical treatments associated with clinical pathways, and prevention strategies or contemporary health trends. The demand is usually defined as a flow of patients classified by age, gender, and medical history, who are included in the simulation when they acquire the particular disease. The simulation is run within a sub-group of the whole population, pre-defined according to the goals of the study (Cooper et al. 2008), (Mar, Arrospide & Comas 2010), (Neovius et al. 2010).

Group 3 (strategic) models concentrate on the population needs and help the decision makers in implementing the capacity planning process at the regional or national level (Gupta et al. 2007), (Vissers, Adan & Dellaert 2007), (Desai et al. 2008). The models are also used to study the influence of the changes in demand on the healthcare units' standards of service. In these "what-if" type models, the "if" refers to the fluctuations in the demand intensity and patterns, and their impact on the system's operation.

This paper builds up on our previous study that reports on the use of combined simulation methods to support healthcare demand predictions (Mielczarek and Zabawa 2016a). In that paper we discussed the implementation of cohort modeling approach using the system dynamics method. The projections of long-term population evolutions were performed on the aggregated data and the analysis was focused on pre-specified age-gender cohorts. The demographic groups were described using parameters such as birth and death rates, life expectancy, and migration descriptors. This paper deploys the discrete event simulation model to study the demand for emergency hospital care. The model considers the demographic changes observed in the structure of the population inhabiting the region, the geographical factors that determine the selection of the preferred hospital, and the temporal factors that reflect the variability of the demand that is related to the season, month and time of the day.

SYSTEM DESCRIPTION

In Poland, the entry point for elective as well as emergency patients arriving to hospitals is an admission unit (AU) or a hospital emergency ward (EW), where the patient is qualified for hospital care and, if necessary, receives some medical treatment. A similar admission procedure applies to emergency patients without referral, emergency patients with referral, and elective patients. The AU provides consultancy and basic medical intervention and qualifies the patient for further hospital treatment. The EW performs an initial diagnosis, offers the medical treatment necessary to stabilize vital functions, and establishes the need for further hospital care. Both the AU and the EW may, after a consultancy, recommend the admission of the patient at the hospital ward or refer her/him for further treatment with the family doctor.

Our research is performed for two administrative districts of Lower Silesia, the fourth largest province in Poland. These two districts are called the Wrocław Region (WR) and encompass nine counties: the capital of Lower Silesia, Wrocław and eight other counties that are close to the capital. There are 17 AU/EW units located in the WR and they serve, in the first place, the inhabitants of the WR. The hospital care may, however, be also delivered to patients coming from other Polish provinces. At the same time, people living in the WR may receive medical treatments from the AUs and EWs located in other Polish sub-regions. The algorithm of forecasting the future amount of services delivered in the WR has to be therefore based not only on the demand coming from the WR population and covered by the WR hospitals but also on the demand generated by the inhabitants of the neighboring regions who arrive at one of the 17 AU/EW

units located in the region. Additionally, a certain number of patients inhabiting the WR may select the hospital outside the WR and this may result in a slight decrease in the demand covered by the WR hospitals.

MODEL CONSTRUCTION

Sub Models

The main model consists of two connected sub-models (Figure 1), each developed using a different simulation method. The discrete event (DES) sub-model of the WR healthcare emergency system was built using Rockwell Automation Technologies' Arena Simulation software, version 15.0. The model generates batches of emergency patients on a daily basis according to month-dependent arrival patterns. The second model was constructed using ExtendSim 9 by Imagine That Inc. to simulate demographic changes of the WR region. The sub-population model uses the system dynamics (SD) method and an aging chain approach to forecast the demographic changes that will be observed within the WR population over next 20 years (Mielczarek, Zabawa 2016b). The age-gender cohort simulation is performed using the deterministic approach, and hence there is no need to repeat the simulation runs. Only one replication per simulation experiment is performed and the output values describing the quantitative status of all age-gender cohorts, as registered in the subsequent years, are exported and kept in the external databases. The population data is then imported to the DES model to perform stochastic simulation for generating patient arrivals to the WR healthcare system.

More information about the aging chain population model may be found in Mielczarek and Zabawa (2016a).

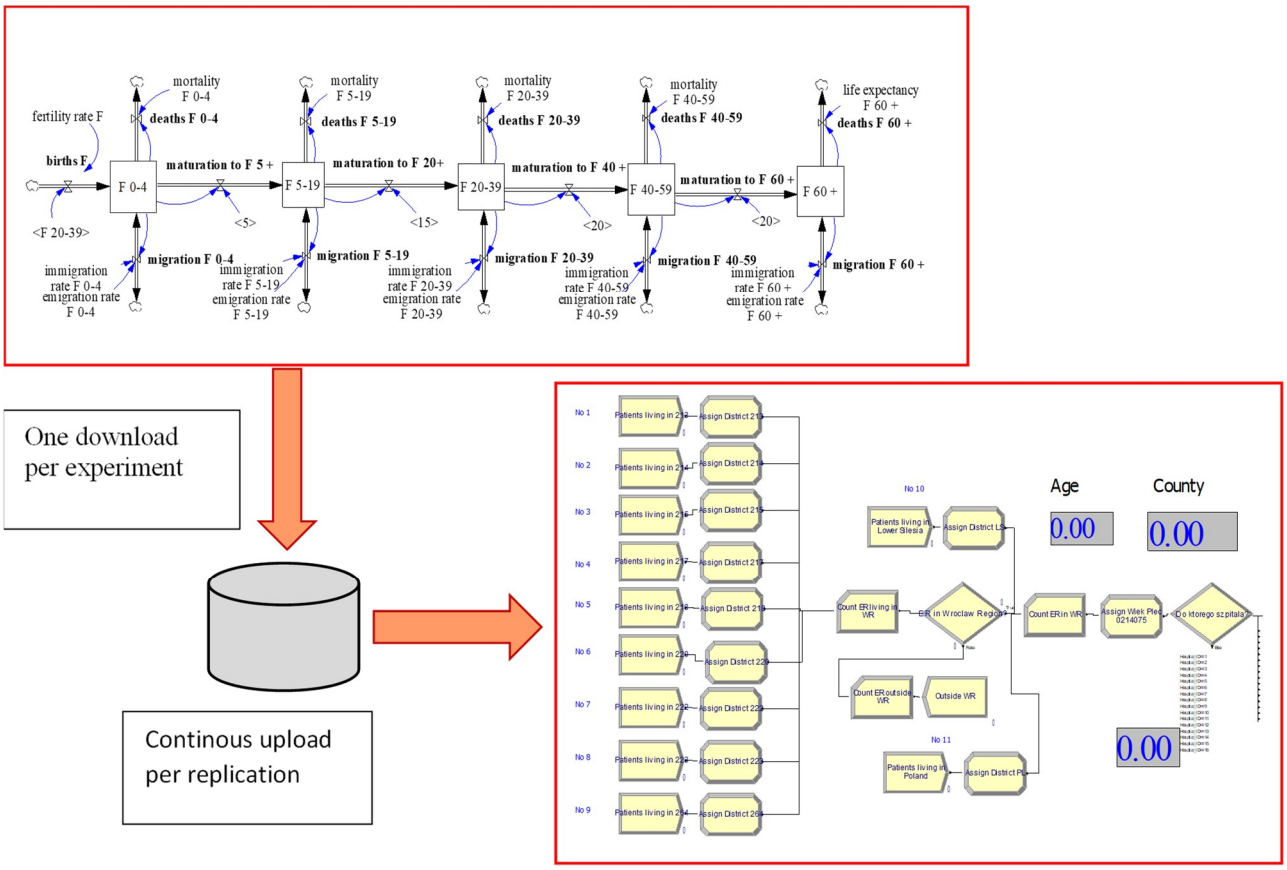


Figure1: Layout of the SD (upper) - DES (lower) model (fragments)

Simulation

The basic input flow is modeled by Arrivals Inflow 1 (Figure 2). The DES model samples the patients' arrivals from Poisson distributions with mean rates changing according to the county and the calendar month. The Inflow 1 describes the individuals whose place of residence is one of the counties of the WR and who express the need for hospital treatment. The majority of these calls will be served by the hospitals located in the WR; however, some patients will receive the treatment outside the region. This leakage is expressed by the Outflow (Figure 2) and represents the WR patients who decided to select the hospital outside the WR. The main Inflow is, however, expanded by the arrivals from other Lower Silesian counties (Arrivals Flow 2) or from the other Polish provinces (Arrivals Flow 3). The final flow of patients is then divided into 17 mini-flows that describe the demand registered in 17 hospitals located in the WR. Every admitted patient receives an individual

treatment in the particular AU/EW and the main diagnosis is formulated. The decision as to whether to send the patient home or start treatment at a hospital ward is usually made within few hours. After a consultation and the medical treatment, the patient is sent home or is referred to a hospital ward.

Simulation starts with the empty and unbiased system and lasts for 365 days. Every experiment is replicated 10 times without the warm-up period. A different stream of random numbers was used in each run. Output metrics include the total daily demand, i.e., the daily number of patients arriving at all WR hospitals, and annual demand as registered by each of the 17 hospitals located in the WR.

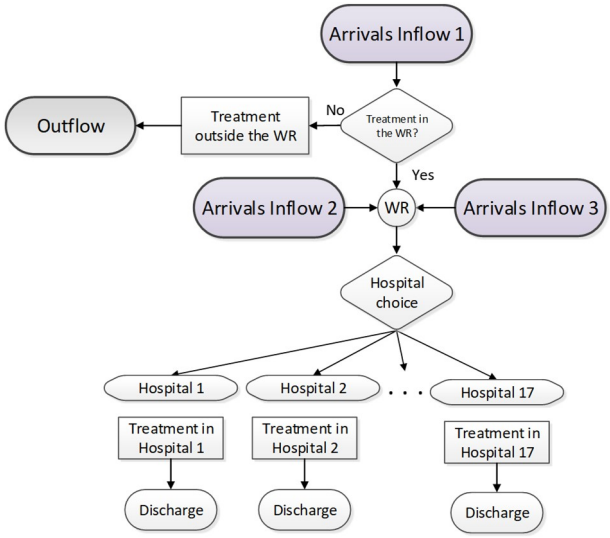


Figure 2: Flowchart of discrete event simulation model

RESULTS

Model Validation

To test the model, a historical validation was performed and a comparative analysis between the model output and actual performance of the system was carried out. Two output measures were calculated: the total annual number of patients as registered in each hospital in 2010 (Table 1) and the total monthly number of patients as registered in the WR (Table 2) in 2009 and 2010.

The results of simulation are consistent with the historical data. There are very small differences between historical and simulation data when comparing the total number of patients arriving at a particular hospital (Table 1), and the MPEs (Mean Percentage Errors) that relates to the whole region is also acceptable. The temporal distribution of patient arrivals exhibits, in some months, certain discrepancies compared to historical demand (Table 2); however, the averages calculated for each month for two validated years (2009 and 2010) are highly convergent. The values of MPEs indicate that the simulation model provides, on average, acceptable results for the estimation of the WR demand.

Table 1: Historical Validation – Total Annual Number of Patients as Registered in Each Hospital

Hospital no.	Historical data	Simulation data	MPE (%)
1	11 133	10 944	-1.7
2	2 919	2 830	-3.1
3	30 956	30 411	-1.8
4	18 385	17 918	-2.5
5	34 299	34 092	-0.6
6	6 066	5 857	-3.4
7	2 282	2 251	-1.3
8	12 317	12 849	4.3
9	6 826	6 816	-0.1
10	20 363	20 068	-1.4
11	13 188	12 748	-3.3
12	2 808	2 799	-0.3
13	27 888	27 242	-2.3
14	4 300	4 254	-1.1
15	1 218	1 194	-2.0
16	5 187	5 129	-1.1
17	3 950	3 921	-0.7
WR	204 085	201 324	**-1,35**

Table 2: Historical Validation – MPE (%) for Total Monthly Number of Patients as Registered in the WR

Month	2009	2010
January	-0.66%	-2.98%
February	0.52%	-4.25%
March	-0.45%	-7.54%
April	-0.40%	6.36%
May	-0.31%	-5.21%
June	0.36%	-15.12%
July	0.54%	-16.51%
August	1.65%	-5.52%
September	0.01%	4.00%
October	-0.45%	-9.46%
November	-0.54%	0.03%
December	0.32%	0.30%
Monthly average	**4.71%**	**0.36%**

Simulation Experiments

Under the base case scenario, we first run the SD sub-model to trace the evolution of the WR population from 2002 to 2020 (Figure 3 and 4). The next step is to evaluate the influence of the observed and predicted demographic trends on the forecast level of demand for hospital services exhibited by the WR population.

The simulation of the demographic changes begins in 2002 and runs through 2014 according to parameters extracted from the historical values, and then the SD sub-model is input with the coefficients calculated based on the demographic trends described in the official forecasts published by the Polish Government Population Council for 2014–2050 (Waligórska *et al*. 2014). The simulation

predicts that the WR population is gradually increasing and the population aging is an irreversible phenomenon. Between 2011 and 2020, the WR population is predicted to grow by 4.57%; however; over the same period the population aged 60+ will increase by as much as 16.20% (Figure 4).

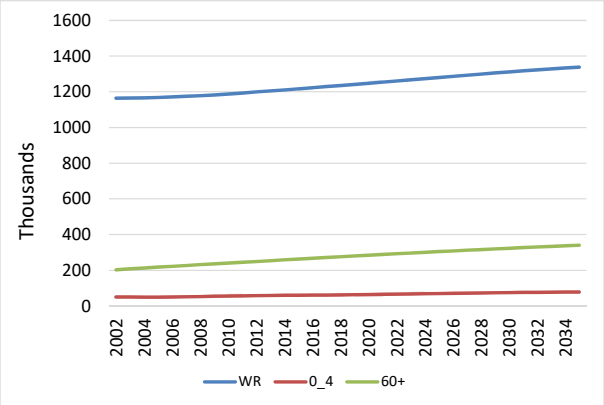


Figure 3: Projected status of the WR population

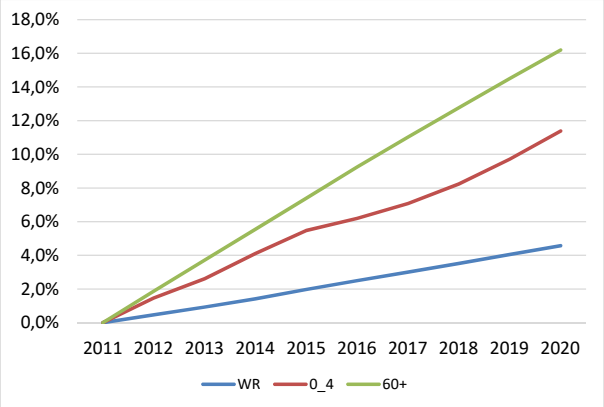


Figure 4: Projected percentage changes of the WR population in relation to 2011

The results of the aging chain simulation are then entered into the DES sub-model to observe the impact of population changes on the total number of patients arriving at the WR hospitals. As expected, demographic trends results in a noticeable growth, between 2011 and 2020, in the numbers of patients registering at the WR hospitals (Figure 5). The analysis of percentage changes of demand between 2011 and 2020, in relation to the quantities registered in 2011, shows that the level of demand will increase between 2011 and 2016. Next, it will remain stable to a year 2018, however between 2018 and 2020 the number of arriving patients will increase by as much as 1.52% compared to 2011 (Figure 5). It means that more than 3000 visits will be registered in the WR hospitals in 2020 relative to the year 2011.

Simulation experiments enable us to observe the demand directed to particular hospitals located in the WR. The detailed analysis shows that there are noticeable differences in the number of arrivals that the particular hospitals will register in the next few years (Figure 6). Although, between 2011 and 2020, the numbers of patients arriving at the WR hospitals will increase on average by about 1.52%, some hospitals will have to deal with the substantial growth in the number of arriving patients (i.e., Unit No 2: 2.54% and Unit No 12: 2.52%), while others will experience only a modest increase (i.e., Unit No 6: 0.53% and Unit No 11: 0.71%). This observation may have important implications for the future distributions of the resources on the regional level.

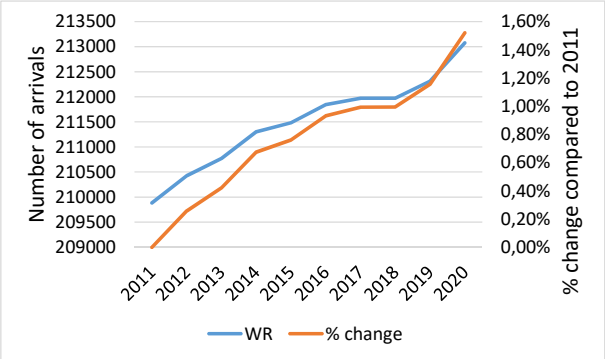


Figure 5: Predicted number of patients arriving at hospitals between 2011 and 2020 as a result of the simulated demographic changes in the WR population

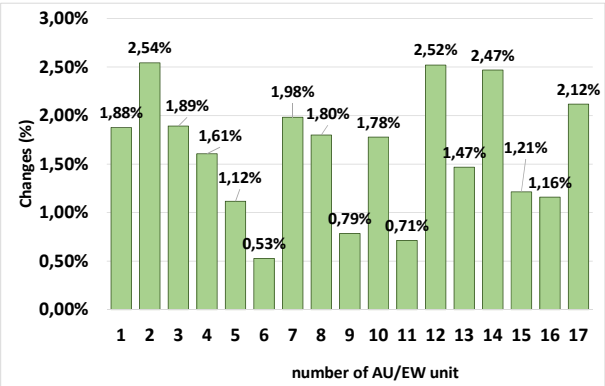


Figure 6: Percentage changes in the number of individuals arriving to the WR hospitals in 2020 compared to 2011

DISCUSSION AND CONCLUSSIONS

Healthcare policy planning at the regional level strongly relies on the proper and accurate estimation of future demand for services as expressed by the population inhabiting the area. The common approach is to focus on the past demand and the current levels of resource utilization (Cardoso *et.al.* 2012). Demographic trends have, however, a significant and stable effect on healthcare demand. The ongoing aging phenomenon, migrations, the declining children's population are the main reasons that the demand for the healthcare services changes gradually but continuously. The approach that neglects the influence of the demographic trends does not ensure an adequate identification of people's needs in the

future and may lead to an non optimal resource allocation. The health policy planning processes require more reliable forecasts of the intensity and the structure of the demand. This might improve the equity of access to health services across the region and adjust the future regional budget to the changing needs of people inhabiting the area.

The present work attempts to develop a conceptual framework of demand simulation models, where the estimation of demand for healthcare services is based not only on the historically registered population needs but also on demographical aspects. We suggest an original approach based on two connected simulation models: the system dynamics model to forecast the demographic changes of the WR population and the discrete simulation model to predict the hospital demand.

Our simulation experiments confirmed the strong impact of the ongoing demographic trends on the volume and geographical distribution of healthcare demand. The forecasts generated by the simulation model are convergent with official projections prepared by Polish Ministry of Health (Ministry of Health 2016). According to this analysis the intensity of the aging process in the region will be stronger than on the national scale and the demand for hospital beds will be increasing. When disregarding this influence, the unreliable forecasts may drive the health planning policies. We have also demonstrated the usefulness of integrating the SD and DES approaches to better explore the relationship between projections of population dynamics and forecasted demand for healthcare services. The complementary use of two simulation methods adds a new value to the process of predicting the future needs by considering a range of characteristics that describe both the population and the region. The demand for healthcare services is strongly driven by uncertain factors, and some of these factors are closely related to on-going changes in age-gender population profiles.

The discussion presented in this paper is a first step toward more comprehensive studies, and several research topics seem to be worth pursuing. First, the disaggregation rate of the aging chain into age-gender cohorts should be higher in both models to enable an even deeper analysis of the demographic trends and their impact on the healthcare demand. The same disaggregation pattern should be applied in the case of SD and DES sub-models to facilitate better communication and bilateral exchange of information. Next, we would like to extend the population model with some external/indirect incentives, such as economic growth, development of education or transportation infrastructure, and influence of the pro-demography policies. Lastly, the changes in population demography directly affect the epidemiological parameters and the risk factors. The changing risk factors modify the morbidity trends and thus influence the prevalence of diseases. The demand for healthcare services changes accordingly.

ACKNOWLEDGMENTS

This project was financed by the grant *Simulation modeling of the demand for healthcare services* from the National Science Centre, Poland, which was awarded based on the decision 2015/17/B/HS4/00306.

REFERENCES

Bayer, S.; C. Petsoulas; B. Cox; A. Honeyman; and J. Barlow. 2010. "Facilitating Stroke Care Planning through Simulation Modelling". *Health Informatics Journal* 16(2), 129–143.

Bowers, J. and G. Mould. 2004. "Managing Uncertainty in Orthopaedic Trauma Theatres". *European Journal of Operational Research* 154(3), 599–608.

Cardoso, T.; M.D. Oliveira; A. Barbosa-Povoa; and S. Nickel. 2012. "Modeling the Demand for Long-term Care Services under UncertainIinformation". *Health Care Management Science* 15, 385–412.

Cochran, J.K. and A. Bharti. 2006. "Stochastic Bed Balancing of an Obstetrics Hospital". *Health Care Management Science* 9(1), 31–45.

Comas, M.; X. Castells; L. Hoffmeister; R. Román; F. Cots; J. Mar; S. Gutiérrez-Moreno; and M. Espallargues. 2008. "Discrete-Event Simulation Applied to Analysis of Waiting Lists. Evaluation of a Prioritization System for Cataract Surgery". *Value in Health* 11(7), 1203–1213.

Cooper, K.; R. Davies; J. Raftery; and P. Roderick. 2008. "Use of a Coronary Heart Disease Simulation Model to Evaluate the Costs and Effectiveness of Drugs for the Prevention of Heart Disease". *The Journal of the Operational Research Society* 59(9), 1173–1181.

Desai, M..; M.L. Penn; S. Brailsford; and M. Chipulu. 2008. "Modellin of Hampshire Adult Services–gearing up for Future Demands". *Health Care Management Science* 11(2), 167–76.

Gupta, D.; M.K. Natarajan; A. Gafni, L. Wang; D. Shilton; D. Holder; and S. Yusuf. 2007. "Capacity Planning for Cardiac Catheterization: A Case Study". *Health Policy* 82(1), 1–11.

Mar, J.; A. Arrospide; and M. Comas. 2010, "Budget Impact Analysis of Thrombolysis for Stroke in Spain: A Discrete Event Simulation Model". *Value in Health* 13(1), 69–76.

Matta, M.E. and S. Patterson. 2007. "Evaluating Multiple Performance Measures across Several Dimensions at a Multi-Facility Outpatient Center". *Health Care Management Science* 10(2), 173–194.

Mielczarek B. 2014. "Simulation Modelling for Contracting Hospital Emergency Services at the Regional Level". *European Journal of Operational Research* 235(1), 287–299.

Mielczarek, B. and J. Zabawa. 2016a. "Modelling Population Growth, Shrinkage and Aging using a Hybrid Simulation Approach: Application to Healthcare". In *Proceedings of the 6th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, edited by Yuri Merkuryev, T. Oren, and M. S. Obaidat. 75–83. Lisbon, Portugal: SCITEPRESS – Science and Technology Publications, Lda.

Mielczarek, B. and J. Zabawa. 2016b. "Modeling Healthcare Demand Using a Hybrid Simulation Approach". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick. 1535–1546.

Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
Ministry of Health. 2016. *Mapy Potrzeb Zdrowotnych (in Polish, Maps of Health Needs).* (March 2018) http://www.mpz.mz.gov.pl/wp-content/uploads/2016/04/01_podsumowanie_dolnoslaskie-1.pdf
Neovius, K.; F. Rasmussen; J. Sundström; and M. Neovius. 2010. "Forecast of Future Premature Mortality as a Result of Trends in Obesity and Smoking: Nationwide Cohort Simulation Study". *European Journal of Epidemiology* 25(10), 703–709.
Persson, M. and J.A. Persson. 2009. "Health Economic Modeling to Support Surgery Management at a Swedish Hospital". *Omega* 37(4), 853–863.
Rohleder, T.R.; P. Lewkonia; D.P. Bischak; P. Duffy; and R. Hendijani. 2011. "Using Simulation Modeling to Improve Patient Flow at an Outpatient Orthopedic Clinic". *Health Care Management Science* 14(2), 135–145.
Testi, A.; E. Tanfani; and G. Torre. 2007. "A Three-Phase Approach for Operating Theatre Schedules". *Health Care Management Science* 10(2), 163–72.
Vissers, J.M.H.; I.J.B.F. Adan; and N.P. Dellaert. 2007. "Developing a Platform for Comparison of Hospital Admission Systems: An Illustration". *European Journal of Operational Research* 180(3), 1290–1301.
Waligórska, M.; Z. Kostrzewa; M. Potyra; and L. Rutkowska. 2014. "Population Projection 2014-2050". CSO, Demographic Surveys and Labour Market Department.
Werker, G.; A. Sauré; J. French, and S. Shechter. 2009. "The Use of Discrete-Event Simulation Modelling to Improve Radiation Therapy Planning Processes". *Radiotherapy and Oncology* 92(1), 76–82.

AUTHOR BIOGRAPHIES

BOŻENA MIELCZAREK is an Associate Professor in the Department of Operational Research at Wrocław University of Science and Technology, Poland. She received an MSc in Management Science, a PhD in Economics, and D.Sc. in Economics from Wrocław University of Science and Technology. Her research interests include simulation modeling, health service research, decision support, and financial risk analysis. Her email address is Bozena.Mielczarek@pwr.edu.pl

JACEK ZABAWA is an Assistant Professor in the Department of Operational Research at Wrocław University of Science and Technology, Poland. He received an MSc in Management Science and a PhD in Economics from Wrocław University of Science and Technology. His research interests include financial statement projections and application of simulation in management, with particular emphasis on discrete rate modeling. His email address is Jacek.Zabawa@pwr.edu.pl

THE EFFECTS OF MODEL SELECTION ON THE GUARANTEES ON TARGET VOLATILITY FUNDS

Gábor Kondor
Department of Finance
Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest, Hungary
E-mail: gabor.kondor@uni-corvinus.hu

KEYWORDS
Target volatility fund, stochastic model, option, guarantee costs

ABSTRACT

Target Volatility Funds are becoming a more and more popular asset class amongst Variable Annuity product designers. After the recent global crisis these funds provided a decent way to assure the guarantees that investors find so attractive. However, pricing of these guarantees highly depends on the modelling assumptions we use. Although this is an exciting and demanding problem, not much attention has been shed on this topic in the academic literature. In my work I extend some of the existing results to the Barndorff-Nielsen–Shephard model and to a Lévy-process with stochastic time.

INTRODUCTION

A Target Volatility Fund (TVF) is a portfolio of a risky and a risk-free asset dynamically rebalanced with the aim of maintaining a stable portfolio level. The price of the guarantees, also called guarantee cost, in question is the cost of providing a minimum payoff for the policyholder. Of course, when it comes to guarantees on TVFs, the guarantee cost equals the value of a European put option written on the TVF. The option price and the extent of how well TVFs can do what they are supposed to may highly depend on the stochastic model we use to simulate the price process of the risky asset.

In the past few years researchers have become more concerned in the pricing of derivatives written on TVFs, however, the earliest studies mainly looked at this asset class from a portfolio management and risk-return analysis point of view. Chew (2011) compares the performance of typical volatility-managed funds and other related market strategies and discusses the advantages of TVFs to investors and insurers. Stoyanov (2011) applies a Heston stochastic volatility model framework calibrated to Asian equity market data and shows that in the long run a target-volatility strategy significantly improves both the downside and the upside of the return distribution relative to a fixed-mix strategy. Contrary to previous researches, Xue (2012) claims that in an SVJD (stochastic volatility jump diffusion) model, because of the possible jumps, volatility targeting may not be superior to fixed allocation in terms of risk-return profile, and investors may favor one strategy over an other based on their own jump risk evaluation. Finally, in such a context, Hocquard et al. (2013) based on empirical researches also emphasizes the benefits of a constant volatility approach, as it can help investors obtain desired risk exposures over the short and long term, reduce tail-risk exposure, and increase the portfolio's risk-adjusted performance.

Morrison and Tadrowski (2013) considers the valuation of guarantees, that is European put options, written on target volatility funds, and the impact of some of the key modeling decisions on the derivative prices. The authors compare three popular stochastic models: 1. constant volatility Black-Scholes model, 2. Heston model which features stochastic volatility, and 3. a stochastic volatility model that incorporates jumps (SVJD), and they also examine the effect of different choices of rebalancing frequencies.

Kim and Enke (2016) proposed artificial neural networks for volatility forecasting to enhance the performance of an asset allocation strategy, and compared it to different volatility forecast methodologies. Last but not least, Torricelli (2017) was the first to develop a continuous-time finance mathematical model for target volatility strategies based purely on observable market inputs, which framework enables an efficient numerical valuation of certain derivatives on TVFs, e.g. of put options.

In my research I follow the path of Morrison and Tadrowski (2013) and expand the set of stochastic models under examination by evaluating European put option prices in the cases of Barndorff-Nielsen–Shephard model and Variance Gamma process with Gamma–Ornstein-Uhlenbeck stochastic clock (which is a Lévy-process with stochastic time), as well.

The rest of the paper is organized as follows. First, I introduce the analysis that Morrison and Tadrowski (2013) used. Afterwards, I describe the calibration methods that I used to determine the model parameters. Then, I interpret the result, and finally, I conclude.

ANALYSIS

To make the way of the analysis clear I briefly describe the method that Morrison and Tadrowski (2013) used.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

First, they derived implied volatilities. To do this, they started off with calibrating their models to the same risky asset that they used in the dynamically rebalanced TVF – they chose EuroStoxx 50 as an underlying product. After that, they applied Monte Carlo (MC) simulation for different maturities and strike prices. In each MC simulation, they generated 5000 paths for the risky asset and at the same time, according to the rebalancing procedure, they derived the price processes of the TVF. Then, they evaluated an option written on the TVF and an option written on the equity underlying the TVF, as well. Finally, they calculated the implied volatilities from the option prices and graphed the results for each model selection, and for three different maturities.

Secondly, they also compared the prices of at-the-money (ATM) put options on TVF and calculated percentage changes.

(As noted before, they examined the effects of rebalancing frequencies, as well, which is not included in this paper.)

Implied volatility is an adaptable tool to compare the prices along the different models because it rescales the results to a more tractable form. We may think that if the rebalancing was perfect, i.e. the stochastic process of the equity was predictable and rebalancing was carried out in continuous time, the implied volatility curves of the options on TVFs would be completely flat. Indeed, seemingly it is the case for the Black-Scholes model but there is no theoretical proof of it. Moreover, it is not even necessarily true, especially for the models where the driving process is not similar to the one of the Black-Scholes model. However, we still can use implied volatilities to compare the differences of the resulted option prices. If we experience some deflection, that indicates the imperfection of volatility forecasting method or the characteristics of the stochastic models (or both).

Unfortunately, we cannot predict perfectly the future volatility, thus we need some kind of estimation. To this end, I use the same Exponentially Weighted Moving Average (EWMA) that Morrison and Tadrowski (2013) did:

$$\left(\hat{\sigma}_t^{equity}\right)^2 = \lambda\left(\hat{\sigma}_{t-\Delta t}^{equity}\right)^2 + (1-\lambda)\frac{1}{\Delta t}\log^2\left(\frac{S_t}{S_{t-\Delta t}}\right),$$

where $\hat{\sigma}_t^{equity}$ denotes the estimated volatility of the equity index over the next period, S_t is the value of the equity at time t, $\Delta t = 1$ business day is the time step for the discretization and λ is the rate of exponential decay. With a value of $\lambda = 0.99$ the mean age of the data being used is $\frac{\Delta t}{1-\lambda} \approx 0.4$ years.

The estimated volatility determines the weight of the equity in the dynamic rebalancing, which is given by

$$w_t^{equity} = \min\left(\frac{\sigma_{target}}{\hat{\sigma}_t^{equity}}, 100\%\right),$$

where σ_{target} is the prearranged target volatility level. This means that there is no leverage allowed in the risky asset if the estimated volatility is below the target volatility level.

To summarize, I examined the effects of modelling assumptions in the case of the models below.

- Black-Scholes (BS) model (Black & Scholes, 1973)
- Heston model (Heston, 1993)
- Bates model (as SVJD model; Bates, 1996)
- Barndorff-Nielsen–Shephard (BNS) model (Barndorff-Nielsen & Shephard, 1999)
- Variance Gamma process with Gamma–Ornstein-Uhlenbeck stochastic clock (VGGOU; discussed in general in Carr et al. (2003))

CALIBRATION

The risky asset I consider in this research is the S&P 500 Index and as an initial price for the simulation I use the spot price of this equity on October 9th, 2017, that is $S_0 = 2544.73$. The 1Y USD LIBOR interest rate at this date is $r = 1.809\ \%$.

For the calibration of the Black-Scholes model, i.e. to determine the constant volatility σ of the model, I used the 2-year historical data of the index.

To calibrate the other four models I applied a technique described in Kilin (2011). He investigated three different methods to ascertain which one of them accomplishes the fastest calibration of various stochastic models. The winner announced adapts the *direct integration method* which is based on calculating the prices of European call options using the characteristic function of the stochastic models. With this, we can calibrate our models to the underlying product's option surface observed on the market, with the parameters chosen form predefined intervals. The calibration itself consists of three runs of the Differential Evolution algorithm which applies the direct integration method, and with the results of the previous as initial guesses, three runs of the Levenberg-Marquardt algorithm. Because of the second phase, the resulting parameters can be out of the predefined intervals. If they seem reasonable we can let them be. As an object function, the algorithm uses the sum of the squared differences of the market option prices and the ones calculated with the actual parameter set. For some details and references of the algorithms mentioned here see Kilin (2011). The characteristic functions of the four models can be found in Schoutens et al. (2003) or Kilin (2011).

For the calibration I used 90 different options in total, along 10 maturities from 0.15 to 3.18 years, and 9 strikes from 1100 - deep in-the-money (ITM) call option - to 2800 – out-of-the-money (OTM) call option. Table 1

shows the resulted parameters of the calibrated models. We may notice some unexpected parameters. First, the high and positive correlation (ρ) in the cases of Heston and Bates models. Our presumption is a negative correlation between the volatility and stock price processes, thus further analysis is needed to determine the cause of this outcome. Next, in the Bates model the parameters λ, μ_J, σ_J, which are associated with jumps, are taken at the lower bounds. This corresponds to rare and small jumps, which is probably because the underlying equity is the S&P 500 Index. For this, to occur a jump the whole market should sustain a shock. At last, the parameter b of the BNS model is quite large and taken at the upper bound. The reason for this is that, the higher upper bound I chose for the initial interval for parameter b the better the model fitted to the actual data. However, at the same time the larger discretization error I received, especially in the long run and in the case of high strike prices. This upper bound seemed still reasonable, as over this the value of the goodness of fit didn't improve much and I didn't experience significant discretization error.

Table 1: Parameters of the Models Calibrated Using Direct Integration Method. The Predefined Intervals Are Shown in Square Brackets below the Parameters.

BS			
$\sigma = 0.114$			
Heston			
$\sigma_0 = 0.067$	$\eta = 0.070$	$\kappa = 0.019$	$\theta = 0.015$
$\in [0.001, 1]$	$\in [0.001, 1]$	$\in [0.005, 6]$	$\in [0.001, 1]$
$\rho = 1.000$			
$\in [-1, 1]$			
Bates			
$\sigma_0 = 0.067$	$\eta = 0.256$	$\kappa = 0.005$	$\theta = 0.015$
$\in [0.001, 1]$	$\in [0.001, 1]$	$\in [0.001, 6]$	$\in [0.001, 1]$
$\rho = 1.000$	$\lambda = 0.050$	$\mu_J = 0.010$	$\sigma_J = 0.010$
$\in [-1, 1]$	$\in [0.05, 2]$	$\in [0.01, 0.5]$	$\in [0.01, 0.5]$
BNS			
$\sigma_0 = 0.004$	$a = 9.317$	$b = 1500$	$\lambda = 6.871$
$\in [0.001, 0.15]$	$\in [0.01, 100]$	$\in [5, 1500]$	$\in [0.01, 10]$
$\rho = -1.10e-08$			
$\in [-50, 0]$			
VGGOU			
$C = 254.18$	$G = 423.80$	$M = 319.10$	
$\in [1, 100]$	$\in [30, 200]$	$\in [30, 200]$	
$\lambda = 1.896$	$a = 11.376$	$b = 11.990$	$y_0 = 1$
$\in [0.5, 20]$	$\in [2, 20]$	$\in [2, 20]$	$= 1$

RESULTS

I set the target volatility level to 5% by reason of choosing a (much) higher target volatility level would have made the TVF implied volatility curves equal to the equity implied volatility curves in the VGGOU case, and at (much) lower target volatility levels the implied volatility calculation algorithm does not always succeed, especially at lower strikes.

To investigate the effects of model selection on the guarantees on TVFs, first I present the ATM put option prices for maturities 2, 5 and 10 years in Table 2. To determine the prices, I generated 10000 scenarios.

After this, I inspect the implied volatilities calculated using another 10000 trajectories and I also show some sample scenarios (Figure 1-10). Particularly, to present these results I use the following two types of figures. The first one graphs the predefined target volatility level (dashed red line), the implied volatility for the equity put options (solid light blue line) and the TVF put option (solid dark blue line). The second type shows an example scenario for the model, which consists of four subfigures: a realization of the equity index (upper left corner), the model volatility and the one received by the EWMA estimator (upper right corner), the corresponding trajectory of the TVF (bottom left corner) and the TVF volatility with the target volatility level (bottom right corner).

ATM Put Options On TVF

Table 2 shows that for every maturity there is an increase in the put option price relative to the Black-Scholes model. However, in most cases it is below 10% so should be considered as an indicator rather than a decisive fact. To confirm the results a robustness analysis is needed. Yet, we can't let the VGGOU case go by which is rather convincing and where the percentage change increases from 60% to a level over 200%. Apart from VGGOU model, for every stochastic model the put option price decreases in time. Since in the VGGOU case the option price starts at a relatively high level and it even grows higher in the long run, it results in an extremely large percentage change for 10 years' maturity.

Table 2: Prices of ATM Put Options on TVF and Percentage Changes Relative to the Black-Scholes Model. Determined Using 10000 Scenarios.

Model \ years	2	5	10
BS	34.40	32.43	23.22
Heston	37.01	36.22	24.79
	7.58%	*11.67%*	*6.75%*
Bates	36.01	33.63	24.43
	4.68%	*3.67%*	*5.24%*
BNS	38.60	33.90	25.09
	12.21%	*4.51%*	*8.06%*
VGGOU	55.20	68.25	71.40
	60.49%	*110.42%*	*207.55%*

Black-Scholes Model

In the case of the Black-Scholes model the implied volatility on the TVF is almost equal to the target volatility for all strikes and maturities (Figure 1).

The EWMA estimator seems to work well in estimating the constant model volatility, as it is close to it in Figure

2, upper right corner. The TVF process varies on a smaller scale than the equity index (Figure 2, subfigures on the left) and the TVF model volatility fluctuates around the target volatility level (Figure 2, bottom right corner). Overall, the method works well in this case.

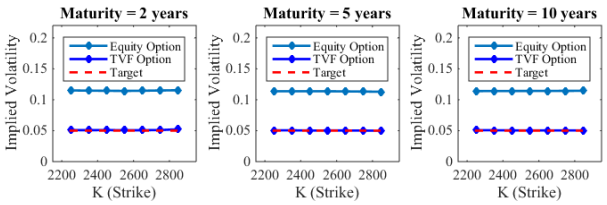


Figure 1: Implied Volatilities for the Black-Scholes Model. Calculated Using 10000 Scenarios.

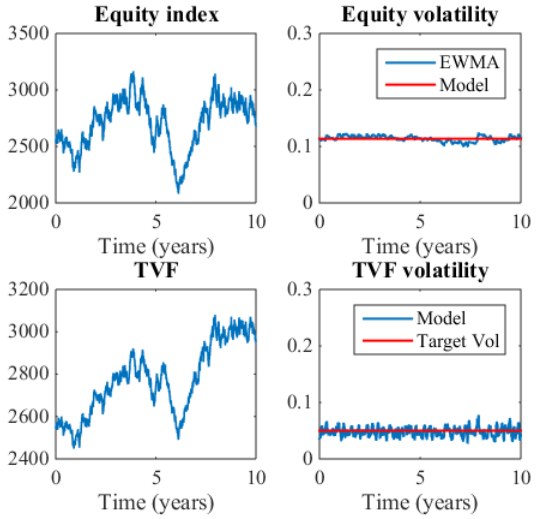


Figure 2: Example Scenario for the Black-Scholes Model.

Heston Model

As for the Heston model, the equity implied volatility curve has a large skew at 2 years' maturity that flattens over time, and the level of the curve increases at the same time (Figure 3).The TVF implied volatility curve tends to fit quite well, although in the short run it has a small skew that seems to disappear over time, as well.

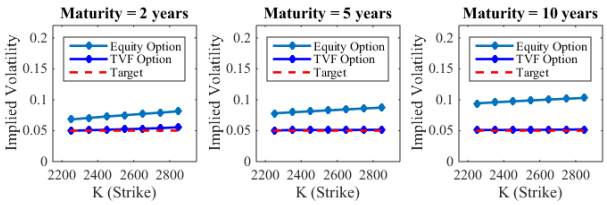


Figure 3: Implied Volatilities for the Heston Model. Calculated Using 10000 Scenarios.

The equity volatility grows towards the long variance as it is higher than the spot variance. The EWMA estimator works well in following it, however, we can see some distortions (Figure 4, upper right corner). The TVF model volatility moves around the target with a small deviation (Figure 4, bottom right corner). This suggests that it can maintain an almost constant volatility level over time, but according to Table 2 the method results in higher guarantee costs.

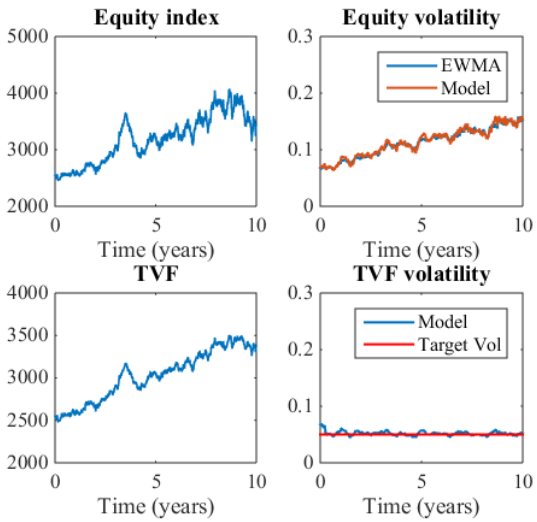


Figure 4: Example Scenario for the Heston Model.

Bates Model

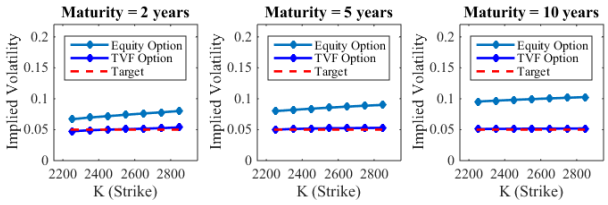


Figure 5: Implied Volatilities for the Bates Model. Calculated Using 10000 Scenarios.

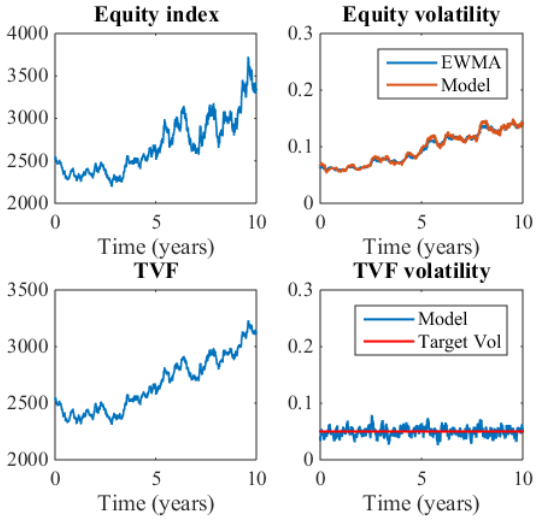


Figure 6: Example Scenario for the Bates Model.

In the case of the Bates model, in many aspects we get similar results to the ones seen at the Heston model

(Figures 5 and 6), the only significant difference seems to be the larger deviation of the TVF model volatility around the target (Figure 6, bottom right corner).

BNS Model

Moving on to the BNS model - the first one that Morrison and Tadrowski (2013) did not consider – we see distinctive results. The implied volatility curves seem to be completely flat, however, the level of the equity implied volatility curve slightly increases over time, but still remains under the implied volatility on equity index of the Black-Scholes model. The implied volatility curves of the TVF are a little bit above the target, and this difference is not likely to disappear over time (Figure 7).

Again, the EWMA estimator tends to do a good job (Figure 8, upper right corner) and the TVF model volatility fluctuates around the target (Figure 8, bottom right corner). The rebalancing method seemingly makes the TVF price process less volatile than the equity price process (Figure 8, subfigures on the left), but yet again, as Table 2 suggests we arrive at higher put option prices.

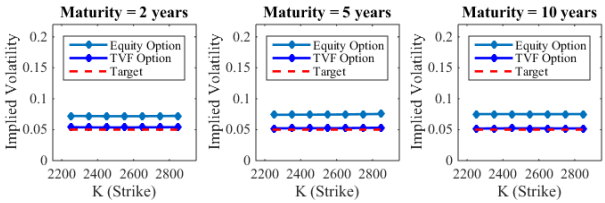


Figure 7: Implied Volatilities for the BNS Model. Calculated Using 10000 Scenarios.

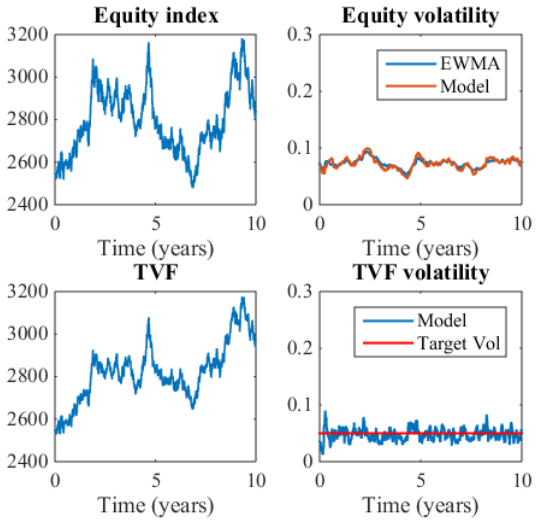


Figure 8: Example Scenario for the BNS Model.

VGGOU Model

Lastly, we investigate the case of the VGGOU model, which is the second one that extends the previous analysis of Morrison and Tadrowski (2013). The implied volatility on the TVF has a skew and is considerably high for all strikes, and even increases over time (Figure 9). This is in connection with the large percentage change in ATM put option prices seen in Table 2.

Although the EWMA estimator seems to work well (Figure 10, upper right corner), the TVF model volatility fluctuates around the target with a much higher volatility than that we have seen in any other case before (Figure 10, bottom right corner). As this figure suggests, the rebalancing method cannot maintain a stable constant volatility for the TVF, but the volatility of the TVF price process seems to be less than the one of the equity price process (Figure 10, subfigures on the left). Furthermore, the smaller target volatility level we chose the lower the level of the implied volatility on TVF will be, however, it will never reach the target level.

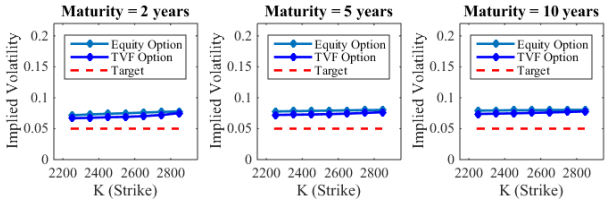


Figure 9: Implied Volatilities for the VGGOU Model. Calculated Using 10000 Scenarios.

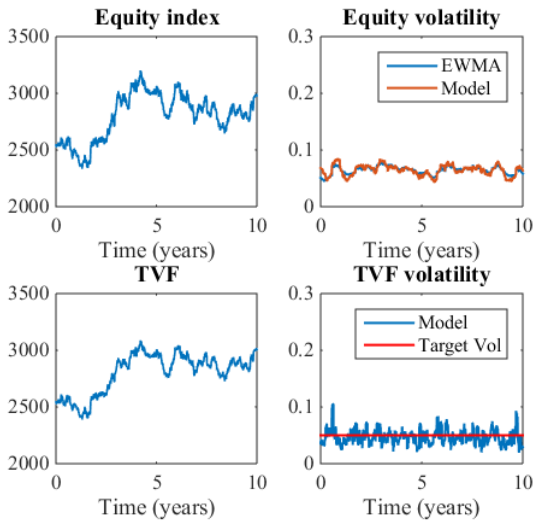


Figure 10: Example Scenario for the VGGOU Model.

CONCLUSIONS

In this paper, I extended the analysis of Morrison and Tadrowski (2013) to the Barndorff-Nielsen–Shephard model and to a Lévy-process with stochastic time, which was the Variance Gamma process with Gamma–Ornstein-Uhlenbeck stochastic clock.

The rebalancing method performs well assuming Black-Scholes model, but probably results in higher guarantee costs if we change to more complex, stochastic volatility models. The case of the VGGOU is the most convincing with extremely dear guarantees relative to the Black-Scholes model. These results are rather interesting,

however, some further examination is needed to investigate the robustness of them. Also, it is necessary to prosecute a more detailed analysis to examine if the results are robust using other, hopefully more volatile underlying products and various target volatility levels. Finally, it is a possibility to inspect the effect of using different volatility estimators like Kim and Enke (2016) did.

REFERENCES

Barndorff-Nielsen, O. E., & Shephard, N. (1999). *Non-Gaussian OU based models and some of their uses in financial economics*. Oxford: Nuffield College.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, *9*(1), 69-107.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, *81*(3), 637-654.

Carr, P., Geman, H., Madan, D. B., & Yor, M. (2003). Stochastic volatility for Lévy processes. *Mathematical Finance*, *13*(3), 345-382.

Chew, L. (2011). Target volatility asset allocation strategy. *Society of Actuaries International News*.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, *6*(2), 327-343.

Hocquard, A., Ng, S., & Papageorgiou, N. (2013). A constant-volatility framework for managing tail risk. *The Journal of Portfolio Management*, 39(2):28-40.

Kilin, F. (2011). Accelerating the calibration of stochastic volatility models. *Journal of Derivatives*, *18*(3), 7.

Kim, Y. és Enke, D. (2016). Using neural networks to forecast volatility for an asset allocation strategy based on the target volatility. Procedia Computer Science, 95:281-286.

Morrison, S. & Tadrowski, L. (2013). Guarantees and target volatility funds. *Moody's Analytics B&H Research series*.

Schoutens, W., Simons, E., & Tistaert, J. (2003). A perfect calibration! Now what?. *The best of Wilmott*, 281.

Stoyanov, S. (2011) Structured Equity Investment Strategies for Long-Term Asian Investors. *EDHEC Risk Institute Publication*.

Torricelli, L. (2017). Assessing target volatility investment strategies using stochastic delayed differential models. SSRN. URL http://dx.doi.org/10.2139/ssrn.2902063.

Xue, Y. (2011). Target volatility: an effiective risk management tool for VA? *Society of Actuaries*.

AUTHOR BIOGRAPHY

Gábor Kondor is a Ph.D. student at the Department of Finance at the Corvinus University of Budapest. His research interests are primarily in volatility derivative pricing and stochastic models. He studied Financial Mathematics and Actuarial Sciences MSc with major in Quantitative Finance at the joint training of Eötvös Loránd University and the Corvinus University of Budapest. His Bachelor study was Applied Mathematics at Eötvös Loránd University as well.

Review of Global Industry Classification

László Nagy,
Department of Finance
Budapest University of Technology and Economics
Magyar tudósok körútja 2, Budapest H-1117, Hungary
E-mail: nagyl@finance.bme.hu

Mihály Ormos
Department of Economics
J. Selye University
Bratislavská cesta 3322, SK-94501 Komárno, Slovakia
E-mail: ormosm@ujs.sk

KEYWORDS
Cluster analysis, market classification, GICS, machine learning

ABSTRACT

This paper introduces the financial market implied industry classification standard. Besides current industry classifications we propose a spectral clustering based quantitative methodology. The main drawback of current standards come from their qualitative classification techniques which can be eliminated in this purely mathematical concept. Calculating the market implied clusters and compering them with global industry classification explores that market implied classification provides better statistical results. However it also turns out that different clustering techniques provide similar classifications, moreover, both methods determine "Real Estate" as a cluster.

INTRODUCTION

Practitioners are using Global Industry Classification Standards (GICS) to keep track of sector based moves. It is widely accepted that firms with similar business activities have similar macro and micro factor sensitivities. Thus, building portfolios with names from the same industry lets us to eliminate specific risk. Sector based bundling gives the opportunity to tailor the macro factor exposure, understand the contamination of macro level shocks and determine regulatory standards. Moreover, it also helps to identify firms which under or outperform the industry hence it helps to separate systematic and non-systematic risks.

Asset managers use industry classification standards for their asset allocation and risk management strategies. Moreover, industry classification is also important for regulators, governments and labour economists to have a deeper understanding of the given state of the economy and be able to implement policies.

GICS is the mainly used classification. The methodology was developed by Morgan Stanley Capital International Inc. (MSCI) and Standard and Poor's (S&P). The categorization combines quantitative and qualitative techniques (MSCI 2015.) to obtain a market oriented economically thoughtful classification standard.

In this article we introduce a spectral clustering based (Nagy and Ormos 2016.) purely quantitative model to unveil the Financial Market Implied Classification (FMIC). Using daily closing prices, we show that the normalized modularity cut (Bolla 2011.) implied clusters and GICS are highly comparable. Moreover, standard Capital Asset Pricing Model based regressions give further evidences for using spectral clustering. The results show that GICS can be looked on as an approximation of the spectral clustering based classification. In addition, the clusters support the latest review of GICS in which Real Estate was added as distinct asset class.

The article structured as follows: Section 2 is a brief overview of industry classification standards. In Section 3 we introduce our spectral clustering based concept. In Section 4 we present the market implied clusters, compare them with groups given by GICS and carry out regression analysis to investigate the risk explanation power of FMIC. Section 5. summarizes the article.

INDUSTRY CLASSIFICATION STANDARDS

Investors are using different classification techniques from long ago to identify industries and characterize their specific behaviour. In 19th century all the market participants had their own classifications which were mainly used to evaluate risk, understand the macro factor sensitivities and build investment strategies (Vose 1916). However, there was no officially accepted framework. Hence, it was unmanageable to aggregate sector level sensitivities and implement economic policies.

Standard Industrial Classification

After the Great Depression in 1937, the US Central Statistical Board (Kolesnikoff 1940.) unified the different industry classifications and established the Standard Industry Classification. SIC was the main

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

guideline of the Federal Government, banks and investors during the 20th century, in addition, the U.S. Securities and Exchange Commission (SEC) still uses it to its industry classification. The system had to be revised several times, in 1958, 1963, 1967, 1972, 1977 and 1987, but it still had several limitations. The main drawbacks of SIC are that it is designed for the economy of early 20th century, it is hard to identify new groups and follow the changes of global economy.

North American Industrial Classification System

NAICS is and industry classification standard which was designed to response to the increasing criticism of SIC (Executive Office of the President Office of Management and Budget. 2017). In 1980s the rapid changes of world economy forced the Office of Management and Budget (OMB) to overhaul SIC. Thus in 1997 the Economic Classification Policy Committee (ECPC) in Nacional the cooperation with Mexico's Instituto de Estadística, Geografía e Informática (now the Instituto Nacional de Estadística y Geografía, INEGI) and Statistics Canada suggested a new industry classification system which supplanted SIC. They agreed in that NAICS should be reviewed in every five years to reflect economic changes. The current form of NAICS defines 20 sectors and 1,057 industries. The classification standard is used for various administrative, regulatory, and taxation purposes.

Global Industry Classification Standards

Besides the SIC and NAICS in 1999 Morgan Stanley Capital International Inc. and Standard and Poor's created the widely used Global Industry Classification Standard. The primary goal of S&P was to enhance its business with introducing sector indices of S&P 500 index. In order to achieve this ambition an adequate industry categorization rule was needed. The classification has a market oriented nature, which incorporates quantitative and qualitative techniques. As GICS is used in the sub-index decomposition of S&P 500 thus it should follow all the changes of the market, hence it has to be reviewed at least annually. At first GICS introduced 10 sectors, 23 industry groups, 59 industries and 123 sub-industries. The classification was revised several times, currently it stands from 11 sectors, 24 industry groups, 68 industries and 157 sub-industries. Each sector, namely Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunication Services and Utilities represents an economically understandable market segment which is the key of the popularity of the classification. GICS is the most famous actively used standard which is assiduously reviewed by the financial market. Because, almost all U. S. market participants benchmark their positions against the performance of GICS sector indices. Moreover, the most liquid products in the world are the Standard & Poor's Depositary Receipts (SPDR) ETFs which tracks the S&P 500 index and sub-indices. Thus, the behaviour of GICS must be in line with the market because like a stock it is continuously reviewed by the market and like earning season at least annually revised by MSCI and S&P.

In this article we propose a purely quantitative technique (FMIC) to analyse the connections between financial market and GICS and study the efficiency of Global Industry Classification Standards.

SPECTRAL CLUSTERING

The original concept of classifying the market, defining sub-groups, distinguishing sectors raises the fundamental question: Who can judge the market?

If we do not want to make any a priori assumptions then we have to look at the data and dismiss other subjective classification guidelines. Mathematically it is possible to represent a datasets as a graph, hence, we can construct an abstract network of stocks. The most straightforward method would be representing stocks with nodes, connection strengths with weights. Thus, we can define the network with $\mathrm{G}(V, W)$ graph where V represents the set of stocks and W contains the connection information. It is widely used that the adjacency matrix W represents the graph, thus all the structure information is embedded in the matrix. Note that if we normalize the sum of each row to one, then we get the transition matrix of the random walk on the graph (Luxburg 2007.);

$$P = D^{-1}W \tag{1}$$

where D represents the diagonal matrix of the sums of rows. Studying stopping times of random walks on graphs sheds some light on the structure of the graph, because, if it takes a long time to reach a subgraph of the graph from a given node then it would mean that the node and the subgraph are well separated. Moreover, the largest eigenvalue of the submatrix of P which belongs to the subgraph controls the distribution of the stopping time;

$$\mathrm{Prob}(\tau \geq n) = 1 - \pi_0 \sum_{i=1}^{n} Q^{i-1} R \cdot 1 \tag{2}$$

where τ represents the stopping time, π_0 the initial distribution, Q the submatrix of transient points, R defines the transition probabilities from transient to recurrent points.

Using symmetricity and spectral theorem

$$\mathrm{Prob}(\tau \geq n) = 1 - \pi_0 V \sum_{i=1}^{n} \Lambda^{i-1} V^T R \cdot 1 \tag{3}$$

Equation 3. shows that the spectrum of adjacency matrix, stopping times and clustering properties are strongly connected.

In addition, the problem can be looked at from an analytical point of view. Fourier analysis is a widely used tool in pattern recognition theory (Shi and Malik 2000). Considering an arbitrary real valued function on the vertexes and defining the below incidence matrix led

some colour to the connections between Laplace operator and graph theory (Chung 1997.);

$$B_{ev} = \begin{cases} 1 & \text{if v is the initial vertex of e} \\ -1 & \text{if v is the terminal vertex of e} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

then $B^T B$ would be exactly the negative discrete Laplace operator because:

$$B^T B f(v_i) = \sum_{v_i \sim v_j} f(v_i) - f(v_j)$$

Notice that if we subtract the adjacency matrix from the diagonal matrix of row-sums then we get the same operator. Thus Laplace matrix can be defined as follows;

$$L = D - W \quad (5)$$

If we think of the adjacency matrix as a noisy matrix and would like to maximize the information content, then we get the modularity matrix;

$$M = W - dd^T \quad (6)$$

where d represents the vector of row-sums. Normalizing the matrix gives as the normalized modularity matrix which helps us to cluster tightly connected networks (Bolla 2011.).

$$M_D = D^{-\frac{1}{2}}(W - dd^T)D^{-1/2} \quad (7)$$

Equation 5. shows that L is the negative discrete Laplace operator hence its eigenvectors would be sine and cosine functions. Analogously to noise filtering techniques we can calculate a Fourier approximation. However, we would like to optimize the normalized modularity cut thus we have to use the normalized modularity matrix. The optimal representation of the original vertices are the rows of that matrix which contains its columns the eigenvectors of in absolute sense largest eigenvalues of the normalized modularity matrix (Bolla 2013.);

$$\left(D^{-\frac{1}{2}}u_1, \dots, D^{-\frac{1}{2}}u_k\right)$$

where $u_1, \dots, u_2$ are the corresponding eigenvectors of $|\lambda_1(M_D)| \geq \cdots \geq |\lambda_k(M_D)| \gg |\lambda_{k+1}(M_D)|$.
To unveil the market implied classification we should calculate the similarity matrix, then determine the normalized modularity, identify the spectral gap in the spectrum of the normalized modularity matrix and finally clustering the optimal representation with k-mean algorithm.
The only hurdle is the calibration of edge weights. Note that a spectral clustering method is effective if and only if in absolute sense decreasing sequence of eigenvalues goes to zero and gap appears in the spectrum. Otherwise, spectral based methods cannot give appropriate classification. Hence, we should test different similarity measures thus we have to calculate the spectrum of different similarity measures implied normalized modularity and Laplace matrices and choose that which provides the best spectral properties. Ormos and Nagy showed tightly connected financial dataset can be analysed with Gaussian square distances.

$$W_{ij} = e^{-||s^i - s^j||^2} \quad (8)$$

Combining Fourier analysis, pattern recognition techniques, random walk and graph theory provides us the optimal cluster property, thus spectral clustering gives us the opportunity to find an approximation of the market implied classification.

FINANCIAL MARKET IMPLIED CLASSIFICATION

Spectral clustering can be used to unveil the hidden market structure of stock indices. The purely quantitative approach needs only closing prices so that the market segments are formed by stock prices. The fundamental assumption behind the model is that at least the weak form of market efficiency holds on a daily scale. This assumption allows intra-day inefficiencies, but accepts that the daily closing auction forces the market into the equilibrium.

Data

The current study identifies FMIC based on stock splits and dividends adjusted daily closing prices between 01/01/2007 and 01/03/2017 of current S&P 500 constituents. The data is provided by Yahoo! Finance.

FMIC and GICS

Calculating the spectrum of normalized Laplace and modularity matrices of Gaussian based similarities shed some light on the structure of the network. Bolla showed that if a graph is dense then the normalized Laplacian matrix cannot be used because the norm of eigenvalues slowly converge to zero. Normalized modularity matrix, however, provides better spectrum properties.

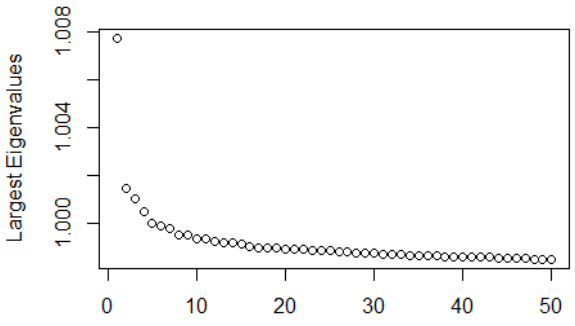


Figure 1: Spectrum of Gaussian based normalized Laplace matrix

Calculating the spectrum of normalized Laplace matrix displays that the eigenvalues converge slowly to zero thus Fourier based approximation techniques cannot be used (Figure 1.). It also suggests that the network structure is not scarce, different clusters should be connected. Hence, we should optimize the information theory based Newman Girvan cut thus we have to calculate the spectrum of normalized modularity matrix.

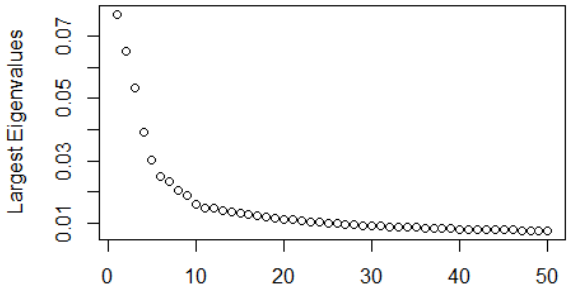


Figure 2: Spectrum of Gaussian based normalized modularity matrix

The spectrum of normalized modularity matrix provides appropriate spectral properties. Figure 2. shows that it has several large eigenvalues and the decreasing sequence of them converges to zero. Nevertheless, the normalized Laplacian cannot be used for clustering. All these implies that the equity index network is dense, most of the stocks are connected.
Identifying spectral gaps highlights that the optimal number of market implied clusters would be 5, 7, 9 or 12.
Note that GICS distinguishes 11 sectors. If we calculate FIMC with 11 clusters then we get controversial results.

Table 1. shows that the 7th cluster dominates the graph, most of the nodes are put into that cluster despite of the cluster size penalty. Moreover, GICS and FIMC 11 classifications are fundamentally different.

Table 1: Frequency table of GICS and FMIC 11

GICS/FMIC 11											
	1	2	3	4	5	6	7	8	9	10	11
D	10	3	10	5	4	8	21	10	7	4	4
S	4	3	0	2	3	2	7	4	3	4	4
E	1	0	4	1	3	2	13	2	5	4	0
F	9	3	2	3	8	5	9	6	9	0	10
H	8	4	0	3	4	1	17	13	6	1	2
I	6	4	5	3	4	4	24	2	8	4	2
IT	8	2	9	4	7	2	15	4	10	1	6
M	3	2	2	5	0	2	6	1	4	0	0
R	4	4	1	3	0	3	8	1	3	2	1
T	0	0	0	0	0	0	2	0	1	2	0
U	2	1	5	1	2	1	10	1	2	1	2

Notes: This table is the contingency table of GICS and FMIC 11; D denotes Consumer Discretionary, C: Consumer Staples, E: Energy, F: Financials, H: Health Care, I: Industrials, IT: Information Technology, M: Materials, R: Real Estate, T: Telecommunications Services and U: Utilities.

However, spectral clustering based methodology proposes to use 12 clusters. If we calculate FMIC 12 then we could see that GICS is in line with market implied classification, see Table 2.

Table 2: Frequency table of GICS and FMIC 12

GICS/FMIC12												
	1	2	3	4	5	6	7	8	9	10	11	12
C	0	8	33	3	0	0	0	0	0	42	0	0
S	0	0	5	1	0	0	0	0	0	4	0	26
E	31	2	2	0	0	0	0	0	0	0	0	0
F	0	5	9	0	0	50	0	0	0	0	0	0
H	0	0	5	0	0	0	0	39	0	0	16	0
I	0	5	55	0	0	0	0	0	6	0	0	0
IT	0	6	15	47	0	0	0	0	0	0	0	0
M	4	1	19	0	0	0	0	0	0	1	0	0
R	0	1	5	1	23	0	0	0	0	0	0	0
T	0	1	4	0	0	0	0	0	0	0	0	0
U	1	0	1	0	0	0	26	0	0	0	0	0

Notes: This table is the contingency table of GICS and FMIC 12

Clusters 1, 4, 5, 6, 7 and 12 cover Energy, Information Technology, Real Estate, Financials, Utilities and Consumer Staples respectively.
The first cluster is dominated by Energy companies. It also contains four Utilities and NRG Energy. These firms are closely related to energy business. Thus, we could say that GICS Energy sector can be quantitatively supported.
Information Technology names are put into Cluster 4. Checking non Information Technology names in Cluster 4 we see that the most of them are highly technological. The three Consumer Discretionary companies are Amazon Inc., Expedia Inc. and Garmin Ltd. that all carry out strongly technology based business. Equinix specialized in data and cloud business thus putting it to a technological cluster is also in line with economic thinking. Monster Beverage could be surprising, however, if we compare it with its peers (Table 3.) we could see that Monster Beverage differs from them.

Table 3: Peer review of Monster Beverage Corporation (Reuters)

Company Name	EV/ Revenue	Price/Book Value Per Share	Price / Revenue	Dividend / Share Yield %
Monster	7.97	7.9	8.19	0.00%
Pepsi	2.95	12.98	2.6	2.70%
Dr Pepper	3.17	8.11	2.52	2.50%
Starbucks	3.85	14.43	3.78	1.70%
Cott	1.18	2.15	0.5	1.80%

Notes: This table contains the peer review of of Monster Beverage Corporation.

Cluster 5 gives mathematical evidence to handle Real Estate as a different sector. Moreover, Financials, Utilities and Consumer Staples are also separated in Cluster 6, 7 and 12 respectively.
Clusters 8, 9, 10 and 11 incorporates Health Care, Industrials and Consumer Discretionary firms. Table 2. shows that these clusters split GICS sectors into two parts. Scrutinizing the results we could see that names are divided along GICS sub industries (Appendix 1).
Cluster 9 is the smallest cluster which encompasses 6 Industrial firms. Studying the sub industry classification we see that 5 companies out of 6 are Airlines and the outlier is Fortive Corporation which has large exposure to aviation business.
In conclusion, we can say that FIMC 12 is in line with GICS sectors and sub-industries, however it gives us a purely quantitative technique to identify clusters.

The spirit of CAPM

Understanding systematic risk is essential part of asset allocation because prices can be scrutinized only within an equilibrium model. Researchers, investors, regulators and banks are seeking models which could distinguish systematic and non-systematic risk.
Following a purely market oriented prospect leads to the Capital Asset Pricing Model (CAPM) which incorporates the systematic risk into the market portfolio. Several empirical studies concluded that CAPM can be used as a benchmark, but has to be made it more precise.
Analysing risk and reward in different frameworks explores different aspects of risk. Note that while standard deviation counts all the moves, β takes into account only that moves which can be explained linearly with the fluctuation of market portfolio. However, investors are sensitive to losses, filtering out therefore gains leads to Expected Downside Risk (Ormos and Timotity 2016.), in addition, various information theory based measures (Ormos and Zibriczky 2014.) can be defined which are strongly connected to log-optimal portfolio theory (Urban and Ormos 2013.).

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta \cdot SRM + \varepsilon \qquad (9)$$

All the linear models explain different aspects of systematic risk (SRM) but they can be compared with regression statistics. Correspondingly, if we add GICS and FMIC 12 to the regressions then we could compare them.

$$\begin{cases} \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 SRM + \beta_2 GICS + \epsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 SRM + \beta_2 FMIC12 + \theta \end{cases} \qquad (10)$$

Table 4. shows that the FMIC 12 outperforms GICS, except when risk is characterized by entropy and semi-variance.

Table 4: Regression statistics of GICS and FMIC 12

	p-value risk	R2	p-value GICS	R2	p-value FMIC 12	R2
Var	0.000	0.110	0.000	0.173	0.000	0.185
Sig	0.000	0.104	0.000	0.169	0.000	0.179
Semi-var	0.151	0.004	0.000	0.111	0.000	0.106
CAPM	0.000	0.187	0.002	0.232	0.002	0.234
EDR	0.000	0.143	0.000	0.210	0.000	0.227
H	0.000	0.026	0.000	0.122	0.000	0.117

Notes: Table 4. contains the regression statistics of different systematic risk models with GICS and FIMC 12.

Notice that semi-variance is not significant and entropy could explain only 2.6% of total variance. Hence, we could say that risk measures with high R^2 values can be used with cluster-variable like FIMC or GICS. Otherwise, there is no need adding cluster-variable to regression, because linear terms are not explained, thus the risk measure captures the same non-linear effect. If we would like to compare the behaviour of entropy (**H**) with FIMC and GICS we have to look the following models:

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 \cdot SRM + \beta_2 \cdot \mathbf{H} + \varepsilon$$

Calculating the regressions we get Table 5.

Table 5: Comparison of risk measures

Entropy	p-value of entropy	R2
Variance	0.164	0.114
Sigma	0.161	0.107
Semi-variances	0.000	0.029
CAPM Beta	0.085	0.191
EDR	0.090	0.148

Notes: This table presents the regression statistics of variance based systematic risk models with entropy.

The results show that entropy works differently, because using it as a cluster-variable adds only little explanatory power. However FIMC and GICS split the date such that the cluster-wise risk is linear in the main risk factor.

Figure 3. highlights that spectral clustering identifies concave clusters. Thus, adding the industry classification to the regressions filters out non-linear effects.

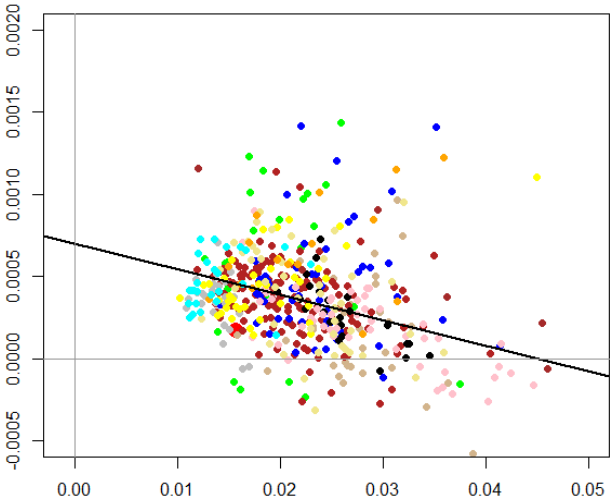


Figure 3: Standard deviation, mean return plot of FMIC 12

Nevertheless, the outcomes imply negative connection between risk and reward which is controversial with economic sense.
Analysing the market implied risk free rates (the interception of the regressions) we see that entropy is the only risk metric whose regression estimation (Table 6.) is in line with the market conventions.

Table 6: Estimated interceptions and p-values

Risk measure	Interception	p-value	Implied 1-year rate
Variance	0.001	0.000	0.207
Sigma	0.001	0.000	0.291
Semi-variances	0.000	0.000	0.148
CAPM Beta	0.001	0.000	0.283
EDR	0.001	0.000	0.341
Entropy	0.000	0.255	0.032

Note: Table 6. summarizes the regression statistics of the constant term.

The source of the problem could be the extreme low interest rate environment. If we calculate regressions without the constant term,

$$\begin{cases} \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i(SRM)^i + \beta_{GICS}GICS + \varepsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i(SRM)^i + \beta_{FMIC\ 12}FMIC\ 12 + \theta \end{cases} \quad (11)$$

we get positive connection between risk and reward.

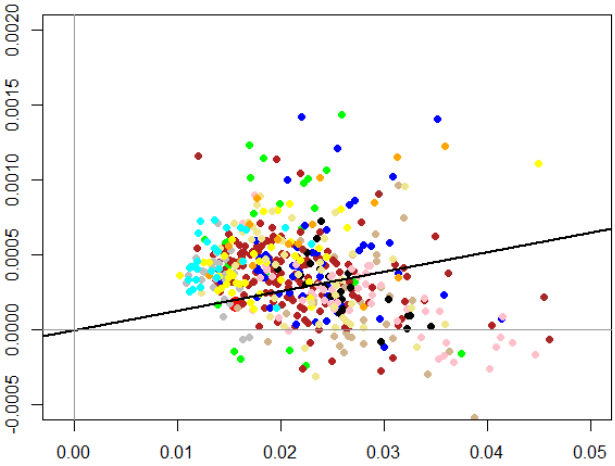


Figure 4: Standard deviation, mean return plot of FMIC 12 with zero interception

Setting the constant coefficient zero and calculating the regressions gives Table 7. which is in line with the 2007-2017 market conditions. Moreover, comparing the position of S&P 500 in Figure 3. and 4. gives further evidence to use regressions without constant, because, the benchmark portfolio (red dote) is on the regression line in Figure 4, meanwhile in Figure 3. is far away from the model expectations.

Table 7: Regression statistics of GICS and FMIC 12 with zero interception

	p-value risk	R2	p-value GICS	R2	p-value FMIC 12	R2
Var	0.000	0.203	0.000	0.595	0.000	0.601
Sigm	0.000	0.374	0.000	0.593	0.000	0.598
Semi-var	0.000	0.253	0.000	0.565	0.000	0.562
CAPM	0.000	0.277	0.000	0.624	0.000	0.625
EDR	0.000	0.375	0.000	0.613	0.000	0.622
H	0.000	0.522	0.000	0.571	0.000	0.568

Notes: This table contains the regression statistics of Equation 12.

It also can be seen that different linear based risk factors explain 20%-52% of the total variance, but adding cluster-variables the explanatory power of the model jumps to 60%. This means that cluster specific and linear risks explains 60% of the fluctuation.

If we generalize the baseline linear model we could specify the following regression;

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i \cdot (SRM)^i + \varepsilon \qquad (12)$$

Risk / R2 of n-th order	1st	2nd	3rd	4th	5th	10th
Var	0.203	0.365	0.495	0.537	0.567	0.582
Sig	0.374	0.547	0.568	0.568	0.569	0.581
Semi-var	0.253	0.348	0.411	0.454	0.477	0.518
CAPM	0.277	0.559	0.574	0.575	0.591	0.626
EDR	0.375	0.587	0.593	0.594	0.595	0.598
H	0.522	0.524	0.526	0.527	0.532	0.535

Expanding the linear model with higher order terms could shed more light on non-linear dependencies, see Table 8.

Table 8: Estimated R2 statistics of polynomial

Notes: Table 8. highlights the non-linear connections between returns and systematic risk factors.

The results are in line with our expectations, because, adding higher order terms of linearly inspired risk metrics to the regression increases the explanatory power, while entropy shows different behavior.
Generalizing Equation (11) we could get the following models;

$$\begin{cases}\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \beta_1 SRM + \beta_2 GICS + \varepsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \beta_1 SRM + \beta_2 FMIC\ 12 + \theta\end{cases} \qquad (13)$$

Equation (13) lets us to distinguish cluster and non-cluster specific higher order connections. The results (Table 9.) show that polynomial terms explain similar effects like GICS and FMIC 12.

Table 9: Estimated R2 statistics of polynomial regressions with FMIC 12

Risk/ R2 of n-th order model with FMIC 12						
	1st	2nd	3rd	4th	5th	10th
Var	0.601	0.602	0.605	0.606	0.609	0.616
Sig	0.598	0.601	0.604	0.605	0.608	0.614
Semi-var	0.562	0.563	0.564	0.565	0.566	0.569
CAPM	0.625	0.636	0.637	0.640	0.640	0.652
EDR	0.622	0.634	0.634	0.634	0.635	0.636
H	0.568	0.569	0.569	0.573	0.573	0.577

Notes: This table summarizes the non-linear filter behaviour of the cluster variable (FMIC 12).

Tables 8, 9. and 10. show that higher order terms do not increase the explanatory power of the regressions. Thus, the remaining 30% part of the variance cannot be explained by liner-based risk factors and clusters.

Table 10: Estimated R2 statistics of polynomial regressions with GICS

Risk measures/ R2 of n-th order model with GICS						
	1st	2nd	3rd	4th	5th	10th
Var	0.595	0.596	0.597	0.598	0.600	0.607
Sig	0.593	0.595	0.597	0.597	0.599	0.606
Semi-var	0.565	0.566	0.566	0.566	0.567	0.570
CAPM	0.624	0.631	0.632	0.634	0.634	0.646
EDR	0.613	0.619	0.620	0.621	0.622	0.624
H	0.571	0.571	0.572	0.574	0.574	0.578

Notes: Table 10. describes the non-linear filter behaviour of GICS.

Analysing returns with linear and polynomial regressions show that FMIC 12 outperforms GICS, however, cluster variables explain similar non-linear connections.

Optimal number of clusters

Analysing the structure of FIMC 12, GICS industries and subindustries the natural question arise; how many clusters do we need?
In spectral clustering and signal analysis there are different analytical and simulation based techniques. The most widely used methodology is spectral gap analysis which is a Fourier based technique. The goal is to approximate the norm of the object with as few elements as possible.
Calculating the spectrum of normalized modularity matrix (Figure 2) shows that the GICS and FIMC 12 could be too elaborated and 5 clusters would be enough to explain non-linear cluster specific effects.
Comparing FMIC 12 with FMIC 5 it shows that FIMC 12 is a more detailed subdivision of the stock market graph. Because, FIMC 5 bundles the FIMC 12 clusters into bigger groups (Table 11.).
However, Real Estate remains a single cluster and Consumer Staples with Utilities are put into FIMC 5 Cluster 3. Other FIMC 5 clusters are also dominated with FIMC 12 clusters.

Table 11: Frequency table of FMIC 12 and FMIC 5

FMIC 12/FMIC 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	36	0	0	0
Cluster 2	4	13	0	0	12
Cluster 3	141	7	2	0	3
Cluster 4	46	5	0	0	1
Cluster 5	0	0	0	23	0

Cluster 6	3	0	0	0	47
Cluster 7	0	0	26	0	0
Cluster 8	5	33	1	0	0
Cluster 9	6	0	0	0	0
Cluster 10	43	4	0	0	0
Cluster 11	0	16	0	0	0
Cluster 12	3	7	16	0	0

Notes: This table is the contingency table of FIMC 12 and FMIC 5.

Analysing FMIC 5 we saw that it is closely related to GICS and FMIC 12. Regression statistics (Table 12.) are in line with Table 7, relative entropy and semi-variance show different behaviour and CAPM has the strongest explanatory power.

Table 12: Regression statistics of FMIC 5 with zero interception

Risk	p-value risk	R2	p-value GICS	R2	p-value FMIC 5	R2
Var	0.000	0.203	0.000	0.595	0.000	0.578
Sig	0.000	0.374	0.000	0.593	0.000	0.579
Semi-var	0.000	0.253	0.000	0.565	0.000	0.536
CAPM	0.000	0.277	0.000	0.624	0.000	0.614
EDR	0.000	0.375	0.000	0.613	0.000	0.604
H	0.000	0.522	0.000	0.571	0.000	0.543

Notes: Table 12 sheds some light on the optimal, lower dimensional market classification.

Regression and frequency statistics support the theoretically proposed five cluster model which incorporates roughly the same information like FIMC 12 and GICS.

Conclusion

Spectral clustering is an adequate technique to unveil the embedded market structure, filter out non-linear effects and make CAPM more precise. The purely quantitative method gives us the opportunity to categorize firms based on their stock market returns, in addition, it lends some colour to Global Industry Classification Standards.

ACKNOWLEDGEMENTS

Mihály Ormos acknowledges the support of the János Bolyai Research Scholarship of the Hungarian Academy.

REFERENCES

A. Urban and M. Ormos. 2013 “Performance analysis of log-optimal portfolio strategies with transaction costs” *Quantitative Finance*. 13: (10) pp. 1587-1597.

D. Timotity. And M. Ormos 2016. “Generalized asset pricing: Expected Downside Risk-based equilibrium modeling” *Economic Modelling* 52:pp. 967-980.

D. Zibriczky. and M. Ormos. 2014. “Entropy-Based Financial Asset Pricing” *PLoS* ONE 9(12): e115742.

Fan R. K. Chung. 1997. “Spectral Graph Theory” *American Mathematical Society*

J. Shi and J. Malik. 2000. “Normalized cuts and image segmentation” *Pattern Analysis and Machine Intelligence* IEEE, Transactions, N.J., 888-905.

L. Nagy and M. Ormos. 2016. “Friendship of Stock Indices” *30th Conference on Modelling and Simulation*

M. Bolla. 2011. “Penalized version of Newman-Girvan modularity and their relation to normalized cuts and k-means clustering” *Physical review E*, Vol. 84, 016108

M. Bolla. 2013. “Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables” *Wiley*

MSCI. 2015. “S&P Dow Jones Indices and MSCI announce Further Revision to the Global Industry Classification Standards (GICS®) Structure in 2016” *Working Paper*.

Executive Office of the President Office of Management and Budget. 2017. “North American Industry Classification System” *Working Paper*.

E. N, Vose. 1916 “Seventy-five years of the Mercantile agency, R.G. Dun & co., 1841-1916” *Brooklyn, N.Y., Priv. Print. at the Printing House of R.G. Dun & Co.*

U. von Luxburg. 2007. “Tutorial on Spectral Clustering” *Statistics and Computing* Vol. 17, 395-416.

V. S. Kolesnikoff. 1940. “Standard Classification of Industries in the United States” *Journal of the American Statistical Association*, Vol. 35, No. 209, pp. 65-73

AUTHOR BIOGRAPHIES

László Nagy is a PhD student at the Department of Finance, Institute of Business at the School of Economic and Social Sciences, Budapest University of Technology and Economics. His main area of research is financial risk measures and asset pricing

Mihály Ormos is a Professor of Finance at Eötvös Loránd University and at J. Selye University. His area of research is financial economics especially asset pricing, risk measures, risk perception and behavioral finance. He serves as one of the contributing editors at Eastern European Economics published by Taylor and Francis. His teaching activities concentrate on financial economics, investments and accounting.

SUPPLEMENTATION OF THE REGULATION OF ANTI-PROCYCLICAL MARGIN MEASURES

Csilla Szanyi
KELER CCP
Rákóczi street 70-72. Budapest, 1074, Hungary
E-mail:
szanyi.csilla@kelerkszf.hu

Melinda Szodorai
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
KELER CCP
Rákóczi street 70-72. Budapest, 1074, Hungary
E-mail:
szodorai.melinda@kelerkszf.hu

Kata Váradi
Department of Finance
Corvinus University of Budapest
Fővám square 8. Budapest, 1093, Hungary
E-mail: kata.varadi@uni-corvinus.hu

KEYWORDS

Margin, central counterparty, EMIR, procyclicality

ABSTRACT

Our paper focuses on the procyclicality of margin requirements of central counterparties (CCP). The role of the central counterparties on the market is to take over the counterparty risk during the trading on stock exchanges. CCPs use a multilevel guarantee system to manage this risk, and the margin is one level of this system. The regulators have recognized, that the margin has procyclical effect on the market – in case of a stress event the margin has to be increased as a consequence of the increased volatility of the market –, so they have to take action in order to avoid this procyclicality as much as possible, not to punish the market in a stressed market situation. In this paper we will introduce those anti-procyclicality methods that were offered by the regulators in the so called EMIR (European Market Infrastructure Regulation) regulation, and point out that how the regulation should be extended in order to apply the anti-procyclicality methods properly and efficiently by the CCPs.

THEORETICAL BACKGROUND

Since 2008 the role and the systematic importance of the central counterparties gradually became essential on the financial markets. Bilateral relationships that prevailed between two counterparties are now by being settled through central counterparties, assuring the markets to be more secured. This guarantees the trade's fulfilment in case one of the traders default. In order to do so, CCPs must have a waterfall system of guarantees, in which margin has a notable weight.
Risk models are used by central counterparties to estimate the margin requirements of portfolios of financial instruments as one part of the guarantee system they have to use. Supervisors have recognized that in order to assure the smooth working of the financial markets, CCPs shall deal with procyclicality as an additional concern besides risk models. Generally speaking, procyclicality is defined as the tendency of any financial variable to move with the economic cycles. This is an undesirable property when the variable acts to intensify financial stress (Financial Stability Forum (2009)). The main negative result of highly procyclicality movement is the difficulty of funding tied with market liquidity risk (Brunnermeier and Pedersen (2009), Heller and Vause (2012)).
Authorities designated under Article 22 of EMIR that supervise CCPs authorized under Article 14 of the EMIR are applying the margining requirements to limit procyclicality pursuant to Article 41 of EMIR and Article 28 of the RTS (Regulatory Technical Standard). Currently the applicable articles for CCPs' referring to procyclicality seemed to be vague consequently its application lied on several presumptions.
Murphy et. al. (2016) follows the mitigation tools for procyclicality as per EMIR standards. Their findings indicate that all of the five tools – three models are based on the EMIR requirements, and two models are being built by themselves – are useful in mitigating procyclicality to some extent, but that the optimal calibration of each tool in a particular situation depends on the relative weights placed by the modeller. Glasserman and Wu (2017) examine the extent of margin buffer needed to offset procyclicality, their findings pointing to the important features of price time series that should inform 'anti-procyclicality' measures but are missing from current rules. Duffie et al. (2015) and Heller and Vause (2012) address the issue of margin requirements following the new regulations. Berlinger et al. (2017) suggests well-chosen margin strategies, concluding that in most cases, it lies inside the set of feasible strategies and represents a delicate compromise between different forces. Their main result is that the anti-cyclical margin is not only the interest of the regulator, but the CCPs as well, since it is decreasing the risks they have to face. They have pointed out, that there exists an optimal margin level,

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

and this margin level is in most of the cases not a risk-sensitive margin.
In order to ensure common, uniform and consistent application of the EMIR provisions in the context of limiting procyclicality of margins without under- or overestimate procyclical margin, placing great burden on market participants, the high level subjectivity and presumptions shall be diminished from the legal content.

MARGIN CALCULATION METHODS

Our paper focuses on highlighting the discrepancies of the legal background on a theoretical and empirical level. Our main goal in this paper is to show how the regulation should be extended in order to reach their goal to build an anti-cyclical margin by the CCPs. In other words the margin requirement, the regulators would like to be achieved by the CCPs, is that the margin should be prudent, stable and reproducible by market participants. We will show that without specifying the method in more details, and applying the EMIR and the RTS regulation without any further assumptions, it can lead to a procyclical margin, or to an unreasonably high margin requirement.
We will analyse in the following three subchapters the three possible anti-cyclicality method for margin calculation, from which a CCP can freely choose, based on the regulation. The only thing we will focus on is the handling of anti-cyclicality, so we will not analyse the effect of any other margin parameter – e.g. the length of the lookback period, or other buffers than procyclicality.
The following assumptions will be unified for all of the three models:

- Risk measure will be the Value-at-Risk (VaR) model, which gives the answer to the following question: what is the possible maximum loss on a given interval, on a given significance level on a portfolio (Jorion, 2007). The model will be calculated was the historical method in two cases, and twice with the delta normal method. The historical method means, that the value of the VaR will be based on the historical prices of the financial asset, while in case of the delta normal method, the VaR will be calculated based on the assumptions, that the logreturn of the financial asset is normally distributed (in more details: Jorion, 2007).
- Significance level: 99%.
- Liquidation period: 2 days.
- Product, the model is calculated to: OTP stock, one of the most liquid central European stock.
- Lookback period: 250 days.
- Only risk factor in the model is the change of the price of OTP stock.
- Calculations are based on logreturns.
- The regulator's general requirement regarding procyclicality: EMIR Article 41.: '…A CCP shall regularly monitor and, if necessary, revise the level of its margins to reflect current market conditions taking into account any potentially procyclical effects of such revisions…' and RTS 28.1: 'A CCP shall ensure that its policy for selecting and revising the confidence interval, the liquidation period and the lookback period deliver forward looking, *stable and prudent* margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected. This shall include *avoiding* when possible disruptive or *big step changes in margin requirements* and establishing *transparent and predictable procedures* for adjusting margin requirements in *response to changing market conditions*…' and finally RTS 28.2: 'When a CCP revises the parameters of the margin model in order to better reflect current market conditions, it shall take into account any potential procyclical effects of such revision.'

The three anti-procyclical methods the regulator offers are being analysed in the following chapters.

Application of a 25% procyclicality buffer

According to EMIR (648/2012) and the RTS's (153/2013) 28.1a) paragraph, the first option for handling procyclicality is to assign a 25% procyclicality buffer as an extra value added to the margin:

a) 'Applying a margin buffer at least equal to 25% of the calculated margins which it allows to be temporarily exhausted in periods, where calculated margin requirements are rising significantly (RTS, Article 28.1a, 2013)'.

This method requires the exhaustion of a 25% margin buffer in case the margin would increase notably. According to Murphy et al. (2016) if the volatility is increasing on the market, the risk is higher, so the margin is higher as well, so this should be the time to exhaust the buffer. In our viewpoint this is true, if we measure the margin on the level of the logreturn, but as a final result we measure the margin on the level of prices. Usually in case of a stress event, the volatility is increasing, but the prices are decreasing, so as a result the VaR will decrease, consequently the margins, too. Therefore, if we want to stabilize the margin – to not to be procyclical – it is not sufficient to exhaust the buffer in case of stress – because the volatility increases –, since it not necessarily means that the margin will increase as well. In Figure 1 and 2 we show this phenomenon. On Figure 1 the VaR can be seen, calculated for logreturns. It shows that between the price and the VaR the correlation is negative. Especially during stressed periods, namely the crisis of 2008, it can be seen, that the VaR has increased notably. But looking at Figure 2 when the prices were falling in 2008, the VaR was decreasing, too. A very notable positive correlation can be seen in the evolution of the price and VaR calculated for prices.

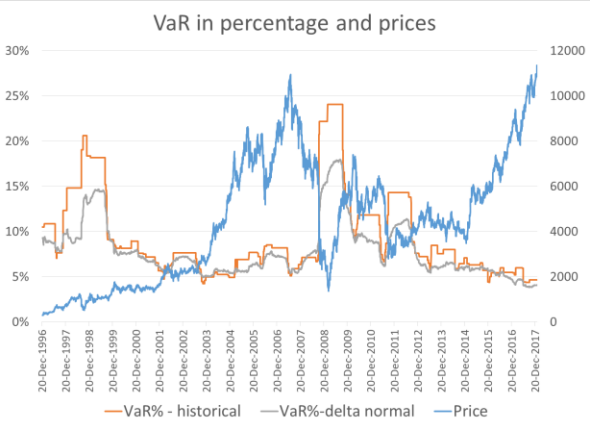


Figure 1: OTP stock's historical and delta normal VaR calculated for logreturns in percentage

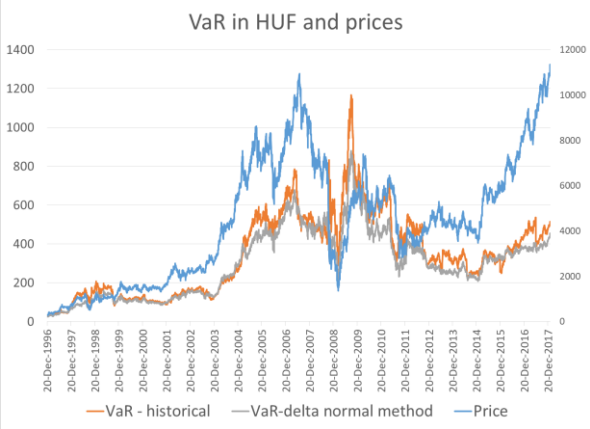


Figure 2: OTP stock's historical and delta normal VaR calculated for prices in HUF

Murphy et al. (2016) defines the time of buffer exhaustion, when there is stress on the market, since the volatility increases in that case. This is the common approach for defining the beginning of the exhaustion, but this is not the proper process, since the regulation does not say that in case of stress the buffer can be exhausted, but in case the margin would increase significantly. This can easily happen – if we measure the margin on HUF basis – when there is a boom on the market. This goes against the will of the regulators from our opinion.

For applying the 28.1 a) approach for margin calculation, we need to use some assumptions, since the regulation says nothing about the following:

- How much is the significant rise in margin requirement, when the buffer can be exhausted?
- How should the buffer be exhausted, and how it should be built back?

We will use the method of Béli and Váradi (2017), namely to exhaust the buffer gradually in the case the EWMA weighted standard deviation – EWMA standard deviation estimation is mainly used in margin models by participants in the OTC derivatives markets (Murphy et al. (2016)) – is higher, than the equally weighted standard deviation and start to build it back when it is the other way around. Moreover, the VaR is defined by a delta normal method, where the standard deviation parameter is either the equally weighted or the EWMA weighted standard deviation, whichever is the lower. We will not explain the method in more details, further we will just apply their method regarding the handling of the procyclicality buffer. Other parts of the model will not be applied. The results can be seen in Figure 3.

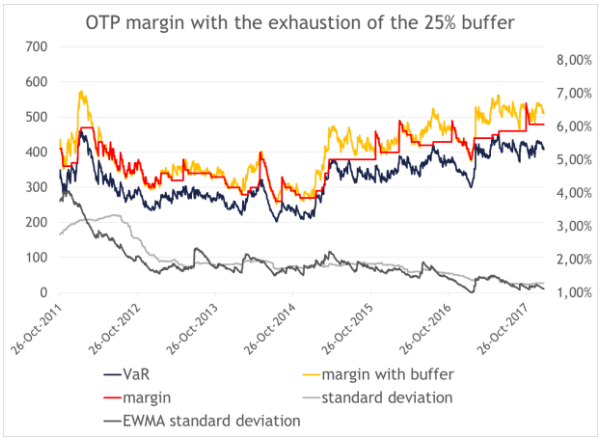


Figure 3: OTP stock's margin with the exhaustion of the 25% procyclicality buffer

According to Figure 3, the margin is being stabilized, based on the value of the changing volatility, and when the risk is increasing – quantified by the method of Béli and Váradi (2017) – on the market, the buffer is exhausted, and when the markets are getting calm, it is being built back. So it fulfils the requirement of the regulator. If we compare the margin to the price evolution of the stock, we can see that they are strongly moving together.

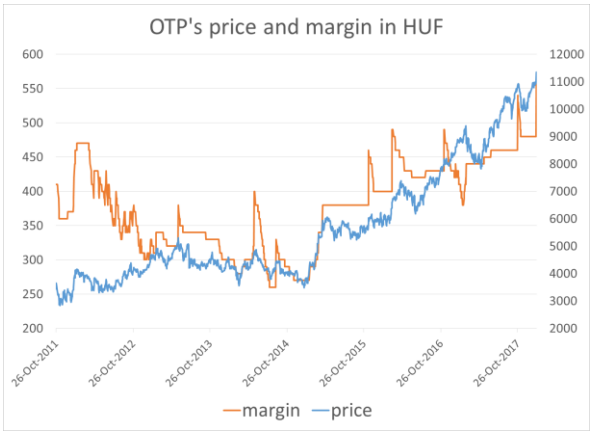


Figure 4: OTP stock's margin and price

Application of a stressed observation period

According to EMIR (648/2012) and RTS's (153/2013) 28.1b) paragraph the second option for handling

procyclicality is to assign 25% weight to stressed observations in the lookback period used to calculate margins.

b) 'Assigning at least 25 % weight to stressed observations in the lookback period calculated in accordance with Article 26 (RTS, Article 28.1b, 2013).'

However the mentioned article and regulation does not detail the definition of stress and does not provide any additional input on how to determine stress periods. ESMA issued a draft guideline[1] for providing an approach to implement the regulation in a harmonized way.

According to ESMA draft guidelines a CCP shall apply consistent approach for identifying a stress period and since every CCP is required to define stress scenarios to size its default fund(s) – which is also a notable part of the multilevel guarantee system, but introducing it goes beyond the topic of this paper – the CCP shall implement the observations defined in its stress testing methodology to its margining methodology.

According to ESMA if a CCP uses short lookback period (1 year is the regulatory minimum), the stressed observations may be limited and it can cause the instability of margin requirements. The inclusion of historical and hypothetical stress scenarios to the margining methodology would provide a consistent approach and the definition of stress would be in line with the CCP's own stress testing methodology. ESMA does not require calibrating the margins to stress levels, the CCPs shall assign 25% weight to its stress observations which are identified within its stress testing framework. ESMA guideline provides additional help to comply with the regulations but can it be implemented in a harmonized way? Does it really help the CCP's clearing members and clients to forecast their own margin requirement in periods of stress?

Implementation of the guideline is not obvious, there are still open questions. CCPs have several stress scenarios across various cleared products. For example there is a CCP that clears commodities (e.g. agricultural commodities like wheat, soy beans, etc) and also clears equity futures and options. While the CCP identifies stress periods for given product groups e.g.: the prices of the equity futures decrease because of a crisis, in the same time the CCP has to consider the correlation between prices and market movements, meaning that the prices of the agricultural commodities – in that given point of the crisis – are not decreasing, they show a slight increase, because the investors choose other investment forms to make their portfolios diversified. In this case the CCP calibrates a stress scenario based on the negative price changes on the equity futures market but this scenario does not describe stress period for agricultural commodities. The CCP has to calibrate another stress scenario for agricultural products based on the price history. This means that the CCP has several stress scenarios and the CCP has to choose the one which is the basis for calculating the margin requirement. The several numbers of stress scenarios for different cleared product groups does not help the clearing members and the clients understanding about the margining methodology of the CCP, however the method provides a consistent approach within the CCP's models.

In line with the guidelines the CCP has elaborated a model to calibrate the initial margin parameter considering the already applied stress scenarios. Because of the complexity of the CCP's stress testing framework there are several scenarios including the given product groups stress periods. For example in case of equity futures there could be at least two or three scenarios (e.g.'98 Russian crisis, '08 subprime crisis, and some hypothetical scenarios based on the reverse price changes of the historical scenarios) that contains stress period, which one shall be addressed as the one scenario to calibrate the margin model? The latest one or the one with bigger negative price shock? Or maybe the hypothetical one which assumes economic growth and positive price changes? For the sake of conservative approach (which is in line with ESMA's and other Supervisors' intentions) we chose the scenarios addressing the biggest negative and positive price changes.

In the following we are modelling the above mentioned approach to calculate the initial margin requirement for OTP considering the biggest negative price change based on the 2008 crisis. In the model historical VaR was applied. According to this method we have calculated the historical VaR value, which got a weight of 75%, while a stress parameter, based on the 2008 crisis (-22.09%) was considered with the weight of 25%. The results can be seen in Figure 4.

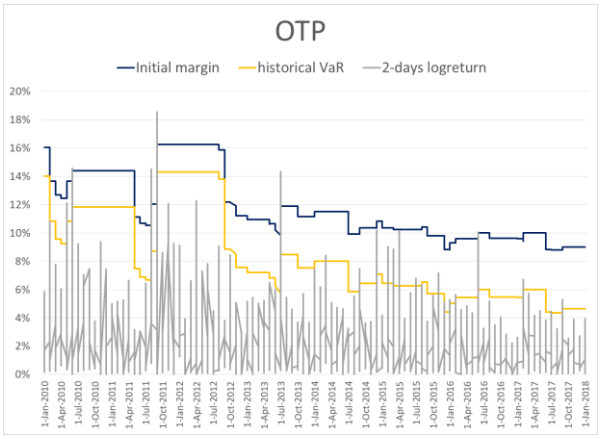


Figure 4: OTP stock's historical VaR and initial margin in percentage

We had 2011 observations in the model and the initial margin was not sufficient in 7 cases, the 2 days log-

[1] https://www.esma.europa.eu/press-news/esma-news/esma-consults-ccp-anti-procyclicality-margin-measures

returns were higher in an absolute value than the calculated initial margin, the model was proper in 99.65% of the cases. If the 2008 stress period had not been included in the model with 25% weight (only historical VaR calculation with 100% weight) then the appropriateness of the model would have been 98.11%. The inclusion of the stress scenario improved the goodness of the model, but analyzing the curves it can be seen that the sudden changes in the log returns result sudden changes in the value of the initial margin, meaning that the additional protection against procyclicality were not proved efficient to provide stable margin requirement. On the other hand the model reacts fast after a significant change in the log-returns but the level of the initial margin remains quite high for a long period of time which puts extra burden on the shoulders of the market participants. ESMA guidelines do not detail the possibility to exhaust the buffer in any case.
Moreover, if we calculate the initial margin in Hungarian Forint terms in Figure 5, we can see, that the anti-procyclicality requirement is not fulfilled, as the initial margin moves together with the price of the stock. The correlation that is calculated to the log-change of the two time series is strong, 0.9192.

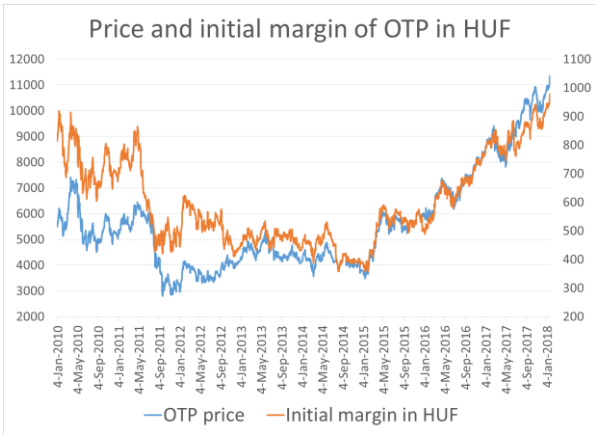


Figure 5: OTP stock's price and initial margin in HUF

Results show that this method is not efficient enough to handle procyclicality, and also in stabilizing the initial margin requirement. Our viewpoints in addition is that taking into account stress in the initial margin calculation is not necessary, since that is the goal of the default fund, which is calculated based on the result of the stress test. There is no need to take stress events twice into account in the multi-level guarantee system.

Application of a margin floor

According to EMIR (648/2012) and RTS's (153/2013) 28.1c) paragraph the third option for handling procyclicality is to define a margin floor, which gives the minimum value of the initial margin:

c) 'Ensuring that its margin requirements are not lower than those that would be calculated using volatility estimated over a 10 year historical lookback period (RTS, Article 28.1c, 2013).'

Based on this requirement we have calculated the initial margin once with the historical method, and once with the delta normal method. We have carried out the calculation exactly as it is written in the regulation, namely that we have calculated with the volatility of the last 10 years, and it gave the minimum value for the margin. The results can be seen in Figure 6 for the historical method and Figure 7 for the delta normal method.

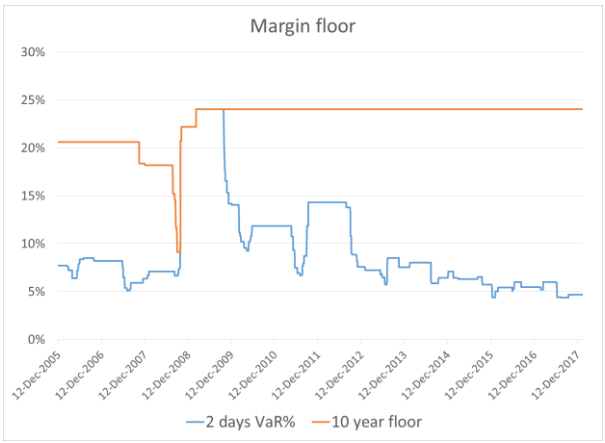


Figure 6: Margin floor based on the previous 10 years with historical method

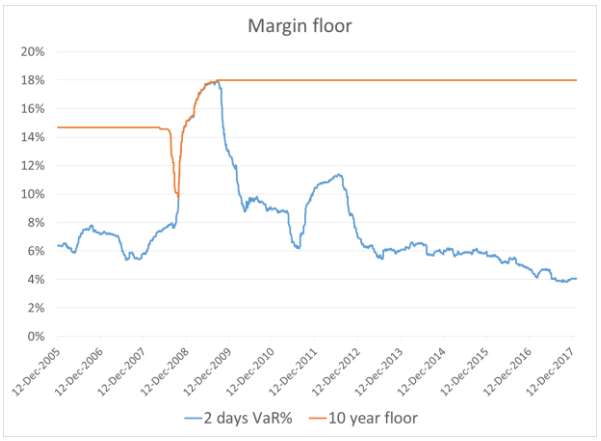


Figure 7: Margin floor based on the previous 10 years with delta normal method

As it can be seen in the figures the margin is high, and too stable, it doesn't follow the market trends. Although it avoids procyclicality, but results an unreasonably high margin requirement.
Calculating the margin in HUF terms can be seen in Figure 8 with both of the methods.

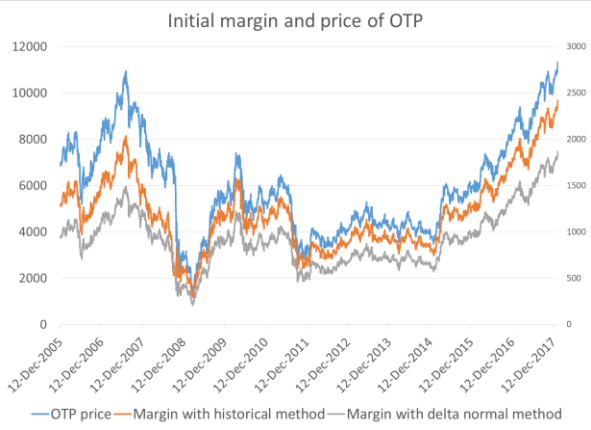


Figure 8: OTP stock's price and initial margin in HUF

We can see the same patterns as in the case of 28.1b) method, so in HUF terms the initial margin is following the market cycles, although in percentage terms it was 'too' stable. Based on Murphy et al. (2016) we run the calculations again, not to have this high margin floor. The approach is to take a certain percentile of the last 10 years' data. We have chosen the 20% according to Murphy et al (2016). The results changed notably as it can be seen in Figure 9, 10 and 11. In Figure 9 and 10 the minimum value of the margin is given by the 10 year floor, and in case when VaR with the 250 days lookback period is higher, then the higher value would be the initial margin. It can be concluded, that we have to take a certain percentile of the last 10 years data, otherwise always the floor would be the effective margin value.

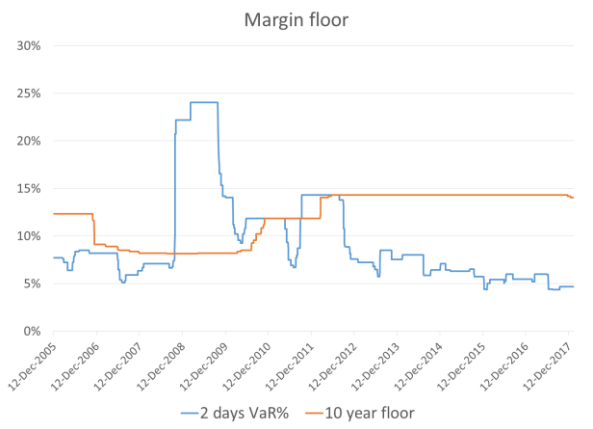


Figure 9: Margin floor based on the previous 10 years with historical method at 20% percentile level

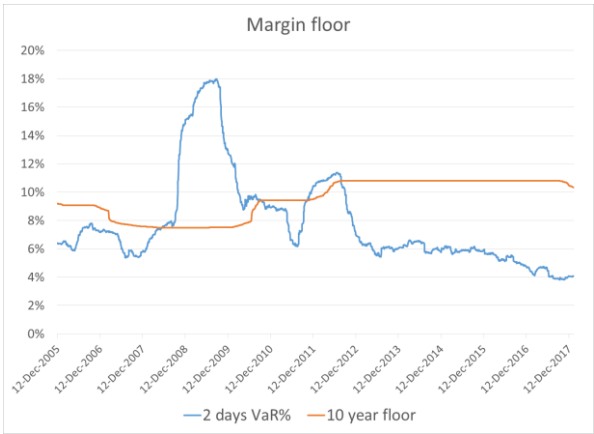


Figure 10: Margin floor based on the previous 10 years with delta normal method at 20% percentile level

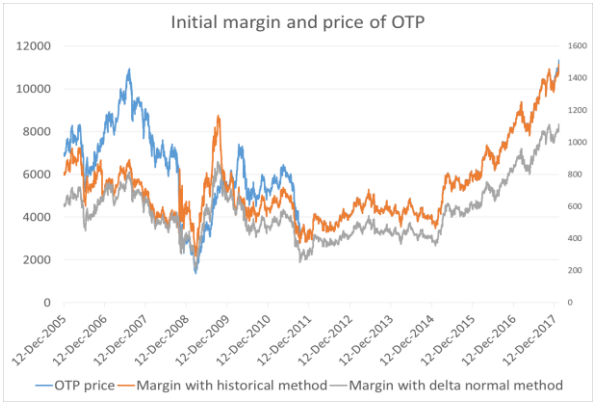


Figure 11: OTP stock's price and initial margin in HUF at 20% percentile

SUGGESTIONS FOR SUPPLEMENTATION

Based on the three models introduced in the previous subchapters, we have the following suggestions for supplementing, or for clarifying the regulations:

28.1a):

- Method of exhausting the margin buffer should be defined: in one step, gradually, etc.
- The main point in exhausting the buffer should be to stabilize the margin, not to decrease it. It is important especially in case of a stress event, for two reasons: 1) it threatens the financial stability of a CCP, 2) in case of stress if a CCP decreases margin, it would cause an increase in the value of the default fund, especially in case of stress. This is not necessarily in the interest of the market participants, since margin can be used by the CCP in case of the market participant's own default, while the default fund contribution can be 'taken away' by the CCP in case of the other clearing members' default.
- Building back the buffer should be explained.
- The regulation should state that the buffer can be exhausted when the risk is increasing – when the

volatility is increasing – not when the margin would increase notable.

- Stabilization should be carried out on margin level, not on the risk measure's – in our case on VaR – level.
- Not the margin increase should be in the focus of the handling of procyclicality buffer, but the stability, the change in the margin value. So not only the notable increase should be prevented, but the decrease as well.

28.1 b):

- Stress definition is also missing.
- Margin should cover losses for market risk in normal market conditions. Losses in case of stress events should be covered by the default funds.

28.1c):

- Definition of market floor should be in the regulation.
- A certain percentile should be applied. The regulator should consider a percentile to be applied uniformly on the markets.
- Further tools shall be introduced to avoid over and under margining due to the less flexible nature of the method.

REFERENCES

Berlinger, E., Dömötör, B. and Illés, F. (2017): *Anti-cyclical versus Risk-sensitive Margin Strategies in Central Clearing,* Working Paper. Corvinus University of Budapest Faculty of Economics, Budapest.

Béli, M. and Váradi, K. (2016): Alapletét meghatározásának lehetséges módszertana [A possible methodology for determining initial margin] *Financial and Economic Review*, Volume 16. issue 2, pp. 119-147.

Brunnermeier, M., Pedersen, L. (2009): Market liquidity and funding liquidity, *Review of Financial Studies*, Volume 22 No. 6

Duffie, D., Scheicher, M., Vuillemey, G. (2015): Central Clearing and Collateral Demand. *Journal of Financial Economics*, Volume 116, No. 2, pp. 237-256.

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 April 2016.

ESMA – European Securities and Markets Authority, Consultation Paper on Anti-Procyclicality Margin Measures. Available: https://www.esma.europa.eu/press-news/esma-news/esma-consults-ccp-anti-procyclicality-margin-measures downloaded: 9th January 2018

Financial Stability Forum (2009): *Report of the financial stability forum on addressing procyclicality in the financial system*, April.

Glasserman, P. and Wu, Q., 2017. Persistence and Procyclicality in Margin Requirements. *Columbia Business School Research Paper No. 17-34*

Heller, D. and Vause, N. (2012): *Collateral requirements for mandatory central clearing of over-the-counter derivatives*, BIS Working Paper No. 373

Jorion, P. (2007): *Value at risk: the new benchmark for managing financial risk.* Vol. 3. New York: McGraw-Hill

Murphy, D., Vasios, M. and Vause, N., (2016): *A comparative analysis of tools to limit the procyclicality of initial margin requirements.* Bank of England Working Paper No. 597.

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://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:052:0041:0074:EN:PDF downloaded: 8th April 2016.

AUTHOR BIOGRAPHIES

CSILLA SZANYI is a senior risk controller at KELER CCP. She majored in finance at Corvinus University of Budapest, at the Department of Finance in 2010. Her main responsibilities at KELER CCP are market risk management on the cleared capital and energy markets, and the evolvement of the risk management framework considering the compliance with EU and Hungarian regulations.

MELINDA SZODORAI is a risk analyst at KELER CCP. Her main responsibilities are operational risk management and regulatory reporting. She majored in finance and management at Babes-Bolyai University, Faculty of Economics and Business Administration (2013). Currently she is also a PhD student at the Corvinus University of Budapest. Her main research areas are market liquidity and market infrastructures.

KATA VÁRADI is an Associate Professor at the Corvinus University of Budapest (CUB), at the Department of Finance. She graduated also at the CUB in 2009, and after it obtained a PhD in 2012. Her main research areas are market liquidity, central counterparties, capital structure and risk management.

SUPPORTED BY THE ÚNKP-17-4-III-BCE-10 (1500000696) NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

Simulation in Industry, Business, Transport and Services

SIMULATION OF AN ORDER PICKING SYSTEM IN A MANUFACTURING SUPERMARKET USING COLLABORATIVE ROBOTS

Fábio Coelho, Susana Relvas, Ana P. Barbosa-Póvoa
CEG-IST, Instituto Superior Técnico
Universidade de Lisboa
Av. Rovisco Pais 1
1049-001, Lisboa, Portugal

KEYWORDS

Industry 4.0, Simulation, Manufacturing supermarket, Order picking

ABSTRACT

Manufacturing industries depend on robust internal logistics operations that enable efficient production strategies, as is the case of an assembly line feed by an internal logistics supermarket. Agile decision making in flexible manufacturing elevates the need for planning the overall shop-floor operations and controlling them. This work explores the existing literature regarding the operation of manufacturing supermarkets and proposes a simulation tool that analyses the order picking activity in a logistics supermarket where the usage of robots is explored in order to feed flexible manufacturing assembly lines efficiently leading to economic savings. This is done using Simio – simulation modelling framework based on intelligent objects. The model outputs suggest that the system performance increases with humans. Although, when uncertainty is considered, the collaborative robots are more flexible, which leads to lower variations of performance.

INTRODUCTION

The manufacturing increasing competitiveness imposes the need of efficient production strategies where the support of internal logistics activities has become a requisite in any manufacturing system. Additionally, the use of flexible robotic solutions is nowadays a reality where the interaction human-robot stands up. Thus, the introduction of collaborative robots (cobots or co-robots) is a path being explored, which however needs an efficient design, planning and control of the overall in-house logistics operations. The automobile sector has adopted the supermarket concept to meet the challenges of just-in-time (JIT) part supply of assembly lines, and decentralized logistics areas were created allowing small deliveries of parts that are immediately needed at assembly line (Emde and Boysen 2012). So, the supermarket is one of the key issues for automobile manufacturers and, subsequently, order picking activities too. Also, market pressure enforces companies to allow customers to customize their products, which leads to mixed-model assembly lines, where more than one model is produced at the same line. Therefore, assembly line feeding is an important aspect to increase efficiency and effectiveness of the manufacturer. To do so an efficient management of such systems is required, where order picking activities are a key concern.

Order picking can be divided into parts-to-picker and picker-to-parts systems (de Koster et al. 2007). The former is typically performed with automated guided vehicles (AGV's) that bring the shelves to the picking station, where the human completes the picking. In the latter, the picker goes throughout the product locations to pick them until completing the orders. It is important to note that parts-to-picker is not appropriate when products have a large volume and have to be stored in individual locations. In addition, kitting is a specific case of order picking, where parts are sorted just-in-sequence (JIS) in dedicated bins (JIS-bins) to be delivered in sequence to the assembly line.

In fact, in industry, the order picking activity continues to be performed by humans because up to now the automatic picking systems have been unable to respond to the requisites of speed, flexibility, precision, and safety necessary to both service level and share the workspace with humans. However, following the trend of industry 4.0, the future goes through collaborative robots that physically interact with each other and with humans in a shared workspace.

In the present work, we use Simio - simulation modelling framework based on intelligent objects - to simulate order picking operations in a manufacturing supermarket. In Simio there is no need to write programming code since it is graphical and intuitive (Pegden and Sturrock 2011). Also, it is simple to present the model to decision-makers once the model looks identical to the real world. Subsequently to the development of the model, several scenarios are studied to determine the most appropriate supermarket configuration that allows a better operational performance.

The main goal of this work is then to offer a simulation tool that analyses the order picking activity in a manufacturing supermarket, allowing one to understand how the system performance responds to the introduction of collaborative robots.

The next section presents a literature review on manufacturing supermarket, order picking and collaborative robots. Then, a model description and its implementation on Simio are described. Subsequently, we explain the model validation and first results are

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

presented. Finally, conclusions are provided, and some future research directions are presented.

LITERATURE REVIEW

According to some authors the concept of manufacturing supermarket is little explored in the research literature (e.g., Emde and Boysen 2012; Saaidia et al. 2014). Still, the papers dealing with the manufacturing supermarkets usually focus on aspects like the definition of routes and scheduling of tow trains or the size and location of supermarkets (Emde 2017).

Related to supermarket locations, (Alnahhal and Noche 2015) tackle this important problem as well as the use of tow trains. The aim of their work is to minimize both transportation and inventory costs. In the (Golz et al. 2012) work, the routing problem of tow trains related to supermarkets is studied based on the automobile industry. The minimization of the number of trips performed by tow trains from supermarket to the assembly line is performed. In the same way, (Emde and Boysen 2012) solve the routing problem of tow trains in order to minimize the number of tow trains used to feed the assembly line. In addition, (Faccio et al. 2013) present a case study of an automotive industry where they intend to optimize the feeding system of mixed-model assembly lines. They also present a framework for design and relate both the number of tow trains and kanban. Furthermore, (Fathi et al. 2014) and (Emde and Gendreau 2017) studied the scheduling problem of tow trains. The former aims to optimize the number of tow trains trips applied to a real case. The latter wants to minimize in-process inventory and, for that, tackle the problem through exact and heuristic solution methods.

Additionally, (Emde et al. 2012) studied the loading problem of tow trains with the objective of minimizing line stocking while accounting for the capacity of the tow trains. Also, (Sternatz 2015) tackle two interdependent problems in parallel, namely assembly line balancing and parts feeding. He found that the worker takes less time to pick parts from a JIS-bin instead of picking from a bulky load unit. In addition, (Battini et al. 2016) studied the same problem. They test direct and indirect parts feeding and report that the application of a combined approach will possibly reduce line stocking and minimize time losses. (Caputo et al. 2015) assign different feeding policies to orders, namely kitting, line stocking and just-in-time, aiming to minimize delivery costs. Two common types of kitting: zone kitting, and batch kitting are often considered. They differ from each other by the fact that in the first one the JIS-bin moves while being filled. On the other hand, in batch kitting the JIS-bin is fixed until it is filled. (Balakirsky et al. 2013) tackle batch kitting processes and demonstrate a case study of robotic kit building. (Hanson et al. 2015) studied the kitting preparation in a real context and refer that some companies prefer to make one kit at a time, while others choose batch picking. Additionally, (Hanson and Medbo 2016) classify different design aspects that can influence this task time.

To perform the picking operation there exist different kinds of autonomous order picking systems. However, (Kimura et al. 2015) remark that fixed equipment is inadequate for high-mix low-volume warehouses. Also, the majority of common mobile robots explored in the literature have only one arm and, therefore, they cannot make order picking as workers. In (Kimura et al. 2015) work, the authors present a mobile dual-arm robot and an autonomous order picking system suitable for high-mix low-volume warehouses. In the same way, (Lemburg et al. 2011) present a robot with two collaborative arms. On the other hand, in (Nieuwnhuisen et al. 2013) work the mobile robot studied is dual-arm but the two arms are not collaborative.

From the above literature review, it can be seen that supermarket stocking and picking problems are neglected in the literature. Additionally, and according to (Nielsen et al. 2017), there is a lack of research in real-world applications of autonomous mobile robots, which leads to poor employment in the industry.

In this context, we intend to develop a simulation model for a real problem of an automobile manufacturing company in order to study the advantages and disadvantages of the collaborative robots' implementation to perform picking operations in a real context.

PROBLEM DESCRIPTION

The reference company for this study is a Portuguese automobile manufacturer that makes use of supermarket concept for line feeding in order to face the distances between the assembly line and the central warehouse.

In order to build a simulation model, it is needed to understand how the real operation is performed. The materials handling begins with the arrival of the products to the factory. Typically, the factory is supplied through the external warehouse or JIT-suppliers, but the materials can also be delivered through external supply. Subsequently, the materials are distributed by tow trains or AGV to both point-of-fit and supermarket locations. Through this system, only the parts needed for a small number of production cycles are available on the line, releasing space at the assembly site and reducing the occurrence of errors.

These supermarkets have a fixed layout, like the one represented in Figure 1, and each supermarket has one dedicated picker.

In the order picking operation under study, when a sequential order arrives (i.e., when an empty dedicated JIS-bin arrives at the supermarket) the assigned picker starts the operation by visiting different supermarket locations to pick the products in order to fulfil a specific kit. After that, the picker leaves the finished kit in a specific place to be later transported in sequence to the assembly line. It is important to note that the picker can perform only one order at a time.

Therefore, the problem developed in this work has to do with order picking in a manufacturing supermarket approached by a simulation model. The objective is that the current operations exclusively developed by humans

can also be developed by robots since this is the intention of the company in order to become more flexible and better face future challenges.

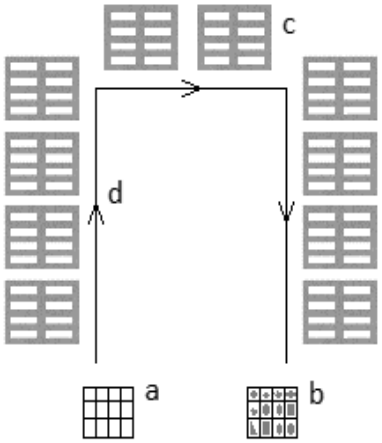


Figure 1: Scheme of Order Picking Operation (a – Empty JIS-bin; b – Fulfilled Kit; c – Different Product Locations; d – Oriented Path)

MODEL DESCRIPTION AND SIMIO IMPLEMENTATION

Based on the real operation described in the previous section a simulation model is developed, which is described below. Such model is then generalized to describe the order picking system of a manufacturing supermarket using collaborative robots.

The supermarket is modelled using an agent-based approach. It consists of humans, cobots, products, storage locations and empty JIS-bins to collect the products.

When an order (i.e., empty JIS-bins dedicated to a given kit type) arrives at the supermarket, one available picker (human or cobot) is assigned to it. Then, it goes through different storage locations in order to fulfil the order (complete the kit) with the requested products. Since the kit is finished, it goes to the final location waiting for transport to the assembly line. At this point, the picker is released and can start the next order.

The purpose of this work is to simulate the order picking operations performed in a manufacturing supermarket in order to calculate the number of kits fulfilled per minute.

In order to build the model in Simio, we use some standard software objects, such as source, worker, vehicle, server, and sink (for readers unfamiliar with Simio, please see Pegden and Sturrock 2013). Therefore, the order picking operation has been modelled as follows:

- Empty JIS-bin – entities sequentially generated by a "source" object that are processed in the supermarket. The empty JIS-bin waits for an available picker that transports it throughout the supermarket locations and completes it according to the kit type;
- Picker – there are two different types available (human, modelled as "worker"; or cobot, modelled as "vehicle") and its function is to fulfil the orders according to each kit type;
- Product locations – modelled as "combiner" object (without processing time), that represents the storage locations of the different products in the supermarket;
- Products - entities generated by a "source" object that are waiting for picking in the storage locations;
- Kit – an entity that represents a fulfil order ready to feed the assembly line.

In order to understand the way the model works some pictures about the properties of some entities and the correspondent symbols are presented. Figure 2 shows the representation of human picker and cobot in Simio. The main properties of these two entities are depicted in Figure 3 and Figure 4.

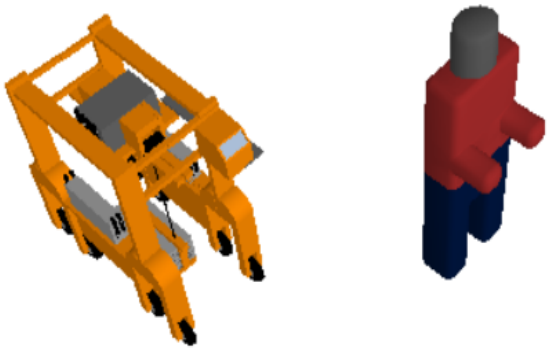


Figure 2: Symbols of the Cobot and Human Picker

Regarding the picking routes, the picker knows exactly which locations he must visit, taking into account the type of kit (order) he receives. In addition, each storage location owns an indication about the quantities of that specific product needed for each type of kit.

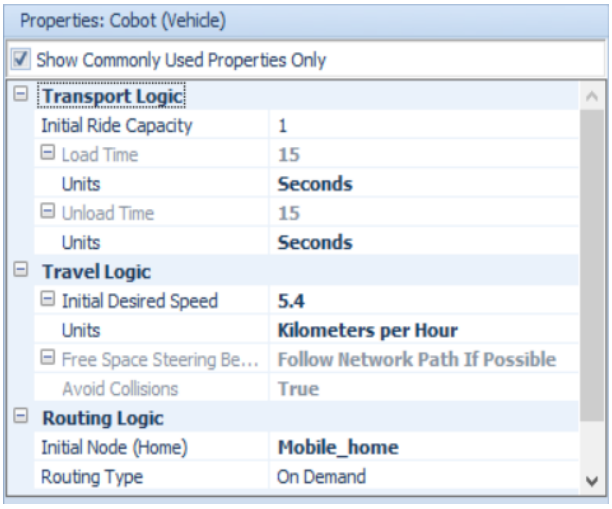


Figure 3: Properties of the Cobot

Some parameters are adjustable, like speed and load/unload times for both cobot and human picker depending on whether the analysis is deterministic or not.

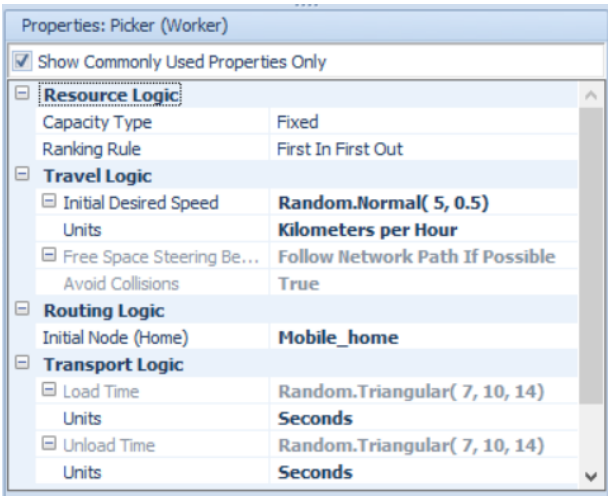


Figure 4: Properties of the Human Picker

MODEL VALIDATION AND PRELIMINARY RESULTS

As mentioned the layout and mode of operation considered in the model developed are representative of the referred automobile company reality.

On the other hand, the operational data introduced in the model has been changed due to confidentiality reasons. However, they are representative and illustrative of the day-to-day operations of the company. For this reason, it enables the execution of the model in order to test different scenarios and analyse both how the model works, and the results obtained.

The different scenarios under analyses are the following:

- Scenario 1 represents the current system configuration with one human picker;
- Scenario 1.1 represents the current system configuration but replacing the human by the cobot;
- Scenario 1.2 represents the current system configuration but human and cobot share the same workspace;
- Scenario 1.3 represents the current system configuration, but now with two human pickers;
- Scenario 1.4 represents the current system configuration, but replacing the human by two cobots;
- Scenarios 2, 2.1, 2.2, 2.3 and 2.4 consider uncertainty in terms of speed and load/unload times of human picker, and uncertainty related to expected orders in scenarios 1, 1.1, 1.2, 1.3 and 1.4, respectively.

For scenarios dealing with uncertainty, we consider for the cobot speed a fixed value of 5.4 kilometres per hour and for the human picker a random value that follows a normal distribution with a mean of 5 and standard deviation of 0.5 kilometres per hour. In the same way, load/unload times (that represents order picking operation) for human picker has a random value that follows a triangular distribution (with a lower limit of 7 seconds, an upper limit of 14 seconds and a mode of 10 seconds), while the cobot has a fixed value of 15 seconds. Regarding the arrival of orders, it follows a random exponential distribution (with a mean of 1 minute) to simulate the uncertainty across different work days. It is important to note that the choice of distributions was made based on i) application to real cases demonstrated in the existing literature (e.g., Liong and Loo 2009), and ii) distribution that best fits the real data observed. Additionally, the user can change the number of both humans and cobots in the model to see how the system responds in terms of service level.

The simulation duration was 24 hours to meet the uninterrupted work of the company under study (24 hours a day). The model was run terminating since the company wants to know how many kits are completed at the end of a 24 hours cycle. The break times are excluded from the model as work always continues exactly as it left off before the break. Regarding experiments, 100 replications were performed with an average computational time of 3,1 minutes each. It is important to note that this number of replications is acceptable since we are interested in estimating the mean of the number of kits per minute. However, if we want to estimate a maximum value the number of replications would have to be higher.

Table 1 shows the comparison of the average number of kits per minute and the distance travelled by the pickers for the six scenarios under analysis.

Table 1: Scenarios' preliminary results

Scenarios	Number of kits per minute	Total number of kits	Number of pickers	
			Human	Cobot
Without uncertainty				
1	0,40	576	1	-
1.1	0,31	446	-	1
1.2	0,70	1008	1	1
1.3	0,81	1166	2	-
1.4	0,63	907	-	2
With uncertainty				
2	0,30	432	1	-
2.1	0,30	432	-	1
2.2	0,66	950	1	1
2.3	0,67	964	2	-
2.4	0,62	893	-	2

The results show that comparing human versus cobot, the former achieves a better performance, especially in deterministic scenarios. This can be explained by the fact that cobot has higher load/unload times than the human picker. However, if we look more closely at the results, we can see that when uncertainty is considered this conclusion is not so obvious. On the one hand the human picker obtains a better performance compared to the robot, but on the other hand presents a greater variation of performance than the robot when the uncertainty is considered.

For example, scenarios 2, 2.2 and 2.3 – where uncertainty in terms of speed and load/unload times of human picker, and uncertainty related to expected orders are considered – reaches a slightly lower performance compared with scenario 1, 1.2 and 1.3, respectively. Again, this can be explained by the fact that all the referred scenarios involve human picker, which is more affected by uncertainty, unlike the cobot. In contrast, scenarios 2.1 and 2.4 present almost the same performance of deterministic scenarios 1.1 and 1.4, respectively.

DISCUSSION AND FUTURE RESEARCH

This paper presents a simulation modelling tool to analyse the order picking activity in a manufacturing supermarket, as well as to analyse how the system performance answers due to the implementation of collaborative robots.

According to simulation results, the performance increases when the human picker is considered. Although, when uncertainty is considered, the collaborative robots are more flexible, which leads to lower variations of performance in these cases. This result can be generalized to several types of uncertainty.

Given that this is an ongoing work, there are future improvements that will be introduced in the model, such as inventory control, routes of tow trains. In future work, we also intend to improve model's layout and analyse how this influences the operation performance. Through all this, the model can be tested all together and validate it with real data.

The developed tool aims to support the automobile company in the design and planning of the aforementioned supermarkets allowing the testing of several scenarios of order picking operation accounting for the presence of uncertainty.

Finally, the model can be generalized to other companies that use the manufacturing supermarket concept.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from Portugal 2020 framework, under the project POCI-01-0145-FEDER-016418 financed by EU/FEDER through the programme COMPETE2020.

REFERENCES

Alnahhal, M., B. Noche. 2015. "A Genetic Algorithm for Supermarket Location Problem", *Assembly Automation*, 35:1, 122-127.

Balakirsky, S., Z. Kootbally, T. Kramer, A. Pietromartire, C. Schlenoff and S. Gupta. 2013. "Knowledge driven robotics for kitting applications". *Robotics and Autonomous Systems*, 61:11, 1205–1214.

Battini, D., M. Calzavara, A. Otto and F. Sgarbossa. 2016. "The Integrated Assembly Line Balancing and Parts Feeding Problem with Ergonomics Considerations". *IFAC-PapersOnLine*, 49:12, 191-196.

Caputo, A. C., P. M. Pelagagge and P. Salini. 2015. "A Decision Model for Selecting Parts Feeding Policies in Assembly Lines". *Industrial Management and Data Systems*, 115:6, 974-1003.

de Koster, R., T. Le-Duc and K. J. Roodbergen. 2007. "Design and Control of Warehouse Order Picking: A Literature Review". *European Journal of Operational Research*, 182:2, 481-501.

Emde, S., M. Fliedner and N. Boysen. 2012. "Optimally Loading Tow Trains for Just-in-Time Supply of Mixed-Model Assembly Lines". *IIE Transactions*, 44:2, 121-135.

Emde, S. and N. Boysen. 2012. "Optimally Routing and Scheduling Tow Trains for JIT-Supply of Mixed-Model Assembly Lines". *European Journal of Operational Research* 217, 287-299.

Emde, S. 2017. "Scheduling the Replenishment of Just-in-Time Supermarkets in Assembly Plants". *OR Spectrum*, 39, 321-345.

Emde, S. and M. Gendreau. 2017. "Scheduling In-House Transport Vehicles to Feed Parts to Automotive Assembly Lines". *European Journal of Operational Research*, 260, 255–267.

Faccio, M., M. Gamberi, A. Persona, A. Regattieri and F. Sgarbossa. 2013. "Design and Simulation of Assembly Line Feeding Systems in the Automotive Sector Using Supermarket, Kanbans and Tow Trains: A General Framework". *Journal of Management Control*, 24:2, 187–208.

Fathi, M., M. J. Alvarez, F. H. Mehraban and V. Rodríguez. 2014. "A Multiobjective Optimization Algorithm to Solve the Part Feeding Problem in Mixed-Model Assembly Lines". *Mathematical Problems in Engineering*, 2014:1, 1-12.

Golz, J., R. Gujjula, H-O. Gunther, S. Rinderer and M. Ziegler. 2012. "Part Feeding at High-Variant Mixed-Model Assembly Lines". *Flexible Services and Manufacturing Journal*, 24:2, 119-141.

Hanson, R., L. Medbo and M.I. Johansson. 2015. "Order Batching and Time Efficiency in Kit Preparation". *Assembly Automation*, 35:1, 143-148.

Hanson, R. and L. Medbo. 2016. "Aspects Influencing Man-Hour Efficiency of Kit Preparation for Mixed-Model Assembly". In *Proceedings of the 6th CIRP Conference on Assembly Technologies and Systems*, 44, 353-358.

Kimura, N., K. Ito, T. Fuji, K. Fujimoto, K. Esaki, F. Beniyama and T. Moriya. 2015. "Mobile Dual-Arm Robot for Automated Order Picking System in Warehouse Containing Various Kinds of Products". In *Proceedings of the 2015 IEEE/SICE International Symposium on System Integration*, 332-338.

Lemburg, J., J. de G. Fernández, M. Eich, D. Mronga, P. Kampmann, A. Vogt, A. Aggarwal, Y. Shi, and F. Kirchner. 2011. "AILA - Design of an Autonomous Mobile Dual-Arm Robot". In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation*, 5147-5153.

Liong, C-Y and C. S. E. Loo. 2009. "A Simulation Study of Warehouse Loading and Unloading Systems Using Arena". *Journal of Quality Measurement and Analysis*, 5:2, 45-56.

Nielsen, I., Q-V. Dang, G. Bocewicz and Z. Banaszak. 2017. "A Methodology for Implementation of Mobile Robot in Adaptive Manufacturing Environments". *Journal of Intelligent Manufacturing*, 28:5, 1171–1188.

Nieuwnhuisen, M., D. Droeschel, D. Holz, J. Stückler, A. Berner, J. Li, R. Klein and S. Behnke. 2013. "Mobile Bin Picking with an Anthropomorphic Service Robot". In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation*, 2327-2334.

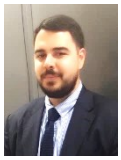
Pegden, C. D., and D. T. Sturrock. 2011. "Introduction to Simio". In *Proceedings of the 2011 Winter Simulation Conference*, 21-31.

Pegden, C. D. and D. T. Sturrock. 2013. “Rapid Modeling Solutions: Introduction to Simulation and Simio”. Pittsburgh: Simio LLC.

Saaidia, M., S. Durieux and C. Caux. 2014. “A Survey on Supermarket concept for Just-in-Time Part Supply of Mixed Model Assembly Lines”. In *Proceedings of the 10th International Conference of Modeling and Simulation*.

Sternatz, J. 2015. “The Joint Line Balancing and Material Supply Problem”. *International Journal of Production Economics* 159, 304-318.

AUTHOR BIOGRAPHIES



FÁBIO COELHO is a PhD student at the Engineering and Management Department of Instituto Superior Técnico (IST), University of Lisbon. He holds a MSc degree from IST. His research focus is on simulation, operations research, logistics and decision support systems. His work develops in close collaboration with industry. His e-mail address is: fabiocoelho@tecnico.ulisboa.pt.

SUSANA RELVAS is an Associate Professor at the Department of Engineering and Management of Instituto Superior Técnico (IST), University of Lisbon. She holds a PhD in Industrial Engineering and Management from IST. Her research focus is on Logistics, Supply Chain Management and Operations Management, by applying state of the art solution methods to real life problems. Through a systems approach, she collaborates with peers and companies in national and international research projects. She is the author of several papers in top impact factor journals and supervises several PhD and MSc candidates. Her e-mail address is: susana.relvas@tecnico.ulisboa.pt.

ANA P. BARBOSA-PÓVOA is a Full Professor in Operations and Logistics, at the Engineering and Management Department of Instituto Superior Técnico (IST), and she holds a PhD from the Imperial College of Science Technology and Medicine. Her research focuses on developing a comprehensive understanding of a variety of problems in supply chains and operations management, supported by novel engineering systems models and techniques. Her work develops in close collaboration with leading companies from a variety of different industries in the decision support and optimization of their operations. She has published widely in international journals and has participated in a large number of conferences where she has presented several plenary lectures. Her e-mail address is: apovoa@tecnico.ulisboa.pt.

STATISTICAL EVALUATION OF EMERGENCY SERVICE DEMAND IN ELECTRIC POWER DISTRIBUTION UTILITIES

Guilherme de Oliveira da Silva, Vinícius Jacques Garcia, Lynceo Falavigna Braghirolli,
Federal University of Santa Maria UFSM, Brazil.
Email: guilhermeos.ep@gmail.com, viniciusjg@ufsm.br, lynceo@gmail.com

KEYWORDS

Electric power distribution, queueing theory, operations research, performance evaluation, service systems.

ABSTRACT

Waiting time in queues constitutes a common problem in customer service systems, where delay is perceived as cost by customers enrolled in the process. In electrical energy maintenance services, this delay is attenuated by the entry of emergency repair orders. This paper aims to evaluate the emergency service demand through a queuing model. For this purpose, emergency order records were analyzed and, after adequate statistical treatment for these data due to their inherent variability, it was possible to calculate some performance indexes. From these results, it was possible to analyze how the variation of some key parameters impacts the system, in order to provide information for the decision making regarding the service system

INTRODUCTION

The performance of a service system deserves a similar emphasis to that of strategic areas within companies, with the damages and implications resulting from low performance being the main causes for such an emphasis (Fitzsimmons and Fitzsimmons, 2010). (Toor, 2008) reports the importance of the existence of a service level agreement, which includes not only the level of service promised, but corrective actions and penalties if this is below the established standard. In this same direction, (Wu, Lee and Cao 2009) highlight a practice used in public services, which consists in the elaboration of regulations for times of interruption of supply of a certain service, in order to ensure the minimization of the time in which the service leaves to be borrowed.

According to (Fitzsimmons and Fitzsimmons 2010), the waiting time to attend a given service is seen as cost by the individuals who are participating in the process. The authors point out that this waiting cost is something that is rarely explained in studies and analyzes, but the experience of being in a certain queue for a service must be the object of attention and observed by the service providers under the physical, behavioral and economic aspects. According to (Mital 2010), similar to the affirmation "zero defects" diffused in the industrial field, in the conjuncture of services the principle of "zero failures" must be followed, since the dissatisfaction of a certain client can go beyond the expected time for the service, arriving the withdrawal and subsequent change of service provider due to the poor performance of the service system, often materialized in the simple mismatch of expectations with the desired levels of service.

The evaluation and proposition of improvements within the field of services and operations, more precisely, in their service systems, are carried out by some works through the use of techniques and concepts traced in queuing theory. As examples, (Yankovic and Green 2011), (Chadha, Singh and Kalra 2012) and (Xu et al. 2014) work such concepts in their researches aimed at optimizing the performance of the service systems. In the context of the electricity distribution sector, the service rendering systems present, according to (ANEEL 2016), attributes such as perceived quality, satisfaction and reliability, turning around 70%, while reliability presents 48.08% as final index, which assumes that the sector as a whole suffers low confidence of its clients.

This paper deals with the development of a statistical analysis on the admission of emergency orders observed from a history series in order to enable a preliminary evaluation of a service delivery system in the electricity distribution sector. In the following section, the existing theoretical basis for the development of the study is presented. The third section shows the methods used to carry out the study and the results obtained from what was drawn. Finally, the fourth and last section contains the conclusions of this study and its main contributions, finalizing the work presenting suggestions for future work.

THEORETICAL REFERENCE

According to (Fitzsimmons and Fitzsimmons 2010), a queuing system corresponds to the waiting of clients who lack one or more services of a certain server (s). The authors comment on the fact that a queue does not necessarily have to be materialized in the form of a line and with people queuing; for example, queues can also be considered as individuals placed on standby by a certain telephone service or people waiting for medical care in an emergency care system.

Applications involving the study of queues may be of different natures. (Mayhew and Smith 2008) present in their work an application of the queuing study to evaluate a government policy, which establishes as goal a maximum limit of 4 hours of service time for 98% of emergency cases that arrive at public hospitals in the city of London, England. The authors demonstrate how the problem can be treated as a queuing case and show that,

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

in the way the process was designed, it is not possible to reach the established time goal.

Another example of application of the queuing study is demonstrated in the work of (Zapata et al. 2010), where the authors present a model for modeling and optimization of repair processes in distribution units. The model, which can be summarized according to Figure 1, follows three major steps in its methodology: (i) adjusting a sample of failure events and service times data to probability distributions; (ii) simulation by Monte Carlo method from the probability distributions adjusted for failures and service times; and (iii) calculation and analysis of maintenance management indicators.

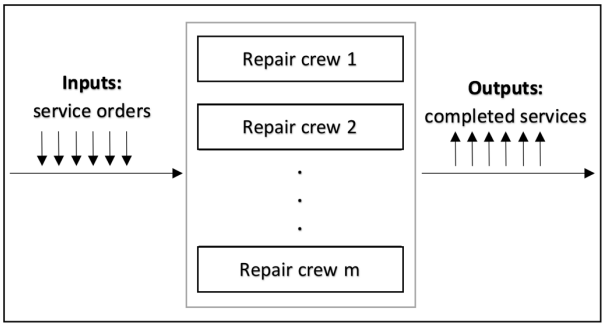


Figure 1: Queue model for the service system considered.

The concept of queuing theory materializes through the construction of models that are able to capture the behavior of the phenomenon in question and translate it into performance indicators. (Taha 2002) presents in his work several models that have application according to the type of queue one wants to study, highlighting the importance of the understanding behind the existing theory on the subject. (Hillier and Lieberman 2014) objectively point to queue models capable of capturing the behavior of the phenomenon and quantifying in the form of performance indicators the same. Within the models brought by the authors and aiming at the current situation in this study, the model of birth and pure death stands out, which is of direct application to the present case.

Assumptions for queuing models and statistical data processing

According to (Mital 2010), a queuing system basically consists of a standard of arrival of those who need a certain type of service, as well as a server standard, queue discipline, system capacity, number of channels services and service stages numbers. (Mital 2010) comments that for the construction of a queuing model capable of understanding the interaction between the elements and generating performance metrics about a system, some assumptions are necessary for such a model to be valid. These hypotheses are as follows:

- Only single orders arrive at the system and there are no mass arrivals;
- The lengths between the arrivals intervals are independent;
- The lengths between the arrivals intervals are identically distributed by a continuous density function;
- The times between arrivals and service times follow an exponential distribution, therefore, the arrival rate and the service rate follow a Poisson distribution;
- The discipline of the queue corresponds to the way customer service is given over time, for example, first-come, first-served (FIFO).

Another determining factor in the construction of queue models is the understanding of the type of data that is being worked on. A previous statistical analysis of the elements available for the study constitutes an important step, which is commonly performed in works with this approach. It is possible to observe such treatment in the work of (Cheevarunothai, Mooney and Wang 2007), in which the authors conduct an investigation of the data through basic descriptive statistics such as median and standard deviation, to then deepen the study through row theory.

(Zavanella et al. 2015) also highlight the relevance of statistical treatment of data prior to the study of queues per se. According to (Zavanella et al. 2015), the evaluation of the probability distribution of service times and the arrival rate are one of the premises for elaboration of the model, however, in cases where there is great variability it is necessary to evaluate how much such oscillation can impact on results. If it is necessary, it is fundamental that different unstable information that does not demonstrate a behavior pattern is considered and treated differently, and the disposal of such information is plausible in some situations.

STATISTICAL PROCEDURE DEVELOPED

The context of this study covers the particularities and constraints that exist in the service rendering process of an electric energy distribution company. It should be noted the complexity of the type of operation carried out by companies in this branch, and the existence of such complexity is attributed to the great variability observed in the entrance of scheduled and emergency services. These services, on the other hand, are processed by teams and need to be balanced with the company's processing capacity so that there is no idle resources or the maximum outdated processing capacity and a likely decrease in the level of service.

An alternative focus may lay on a context where different service levels could be associated with different customers. In that case, a direct correlation between service level and priority may be used to influence the order in which customers are attended, i.e.: customers with high service level will have high priorities, while customers with low service level will have low priorities. A simple way to do this ponderation would be, for example, use multipliers of type 1 (low priority) and 1,5 (high priority), which would result in a privileged service for the customers with higher levels of service.

The steps followed in this research aiming at the traced objective can be summarized as follows:

(i) Adaptation of the matrix formed by all information and data of the emergency orders: calculation of the standard deviations of the service times, per hour, and discard of the values which exceed ± 2 standard deviations, since these values represent possible abnormal variations of the service time, which compromise the modelling of the behaviour of the series;

(ii) From the new matrix formed, elaboration of two data series: accumulated time per hour and accumulated time per hour and day of the week;

(iii) Calculation of the mean and standard deviation of each accumulated series, as well as the Coefficient of Variation (CV) of these series, which corresponds to the division of the standard deviation by the mean;

(iv) Adjustment of probability distribution to lower CV series;

(v) Calculation of performance indexes of the row model.

Steps (i), (ii) and (iii) encompass the statistical examination of the data, while steps (iv) and (v) are part of the queuing model construction process. For the construction of such a model, the performance metrics for average service time, average exit rate and average arrival rate were calculated. Based on these indicators, it was possible to estimate other measures of relevance to the system capable of assisting in decision making regarding its operation.

SIMULATION RESULTS

This work includes the study of an electrical power distribution utility that serves a region corresponding to 99,000 km², serving 118 cities, approximately 1,300,000 customers and with an average of 500,000 services generated annually.

The data used to perform this research constitute the records of emergency orders from July 1, 2014 to July 31, 2015. Stratified by time of day (0 to 23), these records contain data of 392 days which, multiplied by the number of hours, result in 9,408 lines. It is worth mentioning the great variability existing in the data inherent to its emergent nature. Specifically, the service time attached to each hour of the day, which is usually composed of more than one order, presents great variation in the course of any day analyzed.

In this way, it is wrong to investigate the behavior of the emergency orders in a global way, considering as satisfied the hypothesis that such orders have similar behavior in their entirety. Thus, a statistical evaluation is necessary in order to give adequate treatment to the data and, in this way, to guarantee the infallibility of the solution presented later. This treatment of data due to instability composes some of the steps of the method used in this research.

Table 1 shows the header of the matrix formed by the emergency order records provided by the company studied, together with the possible values associated with each parameter. In addition, for purposes of understanding the treated matrix, there are some hypothetical lines of data disposition in the matrix until the last hour of the last day recorded in the matrix. Following what was described as the first stage of the methodology for this research, the first procedure consisted in the calculation of the standard deviation of the hours of service per hour.

Table 1: Matrix records of emergency orders

Year	Month	Day	Day of the week	Hour	Service Time (hours)	# of requests
2014 – 2015	1..12	1..31	1..7	0..23	Variable	Variable
2014	7	1	3	0	0.14	2
2014	7	1	3	1	0.20	3
2014	7	1	3	2	1.5	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2015	7	31	6	23	3.2	7

After this calculation, values that exceeded the lower or upper limits were discarded. Therefore, the new matrix remained with the same aspect presented in Table 1, but now contains 8,938 lines. Such a reduction relates to the lines containing service times exceeding the standard deviation limits and, therefore, have been removed entirely from the matrix.

After the preliminary analysis of the standard deviations, it was possible to observe that the variation of the service times occurs, specifically, under the parameters of day of the week and time of day. Thus, in order to compose the data to enable the calculation of queue performance indices, two data series were elaborated, corresponding to: (a) cumulative service time per hour and (b) accumulated service time per hour and by day of the week. In other words, in (a) we have the sum of all service times of the 392 days recorded for, for example, hour 0, followed by hour 1, hour 2, and so on. In (b), we have the same sum described in (a), but segregating the sum of the time by the 7 days of the week. Table 2 and and Table 3 provide the service time values corresponding to the series discussed.

For both series, the procedure of calculation of mean and standard deviation was repeated. With the upper and lower limits set at ± 2 standard deviations, the corresponding values of the series were plotted in graphs that can be visualized in Appendix A. It is observed the

great similarity between the graphs, being possible to delineate a certain behavior of the times of service throughout the hours of each day.

In (a), a certain stability is observed from hour 0 to hour 8. From time 9, the series gains a variability behavior, which remains approximately until hour 19. From this time on, the time of service decreases, closing hour 23 with a value similar to the values presented in the stable interval between hour 0 and hour 8. This behavior is also noticed when we look at the 7 graphs derived from series (b).

Table 2: Series (a): accumulated service time (hours).

Hour	Accumulated service time (hours)
0	311,30
1	305,94
2	307,60
3	310,74
4	326,53
5	312,81
6	357,27
7	404,79
8	849,69
9	937,03
10	1020,34
11	865,62
12	741,12
13	1142,79
14	836,20
15	875,30
16	885,58
17	794,01
18	878,10
19	760,45
20	638,68
21	506,65
22	417,17
23	317,62

In (b), the stability between hours 0 and 8 is repeated, as well as the decrease and resumption of stability after hour 19. However, the points detected outside the upper control limit in the charts of days 4, 6 and 7, corresponding respectively to Wednesday, Friday and Saturday. In the three graphs, the point that is outside the upper limit is the point that corresponds to hour 13; in the other graphs of the days of the week, the points closest to the upper limit are the hours 9 (day of week 1 - Sunday), 13 (day of week 2- Monday) and 13 (day of week 5 - Thursday).

The last stage of the statistical treatment concerns the calculation of the series CV. For the purposes of analysis, the series (a), which includes the cumulative time of service per hour, was not considered in the CV calculation because, as already noted, its result is shown in a fractional way in the series (b).

Table 4 presents CV values per day for series (b), considering green squares for $CV \leq 40\%$, yellow squares for $40\% < CV < 60\%$ and red squares for $CV \geq 60\%$. It is worth noting that there are only 4 CV values below 40%, 100 values are between 40% and 60% and 64 are above 60%. This represents that 2.38% of the values are in an acceptable range of variability (less than 40%), the rest being in doubt as to the stability and, consequently, the generalization of their results.

Table 3: Series (b): accrued time of service per hour and day of the week

Hour	Day of the week						
	1	**2**	**3**	**4**	**5**	**6**	**7**
0	40.57	49.13	49.81	41.82	48.41	42.26	39.30
1	41.99	45.81	46.38	43.22	44.94	41.74	41.87
2	38.53	52.17	47.26	41.09	47.50	42.25	38.80
3	43.14	46.33	46.72	40.60	48.91	44.21	40.83
4	43.25	49.94	48.17	45.67	48.20	49.70	41.60
5	44.20	45.83	49.02	40.78	45.74	45.11	42.13
6	41.95	54.08	48.12	57.78	49.29	61.21	44.84
7	56.09	57.16	58.53	59.64	57.33	67.25	48.80
8	95.41	122.62	137.93	127.93	134.09	117.65	114.06
9	109.46	140.23	151.12	140.11	142.91	135.67	117.53
10	100.46	146.67	170.11	169.96	163.56	159.17	110.41
11	99.49	125.70	129.56	135.11	128.05	138.23	109.49
12	81.04	98.84	112.03	124.25	114.86	113.14	96.95
13	78.83	161.10	164.27	199.42	184.98	201.13	153.05
14	80.97	130.19	111.96	143.66	119.00	140.34	110.07
15	75.35	141.54	140.97	138.03	137.12	139.27	103.03
16	74.71	142.27	143.03	130.05	138.87	139.44	117.22
17	81.87	122.69	129.58	123.10	130.06	117.05	89.65
18	80.23	132.13	137.24	138.36	144.21	138.30	107.61
19	94.06	115.84	104.28	120.43	117.65	108.70	99.50
20	66.90	105.22	98.36	107.45	96.77	90.51	73.46
21	63.62	70.84	66.11	80.80	88.49	78.50	58.30
22	51.87	62.17	65.58	61.20	60.85	64.91	50.59
23	42.27	42.59	46.36	47.51	51.35	46.24	41.30

Once the values of lower CV within series (b) were detected, the study of performance indicators was limited to these values only for these four new subsets. These subsections correspond to the sum of the individual values of service times for: all hours 9 on all days of week 2, all hours 9 on all days of week 4, all hours 11 on all days of week 4 and all hours 15 on every day of week 2. Figure 2, Figure 3, Figure 4 and Figure 5 present the histograms of the individual service times of subsections, showing that in the four cases we have a strong approximation of the exponential distribution for subsections.

By analyzing the individual time-of-service values recorded per hour, it is possible to notice that each service time registered is linked to a number of occurrences also recorded. Thus, it is possible to estimate the performance measures of average service time queues, average order arrival rate and average order exit rate. Table 5 presents these indicators for the four subseries.

Note that in all subseries the average order entry rate is higher than the average exit rate. It is also possible to notice that the greater the difference between the rates, the longer the service time is. Therefore, a possible measure to improve system performance would be to include more teams at these times these days of the week, or some other action that would be able to meet the demand for order entry, shortening service time.

Table 4: CV values for series (b), in percentages.

	Day of the Week						
hour	1	2	3	4	5	6	7
0	59.25	65.90	61.30	64.03	57.87	67.04	59.70
1	57.87	63.40	58.81	66.77	48.30	67.65	56.97
2	55.63	67.62	61.29	62.81	59.14	67.05	60.33
3	64.27	61.65	58.96	63.57	53.59	67.19	59.19
4	66.66	70.58	60.21	71.11	60.31	77.35	66.39
5	63.75	65.35	61.16	67.15	58.63	65.26	59.14
6	60.40	71.43	63.16	70.77	61.30	70.99	60.80
7	66.76	68.56	69.56	68.98	62.07	64.19	55.85
8	64.34	50.88	53.27	46.18	58.77	50.51	52.81
9	70.43	39.92	42.65	34.15	50.29	52.23	53.37
10	64.59	54.14	45.53	42.92	46.29	48.82	53.16
11	57.29	48.80	45.66	39.61	53.09	51.28	48.90
12	64.96	50.75	56.09	56.11	51.09	57.15	66.01
13	71.26	44.66	56.81	45.34	55.47	47.48	53.19
14	74.74	43.22	51.13	41.15	63.28	42.48	63.75
15	79.28	39.20	51.02	52.82	45.34	42.42	50.10
16	59.70	48.10	53.36	48.15	46.11	49.72	61.47
17	83.51	44.85	40.89	50.44	48.43	49.78	54.85
18	66.56	43.84	46.61	43.74	41.80	49.63	56.20
19	58.34	51.20	50.63	46.62	47.14	53.08	64.94
20	58.41	53.36	58.17	51.81	50.60	55.13	58.11
21	61.64	57.08	57.81	67.38	57.72	69.78	79.71
22	62.63	54.73	51.90	59.92	59.78	56.62	68.84
23	55.87	54.47	61.49	59.72	49.95	66.70	63.55

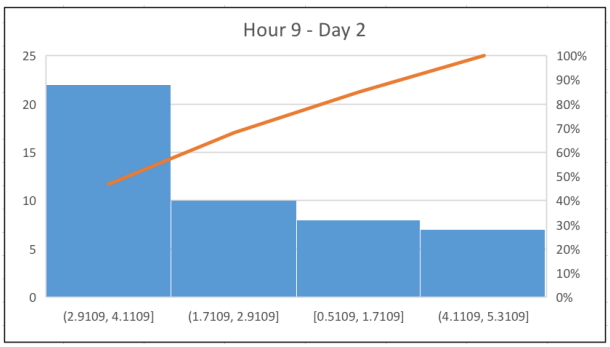


Figure 2: Histogram of the individual service times of the subset corresponding to time 9 and day 2.

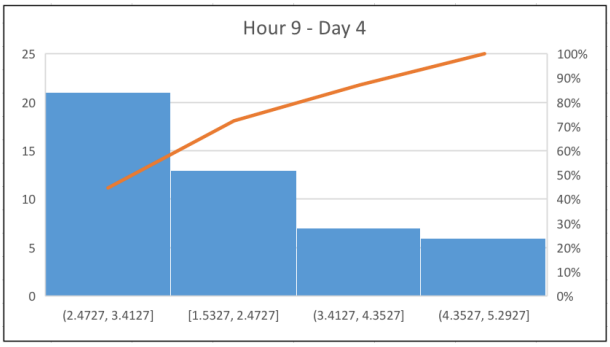


Figure 3: Histogram of the individual service times of the subset corresponding to time 9 and day 4.

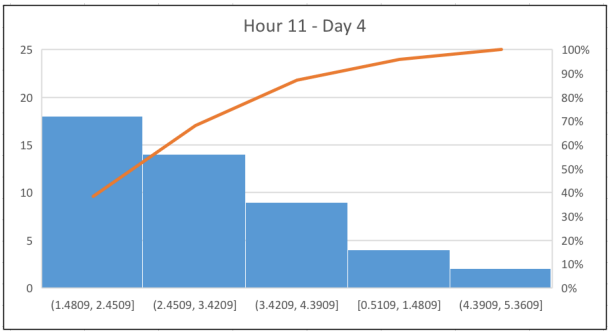


Figure 4: Histogram of the individual service times of the subset corresponding to time 11 and day 4.

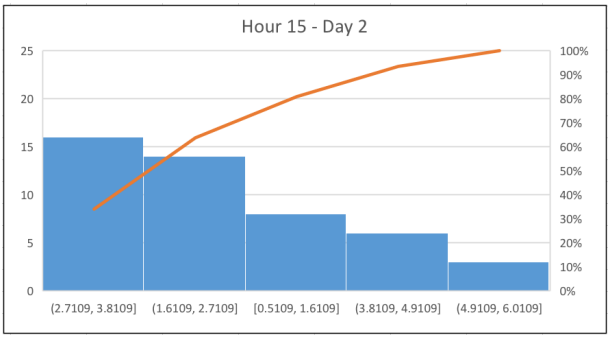


Figure 5: Histogram of the individual service times of the subset corresponding to time 15 and day 2.

Table 5: Performance indicator.

Indicator	Hour of day - Day of the week			
	9 – 2	9 – 4	11 – 4	15 - 2
Average service time (hours)	2.9836	2.9046	2.5807	2.9028
Average output rate (orders/hour)	0.3352	0.3443	0.3875	0.3445
Average input rate (orders/hour)	5.4255	5.4082	4.6981	5.1429

FINAL REMARKS

The performance of service systems is intertwined with the strategic relevance that such performance receives from organizations. Waiting time in queues is one of the consequences of the poor management of such systems,

being seen as cost by some customers, which opens the door for possible service withdrawals or complaints regarding service delays and consequently waiting too long. Some authors defend certain strategies to avoid harm to the consumer by the inefficiency of the system, causing, in some cases, fines or penalties for the service providers.

In this sense, one of the found and diffused ways to study and control this phenomenon is demonstrated in queuing theory. The use of techniques and tools to capture the behavior of different types of queues compose an important measure, which can be seen in a strategic way by companies that are faced with this type of situation. In the electricity distribution sector, the queuing case is presented in the form of scheduling and emergency service orders, which form a kind of non-tangible queue, but which can be studied and analyzed by means presented in queue theory.

This work aimed at evaluating the performance of a service delivery system in the electricity distribution sector through the study of queues. From an initial statistical treatment of the data, due to the great variability inherent to this type of data, it was possible to arrive at the calculation of indexes that mirror the performance of the system as a whole. The final evaluation is that the system studied has a performance that needs improvements that aim to fully meet the demand for the entry of work orders, so that we can work on reducing the total time of service.

As a suggestion for future work is the idea of expanding this study for the remaining hours and days of the week, working, first, the statistical question of inherent variability and then indicators of performance of queues. In addition, other performance metrics can be calculated and presented, mainly in relation to the probability of occurrence of orders, estimating the possible percentage of equilibrium of the system.

ACKNOWLEDGEMENTS

The authors would like to thank the AES SUL Distribuidora Gaúcha de Energia SA for financial support provided to the project “Planejamento dinâmico de Operações”.

REFERENCES

ANEEL, 2010, “AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA”. Resolução normativa 414. 2010. Disponível em: <http://www2.aneel.gov.br/cedoc/ren2010414.pdf>. Acesso em: 25 nov. 2017.

Chadha, R.; A. Singh and J. Karlra. 2012. “Lean and queuing integration for the transformation of health care processes: A lean health care model”. Clinical Governance: An International Journal, v. 17, p.191-199, mar. 2012.

Cheevarunothai, P.; M. Mooney and Y. Wang. 2007. “Statistical and Queuing Analyses of Freeway Incidents in Washington State”. IEEE Intelligent Transportation Systems Conference, Seattle, 2007.

Fitzsimmons, J. A. and M. J. Fitzsimmons. 2010. “Service Management: Operations, Strategy, Information Technology”. McGraw-Hill; 7th Revised edition.

Hillier, F. S. and G. J. Lieberman. 2014. “Introduction to Operations Research”. 10th edition, McGraw-Hill.

Mital, K.M. 2010. “Queuing analysis for outpatient and inpatient services: a case study”. Management Decision. v. 48, p. 419-439.

Mayhew, L. and D. Smith. 2008. “Using queuing theory to analyse the Government’s 4-h completion time target in Accident and Emergency departments”. Health Care Manage Science. v. 11, p. 1-11.

Taha, H. A. 2002. “Operations research”. Prentice Hall, 7th edition.

Toor, T. P. S. 2008. “Building effective service management system”. Business Strategy Series. v. 10, p.61-67.

Wu, J-S.; T-E. Lee and C-H. Cao. 2009. “Intelligent Crew and Outage Scheduling in Electrical Distribution System by Hybrid Generic Algorithm”. The 4th IEEE Conference on Industrial Electronics and Applications, Xi’an.

Xu, X-Y.; J. Liu; H-Y. Li and J-Q. Hu. 2014. “Analysis of subway station capacity with the use of queueing theory”. Transportation Research Part C. v. 38, p. 28-43.

Yankovic, N. and L. V. Green. 2011. “Identifying Good Nursing Levels: A Queuing Approach”. Operations Research. v. 59, p. 942-955.

Zapata, C. J.; J. Díaz; M. L. Ocampo; J. D. Marriaga; J. U. Patiño and A. F. Gallego. 2010. “The Repair Process of Five Colombian Power Distribution Systems”. 2010 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America.

Zavanella, L.; Zanoni, S.; Ferretti, I.; Mazzoldi, L. 2015 “Energy demand in production systems: A Queuing Theory perspective”. Int. J. Production Economics. v. 170, p. 393-400.

AUTHOR BIOGRAPHIES

GUILHERME DE OLIVEIRA DA SILA received his Bachelor's degree on Industrial Engineering from the Federal University of Santa Maria, in 2016. Currently, he is enrolled in his master course on Industrial Engineering at the same University, in Brazil.

VINÍCIUS JACQUES GARCIA received his Bachelor's degree from the Federal University of Santa Maria in 2000, the Master's and Doctor's degree from the State University of Campinas in 2002 and 2005, respectively. Since 2011 he has been professor at Federal University of Santa Maria. His research interests include distribution system planning and operation, combinatorial optimization and operations research.

LYNCEO FALAVIGNA BRAGHIROLLI is a professor of industrial engineering at Federal University of Santa Maria. His research interests include simulation, optimization, lean digital entrepreneurship and engineering education.

APPENDIX

Appendix A: Graphs with the upper and lower limits set at ± 2 standard deviations of (a) cumulative service time per hour and (b) accumulated service time per hour and by day of the week.

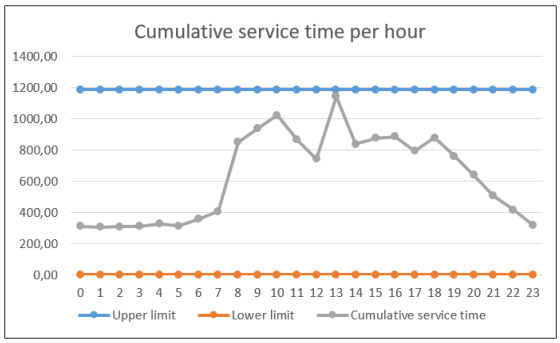


Figure 6: Series (a): Cumulative service time per hour

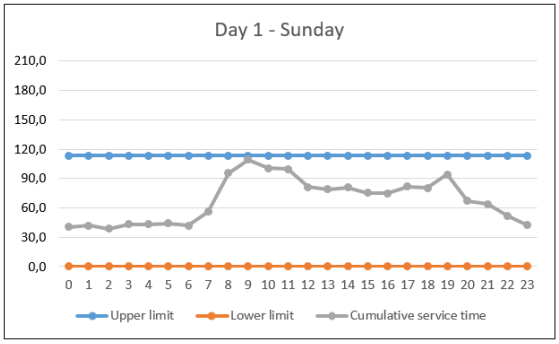


Figure 7: Series (b): Cumulative service time per hour for day 1

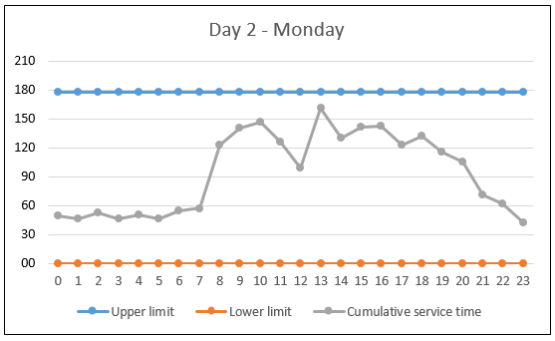


Figure 8: Series (b): Cumulative service time per hour for day 2

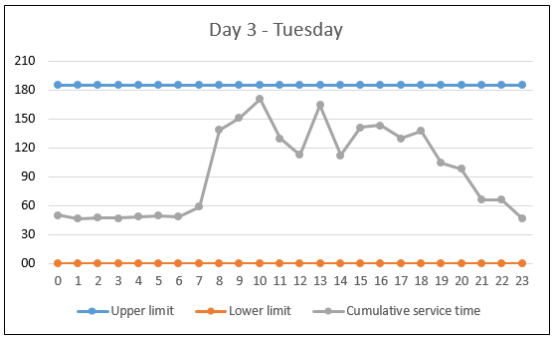


Figure 9: Series (b): Cumulative service time per hour for day 3

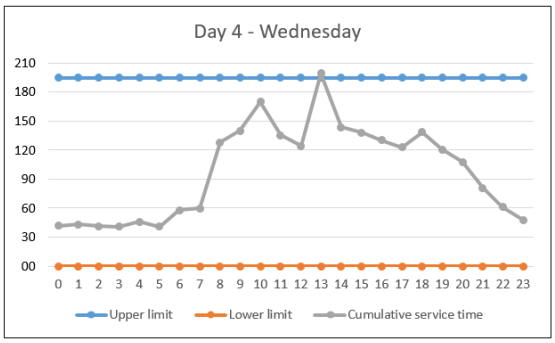


Figure 10: Series (b): Cumulative service time per hour for day 4

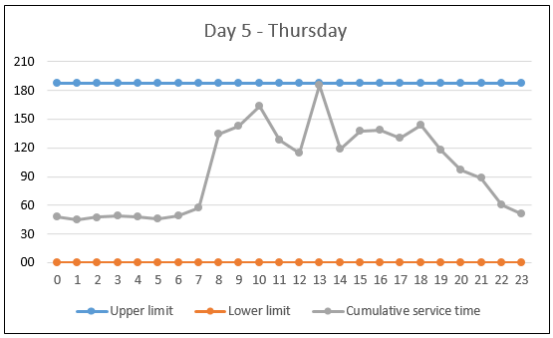


Figure 11: Series (b): Cumulative service time per hour for day 5

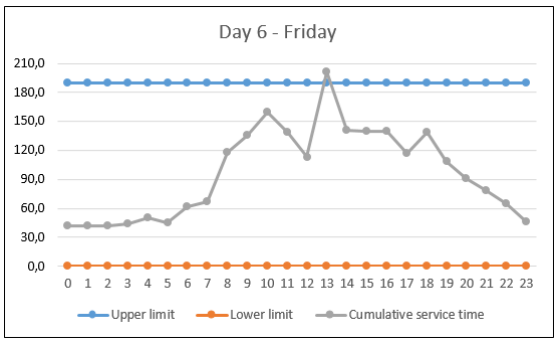


Figure 12: Series (b): Cumulative service time per hour for day 6

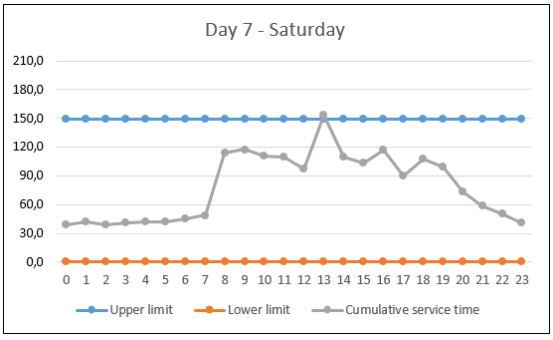


Figure 13: Series (b): Cumulative service time per hour for day 7

SIMULATION BASED ANALYSIS OF ECTOPIC PREGNANCY TREATMENT PROCESS TO SUPPORT PROCESS REDESIGN

Jānis Grabis
Institute of Information Technology
Riga Technical University, Kalku 1
Riga, LV-1658, LATVIA

Zane Grabe
Unit of Obstetrics and Gynecology
Riga East University Hospital
2 Hipokrāta Street, Riga, LV-1038, LATVIA

KEYWORDS
Business process simulation, healthcare simulation

ABSTRACT

Many healthcare delivery processes are managed according to deep-rooted rules and process redesign is challenging. Simulation can provide evidence to support redesign decisions. However, introducing simulation in organizations having limited previous experience with the method requires a lot of effort. In the case study reported, it is proposed to use business process based simulation. A business process is constructed jointly with domain experts and a simulation model is created directly from the business process model, where simulation specific features are represented in sub-processes. That streamlines model validation and communication of modeling results with stakeholders. Business process simulation is used to compare ambulatory treatment and hospitalization in the case of ectopic pregnancy treatment. Simulation results show that the ambulatory treatment should be adopted according to the cost minimization criterion. However, appropriate measures should be taken to mitigate potential negative effects exhibiting in the simulation results as an increase of waiting time.

INTRODUCTION

Simulation is a valuable tool for policy making (Jacobson et al. 2013). However, in many domains such as healthcare stakeholders expect extra assurances that recommendations provided are transparent and risk-averse. Business process modeling has emerged as one of the techniques bringing together various stakeholders in analysis of complex issues and combining managerial and technical perspectives of policy evaluation (Van der Aalst et al., 2003). It is also successfully used in healthcare (McNulty and Ferlie, 2002). BPMN is widely used standard language for representing business processes (OMG, 2011). Many business process management suites support simulation of processes developed (Pereira and Freitas, 2016; Wagner, 2015). However, simulation capabilities of BPM suites are often limited (Wagner et al., 2009; Wagner et al., 2017). These limitations apply to both: 1) inability to simulate all constructs used in modeling of rich business processes; and 2) inability to represented in BPMN some aspects necessary for simulation of complex processes (e.g., queuing, resource management). One solution is to use BPMN for initial representation of the process and implementation of the process simulation model in a fully-fledged simulation tool (Wang et al., 2009). However, that requires extra development effort and might result in detachment of business stakeholders.

This paper investigates application of simulation for evaluation of policy decisions in a healthcare organization unfamiliar with simulation. Simulation is necessary because the organization follows some well-entrenched policies and it is important to combine static analysis, dynamic analysis (e.g., visualization of process dynamics) and numerical analysis to show potential benefits of policy changes. Business processes based simulation is deemed suitable for initial exposure of the organization to these technologies.

The main requirements for simulation modeling in this case are: 1) model should be comprehensible for various stakeholders; 2) model should support evaluation from multiple perspectives; and 3) limited effort can be devoted for a pilot study. From the applied perspective, the objective of this paper is to develop a business processes based simulation model for evaluation of healthcare policy decisions for an organization with limited prior exposure to simulation. The particular case studied is about treatment of ectopic pregnancy and policy decisions should be made about switching from hospitalization to ambulatory care. This case requires that particular attention should be devoted to representation of exceptional events what is supported by BPMN though with limited support for simulation. Therefore, from the theoretical perspective the challenge is to simulate these exceptional events without sacrificing clarity of BPMN models and relying heavily on custom development.

The rest of the paper is organized as follows. Section 2 reviews some of the current research in healthcare simulation with emphasis on process improvement. It also describes the ectopic pregnancy condition and its treatment. Section 3 discusses business process based simulation. A ectopic pregnancy treatment process redesign case study is presented in Section 4. Section 5 concludes.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

BACKGROUND

This section discusses general aspects of using simulation in healthcare and introduces the particular process to be analyzed in this paper, namely, the Ectopic Pregnancy Treatment (EPT) process.

Simulation in Healthcare

Simulation is frequently used for operations management in healthcare. In the context of this paper, the most relevant aspects of healthcare simulation are evaluation of healthcare policies, process improvement and applications of business process modeling based simulation.

Hulshof et al. (2012) classifies types of operations management decisions made in healthcare. They indicate that demand for ambulatory care services is growing. Simulation is often used to investigate these issues. There is an increase of ambulatory surgeries, which are often shorter, less complex and less variable (Hulshof et al., 2012). Improvement of healthcare process requires balancing of efficiency and risk management (Zeigler et al. 2016). The authors have elaborated a simulation based framework for evaluation of healthcare processes to reduce cost and improve service quality. Mielczarek and Uziałko-Mydlikowska (2010) survey different simulation applications, which are categorized as epidemiology, health and car4e systems operation, health and care systems design, and medical decision making. The survey of simulation tools indicate that ARENA, which is a specialized process-oriented simulation tool, is most widely used.

Barjis (2011) identifies managerial simulation as one of types of simulation in healthcare. It concerns strategic planning and policy implementation. The authors identifies user acceptance as one of the main concerns of using simulation in healthcare and stakeholder involvement is critical for adoption of recommendations made as the result of simulation studies.

Antonachi et al. (2016) argue that business process improvement techniques are becoming more common in healthcare and simulation is an essential part of the improvement cycle. A business process model is transformed into a simulation model to enable quantitative and dynamic analysis of healthcare processes. Transformation processes are always associated with some information loss and a need to update the resulting model. Simulation based process reengineering has been successfully used in optimization of sterilization processes (Cassettari et al., 2013). In order to clearly highlight process inefficiencies and to improve the process, as-is and to-be models are developed and evaluated. The evaluation has resulted in development of a new sterilization strategy. Hamrock et al. (2013) point out that simulation is an important tool for improvement of healthcare processes to maximize operational efficiency without sacrificing service quality. Process improvement initiatives often are a result of intricacies of modern cost reimbursement policies. Simulation allows replacement of subjective decisions with decisions motivated by quantitative analysis.

Ectopic Pregnancy and Its Treatment

The EPT process treats an ectopic pregnancy. The ectopic pregnancy is a pregnancy outside of uterine cavity. Management of ectopic pregnancy has changed over the years (Murray et al. 2005). The guiding principle has become an early detection and a medical treatment rather than surgery. However, this condition is still potentially life-threatening and accounts for 4 to 10 percent of all pregnancy related deaths (Creanga et al. 2011). The overall incidence of ectopic pregnancy is approximately 20 per 1000 pregnancies.

The diagnosis of ectopic pregnancy is based upon a combination of risk factors of ectopic pregnancy, clinical presentation, measurement of the serum quantitative human chorionic gonadotropin (pregnancy blood test) and ultrasonography (Murray et al. 2005). The symptoms of ectopic pregnancy most often appear six to eight weeks after the last normal menstrual period. In the case of later detection of ectopic pregnancy sever life-threatening hemorrhage can occur due a tubal rupture.

Any fertile age patient presenting with pelvic pain and vaginal bleeding have to be evaluated for ectopic pregnancy. Main goals and steps of the investigation are confirming pregnancy with the pregnancy blood test, evaluation the patient for hemodynamic instability, ultrasonography for determination the site of pregnancy. Before to start a treatment of ectopic pregnancy, additional blood tests are indicated. These include blood type and screen, complete blood test, liver and renal function tests, in severe cases-coagulation tests.

The three approaches to the management of ectopic pregnancy are surgery, medication or expectant management. Approximately one-third of patients with ectopic pregnancy are candidates for medication – Methotrexate (MTX) therapy (Van Den Eeden et al. 2005). The remaining patients will require surgery due tubal rupture, large ectopic pregnancy, inability to comply with the follow-up of the MTX therapy. Some patients prefer surgery rather than the MTX treatment.

Table 1: Comparison of ambulatory care and hospitalization.

Policy	Advantages	Disadvantages
Ambulatory care	Cost effective No interruption of daily activities	Need for a good medical help access Late response Dependence on patient's self-discipline
Hospita-lization	Easy access Quick treatment for sever conditions	Expensive Necessary to skip daily activities Unnecessary add-on treatments

Most of guidelines advise to perform the MTX treatment in out-patient clinics but there are some hospitals where patients receive hospital based treatment (see Table 1 for comparison of policies). The MTX therapy is noninvasive, safe and cost effective option for ectopic pregnancy treatment. The overall success rate of medical treatment in properly selected patients is nearly 90 percent. During the treatment patients can have abdominal pain. Occasionally pain may be severe, but hemodynamically stable patients often do not need surgical intervention. A patient with severe pain may be further evaluated with transvaginal ultrasonography. There appears to be no clinical benefit from routine serial ultrasound examinations.

SIMULATION APPROACH

The simulation procedure followed combines features of typical simulation modeling projects (e.g., Law and Kelton, 2014) and business process modeling and improvement. Figure 1 indicates the main steps of this procedure with emphasis on aspects characteristic to this project.

Although the simulation procedure described focuses on the particular process redesign case, some of the principles can be applied in other projects as well. In particular, simulation models are developed using the variants based approach frequently used in business process management (Kumar and Yao, 2011), policy decision are explicitly represented in the model, the same model is used for both business process analysis and simulation without transformations and simulation modeling features are encapsulated in business process sub-processes.

Alternative policies to be evaluated are treated as process variants. Business process management literature suggests that business processes variants are either represented in one large model with conditions or developed in separated models (La Rosa et al., 2017). The former approach yields large models, which are more difficult to comprehend while redundancies of activities across multiple model variants are kept at minimum. The latter approach better shows each individual process variant while maintenance of the separate models is more difficult and the same activities are replicated in multiple process variants.

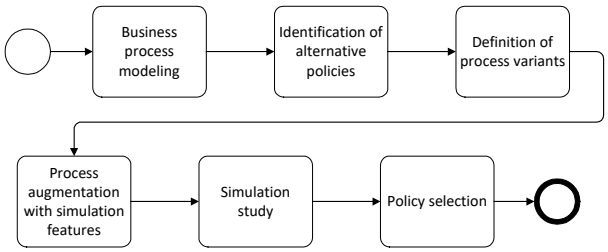


Figure 1: Business process modeling based simulation

Simulation model is implemented by augmenting the business process models. Representation of exceptional events is one of the key aspect in the EPT process. These events imply that exceptions can occur at any moment during the process execution and appropriate measures should be taken to treat them.

It is proposed to model these exceptional event following these steps:

1. Generate entity representing a patients
   a. At this moment generate a Bernoulli random variable indicated whether an exceptional event will occur for this patient;
   b. Generate a random variable representing a time moment when an exceptional event will occur.
2. Create a Collapsed sub-process with a boundary event. This sub-process is used to model exceptions and the boundary event catches exceptions thrown by the exception process. The process execution is interrupted upon caching the event and an exception handling tasks is invoked.
3. Create a sub-process simulating exceptional event
   a. Wait till exception activity delays the sub-process execution till exception should be thrown as indicated by the generated random variable;
   b. If the Bernoulli random variable indicates occurrence of the exceptional event then an exception is thrown.

This procedure is illustrated in Figure 2 showing the main process (left side) and the exception simulation process elaborated in the sub-process (right side).

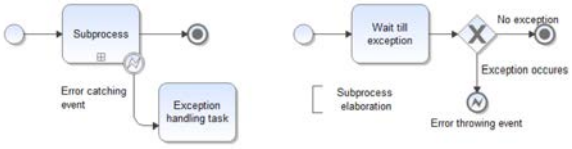


Figure 2: Boundary event in BPMN.

CASE STUDY

The simulation modeling is used to evaluate EPT policies at a large hospital in Latvia. This hospital treats about ectopic pregnancy cases annually. Due to long entrenched policies, patients requiring EPT are hospitalized despite ample evidence that this approach is costly and provides few benefits to the patients. The aim of this simulation study is to provide preliminary evidence that the ambulatory treatment is preferable in majority of cases. Although the hospital is ISO certified, neither business process modeling nor simulation have been previously used in process analysis and improvement

Process Model

The general treatment process introduced in Section 2.1 is mapped using BPMN (Figure 3). The process starts with a doctor visit and regular blood testing and ultra sound investigation for any patient regardless her condition. Ectopic pregnancy is diagnosed as a result of this testing to around 2% patients and the doctor decides on pursuing medical treatment or proceeding with surgery. Surgery is beyond scope of this paper and the

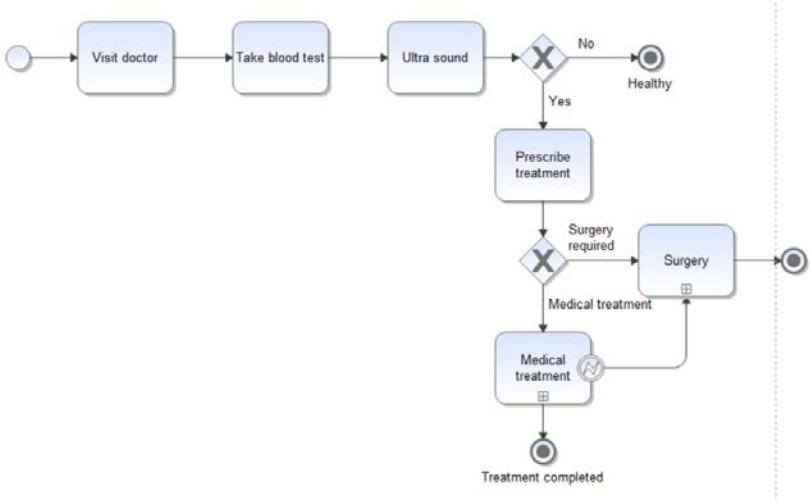


Figure 3: EPT process.

further investigation focuses on medical treatment. The medical treatment is represented as a sub-process in the overall model and it is also indicated that in the case of unsuccessful medical treatment surgery still may be required.

The medical treatment requires continuous observations of patients who are required to have regular blood tests to check treatment efficiency and potential side effects. Medical staff uses test results to decide on further actions. During the treatment process, patients can experience sudden pain. If that is a case and the pain is persistent or severe, the patients seek emergency medical assistance. The medical treatment can be provided either ambulatory or at hospital. Currently, the hospital practices hospitalization what allows for greater responsivity. On the other hand, ambulatory treatment often preferred by other medical institutions might be more comfortable to patients and should be less expensive.

Data and Experimental Planning

The objective of the simulation study is to compare policies of ambulatory treatment and hospitalization. Two main performance measures used are treatment cost and waiting time. The treatment cost is comprised of hourly personnel cost, cost of testing aids and materials and infrastructure cost. The study focuses only on cost factors for which differences between ambulatory treatment and hospitalization might be observed and calculation of actual cost of the whole process is intricate and beyond scope of this study. The infrastructure cost (i.e., cost accrued for having a patient in the hospital) is incurred only in the case of hospitalization. In the case of ambulatory treatment, there could be extra costs of providing emergency treatment in the case of pain. The waiting time is a time spent by patients waiting to see doctor or nurse. Even if appointments are scheduled, there are frequent delays causing patients to experience anxiety. It is assumed that scheduling difficulties are mostly experienced in the case of ambulatory treatment (i.e., if a doctor is late a hospitalized patient does not always need to leave her premises).

Two main experimental factors are considered: 1) medical staff availability; and 2) frequency of pain events. The medical staff availability is associated with the waiting time. The frequency of pain events affects a need for additional and emergency care. The full factorial design combining policy, medical staff availability and frequency of pain events is considered. Simulation is performed for a time-span of one year with random patient arrivals resulting in about 300-400 cases of medical treatment. The business process and simulation models are developed using iGrafx Process.

Data gathering if performed in three steps:

1. Process mapping;
2. Analysis of medical records;
3. Literature data and indirect evidence.

During the process mapping, duration of procedures and their costs are determined. Analysis of medical records yields data about number of patients, frequency of exceptional events (e.g., pain and rupture) and treatments used. These data also yielded interesting insights about medicine taken by different patients and other medical practice what are beyond direct interest of the simulation study. Unfortunately, there were no data about medical staff availability. Indirect observations were used to provide estimates for these data and they were treated as experimental factors to evaluate impact of the assumptions made on the simulation results.

The values of experimental factors considered are reported in Table 2. It is assumed that the low and high values of availability for the ambulatory treatment are lower than for the hospitalization policy. The frequency of pain percentage means the percentage of patients experiencing pain. A single patient can experience pain multiple times.

Representation of Treatment Policies

The medical treatment sub-process of the overall EPT process is further elaborated (Figure 4). Since there are only two policies to be evaluated, it is decided to model

both process variants, namely, ambulatory treatment and hospitalization in a single business process. Initially, a common part of the process is modeled. The general idea of the process is that current condition of a patient is evaluated, medicine is taken and new tests are performed to benchmark treatments outcomes. If treatment is not successful it is repeated but no more than two time. Otherwise the patients enter continuous monitoring process and hopefully successful pregnancy.

Table 2: Values of experimental factors.

Policy	Factor	Low value, %	High value, %
Ambulatory treatment	Availability of medical staff	50	75
Hospitalization	Availability of medical staff	75	90
Both	Frequency of pain events	5	20

The common part is supplemented with the observation branch, which simulates exceptional events occurring during the treatment. These exceptional events are pain occurrences and rupture. The exceptions simulation sub-process is shown in Figure 5. If rupture occurs, this event is captured in the overall EPT process and lead to surgery. If pain is observed, this event is captured in the medical treatment sub-process. The exception simulation process is designed following the approach established in Section 3. For every patient (i.e., transaction in the simulation model), potential occurrence of pain or rupture is evaluated upfront. That includes determining whether the exception should be thrown and when. The rupture and pain activities use this information to decide whether and when to throw an exception and to interrupt the normal flow of the medical treatment process. The exceptions simulation sub-process is not necessary for communicating the business process model with stakeholders but is introduced for simulation modeling purposes.

Refinement of the medical treatment process also includes specifying the pain treatment procedures, which are different depending on the policy. Pain treatment can be accommodated as a part of the overall treatment for hospitalized patients. However, emergence measures might be required for patients having ambulatory treatment. That includes an extra visit to the hospital probably using emergency care services. The pain treatment branch is merged with the main branch at Take Day 4 Blood Test activity. That is a simplification to make model better comprehensible because the pain treatment can occur at every stage of the process.

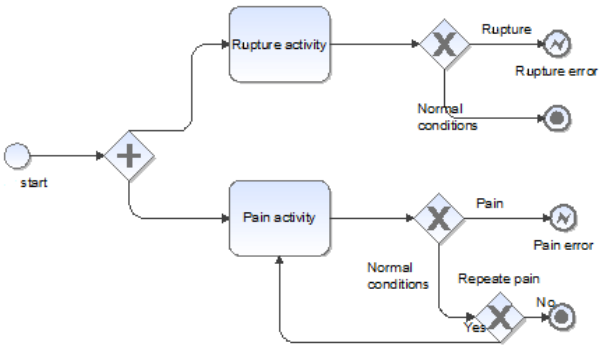


Figure 5: The exception simulation process.

EXPERIMENTAL RESULTS

Simulation experiments are carried out according to the experimental design. The waiting time and cost performance measures for each experiment are reported in Table 3. The cost is expressed in relative units taking hourly rates as the basis. The results indicate that as expected the waiting time is higher in the case of ambulatory treatment and cost is higher in the case of hospitalization. The ambulatory treatment yields slightly higher personnel cost because of involvement of emergency services. However, cost of infrastructure is almost 75% of the total cost making hospitalization an unviable approach according to the cost measure. Medical staff availability has significant impact on waiting time but does not affect cost (because measures to ensure higher availability are not taken into account). The impact is similar for both policies. The pain level influences both waiting time and cost because patients need to take more procedures (and probability of waiting is higher) and additional treatments require

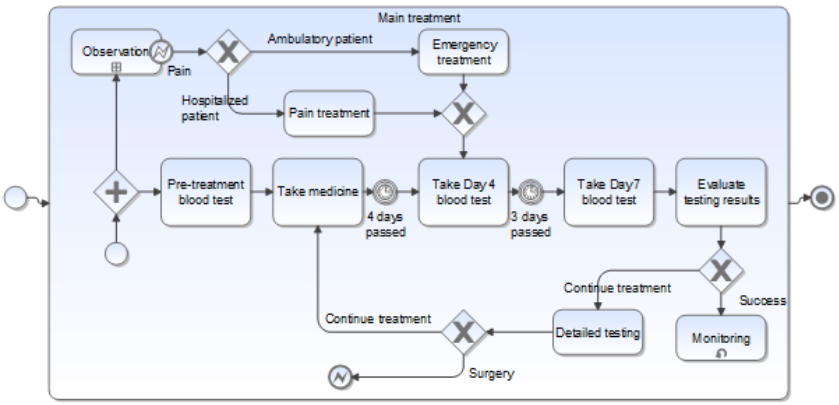


Figure 4: Elaboration of the medical treatment sub-process.

additional personnel hours. Figure 6 shows the waiting time and cost performance measures averaged over all experiments. It confirms the finding that switch to the ambulatory treatment would result in major cost savings. The waiting time increase and the resulting increase in anxiety could be addressed by improving medical staff scheduling and proving extra assurances to patients that their concerns will be addressed in a timely manner.

Table 3: Simulation results.

#	Treatment	Pain level	Availability	Waiting time	Cost
E1	Ambulatory	Low	Low	1.06	57
E2	Ambulatory	Low	High	0.43	57
E3	Ambulatory	High	Low	1.2	63
E4	Ambulatory	High	High	0.51	63
E5	Hospitalization	Low	Low	0.36	188
E6	Hospitalization	Low	High	0.18	188
E7	Hospitalization	High	Low	0.37	191
E8	Hospitalization	High	High	0.19	191

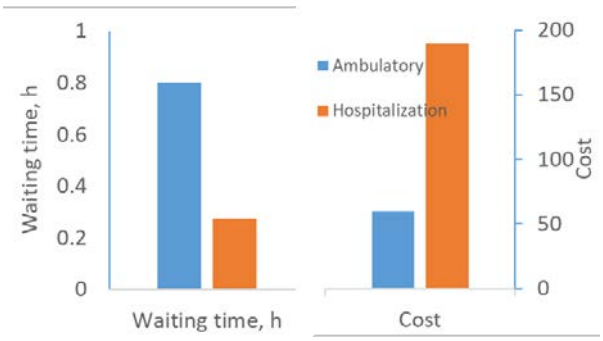


Figure 6: Waiting time and cost depending on policy.

There are some concerns that in the case of the ambulatory treatment some patients would not take medicine as apportioned or violate other treatment requirements. That could result is more frequent pain events. First and foremost this could be alleviated by rising patient awareness and improved communication. From the simulation perspective, this issue is addressed by additional experimenting on waiting time dependence on frequency of pain events (Figure 7). This time availability is 75% for both policies and frequency of pain events is varied from 1% to 30%. It shows that the waiting time increases depending on the frequency of the pain events in the case of the ambulatory treatment. However, this increase is highly variable.

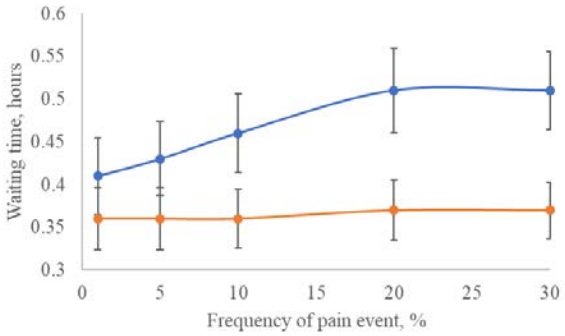


Figure 7: Waiting time dependence on frequency of pain events with 95% confidence bounds provided.

CONCLUSION

The paper has presented business process modeling based simulation of healthcare processes. In order to introduce business process modeling and simulation as a decision-making tool, it is required to use modeling techniques attainable for different stakeholders. The simulation modeling should justify the associated effort and to provide persuading evidence of a positive effect of process improvement initiatives suggested. In order to achieve that business process simulation is used without transforming business process models into platform specific simulation models or relying on custom development. The simulation specific features are encapsulated as sub-processes. The BPMN event handling features enable for representation of exceptional events characteristic to healthcare processes. The particular healthcare process analyzed in this paper is the EPT process. The objective of the analysis is to compare ambulatory treatment and hospitalization. The simulation results show that from the cost perspective the ambulatory treatment is more efficient. However, this policy also yields higher waiting times what might be negative for patients' well-being and treatment efficiency. These results imply that transition from one policy to another should be accommodated with extra effort to provide timely consultations and relief to patients treated ambulatory.

The simulation results confirm conclusions made in medical literature. However, it was important to highlight these differences for the particular hospital to persuade decision-makers. The business process model was developed jointly with experts from the hospital and this approach was confirmed to be beneficial for validating the model and communicating the modeling results.

The paper's limitations are that not enough data are available to assess all parameters of the simulation model and process simulation has limited scope to evaluate treatment outcomes. Particularly, there is limited data available on medical staff availability what is a major parameter in the study. Additionally, the healthcare process redesign is significantly influenced by medical and human considerations, which are not always represented in the simulation model. The

simulation modeling results can be used just as an additional factor for weighting adoption of specific healthcare policies.

REFERENCES

Antonacci, G., Calabrese, A., D'Ambrogio, A., Giglio, A., Intrigila, B. and N.L. Ghiron. 2016. “A BPMN-based automated approach for the analysis of healthcare processes". In *Proceedings of 25th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, WETICE 2016, 124-129.

Barjis, J. 2011. “Healthcare simulation and its potential areas and future trends”. *SCS M&S Magazine*, 2(5): 1-6.

Cassettari, L., Mosca, M., Mosca, R. and F. Rolando. 2013. “An healthcare process reengineering using discrete event simulation". *2013 World Congress on Engineering and Computer Science, WCECS 2013, Lecture Notes in Engineering and Computer Science*, 2;1174-1179.

Creanga AA, Shapiro-Mendoza CK, Bish CL, et al. 2011. “Trends in ectopic pregnancy mortality in the United States:1980-2007”. Obstet Gynecol 117:837.

Hamrock, E., Paige, K., Parks, J., Scheulen, J. and S. Levin. 2013. "Discrete event simulation for healthcare organizations: A tool for decision making". *Journal of Healthcare Management* 58(2):110-124.

Hulshof, P.J. H., N. Kortbeek, R.J. Boucherie, E.W. Hans, P.J.M. Bakker. 2012. “Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS”. *Health Systems* 1(2):129-175.

Jacobson, S.H., Hall, S.N. and J.R Swisher. 2013. “Discrete-event simulation of health care systems”. *International Series in Operations Research and Management Science* 206:273-309.

Kumar, A. and W. Yao. 2011. “Design and management of flexible process variants using templates and rules”. *Computers in Industry* 63 (2):112-130.

La Rosa, M., Van Der Aalst, W.M.P., Dumas, M. and F.P. Milani. 2017. "Business process variability modeling: A survey". *ACM Computing Surveys* 50(1): 2:2-2:45.

McNulty, T. and E. Ferlie. 2002. Reengineering Health Care: The Complexities of Organizational Transformation. *Oxford University Press.*

Mielczarek, B. and J. Uziałko-Mydlikowska. 2012. "Application of computer simulation modeling in the health care sector: A survey". *Simulation* 88(2):197-216.

Murray H, Baakdah H, Bardell T, Tulandi T. 2005. “Diagnosis and treatment of ectopic pregnancy”. CMAJ 173:905.

Law, A. M., and W. D. Kelton. 2014. *Simulation Modeling & Analysis.* 5th ed. New York: McGraw-Hill.

OMG. 2011. “Business Process Model and Notation.” http://www.omg.org/spec/BPMN/2.0/.

Pereira, J.L. and A.P. Freitas. 2016. “Simulation of BPMN process models: Current BPM tools capabilities”. *Advances in Intelligent Systems and Computing* 444:557-566.

Van Den Eeden SK, Shan J, Bruce C, Glasser M. 2005. “Ectopic pregnancy rate and treatment utilization in the large managed care organization”. Obst Gynecol 105:1052.

Van der Aalst, W.M.P., A.H.M ter Hofstede and M. Weske. 2003. “Business Process Management: A Survey”. In *Proceedings of International Conference on Business Process Management, BPM 2003*, 1-12.

Wagner, G. 2015. "Tutorial: Information and Process Modeling for simulation". In *Proceedings - Winter Simulation Conference,* 103-117.

Wagner, G., Nicolae, O. and J. Werner. 2009. "Extending discrete event simulation by adding an activity concept for business process modeling and simulation". In *Proceedings - Winter Simulation Conference*, 2951-2962.

Wagner, G., Seck, M. and F. McKenzie 2017. "Process modeling for simulation: Observations and open issues". In *Proceedings - Winter Simulation Conference*, 1072-1083.

Wang, T., Guinet, A., Belaidi, A. and B. Besombes. 2009. "Modelling and simulation of emergency services with ARIS and Arena. case study: The emergency department of Saint Joseph and Saint Luc hospital". *Production Planning and Control* 20(6):484-495.

Zeigler, B.P., Carter, E.L., Molloy, O. and M. Elbattah. 2016. "Using simulation modeling to design value-based healthcare systems". In *Proceedings OR58: The OR Society Annual Conference*, 33-48.

AUTHOR BIOGRAPHIES

JĀNIS GRABIS is a Professor at the Faculty of Computer Science and Information Technology, Riga Technical University, Latvia. He obtained his PhD from the Riga Technical University in 2001 and worked as a Research Associate at the College of Engineering and Computer Science, University of Michigan-Dearborn. He has published in major academic journals including OMEGA, European Journal of Operational Management, International Journal of Production Research, Computers & Industrial Engineering, IEEE Engineering Managenet Review and others. He has been a guest-editor for two top academic journals and member of the program committee of several academic conferences. His research interests are in supply chain management, enterprise applications and project management. His email address is: `grabis@rtu.lv`.

ZANE GRABE is a gynecologist and ultrasound specialist at Rigas Eastern Clinical University Hospial. She obtained Medical Doctor’s Diploma from the Latvian Academy of Medicine in 1995. She teaches Gynaecology at the Rigas Stradins University, Department of Obstetrics and Gynaecology.

Simulation of Intelligent Systems

BEHAVIOR TREE BASED KNOWLEDGE REASONING FOR INTELLIGENT VESSELS IN MARITIME TRAFFIC SIMULATIONS

Volker Gollücke
Daniel Lange
Axel Hahn
Department of Computer Science
CvO University of Oldenburg
26131, Oldenburg, Germany
E-mail: volker.golluecke@uni-oldenburg.de

Sören Schweigert

Cooperative Mobile Systems
OFFIS e.V.
26131, Oldenburg, Germany

KEYWORDS
Maritime Knowledge Representation, Behavior Modelling, Behavior Trees, Maritime Traffic Simulation, S-100

ABSTRACT

For simulation based verification and validation (V&V) of maritime system designs, the system under analysis is exposed to a variety of traffic scenarios. Usually bridge and shipping simulators do not provide intelligent behavior for the simulated ships. Instead, they use simple route following techniques, or just follow a given direction. In automated V&V scenarios, a lot of different simulation runs must be executed e.g. to test new assistance systems in various situations. To cover the needed number of important situations, an automated behavior of target ships is needed.
This paper presents a technique to configure and calculate realistic and intelligent ship behavior. Each ship has its own knowledge about the environment and uses this knowledge to decide what kind of behavior the ship shows using the Behavior Tree technique.

INTRODUCTION

Many V&V scenarios in the maritime domain need realistic ship behavior. Under different environmental conditions, different ship behavior is required. As in other domains, assistance systems as well as their functional safety are becoming increasingly important.
For verification and validation in engineering assistance systems, German maritime industry and research institutes have launched a test bed for e-Maritime applications (eMIR – eMaritime Integrated Reference Platform) (Hahn and Noack 2016). eMIR provides, among others, services for research and industry projects to ensure the functional safety of new systems. For this purpose, eMIR is divided into a virtual simulation based platform (named HAGGIS) (Schweigert et al. 2014) as well as a physical platform (called LABSKAUS) for field testing.
In this paper, we will address the virtual simulation part of eMIR that is capable of testing new assistance systems like collision avoidance systems. Core element of HAGGIS is a maritime traffic simulation (MTS) (Hahn 2015). For engineering, complex systems, validation and verification we will make use of realistic simulated system environments. For this purpose, HAGGIS and in particular the MTS offer a framework to create a variety of scenarios to be simulated with the system under test. While maritime simulation systems for bridge crew training usually use a simple route following approach for target ships, V&V scenarios need target ships that behave realistically and adopt to 'relevant' parts of the environment for test automation. In terms of realism, the distinct behavior (amongst target ships) such as evasive or overtaking maneuvers, can be simulated. Therefore, a structured way of configuring and simulating target ships' behavior is required.
A target ship's environment e.g. comprises of dynamic terrain, with its bathimetry, currents and waves in water and air, as well as movable or static objects, such as other ships, landmass and wind farms. Maritime safety information is issued from shore based systems. Additionally, regularitions, e.g. anti-collision regulations (ColRegs (Organization 2003)) have to be followed.
This paper proposes to simulate each vessel behavior by an agent which uses an extended Behavior Tree technique (Ogren 2012; Colledanchise et al. 2016).
In the following, an overview on behavior modelling for V&V of maritime systems is given and requirements for behavior simulation are described. Then, we propose the concept of applying Behavior Trees and usage of ships own belief about the environment. An application example in a ship's overtaking condition is covered for evaluation of the concept.

CURRENT STATE OF BEHAVIOR MODELING AND SIMULATION IN THE MARITIME DOMAIN

Like mentioned in the introduction, for many traffic simulations, it is often sufficient to specify the course or a route to be followed. In addition to these simulators, there are other scientific approaches for describing the behavior of target ships.
Köse et al. (Köse et al. 2003) presented the simulation of maritime transport in Istanbul's Strait. Their simulator contains several assumptions. The ship's time of arrival is evenly distributed, it is forbidden to overtake each other

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

and the intended speed for the ship in the strait is fixed at 10 knots.

Fan and Cao (Fan and Cao 2000) presented a model that calculates the throughput of a waterway from the average ship size, the average ship speed, the average separation distance between ships and the probability that each type of ship appears in the waterway. While different parts of a waterway may have different average speeds due to factors like the physical environment or other. The average speed of vessels based on the entire waterway may only be a good estimate for calculating capacity of small sea areas.

The SMARTS (Ship-with-a-captain MARine Traffic Simulation System) of Osaka University is a multi-agent model for the simulation of shipping traffic in the Bay of Osaka and the Bay of Tokyo (Japan) (Hasegawa et al. 2001). It can automatically generate a maritime traffic flow based on statistical data to control the simulation runs. Hasegawa et al., the inventors of the system, used a fuzzy expert system to navigate ships through the waterways.

In addition to these examples from the maritime domain, two newer technologies for describing behavior in the domain of artificial intelligence have emerged in recent years. On the one hand, the Utility AI concept ((Rabin 2013),p. 113ff) which carries out the selection of behavioral patterns via an evaluation function and the Behavior Tree concept, which decides which behavioral pattern is carried out by means of conditions within a tree hierarchy. Since Behavior Trees were developed in the commercial sector (created by the company BUNGIE for the development of HALO 2 (Isla 2005)) and were subsequently adapted by many developers for their own purposes, there is no industrial standard. Due to the flexible possibilities for behavioral modeling, Behavior Trees have already been scientifically investigated and partially defined in different publications (Colledanchise et al. 2016; Ogren 2012).

Both approaches cover different objectives. While the AI Utility concept uses the scoring function to offer a bigger variance in the choice of behavioral patterns, the Behavior Tree concept offers a design option that is easier to understand for non-technicians.

REQUIREMENTS FOR SIMULATION BASED V&V

From the given motivation, different requirements for a V&V simulation system can be derived.

The most obvious requirement is for accelerated execution of simulation runs to be able to create a meaningful coverage of simulated state spaces, required by the V&V methods in reasonable time.

Related to this, there is the demand for a high coverage of the simulated state space by different test vectors. This means that a large number of different scenarios must first be created and then simulated. Classical simulations in the maritime domain are designed for interaction with a human user, who defines the exact scenarios and, if necessary, adjusts the scenario during runtime of the simulation. This can be, for example, manually changing the trajectory of a ship by which it reacts to the behavior of the system under investigation. While this works out fine for training of bridge crews, it is not feasible for automatic verification and validation of new assistance systems, as it limits the number of executed simulations.

To summerize, simulation based V&V requires the simulations to behave as realistically as possible, adapting dynamically to the behavior of the system under test and under consideration of the applicable regulations (e. g. collision prevention rules and maritime traffic regulations). Agent based approaches have been proven to fulfil this requirement. However, within a simulated environment, it can be assumed that all necessary information is available to an agent. This is not the case in reality. Therefore, it has to be possible to filter the existing knowledge to restrict the knowledge of the simulated agents. This initially refers to the declarative knowledge of the agent, but also holds for the methodical knowledge of the agent. An example of this is forgetting or ignoring traffic rules. However, this is not further covered in this paper.

With regard to the modelling and simulation of vessel behavior in a maritime traffic simulation we found some additional requirements to be considered.

Based on the investigation of the collision prevention rules (Organization 2003) and the work "Distribution of debts in case of ship collisions" (Bierwirth 2004),"Collisions and their Causes" (Cahill and Britain) 2002) as well as "Managing Collision Avoidance at Sea: A Practical Guide" (Lee and Parker 2007) we derived three basic behavioral aspects that need to be mapped within a behavioral component.

1.) **General driving behavior**
    Describes the behavior of the vessel with regard to preferences such as efficiency and forward-looking driving.
2.) **Goal achievement behavior**
    Describes the behavior of the goal achievement of set target coordinates. Examples would be a frequent correction of the course.
3.) **Collision avoidance behavior**
    Describes the the ship's rules for collision avoidance. The ship could thus assess, if it is necessary to keep the vehicle in the event of an imminent collision and plan and carry out necessary overtaking maneuvers independently.

These behavioral aspects have in common that they can be expressed with a set of rules, which decide, based on information from the surrounding environment, what kind of actions are executed next. Also, these behavioral aspects may be composed of a series of simpler tasks. The overtaking task, as part of the collision avoidance behavior for example consists of: detecting the demand for overtaking, the overtaking manoeuvre itself and to trace back to the original route. During each of these partial tasks, additional task such as general collision prevention must be considered. This leads to the requirement that different behavior tasks need be

combined and executed in conjunction while maintaining a hierarchical, defined order.
As a maritime traffic simulation consist of many vessels and while they should behave realisticly, they also should show differences when handling a situation like overtaking another ship. Some captains target safety while other target efficiency, in terms of the time to reach their goal. Therefore, another requirement is to have the possibility to give different behavior aspects different weights while designing and executing them.
In addition, the possible large number of vessels leads to the need to setup different behaviors or different weights as easily as possible, for example by reusing previously created behaviors.
Within the next sections, we will present an approach that fulfils those requirements. First, we introduce our maritime simulation system MTS and how it represents the available knowledge about the environment, and how each individual agent build up its internal believes about this environment. Later, we will present, how Behavior Trees can be used to represent the procedural knowledge required to execute intelligent behavior and how they take advantage of the chosen environment representation.

KNOWLEDGE REPRESENTATION FOR MARITIME SIMULATIONS

The Maritime Traffic Simulation (MTS) as part of the virtual testbed HAGGIS, consists of two main components (see Figure 1). Those components are the World component, which represents the descriptive knowledge about the simulated surroundings and a set of vessel agents, where each agent represents one vessel inside the simulation in terms of a multi-agent simulation.

The *World* component can be seen as the ground truth of the simulated scenario that keeps all the static and dynamic knowledge about the environment. Thereby, the *World* is represented using the new hydrodynamic standard S-100, developed by the international Hydrographic Organization (IHO) in 2010 (IHO 2015). This generic standard defines a way to describe (maritime) feature types and information types in a structured and interoperable manner. Features include for example the knowledge from sea charts for static objects, water depths and traffic areas and rules for a specific sea area. Other features expressed with means of a S-100 conform standard are for example weather data and forecasts or maritime safety information (MSI) for nautical specialists. Those MSI in turn indicate among others, floating obstacles or malfunctioning of nautical equipment. Information, expressed as information types, provide additional information for features, like data quality or meta data, providing information about the data capturing process
As a standard with a very strong geographical reference, the IHO S-100 Standard is based on the ISO 19000 standard series and specializes, the product specifications from ISO 19131. Each product specification represents its own standard. For example, the nautical charts are standardized by the Product Specification S-101 (Electronic Navigational Charts), whereas the Standard S-124 will represent important navigational warnings (MSI - Maritime Safety Information) in the near future.

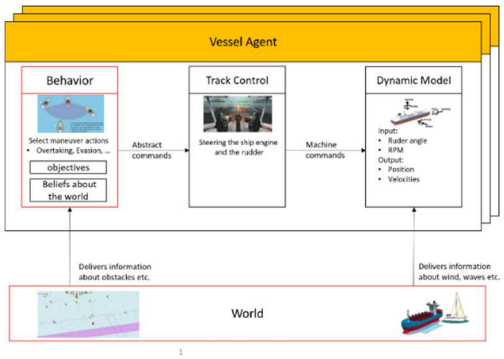


Figure 1: Common structure of vessels agents in HAGGIS

In addition to the geographical reference, the S-100 standard was also developed with the aim of interoperability between various standards in the same family. This includes a global registry in which all standard-compliant data types, associations and attributes (A) can be stored. The specialty about the S-100 is that the data types stored in the registry and especially the attributes and associations stored there can be reused in other standards of the S-100 family. This reuse of standard parts is intended to ensure that the vocabulary and the corresponding semantics are harmonized across the board. In concrete terms, this means that when two standards, as part of the description of a ship, refer to a MMSI (Maritime Mobile Service Identity), the meaning of MMSI is the same in both standards.
To express the knowledge represented by the S-100 standard, we use the following notation:

$$PS = [FT, IT, A] \quad (1)$$

Where FT represents an S100 Featuretype, IT contains the available information types and A presents a set of attributes that can be used to characterize the feature and information types.

$$A = [n, ft] \mid ft \in FT \ \vee \ ft \in IT \quad (2)$$

Those attributes can be described by their name (n) and a given type (ft). Actually, within the S-100 standard, they do contain additional information like multiplicities or descriptions, however those information are not needed for the presented approach.
Using the Product Specification (PS), we can describe the representation of the knowledge as the following tuple:

$$K^{G|A} = [F, I, AB] \quad (3)$$

Where K^G represents the global knowledge, available within the simulation system and thus the Ground truth and K^A represents the knowledge of an agent. Within this knowledge representation, set F represents the set of available features and I the set of available, additional,

information described by an information type. Thereby features and information share the same representation and as we tend to mainly use features within the simulation we will combine both in the following sections.

$$F = [n, p, ft, ab] \mid p \in F \ \wedge ft \in FT \quad (4)$$

$$ab \in AB = [a, v] \ \mid a \in A \ \vee a = f(ab_i, ab_j) \quad (5)$$

Within the S-100 each feature and thus each information is buildup of a name (n), an optional parent (p), the feature type (ft) and a set of attribute bindings (ab), as described in equation 4. Using this representation and in particular by introducing the parent, the knowledge about the vessel can be seen as a tree structure itself.
The attribute binding, as shown in equation 5 is represented through a tuple of attributes (a) and their values (v). This represents the S-100 concept to reuse attributes (A) as well as their semantics at different positions. In addition to the normal attribute binding we also allow to conclude knowledge from two other attribute bindings.
The second part of the MTS is a set of vessel agents.
Each vessel is an independent component that interacts on its own with the environment and possibly with other independent agents. Nevertheless, all vessels agents follow a common structure as shown in the upper part of Figure 2. In the MTS, each agent is composed of three parts, which can be combined and configured almost arbitrarily with each other. This allows a large diversity of ships to be modelled and simulated.
These three parts are

1) The intelligent behavior, which can be compared to the captain of a ship. It uses its internal knowledge about its environment to reach its goals in an intelligent and realistic manner. We will discuss this part of the text in the following sections.
2) The so-called Track-Control, which is a bridge between the abstract, intelligent behavior as well as the technical - physical behavior of the ship and
3) the physical behavior of the ship in its environment. This includes a simulation of the ship's engines and rudders, as well as an application of the induced forces. Those are applied in combination with the physical properties of the water and other environmental factors such as wind and current.

As already mentioned, in this approach, the track control represents an abstraction layer between the intelligent behavior of the ship and the possibly complex physical interactions of the ship. It translates abstract commands, such as those given by the captain, into concrete machine and rudder actuations and is also able to distribute them to several machines and rudders if required.
Similar to a single ship, each behavior specifies a certain structure as shown in Figure 2. According to Figure 2, each ship has its own representation of its environment (K^A). That is usually a subset of the knowledge from the World Component (see equation 6).

$$K^A = (K^G \cap K^L) \cup K^S \cup K^E \quad (6)$$

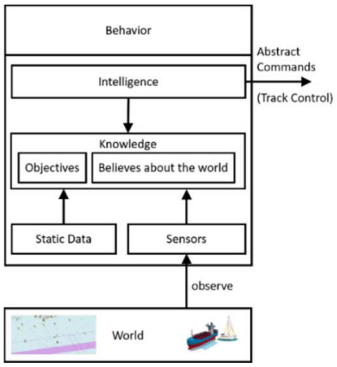


Figure 2: General structure of the behavior component

The agents knowledge is usually captured by the sensors of the agent as in the classic agent models (see (Russell and Norvig 2003)). In the maritime environment, these sensors include, in particular, radar sensors and the automatic identification system AIS, as well as sensors that monitor the internal condition of the ship. Those sensor readings are expressed through K^L. Together with the static knowledge about the ship (K^S), such as its dimensions or the knowledge from nautical charts, it represents the agent's beliefs about its current environment.

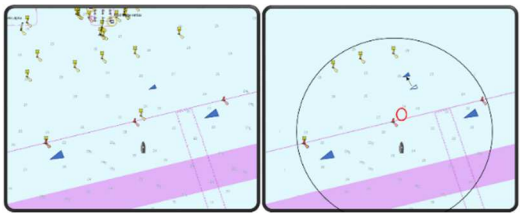


Figure 3: Ground Truth Environment (left) and believe about the current environment (right).

It is important to mention that the vessel's internal representation does not necessarily have to be a correct assumption about the surrounding, but may contain misfits (K^E) based on faulty sensor readings (as described in (Schweigert et al. 2014)) or outdated data (e. g. missing updates of the charts), and can thus deviate from the Ground Truth of the Simulation, as shown in Figure 3.
The figure displays the Ground Truth on the left side and the belief about the current environment on the right side. It shows objects like other vessels, buoys and depth information. The black circle in the figure represents the range of the belief, the red circle a missing object from the Ground Truth environment and the arrow a wrong assumption of a position of another vessel.
For the internal representation of the assumptions, the already presented S-100 based data model is also used within the vessel agent.

USING BEHAVIOR TREES AND ENVIRONMENTAL BELIEVES TO MODEL AND SIMULATE SHIP BEHAVIOR

Behavior Trees use structures in the form of directed, cycle-free graphs. Within the scope of this work, the graphical notation from the work of Ögren and Millington (Ogren 2012) are used.
Two nodes connected by edges are related to each other. A distinction is made between six types of nodes. If it is a node that is not a leaf, it can be of type Selector (executes its child nodes one by one until one of the children had success with its execution), Sequence (also executes its child nodes one by one but if one child node fails the following are not executed anymore), Parallel (Nodes of the type Parallel run all child nodes concurrently). If it is a leaf, it is either of type Action (executes a predefined action e.g. start overtake maneuver) or Condition (check whether a condition has been entered e.g. is a ship in front of us). The last type of a node, the Random Node Selector will be introduced in the section "Introduction of probabilistic selection strategies".
At runtime, the root of the used Behavior Tree generates a signal, also called tick, and sends it through the tree. The tick follows the specifications of the depth search and thus only traverses into depth, whereby the nodes can be given a fixed hierarchy. In the leafs, calculations are finally made and the defined behavior is executed.

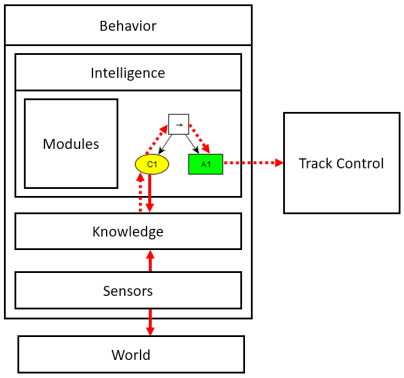


Figure 4: Behavior Tree access to ships belief

Considering the structure of the Behavior component of the Vessel Agent, Behavior Trees are the agent's rule set and are therefore classified under "Intelligence", as shown in Figure 4.
Since conditions are used to direct the flow of control, they rely on ship beliefs (solid red arrows) to make decisions based on returned information (dotted red arrows). Condition C1 uses information from the knowledge of the vessel. The information is transferred and evaluated in C1 to a decision. If the decision is negative, the subtree is canceled. Otherwise the corresponding action A1 is executed and commands are delegated to the track control component.
Using this, a Behavior Tree can be expressed using the following recursive equation 7, with $t \in$ *{Selector, Sequence, Parallel, Action, Condition, Random}* describing the type of the node, and $r \in \{Succsess, Failure, Running\}$ describing the result of the tree.

$$BT = [t, r, BT] \tag{7}$$

At this point the Behavior Tree and especially the condition can take advantage of the modular description of the used S-100 based data model and it attribute bindings, e.g. reusing attributes in features.
For this purpose, we can further specify the conditional as well as the action node types, and how they use the agents' as well as the globally available knowledge. In case the Behavior Tree represents a conditional tree (BT_C), the result is written as a function, defined over attributes available within the agent's knowledge.

$$BT_C = [t, r(f(a^A) \rightarrow \{true, false\}), \emptyset\ ] \mid a^A \in A^A \tag{8}$$

That is: Within a Feature Catalogue, we have defined the two attributes Speed Over Ground and Course over ground which are used in different types of vehicles, like vessels, airplanes, helicopters but also within floating obstacles. Since the semantic is known for those two attributes, we do not have to care, whether we are looking onto a vessel or a floating obstacle to determine if we have to avoid this obstacle but select all objects we could find in a certain radius around the vessels current position and search for those attributes.
Naturally, a realistic behavior must consider more than just the course and speed of a possible target but also its size and possibly it's mass to determine if it could result in a thread but that information can be determined the same way, without knowing the object but knowing that it is characterized with those attributes.
If the Behavior Tree on the other hand represents an action (BT_A), the tree is formalized as follows.

$$BT_A = [Action, r(f(a^G, \Delta t) \rightarrow AB^G), \emptyset\ ] \mid a^G \in A^G \tag{9}$$

With Δt being the tick's time and a^Gan attribute from the global knowledge that is actually changed by the action. By modifying the global knowledge this may also affect other agents, as they do observe their surrounding and update their internal knowledge in every time step.

Modeling and simulating a collision regulation behavior

A possible application for the Behavior Trees is the modelling of the International Regulations for Preventing Collisions at Sea (ColRegs). In the following, we present a modeling approach for a simple "drive on starboard side" behavior in the form of a Behavior Tree and then add additional parts to the tree. Attention will be given to the aspects of modularity and complexity.
In general, according to the ColRegs, ships are obliged to drive on their trajectory as far to the right as possible. This procedure is an easy way to avoid collisions. To realize this behavior, it is assumed that the ship can orientate

itself along the coastline and position itself accordingly in the fairway.

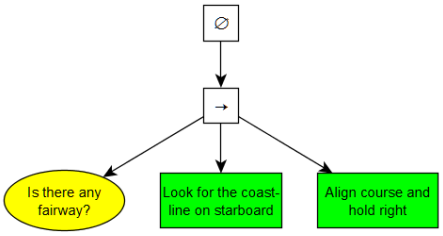


Figure 5: Behavior Tree for the "drive on starboard side" behavior

Figure 5 shows the Behavior Tree. The desired behavior is split into logical units, which can be numerically larger or smaller depending on the initial design decision. This decomposition of the behavior is already based on a divide and conquer approach during modeling and allows the developer to think in small modules and to consider sub-problems. Using our notation, we can express the condition "Is there a fairway" as follows:

$$\exists ab_f \in AB^A \wedge ab_f = [a_f, v_f] \wedge a_f = [n_f, ft_f] \wedge ft_f = Fairway \quad (10)$$

Which can be read as: There exists an attribute binding within the knowledge of the agent, whose attributes type is of type Fairway.
Since the tree first traverses into depth, it is checked if the fairway exists. If so, the condition returns a positive return value to the sequence-node and performs the first action on looking for a coastline for orientation on starboard. Once the coastline was recognized, the action returns a positive return value and the last action would start. This action is used to calculate the course based on the information collected in the last action.

Extending the Tree with additional behavioral modules

The usual driving behavior rules not only regulate the positioning in the fairway, but also describe the use of orientation aids such as buoys. Therefore, the following example will explain how Behavior Trees can be extended with additional behavioral modules.
Figure 6 shows what an extended model of the right-hand driving regulations could look like. First of all, the buoys are checked to see if they are in sight. If this is not the case, the ship orients itself at the coast and calculates the appropriate distances, as shown in Figure 5.
When buoys are in sight, the tree distinguishes between two cases. If the ship comes from the high seas, it is oriented towards the red buoys on the starboard side. If the ship is moving towards the high seas, it calculates distances to green buoys on the starboard side.

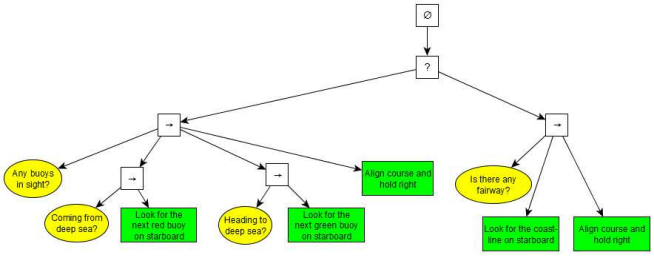


Figure 6: Navigate on starboard-side

It was shown, how simple existing rules can be supplemented by further conditions and actions without having to change the existing model. Since all ColRegs can probably be modeled and implemented as isolated rules, it should be possible to combine the behavior for arbitrary scenarios from any number and combination of ColRegs later on.
According to this introductory example, we would like to propose extensions to the Behavior Tree concept to allow more dynamic and randomness in the simulation environment. For this reason, some probabilistic selection strategies and the efficiency parameter are introduced.

Introduction of probabilistic selection strategies

For the purpose of V&V research, it is helpful to randomize behavior, to obtain a sufficient coverage of test cases. Assume a ship is supposed to overtake another ship on the left, on the right or not at all. Then it is recommended to use a "Random Node Selector" as described in (Millington and Funge 2009) and as shown in Figure 7-A. An action to be executed is selected randomly in each iteration step. However, there are also cases where some more control over the random selection is necessary. For example, if the ship should overtake on the left side with a very small probability, on the right side with a greater probability, and not at all with an even greater probability. To describe such situations, the ~? Operator, in conjunction with edge weights, is introduced as a "simple weighted random selection" (see Figure 7-B). The edge weights correspond to the probability with which the subsequent subtree is to be selected. Since the node always has to return a return value, the overall probability of all edge weights of a ~? Node is always 1.
Since the Behavior Tree iterates at regular intervals, the question arises as to how a similar situation, at the same runtime, should be treated. The vessel can now either select an action again by simple weighted random selection, or the result of the last selection can be memorized, so that the vessel reacts as before.
In order to be able to describe both variants, the "+?" Node is introduced as a "disposable weighted random selection" of the "+?" Nodes (see Figure 7-C). The "+?" Node stores the previously result and re-executes its related subtree at each iteration step.

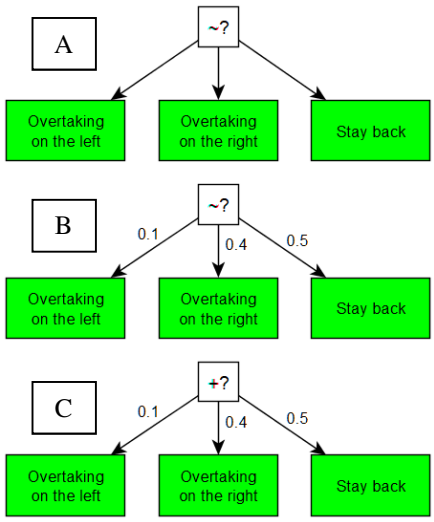


Figure 7: Different types of random selection

This additional expansion would lead to a unique behavioral profile for each vessel in the course of the simulation.

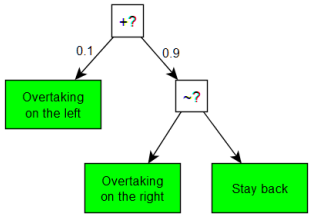


Figure 8: Combination of selection strategies

Suppose there are some ships with a behavior as shown in Figure 8. Then a small number of ships will always overtake on the left side if they encounter an obstacle and all other ships will either overtake on the right side or stay behind. These combinations allow complex dynamic behaviors to be generated even though only one tree is used. However, this will affect the readability of more complex trees.

Introduction of the efficiency parameter

As a further approach, the efficiency parameter will be introduced. It is intended to change the behavior of the ship at runtime within a predefined range. The parameter refers to the captain's driving style and indicates the tendency to have a safe, neutral, or efficient driving style. For further understanding, the overtaking behavior outlined in Figure 9 is considered.
The green vessel keeps its course and is overtaken by the blue vessel. The blue vessel is aware of the slower green one, which runs in the same direction and thus can be overtaken. It calculates an overtaking course, parallel to its own route and resumes the old course once it has gained sufficient distance to the green vessel. Applied to the overtaking behavior, the value of the efficiency parameter could affect the distance to the preceding ship, which is needed until the captain initiates an overtaking maneuver.

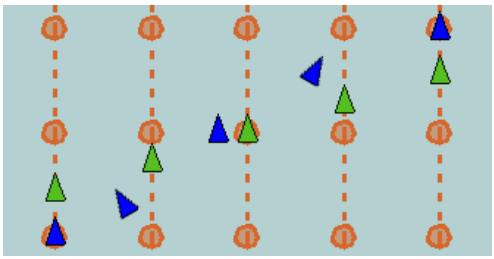


Figure 9: Different phases of overtaking with two vessels

In case of a safer driving behavior, the captain would start overtaking much earlier than a captain with efficient intentions would do. Since it depends on the application and desired behavior, in which situation a captain behaves safe or not, no general recommendations can be given here. Nevertheless, the example of the overtaking behavior is intended to show what such a parameter might look like.
The efficiency parameter is defined as the interval between [-1, 1], where -1 is the safest and passive, 0 is the neutral, and 1 the most efficient tendency. In the following example, the required distance to the preceding vehicle, so that an overtaking maneuver is initiated, is to be bound to the efficiency parameter. To achieve this, a second interval is required, on which the interval of the efficiency parameter is mapped. This additional value interval is the distance from the preceding vehicle with the minimum as the most efficient and the maximum as the safest value.

$$f(x) = (\frac{x+1}{2} * (\min - \max)) + max \qquad (11)$$

In addition, a mapping function for two intervals like shown in equation (11) is necessary, which returns a corresponding distance y for an efficiency parameter x.

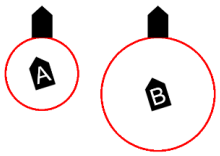


Figure 10: One vessel starting an overtaking maneuver with different efficiency parameters

Figure 10 shows a ship in two different situations A and B. In situation A, the efficiency parameter is close to 1 and the distance to the preceding ship is relatively small, as it begins to overtake. In situation B, the efficiency parameter is lower and the distance is higher. This concept is realized by defining the value intervals wherever they are used. In this case, in a condition that checks the distance to the preceding vessel using the own believe of the overtaking vessel. The efficiency parameter can be changed via an external parameter in the behavior

class object. At each tick, the affected leaf node checks the parameter value and makes a corresponding decision. In this way, many ships could be equipped with the same behavior, but they would still be able to behave differently.

USE CASE

The functionality of the requirements and the efficiency parameter are shown by configuring and performing an overtaking maneuver in a maritime traffic simulation. For this purpose, a scenario with two vessels, which corresponds to Figure 9, is created in the Maritime Traffic Simulator. These are two similar vessels with the same Behavior Tree, but the ship that is passing through is faster.

Figure 11: Trajectories of an overtaking maneuver

The trajectories of the simulation are shown in Figure 11. Course t1 shows an overtaking maneuver with an efficiency parameter of -1. There are different points that can be affected by changing the efficiency parameter. Distance d1 describes the distance between the two ships before the start of the overtaking. Distance d2 describes the distance during the overtaking operation, and d3 is the distance before the overtaking maneuver is completed.
Our configuration provides identical intervals for d1 and d3, thus changing to identical values depending on the efficiency parameter. Distance d2 varies only with a very small interval. At this point, an advantage of modularity is demonstrated. By adding another node in the Behavior Tree and corresponding another class in the implementation, d1 and d3 could be separated from each other. This would make it possible to add more variability to the scenario in a simple way. Additionaly it should be mentioned that d1 and d3 might vary due to different ship sizes. Course t2 shows the same scenario with the maximum efficiency parameter of the overtaking vessel. The distances for initiating and terminating the overtaking maneuver are significantly shorter and the entire overtaking process is carried out at a shorter distance. However, the ships are getting closer, which increases the risk of collision. Since in both cases the overtaking ship first follows a route and then changes to the overtaking maneuver, it is proven, that several rules can be used and a hierarchy is followed. In this case: follow the Overtaking Behavior before the Route-Follow Behavior.
The Behavior Tree can be used in several ships, and even if the configuration is identical, the efficiency parameter can be modified in a way to create different behaviors.

CONCLUSION

The attempt of modelling the overtaking scenario within the Behavior Tree concept gave an insight that the requirements towards a more realistic vessel behavior can be achieved. The strengths of the investigated approach come from the modular structure of the Behavior Tree concept, and the possibility to decide what kind of behavior is executed using the knowledge about the environment from each vessels perspective. It became visible that the S-100 based model for describing environmental knowledge and the Behavior Tree concept could be well combined. The V&V of maritime scenarios, would gain from the manifold parameterization and the ability to still maintain a non-deterministic behavior for various vessel agents. The modularity and easy access to the environmental knowledge allows the quick composition of different behaviors for many vessels. Whereas the parameter space in this attempt is constrained between safe and efficient behavior of an artificial captain.
By using global behavior-determining parameters and the possibility to set these randomly distributed, many simulated vessels can be equipped with similar Behavior Trees without these vessels behaving in the same way.
The modular structure of the Behavior Trees makes it possible to modify different behaviors. Thus behavioral building blocks, in the sense of different behavior trees that represent a sub-behavior, can be combined to model new behavior.
While Behavior Trees are easier for the designer to understand for individual behavioral aspects, they can quickly become confusing in case of larger or complicated behavior patterns. Next, we will approach the handling of large behavior patterns looking into the modelling of behavior aspects via Behavior trees and choosing the execution of these Behavior Trees via scoring functions used in the Utility AI concept.

ACKNOWLEDGMENTS

The work presented in this paper is supported by the Center for Critical System Engineering for Sociotechnical Systems at the University of Oldenburg, OFFIS and DLR. The center is funded by the Federal State of Lower Saxony, Germany under grant numbers VWZN3237 and VWZN3270.

REFERENCES

Bierwirth, Michael. 2004. “Schuldverteilung bei Schiffskollisionen.” Diplomarbeit (FH). Hochschule Bremen. https://www.hs-bremen.de/internet/hsb/projekte/maritime/studium/nautikseeverkehr/diplombachelor/diplarb_bierwirth.pdf.

Cahill, R.A., and Nautical Institute (Great Britain). 2002. *Collisions and Their Causes*. Nautical Institute. https://books.google.de/books?id=6RpFQgAACAAJ.

Colledanchise, Michele, Alejandro Marzinotto, Dimos V. Dimarogonas, and Petter Oegren. 2016. "The Advantages of Using Behavior Trees in Mult-Robot Systems." In *ISR 2016: 47st International Symposium on Robotics; Proceedings of*, 1–8. VDE. http://ieeexplore.ieee.org/abstract/document/7558452/.

Fan, Henry SL, and Jia-Ming Cao. 2000. "Sea Space Capacity and Operation Strategy Analysis System." *Transportation Planning and Technology* 24 (1):49–63.

Hahn, Axel. 2015. "Simulation Environment for Risk Assessment of E-Navigation Systems." In *ASME 2015 34th International Conference on Ocean, Offshore and Arctic Engineering*, V003T02A072–V003T02A072. American Society of Mechanical Engineers. http://proceedings.asmedigitalcollection.asme.org/pdfaccess.ashx?url=/data/conferences/asmep/86128/v003t02a072-omae2015-41498.pdf.

Hahn, Axel, and Thoralf Noack. 2016. *EMartitime Integrated Reference Platform*. Deutsche Gesellschaft für Luft-und Raumfahrt-Lilienthal-Oberth eV.

Hasegawa, Kazuhiko, Go Tashiro, Seiji Kiritani, and Koji Tachikawa. 2001. "Intelligent Marine Traffic Simulator for Congested Waterways." In *7th IEEE International Conference on Methods and Models in Automation and Robotics*, 632–636.

IHO. 2015. "S-100 – UNIVERSAL HYDROGRAPHIC DATA MODEL." Edition 2.0.0. MONACO: International Hydrographic Organization.

Isla, Damian. 2005. "Handling Complexity in the Halo 2 AI." In *Game Developers Conference*. http://www.gamasutra.com/gdc2005/features/20050311/isla_01.shtml.

Köse, Ercan, Ersan Başar, Emrullah Demirci, Abdulaziz Güneroğlu, and Şebnem Erkebay. 2003. "Simulation of Marine Traffic in Istanbul Strait." *Simulation Modelling Practice and Theory* 11 (7):597–608.

Lee, G.W.U., and J. Parker. 2007. *Managing Collision Avoidance at Sea: A Practical Guide*. Nautical Institute. https://books.google.de/books?id=ZrwNNQAACAAJ.

Millington, Ian, and John David Funge. 2009. *Artificial Intelligence for Games*. 2nd ed. Burlington, MA: Morgan Kaufmann/Elsevier.

Ogren, Petter. 2012. "Increasing Modularity of UAV Control Systems Using Computer Game Behavior Trees." In *AIAA Guidance, Navigation and Control Conference, Minneapolis, MN*.

Organization, International Maritime. 2003. *COLREG: Convention on the International Regulations for Preventing Collisions at Sea, 1972*. IMO Publication. International Maritime Organization. https://books.google.de/books?id=_ZkZAQAAIAAJ.

Rabin, Steven. 2013. *Game AI pro: Collected Wisdom of Game AI Professionals*. CRC Press.

Russell, Stuart J., and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*. 2nd ed. Pearson Education.

Schweigert, Sören, Volker Gollücke, Axel Hahn, and André Bolles. 2014. "HAGGIS: A Modelling and Simulation Platform for E-Maritime Technology Assessment." In , 10. Istanbul, Turkey.

AUTHOR BIOGRAPHIES

VOLKER GOLLÜCKE started to work at the University of Oldenburg in 2012. He received his PhD (Dr.-Ing.) in 2016 at the University of Oldenburg. Currently, he is working within the project CSE, where he focuses on simulative risk assessment in the maritime sector. His e-mail address is : `volker.golluecke@uni-oldenburg.de` and the homepage of his group can be found at `https://www.uni-oldenburg.de/sao/`.

SÖREN SCHWEIGERT received his PhD (Dr.-Ing.) in 2016 at the University of Oldenburg. Since 2009 he is working at the OFFIS research institude in the Transportation devision. Currently he is working in the ENABLES-S3 project, where he focuses on maritime co-simulation and scenario based V&V. His E-Mail address is: `soeren.schweigert@offis.de` and the homepage of his group is `http://www.offis.de/`.

DANIEL LANGE received his Bachelor of Science in 2017 at the University of Oldenburg and started to work at offis in the research group CMS (Cooperative mobile systems) as student research assistant. His E-Mail address is: `daniel.lange@uni-oldenburg.de`.

AXEL HAHN is full professor at the University of Oldenburg and leads the working group Systemanalysis- and Optimisation and is board member of the research institute OFFIS. His research topics are safety and efficiency in marine transportation systems. His E-Mail address is: `axel.hahn@uni-oldenburg.de` and the homepage of his group is `https://www.uni-oldenburg.de/sao/`.

COMPARATIVE ANALYSIS OF METAMODELING TECHNIQUES BASED ON AN AGENT-BASED SUPPLY CHAIN MODEL

Mert Edali
Department of Industrial Engineering
Yildiz Technical University
34349, Besiktas, Istanbul, Turkey
E-mail: medali@yildiz.edu.tr

Gonenc Yucel
Department of Industrial Engineering
Bogazici University
34342, Bebek, Istanbul, Turkey
E-mail: gonenc.yucel@boun.edu.tr

KEYWORDS
Agent-Based Modeling, Beer Game, Metamodeling, Input Sampling.

ABSTRACT

Agent-based models comprise interacting autonomous entities, and generate both individual and emergent system-level outputs. These models generally have a large set of free parameters, whose impact on output needs to be explored. Considering also the need for replication due to stochasticity, a proper analysis requires a very large set of simulation runs. Therefore, obtaining a simpler representation of a simulation model (e.g., metamodel) can prove useful. We primarily focus on the potential utilization of various metamodeling approaches, namely Decision Trees, Random Forests, *k*-Nearest Neighbor Regression, and Support Vector Regression in predicting the two different types of outputs of an agent-based model. Results show that system-level outputs are predicted with higher accuracy compared to individual-level outputs under equal sample sizes. Although there is no single metamodeling technique performing best in all cases, we observe that support vector regression is more robust to the increase in the dimension of the problem.

INTRODUCTION

Agent-based modeling is a modeling approach used to analyze complex adaptive systems from various domains such as archaeology (Axtell et al. 2002), sociology (Edmonds and Hales 2005), and economics (Tesfatsion 2002). The main building block of these models is agents; individual entities acting autonomously based on their objectives, preferences, internal states and perceptions about their environments. The interactions of these micro-level autonomous entities drive the macro-level system dynamics in agent-based simulation models (Gilbert 2008; Wilensky and Rand 2015).

Agent-based modeling is a "bottom-up" approach since the system of interest is modeled by defining the decision rules of individual entities (i.e., agents), and the system-level outcome is a result of the interactions of these agents at the individual-level (Macal 2010). Therefore, agent-based models generate both individual and system-level outputs (Wilensky and Rand 2015). For example, in the Sugarscape model (Epstein and Axtell 1996), the wealth of an individual is an important indicator to check how metabolism and vision attributes of an agent affect its wealth. Besides, wealth distribution of agents is of interest to monitor the economic inequality among the members of the society.

After verification and validation steps, the analyst can use the model as an experimental platform for policy/scenario analysis/testing and model exploration. While the execution time of a single run of an agent-based model is generally measured in seconds, if not in minutes, many replications for each parameter combination are needed due to stochasticity. Additionally, these models are notoriously known for the large number of parameters to be specified, which expands the set of parameter combinations to be explored. As a result, policy/scenario analysis/testing and model exploration turn out to be time-consuming tasks. Therefore, obtaining a simpler representation of a simulation model can prove useful. In this study, we primarily focus on the potential utilization of various metamodeling approaches in the context of agent-based modeling studies.

A metamodel is as an approximate representation of the input-output relationship of a simulation model (Kleijnen and Sargent 2000; Kleijnen et al. 2005). Instead of running a complex simulation model, it can be time-saving to generate estimates of simulation outputs from a simplified representation of the model. Therefore, metamodels are extensively used in simulation literature (Kleijnen and Sargent 2000). The use of metamodels comes into prominence especially when the required time to run a simulation model takes days instead of minutes (Kleijnen and van Beers 2004). Besides time-saving benefits, metamodels also give the analyst more insights into the model. For example, the analysis of a linear regression metamodel may reveal the interactions between simulation model parameters, or the significance of the effect of simulation model parameters on model output.

Fitting a metamodel requires input and output data obtained from the simulation model for training. Therefore, an appropriate input sampling method which helps the analyst capture the behavioral richness of the model outputs should be selected. However, performing input sampling for agent-based models becomes a major challenge since sampling methods offered by classical

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

design of experiments (DoE) literature may not be applicable due to underlying assumptions of these methods such as linear relationship between inputs (model parameters) and output, normally distributed errors, and small number of inputs (Sanchez and Lucas 2002; Kleijnen and van Beers 2004; Lee et al. 2015). Besides, agent-based simulation models have some special characteristics which may render classical DoE-based sampling techniques inapplicable for metamodel fitting. For example, tipping point behavior and other emergent properties of agent-based models such as adaptation and path dependency may imply potential nonlinear relationships between model inputs and outputs (Wilensky and Rand 2015; ten Broeke et al. 2016). Kleijnen et al. (2005) emphasize that matching metamodeling methods and sampling techniques to simulation approaches will be a significant contribution to the existing literature.

Although metamodeling approaches coupled with appropriate input sampling methods (for training purposes) has the potential to be valuable additions to the toolbox of agent-based modelers in principle, very limited work has been done to explore this potential. In that respect, we aim to present the results of our experimental evaluation of several metamodeling approaches in predicting the output of an agent-based simulation model. In that context, we provide a conceptual classification for output types that one can get as a consequence of a simulation run with an agent-based model. Then, we compare the performances of a set of well-known metamodeling approaches in predicting two different output types.

The remainder of this paper is organized as follows: Section 2 summarizes the sampling and metamodeling techniques used in this study. Section 3 presents experimental setting. Section 4 contains results and discussions. Finally, Section 5 concludes the study.

BACKGROUND

As mentioned earlier, one requires a training set that includes tuples of model inputs and resulting outputs in order to develop a metamodel for a simulation model. Determination of this training set is of primary importance for the performance of the resulting metamodel. One can employ various sampling approaches to identify the input combinations to be included in the training set. In this study, we consider four different sampling techniques to generate input and output data for metamodel training: (i) full factorial design (FFD), (ii) random latin hypercube sampling (RLHS), (iii) maximin latin hypercube sampling (MLHS), and (iv) random sampling (RS). To approximate the input-output relationship of a simulation model, we consider four well-known machine learning techniques for metamodel development: (i) decision trees (DT), (ii) random forests (RF), (iii) support vector regression (SVR), and (iv) *k*-nearest neighbor regression (*k*-NN R).

Sampling Techniques

In full factorial designs, all combinations of predefined parameter values (i.e. levels) are considered. For example, in a two-factor factorial design with parameters X and Y, there will be $x \times y$ number of design points if there are x different values of X and y different values of Y. One of the mostly utilized factorial design is 2^k when there are k model parameters (factors) each having two different levels (parameter values) denoted by - (low) and + (high) (Montgomery 2013). If the analyst assumes a more complex metamodel, it is possible to use 3^k or even more detailed factorial design in the form of m^k where m denotes the number of factor levels (Kleijnen et al. 2005).

Although full factorial designs are easy to construct, number of sample points exponentially increases as the number of factors, k, increases. To overcome this problem, the analyst can utilize random sampling to generate samples with desired number of points, where samples are generated from the joint probability distribution of the factors. The drawback of random sampling strategy is that it does not ensure evenly distributed sample points, especially in the presence of small sample sizes (Crombecq et al. 2009). Therefore, the analyst can utilize random latin hypercube sampling (RLHS), which is a sampling strategy with space-filling property and aims to ensure that all parts of the input space are sampled (McKay et al. 1979; Montgomery 2013). This is achieved by dividing each factor into n distinct intervals of equal probability and samples are randomly drawn so that there is only one sample in each interval for each factor (Keane and Nair 2005). RLHS can be used to develop complex metamodels involving many quantitative factors (Sanchez and Lucas 2002; Kleijnen et al. 2005).

Although RLHS has space-filling property, it does not always guarantee that parameter space is evenly sampled. Therefore, there are many extensions to the generic RLHS algorithm to ensure the space-filling property. Maximin LHS (MLHS) is one of these extensions and aims to maximize the minimum distance between sample points. The reader is referred to Beachkofski and Grandhi (2002) and Deutsch and Deutsch (2012) for the details of MLHS algorithm.

Metamodeling Techniques

A decision tree (DT) is a hierarchical classification and regression technique (Alpaydin 2014). The method generates axis-parallel binary splits on the input space (model parameter space) by using only one input variable at each iteration (Bishop 2006). The number of splits is generally determined by node impurity, a measure of heterogeneity of a node of the tree (Loh 2011). In regression, split decisions are based on the comparison of sum-of-squared errors (SSE) and a threshold value which represents the maximum allowable SSE. Each terminal (leaf) node calculates the mean (or median) of the outputs it includes. Several algorithms such as C4.5 (Quinlan 1993) and CART

(Breiman et al. 1984) are proposed to construct decision trees for classification and regression. Another important attribute of the decision tree approach is that it allows to capture the most important input variables (e.g., simulation model parameters) (Breiman et al. 1984). Decision trees can also handle variable interactions. In addition, Sanchez and Lucas (2002) and Kleijnen et al. (2005) state that regression trees are more interpretable compared to traditional regression models since predictions for model outcomes can be easily obtained by simply following the discriminating rules at each node.

A random forest (RF) is an ensemble of individual decision trees each using a random subset of training set or input variables. In regression problems, the mean of the outputs obtained from each tree is returned as predictions. The random selection procedure and combining the predictions of trees lead to significant accuracy improvements in regression problems (Breiman 2001; Alpaydin 2014). Compared to an individual decision tree, results obtained from a random forest are not directly interpretable since random forests combine outputs obtained from a high number of trees (e.g., 1000 trees). However, it is possible to visualize the individual trees of a forest to observe the discriminating rules. In addition, it is possible to determine the most important variables (in our case, input parameters of a simulation model) for prediction (Breiman 2001; Chen et al. 2011).

Support vector regression (SVR) is an extension of support vector machine (SVM) classification technique to the problems with continuous outputs. Contrary to traditional regression models aiming to minimize SSE, SVR model incorporates ε-insensitive loss function, which yields regression models robust to noise (Alpaydin 2014). SVR is formulated as a quadratic optimization problem. In its simple form, one can use SVR for linear regression. However, the dual formulation allows one to conduct nonlinear regression by employing kernel functions such as polynomial kernels and radial-basis functions. Besides these well-known kernels, one can also develop specific kernels for different applications (Alpaydin 2014). SVR technique has been successfully implemented as metamodels to approximate highly nonlinear systems (Zhu et al. 2009, 2012). Since agent-based models capture nonlinearities embedded in a system, SVR stands as a promising tool for metamodeling.

k-nearest neighbor regression (*k*-NN R) method simply predicts the output of a test instance by averaging the outputs of *k*-nearest neighbors of that test instance. Contrary to abovementioned regression methods, *k*-NN method does not require an explicit training step; we only store the training set to determine *k* closest neighbors and their output averages (Chen et al. 2009; Hu et al. 2014). To quantify closeness, Euclidean and Manhattan distance can be used (Juutilainen and Röning 2007).

For a more detailed background on the abovementioned machine learning methods, the reader is referred to Hastie et al. (2009).

EXPERIMENTAL DESIGN

Beer Game

In this paper, we use the Beer Game to conduct experiments (Sterman 1989; Edali and Yasarcan 2014). Beer Game is a four-agent supply chain simulation. In this game, each agent controls the inventory level of one of the four echelons, which can be listed as a retailer, a wholesaler, a distributor, and a factory. Although the agents are not allowed to communicate and share information, they aim to minimize the *team total cost* at the end of the game. *Team total cost* is the sum of the individual costs of echelons. The individual cost of an echelon is calculated by summing up inventory holding and backlog costs generated at each simulated week. We run the model for 520 simulated weeks. The main outcome of interest of the Beer Game is a terminal value, *team total cost*. Ordering behaviors of agents can be represented by the anchoring and adjustment heuristic (Tversky and Kahneman 1974). This heuristic has two main parameters, namely stock adjustment fraction (α) and weight of supply line (β) and each agent implicitly uses these two parameters in ordering decisions (Sterman 1989).

Beer Game is a suitable model as an experimental platform since (i) it is a simple model with four agents and eight parameters in total, (ii) it is highly nonlinear, (iii) it can produce a rich set of outputs including (unpredictable) chaotic behavior. In other words, it is a simple model with complex behavior capabilities that has the potential to challenge the predictive capabilities of a metamodel to be trained based on limited input-output tuples from this model. In that respect, it stands as a very good experimental ground for our comparative analysis on metamodeling approaches.

Output Type Categorization

Independent of the selected model, we can categorize outputs of an agent-based simulation model. First criterion is based on whether the output is related to a single agent (individual-level) or to the population (system-level). Second criterion considers the temporal aspect of an output: (i) the output can be measured at a single time point in the simulation horizon or (ii) it can be a function of a (large) set of instantaneous measurements over the simulation horizon (e.g, average over time, total over time, maximum/minimum value during a simulation run). This kind of output categorization will give an idea about the difficulty of predicting model outputs; we claim that predicting over-time values are easier than predicting instantaneous values and predicting system-level outputs are easier than predicting individual-level outputs.

Table 1: Output Type Categorization for the Beer Game

	Instantaneous Values	Over-Time Values
Individual-Level Output	Inventory level of the retailer at week *t*	Maximum inventory level of the retailer
System-Level Output	Team total cost at week *t*	Team total cost at final time

In this study, we generate metamodels for the prediction of two different outputs of the Beer Game: The first one is a system-level output and is an over-time value, *team total cost*. The second output is individual-level and an over-time value, *maximum inventory level of the retailer*. We consider two different sets of model input parameters: (i) α_R (stock adjustment fraction of the retailer), (ii) α_R and β_R (stock adjustment fraction and weight of supply line of the retailer) as metamodel (and simulation model) inputs. All of the parameter values of remaining echelons are set to the average values of these parameters used by the participants in the board version of the game (Sterman 1989).

Datasets

We use two different datasets in the context of this study: a training set and a test set. In each dataset, sample points have two main components; the first one is parameter values and the second component is the simulation model output obtained by running the model with these parameters. Training set is used to fit a metamodel. We employ the sampling techniques mentioned in the "sampling techniques" section for training set generation. Since RLHS, MLHS, and RS techniques generate different samples due to randomness, we generate 30 sets of samples by using each technique and fit a metamodel with each one of these 30 sets. For the experiments where the input parameter is only α_R, we generate 21 sample points with each sampling technique for metamodel fitting. In the two-parameter case (i.e., α_R and β_R), we generate 25 sample points. For one- and two-parameter cases, we use test sets each having 5,000 instances to assess the prediction performance of the metamodels.

Hyperparameter Optimization

The metamodeling techniques that are used in this study have some hyperparameters to be optimized. These hyperparameters are *C* (penalty factor), ε (parameter of the epsilon-insensitive loss function), and γ (spread parameter of the Gaussian kernel) in SVR; *ntree* (number of trees in the forest) and *mtry* (number of randomly selected candidate variables at each split) in RF; *k* (number of neighbors) in *k*-NN R; *minsplit* (minimum number of instances in a node for splitting), *minbucket* (minimum number of instances in a terminal node), and *cp* (complexity parameter) in DT.

To optimize the hyperparameters of SVR, RF, and *k*-NN R, we perform a grid search on the selected subset of hyperparameter space of each technique. For each hyperparameter combination, we perform leave-one-out cross-validation on the training set. Then, the metamodel with the hyperparameter combination yielding the minimum leave-one-out cross-validation error is selected. Finally, the metamodel with the selected hyperparameters is used to predict the instances on the test set. Hyperparameter subsets of each metamodeling technique considered in optimization are given in Table 2. However, in the DT method, we follow a different procedure: We first fully grow a tree by setting *minsplit* = 2, *minbucket* = 1, and *cp* = 0. Then, we prune the tree. For tree pruning, the reader is referred to Breiman et al. (1984).

Table 2: Hyperparameter subset of each metamodeling technique

Metamodeling Technique	Hyperparameters
Support Vector Regression	$C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$
	$\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$
	$\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$
Random Forest	*ntree* ∈ {50, 100, 150, 200, 500, 1000, 2000}
	mtry ∈ {1} (one-parameter case), *mtry* ∈ {1, 2} (two-parameter case)
k-NN Regression	$k \in \{1, 3, 5, 7, 9\}$

As we mentioned, we generate 30 sample sets for RLHS, MLHS, and RS. We perform hyperparameter optimization on each sample set individually.

Metamodel Performance Evaluation Criteria

The main performance criteria for metamodel evaluation is Mean Absolute Percentage Error (MAPE), which is given as $MAPE = (1 / N) \times \Sigma\, |\hat{y}_i - y_i| / y_i$, where $\hat{y}_i$ and y_i are metamodel prediction and simulation model output, respectively. Besides MAPE, we also report Percentage Distribution of Relative Prediction Error ($PDRPE_{x\%}$) (Alam et al. 2004), which calculates the percentage of the metamodel outputs whose Relative Prediction Error ($RPE_i = \hat{y}_i / y_i$) are within ±x% error.

RESULTS AND DISCUSSIONS

In this section, we present the results of the experiments and some discussion on the results. In each table, $PDRPE_{10\%}$ and MAPE values show the performance on the test set. MT stands for Metamodeling Technique and ST stands for Sampling Technique. Total Time (TT) is the sum of simulation time (for running the simulation model with the parameter values obtained from sampling), training time (for training the metamodel and

hyperparameter optimization) and test time (runtime of the metamodel for the prediction of the instances in the test set). All the reported times are in seconds. For RLHS, MLHS, and RS, we give the averages of the performance measures since they are replicated 30 times.

Case 1: One Parameter – System-Level Output

Case 1 consists of experiments where the model output is *team total cost* and the only model input is α_R. Detailed results are given in Table 3. The first observation is that all the methods yield similar MAPE values around 11%, which is a satisfactory result with a considerably small training set size. The lowest MAPE, which is 9.96%, is achieved when we use *k*-NN regression with full factorial sampling design. Besides, regardless of the metamodeling and sampling technique, 83% of the test set instances are within ±10% error on average. Another clear observation is that random sampling method yields slightly higher MAPE values for each metamodeling technique compared to the other sampling techniques. Random forest is the most time-consuming method since we take 30 replications due to the random training and parameter subset selection of the method. Decision tree stands as the fastest technique compared to the other techniques.

Table 3: Results of experiments when model output is *team total cost* and model input is α_R (Case 1)

MT	ST	$PDRPE_{10\%}$	MAPE	TT
SVR	FFD	85.66	11.66	7.79
	RLHS	**87.14**	11.22	7.66
	MLHS	84.99	11.59	7.71
	RS	85.03	12.03	7.54
DT	FFD	85.28	12.36	2.41
	RLHS	78.95	12.26	**2.22**
	MLHS	78.98	12.30	2.33
	RS	77.46	13.17	2.23
RF	FFD	84.74	10.38	31.24
	RLHS	83.58	11.38	28.48
	MLHS	82.61	11.44	27.38
	RS	80.08	12.26	27.37
k-NN R	FFD	85.96	**9.96**	4.44
	RLHS	81.16	11.70	4.26
	MLHS	81.62	11.84	4.37
	RS	79.39	12.70	4.27

Case 2: Two Parameters – System-Level Output

In Case 2, the model output is *team total cost* and the model input parameters are α_R and β_R. We observe that performance of each method significantly deteriorates (with an average 28% increase in MAPE) compared to one-parameter case (Case 1). However, support vector regression is the least affected method (with an average 20% increase in MAPE) by the increase in the dimension and performs best in all sampling techniques. The lowest MAPE, which is 29.3%, is achieved when we use support vector regression with full factorial sampling design. Besides, this metamodeling and sampling technique combination yields significantly high $PDRPE_{10\%}$ (65.10%) value compared to the other results in this case. *k*-NN regression is the second best method (Table 4).

Table 4: Results of experiments when model output is *team total cost* and model inputs are α_R and β_R (Case 2)

MT	ST	$PDRPE_{10\%}$	MAPE	TT
SVR	FFD	**65.10**	**29.30**	10.45
	RLHS	49.15	33.03	10.05
	MLHS	55.54	31.15	10.23
	RS	50.39	34.02	10.09
DT	FFD	46.88	37.43	2.90
	RLHS	34.23	46.48	**2.63**
	MLHS	28.93	56.78	3.64
	RS	31.82	49.01	2.63
RF	FFD	38.38	47.62	79.36
	RLHS	41.76	40.70	89.02
	MLHS	43.71	41.68	89.61
	RS	39.70	42.20	89.37
k-NN R	FFD	49.40	36.60	5.42
	RLHS	41.80	35.01	5.12
	MLHS	45.63	34.97	5.17
	RS	41.03	36.59	5.13

Case 3: One Parameter – Individual-Level Output

In Case 3, the model output is *maximum inventory level of the retailer* and the model input is α_R. Detailed experimental results are given in Table 5.

Table 5: Results of experiments when model output is *maximum inventory level of the retailer* and model input is α_R (Case 3).

MT	ST	$PDRPE_{10\%}$	MAPE	TT
SVR	FFD	**51.40**	48.00	7.70
	RLHS	42.49	49.28	7.51
	MLHS	43.99	48.61	7.69
	R	38.53	52.98	7.46
DT	FFD	46.16	53.31	2.30
	RLHS	41.34	51.92	**2.21**
	MLHS	40.62	51.01	2.40
	RS	36.84	55.14	2.21
RF	FFD	46.62	**46.96**	30.56
	RLHS	44.61	48.27	28.70
	MLHS	43.68	48.69	28.56
	RS	39.67	51.52	28.02
k-NN R	FFD	48.14	48.85	4.27
	RLHS	43.55	51.81	4.21
	MLHS	42.38	50.62	4.41
	RS	38.73	55.75	4.22

Compared to Case 1, MAPE values much higher since the coefficient of variation of the outputs is larger in Case 3. All of the methods perform similar in terms of

MAPE. However, differences between MAPE values are much higher compared to Case 1. The minimum MAPE (46.96%) is achieved when we use random forest with full factorial design. Random sampling method gives slightly worse results in terms of MAPE compared to the other sampling techniques for each metamodeling technique.

Case 4: Two Parameters – Individual-Level Output

In Case 4, the model output is *maximum inventory level of the retailer* and the model input parameters are α_R and β_R. In this case, we obtain very high MAPE values (see Table 6), even larger than 100% (DT, RF, and *k*-NN R). The results indicate that the output is very hard to predict with a limited number of training points. The minimum MAPE, which is 66.10%, is achieved when we use support vector regression with maximin LHS. SVR gives the minimum MAPE values for all of the sampling techniques.

Table 6: Results of experiments when model output is *maximum inventory level of the retailer* and model inputs are α_R and β_R (Case 4).

MT	ST	$PDRPE_{10\%}$	MAPE	TT
SVR	FFD	**28.00**	78.46	10.54
	RLHS	25.28	74.71	10.10
	MLHS	26.74	**66.10**	10.15
	RS	24.66	78.32	10.03
DT	FFD	13.62	131.38	2.89
	RLHS	14.22	123.38	2.64
	MLHS	13.43	141.18	2.65
	RS	15.30	116.53	**2.64**
RF	FFD	16.76	114.14	81.42
	RLHS	19.15	100.37	88.80
	MLHS	21.93	93.76	90.58
	RS	17.62	106.91	87.11
k-NN R	FFD	14.96	131.08	5.44
	RLHS	19.76	80.97	5.15
	MLHS	19.94	82.22	5.17
	RS	18.12	93.78	5.15

CONCLUSIONS AND FUTURE WORK

In this study, we employ four different regression techniques from machine learning domain for metamodeling. For each metamodeling technique, we consider four different sampling techniques for training purposes. Results show that the analyst can obtain predictions using a metamodel trained with a small training set instead of running the simulation model in a relatively short time (e.g., 80% shorter than running the simulation model). However, to increase the prediction accuracy of a metamodel, the analyst should expand the training set, which naturally increases training time.

Although the metamodeling techniques used in this study are time-saving, the analyst should use them with caution since metamodel accuracies depend on output types. Results show that *team total cost*, which is a system-level output, is predicted with higher accuracy compared to *maximum inventory level of the retailer* under equal sample sizes. We can conclude that system-level outputs are easier to predict compared to individual-level outputs. However, this claim should be validated by further experimenting with other agent-based models. We also observe that we should increase the training set size when the model output is individual-level or system-level with high dimensions to obtain better metamodel predictions.

Experimental results show that there is no single metamodeling or sampling technique performing best in all cases. However, we observe that support vector regression is more robust to the increase in the dimension of the problem. Besides, in all of the four cases, highest proportion of instances predicted with maximum 10% error are realized when we use support vector regression method.

Although the error values are very high in some cases, metamodels can guide the analyst to explore and focus on the parameter subspaces where model output deviates from the regular form captured by the metamodel. These subspaces will potentially be in the neighborhood of the sample points where the error values obtained by leave-one-out cross-validation process are high. In that respect, the added value of metamodels may be more about guiding the model exploration process, rather than substituting the model. A metamodel that is trained with a small training set (e.g., 25 model runs) may narrow down the parameter space that needs to be explored significantly, and reduce the time and effort required in exploring the behavior space of an agent-based model.

As a continuation of this study, we are planning to increase the input parameter set gradually up to eight with the Beer Game, and observe the performance deterioration as well as the required increase in the training set to compensate that. Furthermore, we plan to expand the study by following a similar procedure with other agent-based models.

ACKNOWLEDGEMENTS

This research is supported by Bogazici University Research Fund (Grant No: 12560 - 17A03D1).

REFERENCES

Alam, F. M., K. R. McNaught, and T. J. Ringrose. 2004. “A Comparison of Experimental Designs in the Development of a Neural Network Simulation Metamodel”. Simulation Modelling Practice and Theory 12 (7): 559–578.

Alpaydin, E. 2014. Introduction to Machine Learning. MIT Press.

Axtell, R. L., J. M. Epstein, J. S. Dean, G. J. Gumerman, A. C. Swedlund, J. Harburger, S. Chakravarty, R. Hammond, J. Parker, and M. Parker. 2002. “Population Growth and Collapse in a Multiagent Model of the Kayenta Anasazi in Long House Valley”. Proceedings of the National Academy of Sciences 99 (suppl 3): 7275–7279.

Beachkofski, B., and R. Grandhi. 2002. “Improved Distributed Hypercube Sampling”. In Proceedings of the 43rd

AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 1274.
Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer.
Breiman, L. 2001. “Random Forests”. Machine Learning 45 (1): 5–32.
Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and Regression Trees. CRC Press.
Chen, X., M. Wang, and H. Zhang. 2011. “The Use of Classification Trees for Bioinformatics”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (1): 55–63.
Chen, Y., E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. 2009. “Similarity-Based Classification: Concepts and Algorithms”. Journal of Machine Learning Research 10 (Mar): 747–776.
Crombecq, K., L. De Tommasi, D. Gorissen, and T. Dhaene. 2009. “A Novel Sequential Design Strategy for Global Surrogate Modeling”. In Proceedings of the 2009 Winter Simulation Conference, 731–742. IEEE, Piscataway, N.J.
Deutsch, J. L., and C. V. Deutsch. 2012. “Latin Hypercube Sampling with Multidimensional Uniformity”. Journal of Statistical Planning and Inference 142 (3): 763–772.
Edali, M., and H. Yasarcan. 2014. “A Mathematical Model of the Beer Game”. Journal of Artificial Societies and Social Simulation 17 (4).
Edmonds, B., and D. Hales. 2005. “Computational Simulation as Theoretical Experiment”. Journal of Mathematical Sociology 29 (3): 209–232.
Epstein, J. M., and R. Axtell. 1996. Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press.
Gilbert, N. 2008. Agent-Based Models. Number 153. Sage.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.
Hu, C., G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka. 2014. “Data-Driven Method Based on Particle Swarm Optimization and k-Nearest Neighbor Regression for Estimating Capacity of Lithium-Ion Battery”. Applied Energy 129:49–55.
Juutilainen, I., and J. Röning. 2007. “A Method for Measuring Distance from a Training Data Set”. Communications in Statistics–Theory and Methods 36 (14): 2625–2639.
Keane, A., and P. Nair. 2005. Computational Approaches for Aerospace Design: The Pursuit of Excellence. John Wiley & Sons.
Kleijnen, J. P., S. M. Sanchez, T.W. Lucas, and T. M. Cioppa. 2005. “State-of-the-Art Review: A Users Guide to the Brave New World of Designing Simulation Experiments”. INFORMS Journal on Computing 17 (3): 263–289.
Kleijnen, J. P., and R. G. Sargent. 2000. “A Methodology for Fitting and Validating Metamodels in Simulation”. European Journal of Operational Research 120 (1): 14–29.
Kleijnen, J. P., and W. C. van Beers. 2004. “Application-Driven Sequential Designs for Simulation Experiments: Kriging Metamodelling”. Journal of the Operational Research Society 55 (8): 876–883.
Lee, J.-S., T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, J. G. Polhill, Z. Sun, and D. C. Parker. 2015. “The Complexities of Agent-Based Modeling Output Analysis”. Journal of Artificial Societies and Social Simulation 18 (4).
Loh, W.-Y. 2011. “Classification and Regression Trees”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (1): 14–23.
Macal, C. M. 2010. “To Agent-Based Simulation from System Dynamics”. In Proceedings of the 2010 Winter Simulation Conference, 371–382. IEEE, Piscataway, N.J.
McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code”. Technometrics 21 (2): 239–245.
Montgomery, D. C. 2013. Design and Analysis of Experiments. John Wiley & Sons.
Quinlan, J. R. 1993. C4.5: Programming for Machine Learning. Morgan Kauffmann.
Sanchez, S. M., and T.W. Lucas. 2002. “Exploring theWorld of Agent-Based Simulations: Simple Models, Complex Analyses”. In Proceedings of the 2002 Winter Simulation Conference, 116–126. IEEE, Piscataway, N.J.
Sterman, J. D. 1989. “Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment”. Management Science 35 (3): 321–339.
ten Broeke, G., G. van Voorn, and A. Ligtenberg. 2016. “Which Sensitivity Analysis Method Should I Use for My Agent-Based Model?”. Journal of Artificial Societies and Social Simulation 19 (1).
Tesfatsion, L. 2002. “Agent-Based Computational Economics: Growing Economies from the Bottom Up”. Artificial Life 8 (1): 55–82.
Tversky, A., and D. Kahneman. 1974. “Judgment Under Uncertainty: Heuristics and Biases”. Science 185 (4157): 1124–1131.
Zhu, P., F. Pan, W. Chen, and S. Zhang. 2012. “Use of Support Vector Regression in Structural Optimization: Application to Vehicle Crashworthiness Design”. Mathematics and Computers in Simulation 86: 21–31.
Zhu, P., Y. Zhang, and G. Chen. 2009. “Metamodel-Based Lightweight Design of an Automotive Front-Body Structure using Robust Optimization”. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 223 (9): 1133–1147.
Wilensky, U., and W. Rand. 2015. An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. MIT Press.

AUTHOR BIOGRAPHIES

MERT EDALI is a Research and Teaching Assistant in Industrial Engineering Department at Yildiz Technical University. He earned his B.S. degree from Yildiz Technical University, Istanbul, Turkey, in 2011. He earned his M.S. degree in Industrial Engineering from Bogazici University, Istanbul, Turkey, where he continues his studies as a PhD student. His e-mail address is `medali@yildiz.edu.tr`.

GONENC YUCEL is an Associate Professor in Industrial Engineering Department at Bogazici University. He received his B.S. and M.S. degrees in Industrial Engineering from Bogazici University in 2000 and 2004. He earned his PhD degree in Policy Analysis from Delft University of Technology. He has been focusing on simulation-supported policy and strategy analysis in his research, utilizing agent-based, as well as system dynamics models. His email address is `gonenc.yucel@boun.edu.tr`.

BLIND SERACH PATTERNS FOR OFF-LINE PATH PLANNING

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

KEYWORDS

Lévy search, path planning, seed spreader, inverse-Lévy patterns, autonomous underwater vehicles.

ABSTRACT

Path planning is crucial for efficient utilisation of autonomous underwater vehicles. The goal of the mission of an autonomous underwater vehicle determines suitable strategies for path planning. Blind search methods can be used for off-line path planning for unknown environments to locate phenomena of interest. Different blind search patterns have been implemented and evaluated in terms of their ability to reach the mission's goal. A novel blind search pattern that is based on a truncated Lévy distribution is also proposed and compared with other search patterns as path-planning algorithms. The simulations show that Lévy search pattern can outperform other search patterns for small size phenomena. On the other hand, the proposed inverse-Lévy pattern can locate large size phenomena more than other search patterns. The simulations show that the probability of locating the most important phenomenon by a single autonomous underwater vehicle using blind search patterns is much smaller than of a swarm of autonomous underwater vehicles in similar conditions. However, Lévy and inverse-Lévy can be used for the worst-case scenario of no communication nor ability to use feedback information.

INTRODUCTION

Autonomous Underwater Vehicles (AUVs) have attracted significant attention in the last several years and have been widely used in marine geoscience, military, commercial and environmental applications (Wynn et al. 2014). Their ability to operate autonomously makes them the best option for fast and accurate exploration, inspection and search in extreme and unknown environments. They are key players for exploring and searching different water systems where the risks are high. They can be used to locate harmful dumped waste, lost ship containers and collect data in inaccessible or dangerous underwater environments.

Efficient path planning is a crucial issue for modern AUVs. Path-planning algorithms produce feasible trajectories for AUVs to reach their goal. They generate geometric paths, without considering any specified time law (Gasparetto et al. 2015). The goal of the mission of an AUV determines the strategy for path planning. Collecting data from underwater environment, for example, may require a strategy that enables the AUV to pass over all the points of the area or volume of interest within a set of constrains. Different Coverage Path-Planning (CPP) algorithms have been used to determine such trajectories (Galceran and Garreras 2013). Finding locations of harmful garbage, groundwater outflow or lost cargo may necessitate using a different strategy that enables finding the optimal path to these locations with different constrains (Nolle 2015). Efficient CPP algorithms can be used for searching since they usually sample the mission's space to have a complete view of the phenomena of interest. However, they may not be as efficient as artificial search algorithms that utilise minimum samples to reach their goal.

Regardless of the mission's aim, path-planning algorithms should be reliable and enable the AUVs to navigate in both known and unknown environments. Preliminary path is usually defined for an AUV based on the available information of the mission environment and its kinematic constraints. Path and trajectory planning algorithms can used to generate this preliminary path. In off-line path-planning algorithms where the environment is assumed to be known in advance, the AUV follows usually this pre-defined path without any modification. For on-line algorithms, also known as sensor-based algorithms (Noborio et al. 2000), the path is continuously updated through a cognition algorithm that use information from the environment taking into account the characteristics of different sensors and kinematic and communication constraints.

In order to develop an autonomous platform for submarine exploration, which is the final aim of this research, off-line path-planning algorithms should be developed for robust navigation. In this research, a small swarm of AUVs share their individual experience during the search to locate the phenomena of interest and a mechanism for optimal utilization of the cumulative experience will be used to direct the swarm towards the location of those phenomena. However, communication and localization problems can be obstacles in sharing and utilising the cumulative experience of the swarm of AUVs (Tholen et al. 2017). Effective and efficient off-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

line path pattern can help to alleviate these problems towards robust search.

Blind Search and Off-Line Path Planning

This paper builds on the results of another study of the same research that recommends investigating the behaviour of other search algorithms to use in a case of facing problems in communication between AUVs (Tholen et al. 2017). Here, the worst-case scenario of off-line path planning is studied. In this context, path planning refers to the process of producing a geometric path to the highest priority phenomenon. The path of the AUV to find a global optimum is to be pre-defined with no information about the environment, except its boundaries. A set of path planning and search algorithms were implemented and evaluated in terms of their ability to define a preliminary path for AUVs to reach their common goal in an unknown environment. The suitability of using these algorithms by each individual AUV in a swarm to guide the search, in the case of no communication and difficulties in controlling the movement based on the available information, is studied.

Search for targets should optimise the chances of reaching them. For a stable environment with prior knowledge, deterministic search patterns can be efficient. However, in dynamically changing or unknown environments, probabilistic search methods can be more efficient. In order to study the worst-case scenario with no prior knowledge, no communication and no way to improve the search path utilising feedback information, the seed spreader algorithm (Galceran and Garreras 2013) has been selected as deterministic algorithm to evaluate its ability to guide the search towards the global optimum.

On the other hand, different probabilistic search algorithms can be used for path planning. However, for this study, the search algorithms those comply with off-line path-planning requirements, as described above, are single-point blind search methods. These search methods iterate a single starting solution to reach the optimal solution and do not use any kind of feedback. Two well-known pure random single-point search algorithms, i.e. Lévy flights (Viswanathan et al. 1999) and random walks, have been selected for this purpose. In addition, a third probabilistic search algorithm based on Lévy flights was proposed.

The seed spreader algorithm is a deterministic path-planning pattern for complete coverage of simple regions. It covers the area of interest in a sweeping motion pattern by moving back and forth. It is also known as lawn mowing. This search pattern can guide the AUV to locate the search goal. The seed spreader algorithm, when used as path-planning mechanism to cover the space of interest, decomposes the space into cells and the efficiency of the algorithm depends on the details of the decomposition process. Since one of the objectives of the research is to evaluate the performance of the seed spreader as a deterministic search algorithm, the motion pattern is simulated without decomposition.

Three probabilistic search patterns were selected for path planning. Random search methods can be a valid option for path-planning problems and have been successfully applied in some floor-cleaning robots (Galceran and Garreras 2013). Different random search methods can be used for solving the path-planning problem. Starting from a position, different path patterns can be generated by generating a sequence of random steps in terms of their lengths and directions. The produced search pattern is controlled by the probability distribution used to generate the length and the direction of steps.

In the random walk search for path planning, the AUV follows randomly pre-generated step sizes in random directions within the specified region. The step length and its direction can be randomly selected based on uniform distributions. The x and y coordinates of each step can also be generated using uniform distributed random numbers. By randomly generating the coordinates, the next step length and its direction can be defined. When applying random walk algorithms, the range of the step lengths need to be set according to the problem at hand. Selecting the range of step lengths can affect the search performance.

Another random and well-known biological search is Lévy flights or walks (Sims et al. 2008). Lévy flights represent an optimal solution to the biological search problem in complex landscapes (Reynolds 2015). This model as a search pattern can suits dynamic and complex environments. Lévy search is a kind of a random walk search with unique characteristics. Its search patterns consist of repeated clusters of relatively short step length and rare longer steps. The step lengths of Lévy flight are generated by a probability distribution with a power-law tail. The step lengths can be generated using formula (1)

$$l = r^{\frac{-1}{\alpha}} \tag{1}$$

where l denotes the step length, r a uniformly distributed random number in the range $[0,1]$ and α is a parameter can have any value in the range$[1,2]$.

The step length formula can be applied to generate the step length and its direction can be chosen based on the density function of a uniform distribution. The formula can also be used to produce x and y coordinates of a Lévy step in two-dimensional space.

Solving a problem using Lévy search patterns requires selecting proper parameters for the search. The value of α can impact the search pattern since it affects the step length. The step length equation (1) can produce very large values that can be outside the search range for random numbers near zero. In order to limit the length of the step size to a specific value (l_{max}) within in the search range, the generated random number can be mapped into

a sub-interval of the interval [0, 1]. The new interval is defined by$[\,r_{min}, 1]$, where r_{min}can be calculated using equation (2).

$$r_{min} = \frac{1}{l_{max}{}^{\alpha}} \tag{2}$$

Limiting the step length of Lévy patterns produced a truncated Lévy search (Xiong and Lam 2010), where $P(x) = 0\,, x > l_{max}\ and\ P(x) = 0, x < 1$. A truncated Lévy search can be more applicable in search, however, it introduces a new parameter that needs to be set and can influence the search patterns.

A third random search pattern, which combines exploring and exploitation with different aspects from that of Lévy search, can help finding an optimal path to the search goal. In Lévy search, very small steps dominate the search pattern, whereas in the proposed search, the pattern is dominated by long steps. In this mechanism, which is referred to as inverse-Lévy in this paper, the search pattern consists of repetition of a cluster of long steps followed by occasional short steps. Having such pattern can promote exploring the whole area with occasional exploitation. This combination of exploring and exploitation can suit finding a path for the phenomenon of interest in unknown environments. The step lengths are based on a Lévy distribution and derived from equation (1). The step length for Inverse-Lévy is calculated according to equation (3).

$$l_{IL} = l_{max} - r^{\frac{-1}{\alpha}} \tag{3}$$

where l_{IL} denotes the step length of inverse-Lévy, l_{max} is a parameter that defines the maximum step length, r a uniformly distributed random number in the range $[\,r_{min}, 1]$, where: r_{min} is defined in equation (2), and α is a parameter that can have any value in the range [1,2].

In inverse-Lévy pattern search, two parameters, as in truncated Lévy, can affect the details of these patterns and can influence its ability to plan for optimal paths. However, since step lengths close to the maximum step length are expected to dominate the search pattern, the search pattern will be more sensitive to the value of l_{max} parameter.

SIMULATION

A set of experiments was conducted using different search patterns to locate phenomena of interest with different priorities and different sizes. The seed spreader, random walk, Lévy flight and inverse-Lévy were implemented, tested and evaluated for their ability to locate phenomena of interest in an area of 400m x 400m. These search patterns were tested with different parameter values for different phenomena. The four algorithms exhibit different search patterns while exploring the mission area. Figure 1 shows the search patterns of these algorithms with a maximum step length of 50 meter.

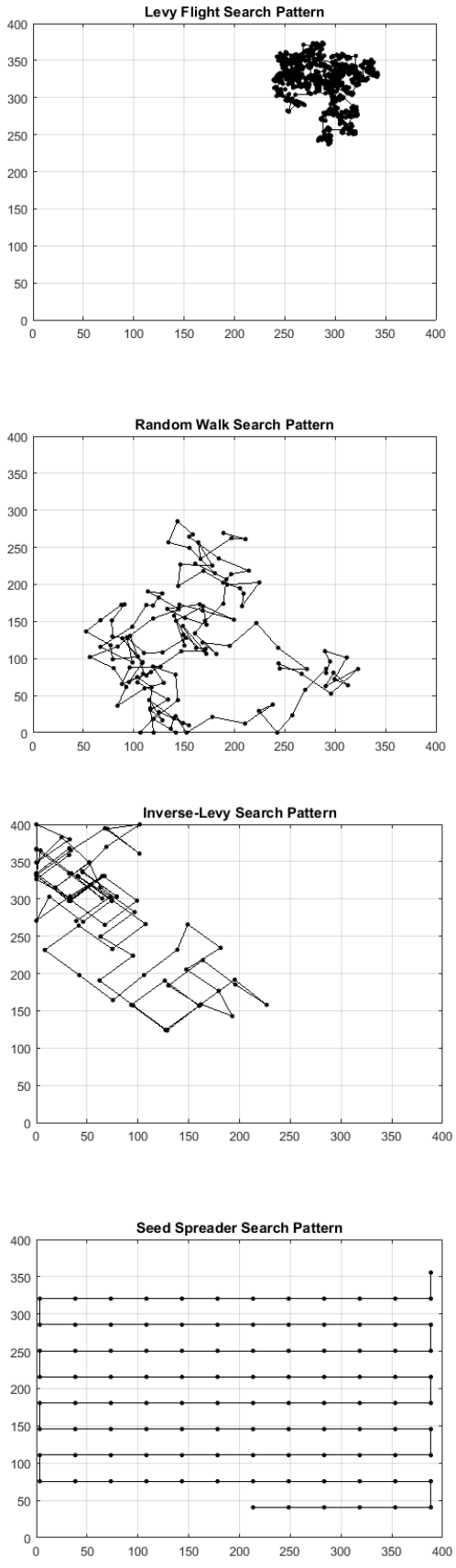


Figure 1: Search Patterns of Different Algorithms

In these experiments to simulate a multimodal search space, three phenomena of interest are assumed having

random locations in the mission area. Since blind search algorithms do not use any gradient information, different priorities instead of fitness score were assigned to different phenomena. The experiments were conducted with phenomena of two different sizes. The performance of each search pattern is evaluated in terms of its ability to find the phenomenon of the highest priority and in terms of finding other phenomena. The ability of the algorithms in locating phenomena of different priorities with the same size were compared.

The stopping criteria for search is either performing the maximum number of steps or travelling for a maximum distance. In all the experiments, the maximum number of steps was set to 1000 and the maximum travelling distance was set to 3600 meters because of the energy constraints of the AUV. Each experiment was repeated for 1000 times.

In the implemented seed spreader algorithm, the step lengths in both directions are equal. The theoretical optimal value of the step length to explore the mission area, given the maximum travelling distance constrain, can be calculated (Tholen et al. 2018) and has a value of about 57 meter. However, using this step length, in the worst-case scenario, the seed spreader algorithms can locate phenomena with a minimum radius of 28.5 meter. The algorithm was tested for different step lengths above and below this value. It was tested for step length values of 5 meters to 100 meters with an increment of 5 meters. The same range of values were also used as maximum step length for random search, Lévy and inverse-Lévy algorithms. In random search, both coordinates of each step are selected randomly from a range of values from zero to a maximum step length. For both Lévy and inverse-Lévy methods, the algorithms were tested using different values of α in the range$[1,2]$.

Walk and Flight Scenarios

While tracking the pattern defined by the algorithm, the AUV can either follow one of two scenarios. The first scenario is to scan the points on its path to the next step location for a phenomenon. The second scenario is to examine only the locations of the steps for phenomena. In the remaining of the paper, when a search pattern uses the first scenario it is referred to as a walk. Random walk, Lévy walk, inverse-Lévy walk and the seed spreader walk scan the path while moving to the next step position. On the other hand, if the search pattern follows the second scenario, it will be referred to as flight. In other words, random flight, Lévy flight, inverse-Lévy flight and the seed spreader flight focus on the location of the next step.

The walk scenario can be applied when the AUV is equipped with sensors that can measure environment data for phenomenon recognition using minimum resources. While the flight scenario can be more suitable for AUVs with sensors that consume considerable amount of resources (Zielinski et al. 2009). However, the experiments were conducting with the assumption that the AUVs are using ideal sensors.

The two different scenarios have been simulated using the different search patterns. In each set of experiments, the search patterns were evaluated for finding phenomena with small and large sizes. The size of large phenomenon is assumed equals to the area of a circle with a radius of 20 meter. The small phenomenon has a radius of one meter.

The starting point of each search pattern is selected randomly for fair comparison between different search algorithms. It is also based on the worst-case situation where the AUV is in the middle of searching and becomes unable to communicate with the rest of the swarm and unable to use environment data to guide the search.

RESULTS

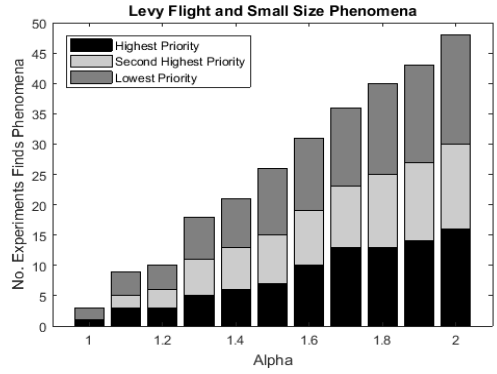
The results of the simulation for finding a path to the mission goal are shown in Figures 2 to 9. These figures show the number of times that each algorithm succeed in finding the phenomenon of highest, second highest and lowest priority. If an algorithm locates more than one phenomenon, it will be assigned to the phenomenon of the highest priority. In other words, the algorithm that managed to locate the highest priority phenomenon may also succeed in locating the second highest and the lowest priority phenomena.

In the graphs that show the performance of the different algorithms, different scale ranges were used to obtain a close view of the performance of the search patterns. For phenomena of small sizes, the maximum value of the scale range is set to 50 for flying scenario and to 180 for walking scenario. Whereas, in the graphs that compare the performance of the algorithms on large sizes phenomena the maximum scale ranges are 700 and 900 for both scenarios, respectively.

Flights to Phenomena

The four search patterns were tested for locating the phenomenon with the highest priority and locating more than one phenomenon. The position of three phenomena have been selected randomly in the search space and tested for the different search algorithms that examine the step locations only for a phenomenon. Search patterns with the concept of flying to the goal were tested with phenomena with equal sizes but different priorities.

The experiments for small size phenomena show that Lévy flight search pattern can find the phenomenon with the highest priority and other phenomena more than inverse-Lévy flights (Figures 2 and 3). On the other hand, Inverse-Lévy flight outperforms Lévy flights for large phenomena sizes. The graphs also show the performance of both search patterns is highly dependent on α.



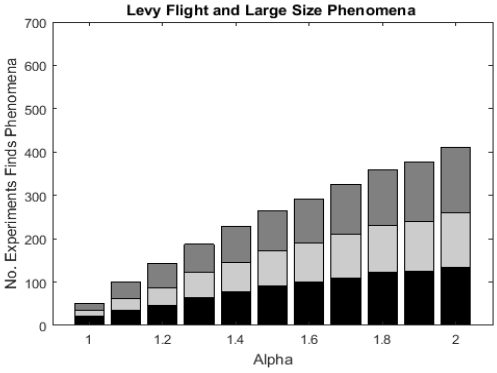
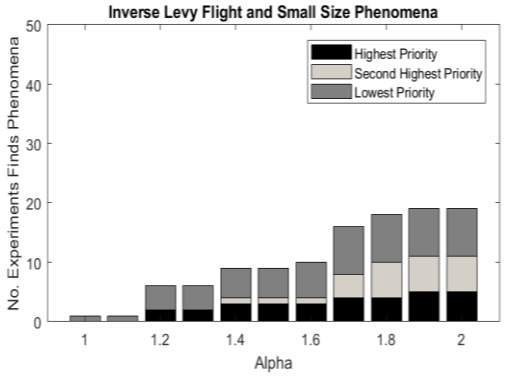


Figure 2: Lévy Flight



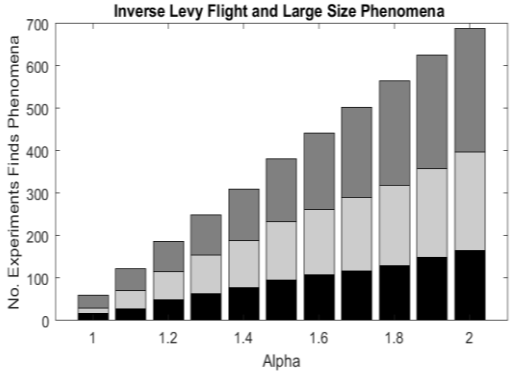


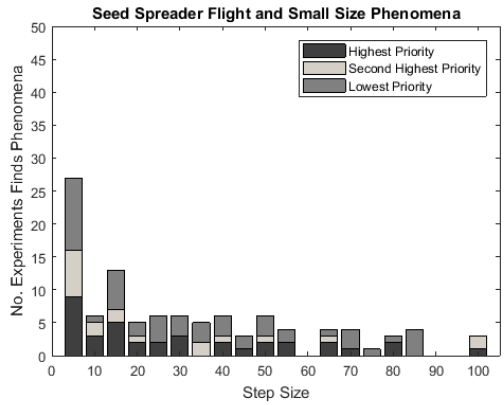
Figure 3: Inverse -Lévy Flight

The simulations also show that Lévy flight outperforms the other two search patterns, i.e. the seed spreader flight (Figure 4) and random flight (Figure 5) for phenomena of small sizes. Nevertheless, for large phenomenon sizes, the seed spreader algorithm performs better than other search algorithms, in most cases, in terms of finding at least one phenomena. However, for specific values of α and maximum step length, the inverse-Lévy method outperforms other algorithms. The experiments show that for large phenomena, the seed spreader algorithm in most cases guides the search to the lowest priority phenomenon. Whereas, each phenomenon has the same chance to be located by other algorithms.

Walks to Phenomena

The same set of experiments was repeated with scanning for phenomena while moving to the next position. In these experiments, an AUV scans the search pattern every meter for a phenomenon.

The simulations show an improvement in locating phenomena for all algorithms as expected (Figures 6 to 9). The simulations demonstrate that for specific values of α and step lengths, inverse-Lévy walk (Figure 6) outperforms other search walks.



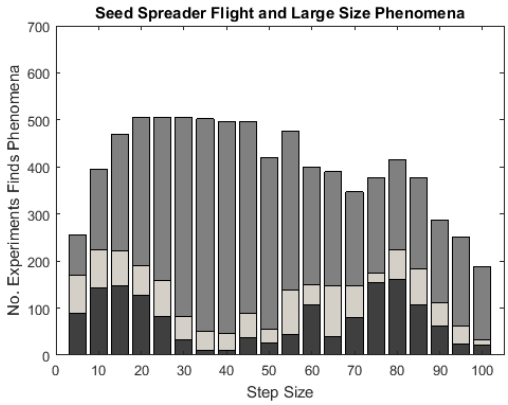
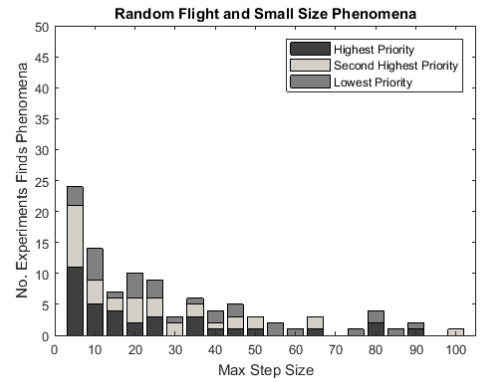


Figure 4: The Seed Spreader Flight

The experiments demonstrate the same trend, noticed in the experiments of fly to the goal, where the seed spreader algorithm (Figure 7) usually guides the search to the lowest priority phenomenon in contrast to other algorithms.



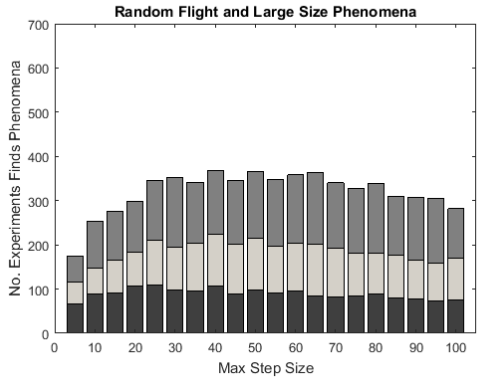
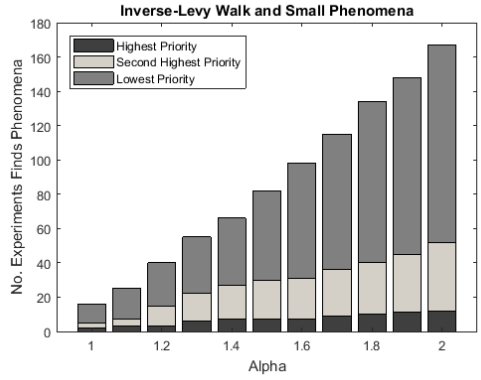


Figure 5: Random Flight



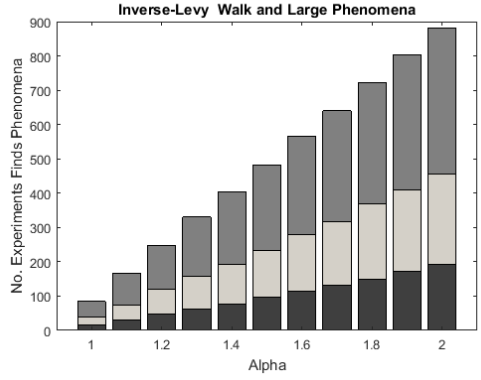
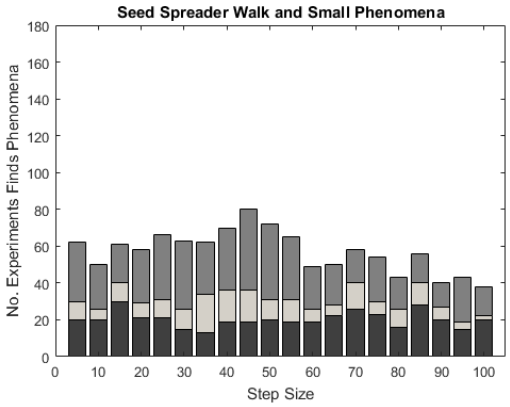


Figure 6: Inverse-Lévy Walk



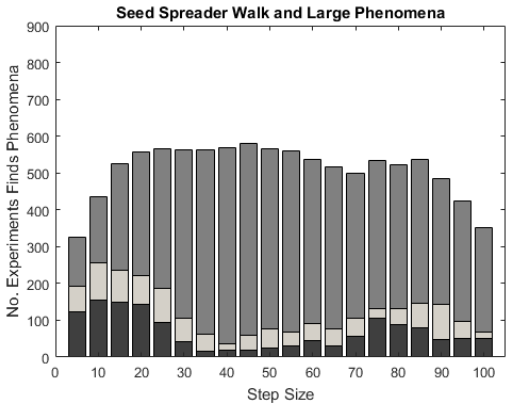


Figure 7: The Seed Spreader Walk

The experiments for locating small size phenomena demonstrate that for Lévy walks with $\alpha \geq 1.4$ Lévy patterns (Figure 8) slightly outperform other search patterns and inverse-Lévy has the worst performance

The performance of random walk (Figure 9) in finding phenomena of small size is much better than random flights (Figure 5). However, for locating phenomena of large sizes there is no significant improvement in the performance compared with random flights.

CONCLUSION AND FUTURE WORK

The experiments clearly show that the probability of locating the phenomenon of interest using a single AUV with blind path-planning algorithms is very small even for phenomena with considerable large sizes. If this probability compared to that of a Particle Swarm Optimisation (PSO) with three AUVs (Tholen et al. 2017), working in perfect conditions as that of the simulation in this paper, blind path-planning algorithms cannot be used alone to locate points of interest.

However, they might be used as backup search in the worst-case scenario. In that case, for small phenomena, Lévy search patterns with specific values of α can be used since it can direct the search toward the most interesting phenomenon compared to other algorithms. For large sizes phenomena, inverse Lévy may be a good

option. However, the performance of Lévy and inverse-Lévy depends on the settings of the control parameters, i.e. the maximum step length and α. Inverse-Lévy search patterns are more sensitive to the maximum step length than Lévy search. Using feedback information about the search progress can help in self-adapting the control parameters and might improve their performance.

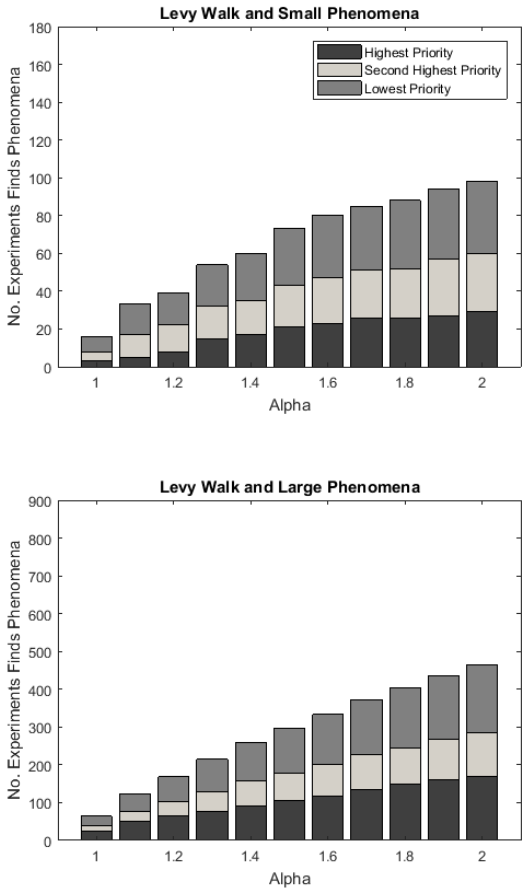


Figure 8: Lévy Walk

Studying different self-adaptive techniques and the possibilities of applying them to Lévy or inverse-Lévy search patterns is the next step of this research. There are different ways to make these search patterns self-adaptive. A possible adaptive technique for inverse-Lévy is to control its maximum step length by increasing it when there is no improvement in the search to promote exploration while reducing exploitation. Adaptation can also be done by extending the search space of the search algorithm to include the two control parameters of inverse-Lévy or Lévy, i.e. l_{max} and α, and redefining the problem as to find the best parameters and the location of the most important phenomenon.

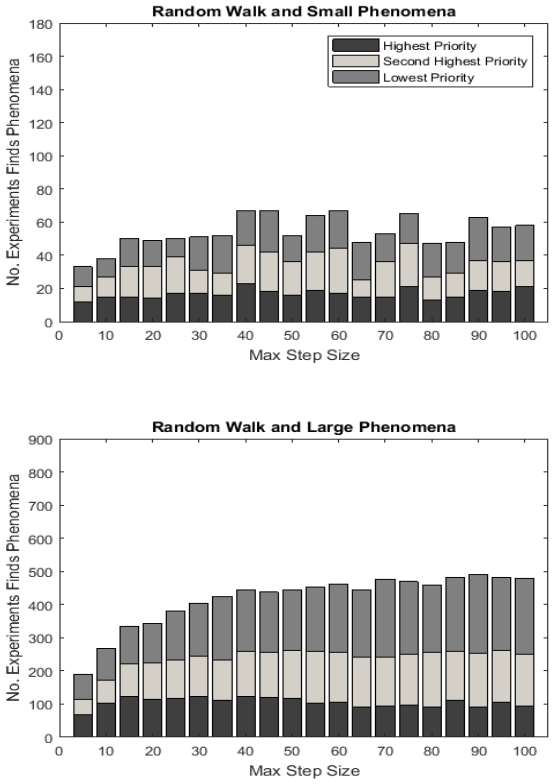


Figure 9: Random Walk

REFERENCES

Galceran, E., & Garreras, M. (2013). A Survey on Coverage Path Planning for Robotics. *Robotics and Autonomous Systems*, 1258-1276.

Gasparetto, A., Boscariol, P., Lanzutti, A., & Vidoni, R. (2015). Path planning and trajectory planning algorithms: A general overview. In E. Carbone and F. Gomez-Bravo, *Motion and Operation Planning of Robotic Systems: Background and Practical Approaches* (S. 3–27). Cham, Switzerland: Springer-Verlag.

Noborio, H., Yamamoto, I., & Komaki, T. (2000). Sensor-based path-planning algorithms for a nonholonomic mobile robot. *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)*, (S. 917-924).

Nolle, L. (2015). On a search strategy for collaborating autonomous underwater vehicles. *Mendel*, (S. 159-164).

Reynolds, A. M. (2015). Extending Lévy search theory from one to higher dimensions: Lévy walking favours the blind. *A Royal Society journal*.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., . . . Shepard, E. L. (2008). Scaling laws of marine predator search behaviour. *Nature* .

Tholen, C., El-Mihoub, T., Nolle, L., & Zielinski, O. (2018). On the robustness of self-adaptive Levy-flight. *OCEANS'18 MTS/IEEE Kobe / Techno-Ocean 2018* (S. In press). Kobe: IEEE.

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. *Mendel.*

Viswanathan, G. M., Buldyrev, S., Havlin, S., da Luz, M., Raposo, E., & Stanley, H. (1999). Optimizing the success of random searches. *Nature*, 911–914.

Wynn, R. B., Huvenne, V. A., Le Bas, T. P., Murton, B. J., Connelly, D. P., Bett, B. J., . . . Hunt, J. E. (2014). Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience. *Marine Geology*, 451-468.

Xiong, J., & Lam, J. (2010). Using truncated Le´vy flight to estimate downside risk. *Journal of Risk Management in Financial Institutions* , 231–242.

Zielinski, O., Busch, J., Cembella, A., Daly, K., Engelbrektsson, J., Hannides, A. K., & Schmidt, H. (2009). Detecting marine hazardous substances and organisms: sensors for. *Ocean Science*, 329-349.

AUTHOR BIOGRAPHIES

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 Sciences. His current research is in the fields of applied computational intelligence and autonomous underwater vehicles.

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.

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.

REALTIME SIMULATION AND 3D-VISUALISATION OF SURFACE AND UNDERWATER VEHICLES FOR MONITORING AND EVALUATING AUTONOMOUS MISSIONS

Tobias Theuerkauff, Yves Wagner, Frank Wallhoff
Institute for Technical Assistive Systems
Jade University of Applied Sciences
26121 Oldenburg, Germany
E-mail: {tobias.theuerkauff,yves.wagner,frank.wallhoff}@jade-hs.de

KEYWORDS
Simulation, 3D-Visualisation, Real Time, Low-Cost-Monitoring, Unmanned Surface Vehicle (USV), Autonomous Underwater Vehicle (AUV), Remotely Operated Vehicle (ROV), Mission Planning, Maritime

ABSTRACT

The use of autonomous underwater vehicles requires stable and reliable algorithms within the control software. A misdirected vehicle can quickly lead to a costly damage or even loss of the vehicle. In this paper, an application is presented that allows an ongoing underwater mission to be tracked in real-time within a 3D simulation and, in the event of a problem, aborts it to set the vehicles into a safe state. It can communicate directionally with the control software of each vehicle, thus providing the basis of a simulation environment. Furthermore, first implementations for evaluating the control algorithms of autonomous vehicles within the simulation environment will be presented.

INTRODUCTION

The exploration and use of underwater regions in the oceans as well as in inland waters are becoming increasingly important. The use of these regions depends on the research of these areas. Furthermore, the already economically exploited regions, e.g. through the construction of offshore wind turbines, has to be constantly monitored and controlled (Tchakoua et al. 2014) as well as the inspection of underwater pipelines (Xiang et al. 2014). The use of divers to solve the problem is often used successfully today. However, the use of human resources is a costly, time-consuming and dangerous solution. To curb these factors, an increased use of autonomous underwater vehicles is sought. The autonomous underwater missions with independent, or as well as the composite of autonomous operating vehicles is currently increased in the development. One problem with the use of autonomous vehicles under water is the lack of visual contact. In order to make an autonomous movement of the vehicle under water possible, the vehicles must be equipped with various sensors to capture their environment. In addition, the algorithms to control the vehicle must run absolutely reliable and error-free to prevent loss or damage to the vehicles. For the interpretable perception of the environment, the cognitive control algorithms must be able to process and evaluate all sensor data in parallel and in real time. In particular, in the test phase of the control algorithms malfunction can quickly occur. In this work a prototypical software is presented, which visualizes the movements of the vehicles in real time in a 3D environment. The entire mission of the vehicle, or multiple vehicles within a compound operation, can be permanently monitored and analysed with the visualization. In the case of occurring errors within the control algorithms or in case of misinterpretations of the multisensor data, the operator can thus perform a manual and situation-related intervention in the mission execution manually. Thus, the risk of costly damage or the loss of vehicles can be greatly reduced. The system is based on a low-cost implementation and can be applied to various underwater and surface vehicles as well as complete overwater scenarios.
In the further course, the visualization environment is to be further developed into a simulation environment for testing and evaluating underwater vehicle control algorithms. First the visualization as well as the communication of the visualization environment to real and emulated vehicles will be presented. Approaches and first implementations for the simulation of vehicles with emulated sensors are also considered.

RELATED WORKS

At the University of Porto in the LSTS institute was developed a 2D-visualization of subsurface, surface and air vehicles for supporting multiple vehicle operations (Dias et al. 2005). The map basis for the visualization of the vehicles are electronic nautical charts (ENC data), which are used in the seafaring for navigation. Within the map all the different vehicles, which are registered in the network, are shown in the visualization. The state of the position is not visualized in a fluid visualization, but based on the latest received data of the vehicles. The connection to the vehicles takes place via receiving the IMC-messages send by different vehicles. The application can receive messages as well as send messages via a xml-console into the network to communicate with each registered vehicle. Furthermore, it is possible to plan a mission for one or more vehicles

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

in the network and to send and execute it within the vehicles (Dias et al. 2006).
In the underwater simulator from the IRS Lab the OpenSceneGraph library is used as the basis for visualizing underwater missions. With the open-source UWSim they provide a possibility for simulation and graphical representation of underwater missions. For the control of virtual underwater vehicles and the spread of simulated sensor data, a network-based interface is applied. The underwater terrain model can be configured by the user via a XML file. For this, parameters such as water surface, underwater visibility and particle density in water can be defined. The underwater robot can also be adapted via an XML file. By default, an underwater vehicle with six degrees of freedom (6 DOF) in motion is provided. The software further offers individually configurable sensors that can be linked to the vehicles. It is possible to control several robots within a simulation. The integrated physics engine osgBullet offers the possibility to simulate highly simplified, physical aspects in the underwater world (Prats et al. 2012).

CONCEPT

Submarine vehicles may be connected by a cable to a surface vessel or to an overwater central facility via a cable. In this case we speak about a Remotely Operated Vehicle (ROV). Furthermore, it is possible to connect the vehicle via a wireless connection with a surface vessel or a central office. For this purpose, an Autonomous Underwater Vehicle (AUV) in a Network, especially with larger distances and depths, the acoustic transmission of data is used. The last scenario is a fully autonomous use of the vehicle over a certain time frame up to several days - during which the vehicle has no connection to a base station or an surface vessel. The recorded data of the vehicle will be transmitted to a base station after the mission. The last scenario will not further be considered in this paper.

Within the acoustic communication a bidirectional data transfer between the AUVs are possible. In case of a wired connection, the underwater vehicles always communicate to each other via an USV (Figure 1). When locating and determining underwater vehicles with the use of acoustic modems, the underwater position of the vehicles is calculated in relation to the surface vessel or a base station. This is equipped with a GPS system so the coordinates of the vehicles can be determined within a world coordinate system. In the considered scenarios, the underwater vehicles are in constant contact with each other and with the surface vessel or the base station. Information and sensor- or steering data can be transmitted between the individual vehicles via a high data rate (cable-bound data transmission up to 1 gb/s) or low a data rate (acoustic data transmission with a few kb/s).

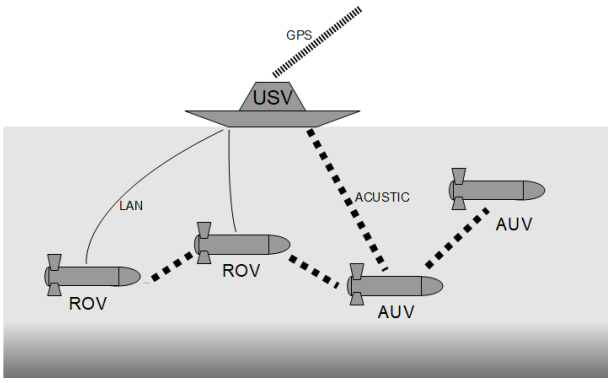


Figure 1: Schematic representation of communication between underwater vehicle and surface vessel

Depending on the transmission rate, simple data such as vehicle ID, position and time stamp up to complex data sets such as camera images can be transmitted. The transmission of data is realized via a uniform transmission protocol. Thereby, each vehicle is able to interpret and evaluate the received data. To do this, the control software of the vehicle must register as an actuator in the communication system. The ongoing communication as well as the generally defined transmission protocol and the registration in the communication network are the base concept of the developed simulations.

The visualization software also registers itself as an actor in the communication network and thus receives all transmitted data of the various vehicles (Figure 2). With every message of a vehicle, we also transmit the ID of the sender, with which the vehicle and the vehicle type can be clearly interpreted. The visualization software can thus automatically emulate and display the corresponding vehicle in the visualization for each vehicle. The software evaluates all received messages and assigns them to the respective emulated vehicle in the visualization. This allows the emulated vehicles to be positioned and oriented within the 3D visualization. Furthermore, all transmitted sensor data of the individual real vehicles can also be visualized on the virtual model. The vehicle status as well as the position, orientation, and sensor values of the vehicles can be permanently monitored by the operator. In the 3D visualization, a georeferenced, digital terrain model (DGM) of the underwater region can be displayed. Thus, by visualizing the position and orientation of the virtual vehicle in combination with the digital terrain model, the operator can detect a hazard, for example, in a foreseeable collision with the terrain. As the visualization software is registered as a real actuator in the communication network of the control software of the vehicles, a bidirectional communication is possible. It is thus able to distribute its own messages within the communication network. In the case of the described hazard scenario, the operator can send a message to the relevant vehicle, to enter into a safe state (e.g., direct arise).

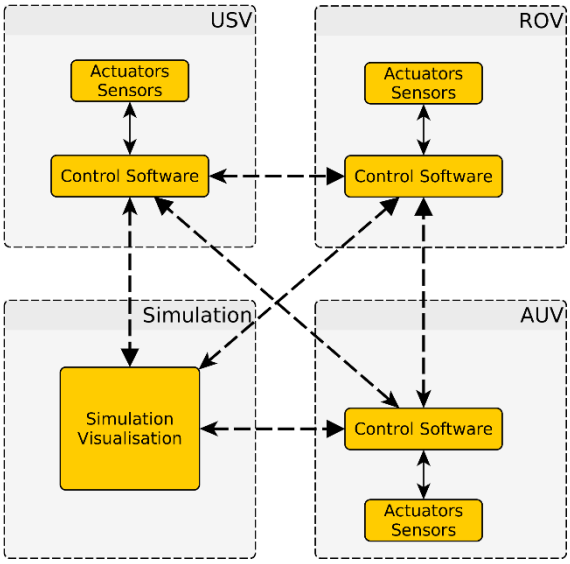


Figure 2: Schematic representation of the communication channels

PROTOTYP COMPOSITION

There are two main components within the development of this monitoring software. The basic hardware components (underwater vehicle and surface vessel) as well as the associated control software have to be considered. The implemented software must accordingly incorporate both components equally. In the experimental environment for the prototype development an underwater vehicle (ROV) was used. As control software the open-source software Dune has been used. The control software is installed on the one hand on the real vehicle to obtain real data. Furthermore, several instances of the control software, without any reference to a real vehicle, were started.

By doing this, several vehicles can be simulated, which are displayed within the visualization. The data of the control software of the real ROV and the data of the simulated ROV are transmitted by the registration in the communication network between all actuators. As a third instance, the visualization environment has been registered as a further actor in the network to receive all transmitted data in the network (Figure 2). For each vehicle, the corresponding, virtual vehicle is automatically added into the visualization. If a real vehicle sends its current position data, it will now be assigned to the corresponding virtual ROV in the visualization. By referring the true coordinates to a basic actuator, which is also displayed in the visualization, since it is registered in the network, the positions of the individual vehicles can be represented in accordance with reality.

Deployed Hardware

Within the project different hardware components were used, which are not based on a project-specific development. BlueRobotic's BlueRov2 underwater vehicle and EvoLogics acoustic communication modems are hardware components purchased for the higher-level project *Development of innovative technologies for autonomous maritime systems* (Entwicklung innovativer Technologien für autonome maritime Systeme - EITAMS).

The underwater vehicle BlueRov2 used in the project is a low-cost vehicle, which in private use is an out-of-the-box vehicle with visual control via PC. But it is also used in research projects for scientific research. It is a small vehicle which is maneuverable only in soft waters. But there it is very maneuverable due to the six specially arranged propellers and can accommodate additional loads and hardware through an expansion set. Additional weight must be compensated by buoyancy, which in turn limits maneuverability. The ROV is equipped with several sensors such as an inertial measuring system, temperature sensors and a camera. The hardware can be addressed freely. The BlueRov2 include a PixHawk controller with integrated IMU, which has its origin in the unmanned aircraft control.

To determine the underwater position of the ROV / AUV acoustic modems from EvoLogics GmbH are used. These modems are also used to transmit various data when the cable connection to the ROV is interrupted or deliberately omitted (AUV). But the data transfer rate of max. 31.2 kbps is very low. Every vehicle must be equipped with a USBL modem. The calculation of the position data is relative to the surface vehicle. This is additionally equipped with a GPS system, so relative coordinates can be transformed into absolute coordinates of a world coordinate system.

Deployed Software

The implemented and used software can be divided into three basic areas. The control of the simulated and the real ROV as well as for the control of the AUV a framework is used. This must be able to address the actuators as well as the sensors of the vehicles and distribute the data of the hardware in the system. The visualization of the vehicle in the 3D environment should be real-time capable, so that at any time the actual state of the vehicles is visible. The third area is a communication protocol that can be interpreted by all components.
The basis for the visualization and simulation is the Unreal Game Engine. We used the very powerful game engine from Epic Games Inc. It is open source available and is constantly being developed. Own source code can be implemented as well as the adaptation of the original code which is needed for the own applications. The Unreal Engine contains the typical core elements of a game engine (sound, graphics, network and physics engine) programmed in C++. Using an in-build editor, terrain models, simple geometries and complex visual effects can be embedded in the visualization environment. Furthermore, it is possible to import your own models and terrain data. The spatial reference is

produced via a metric coordinate system, which can be used as the basis for the georeferencing of the models and terrain data. Own source code can be programmed directly in C++ or via a visual scripting system (Blueprints). The connection between visual scripting and direct programming in source code is achieved through Unreal-Specific Classes (UCLASS) and attributes. Due to certain tags, the functions and variables that are programmed in the source code can also be addressed in the graphical interface with visual scripting. In-depth as well as basic functionalities can be implemented in the C++ code and made available for further development in the visual scripting editor. The use of the functionalities thus provided can then also be assigned by users with less programming experience to the objects in the visualization.

For vehicle control and to address the sensors and other hardware components on the vehicles, the framework DUNE is used. This was developed of the University of Porto by the Institute LSTS in Portugal. DUNE is used in various projects, so it is constantly evolving. It is a software that runs on the vehicles and it is able to interact with the vehicle's sensors and actuators as well as communicating the vehicle data over a network. Furthermore, a variety of algorithms and functions are provided, which can be used for navigation, vehicle control and monitoring as well as for maneuver planning and their execution. The framework is CPU architecture and operating system independent. It has been programmed in C ++ and is open source available. The possibility of their own development and adaptation to their own vehicle is thus given.
The basic concept in DUNE is the implementation of so-called tasks. Each time the sensor value of a sensor of the vehicle has to be picked up and distributed in the network, a task is started which packages the respective information into a special message (IMC) and distributes it on the network. Similarly, a DUNE instance can receive IMC messages that were produced and sent by a task from another DUNE instance. Since communication within a DUNE instance is also done by sending and receiving tasks, each DUNE instance is also able to receive and process self-produced messages. Another major aspect of dune is the using of predefined profiles. When starting a DUNE instance, a configuration file of the vehicle and a configuration file for the actions to be executed are transferred as parameters. The vehicle configuration contains all vehicle-specific basic data such as name, size, extent, weight, etc. as well as any information about the available sensors. Each sensor integrated there has in turn its own configuration file with associated sensor-specific information. The configuration file defines, which so-called tasks the started DUNE instance should execute, or which task it has to respond to.
The implementation and working with DUNE is similar to the Robot Operating System (ROS) middleware. One difference is, that ROS is primary developed for land-based roboters while DUNE is primary developed for surface vessels und subsurface vehicles. So in DUNE are many algorithms, sensors and actors already integrated, wich can be used in this project. Furthermore four of five philosophical priciples (peer-to-peer, tool-based, thin, free and open-source) of ROS (Quigley et al. 2009) are comparable to DUNE. Instead of multi-lingual programming, DUNE only supports C++ and Java.

The used communication protocol in DUNE is the Inter-Module Communication (IMC) protocol, which is also developed at the University of Porto. IMC is an XML-based message format, which is divided into the three areas. Each message consists of a header, message and footer part. The header contains the basic message information. These include the timestamp, message ID, source address and the destination address. The message footer contains a checksum for recheck the messages. The actual information is stored in the message part. For the message part the IMC-protocol provides multiple predefined messages. These predefined messages are divided into different subareas, such as navigation, sensor data or mission planning. Each message contains sensor or actuator-specific values such as GPS coordinates, sensor values or status data. The IMC format ensures the common communication interface between the individual components. The IMC format is completely defined in a version based IMC.xml file, and can be extended according to your own requirements. This file is interpreted by DUNE to generate the source code and must therefore be included by all other applications that use IMC messages for communication.

DEVELOPMENT AND IMPLEMENTATION

In the development of the visualization and simulation software, one focus was placed on the real-time capability and extensibility of the application. Furthermore, the implementation of the interfaces for communication with external applications like DUNE played a major role. This ensures that the visualization can also be operated with other applications. In the following, the individual components of the visualization and the interfaces will be explained in detail. In addition, the sending and receiving of messages with the DUNE framework and IMC is explained.

Sending and Receiving Messages in DUNE

The DUNE Framework has a sophisticated concept for message transmission in the network. The transmission protocol is the IMC format. For this purpose, within the DUNE instances so-called Producer task are created, which produce messages and distribute them in the network. In addition, consumer tasks can be created, which listen to all or to specific messages. The message type within an IMC message is always specified for a data set and contains the specific sensor values in addition to the message ID and the name. For example, with the message ‘Acceleration’, the acceleration values are transmitted in the three axis directions. The axis and the associated value, the data type and the unit of measure

are transferred as parameters. The assignment of the message to the corresponding vehicle is defined in the header of the message. Based on these values, the message can be received and evaluated by another listening instance.

Real Time Visualization

For the implementation of the simulation and visualization the Unreal-Game Engine was used. Particularly with the real-time capability of the engine, very good results could be achieved in previous projects. Several environments have been implemented as test scenarios for this project. On the one hand, these are complex areas with georeferenced digital terrain models (DTM) based on real data. On the other hand, smaller experimental pools were modeled, which are also available to the institute for experiments with the real ROV. The georeferenced terrain models cover an area of up to one square kilometer and are based on multibeam echo sounder and electronic nautical chart (EMC) data. The modeled research basins correspond to the size of the real basins with a size of 3m * 2m * 1m and of 10m * 20m * 5m. The modeling of the experimental environment will not be discussed further here; this is described in more detail in (Theuerkauff et al. 2017).

The digital models of the AUV and the ROV were available as CAD models. By preprocessing, they could be converted into a suitable data format (FBX), which can be imported into the Unreal engine. Smaller components such as lamps and sensors can be modeled directly on the vehicle model in the editor of the game engine via the use of simple geometries. Changes to the vehicle components thus require no complete remodeling and import of the vehicle. Furthermore, the components thus added can be changed with little effort, regardless of the main model, in the individual parameters such as orientation at runtime.

A single vehicle always consists of a main vehicle class and a model object, which in turn can contain one or more geometries or complex models. So it is possible to simple substitute or modify the vehicle model. The main vehicle class is an Unreal specific class (UClass), which ensures that functions are programmable that can be reused in the Engine's Blueprint Editor. The base class contains the basic information of a vehicle. Further, vehicle specific data and functionalities are stored as an additional special vehicle data object in the class. Vehicles can be accessed in real time during ongoing visualization. Thus, parameters such as the position and orientation of the virtual vehicle are customizable according to the data received. For this purpose, the position and orientation data are transferred to the respective vehicle class with the virtual vehicle. The corresponding vehicle class can be determined from the source address passed in the IMC message. The actual positioning and orientation of the vehicles in the simulation takes place with the aid of visual scripting (Blueprint). This allows easy customization of the positioning and orientation routines without working directly in C ++ source code. The received position and orientation data are also interpolated to ensure a fluid visualization of the vehicles.

Connection to the Control Software

The connection of the visualization to other applications takes place via a UDP connection. The interface depends on a DUNE instance which is running in front of the simulation software, so that the simulation only communicates with the DUNE instance, which in turn distributes the data into the network or forwards received data to the simulation. Through this network interface, the visualization software is able to receive messages from the control network as well as to distribute new messages into the control network (Figure 3).

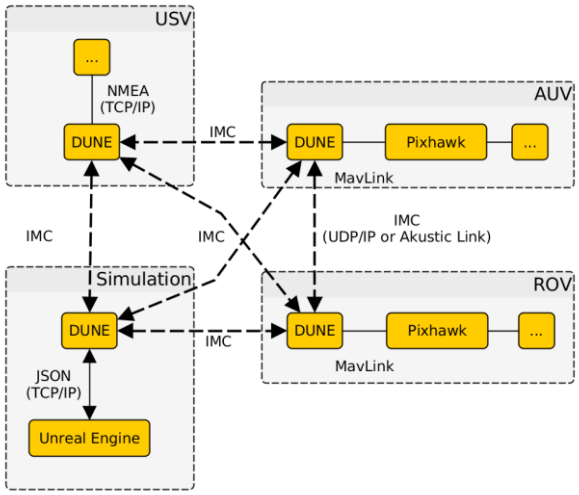


Figure 3: schematic representation of the implemented connection of the simulation environment

So that the DUNE instance serving as simulation interface is not displayed as an additional vehicle in the network, the profile of the DUNE instance has to be set into a special simulation mode. The profile can set at startup via the transmitted configuration file as previously described. With this implemented *UnrealSimulation* profile the interface DUNE instance will receive all communication data and send them to the simulation environment. Furthermore, with the defined profile the DUNE instance can receive messages from the simulation environment to distribute them in the vehicle network.

To do that, the DUNE instance converts all received data into a JSON object, and then forwards it to the visualization environment. On the other side the DUNE instance listen on an UDP interface for messages from the simulation environment. The UDP socket created in the simulation environment constantly listens to the messages of the DUNE instance on a specified port. The in the simulation environment received messages are converted back to the IMC format via a wrapper class. For this purpose, the IMC configuration file (IMC.xml) is imported when the visualization is started and saved in a suitable data structure. The message structure of the incoming JSON messages is determined by the message

name from the IMC data structure. Thus, all messages can automatically be converted from the JSON format back to the IMC structure. Newly defined message structures in the IMC file are automatically converted within the visualization component without further adjustments.

Real Time Simulation

At the virtual underwater vehicles and surface vessels in the simulation, various sensors are connected. At present, an ROV has a depth sensor, an acoustic modem, several laser distance sensors and a camera. The AUV is equipped with an acoustic modem and a GPS system. The sensor data are simulated according to the environment. For this it is possible to set various parameters of the environment within the visualization. These include ambient light, water temperature and turbidity of the water by particles. The influence on the measurement data of the sensors currently consists only in changing the range of the laser distance detection. The sensor data are transmitted in real time to the control software for further processing using the methodology described in chapter *Connection to the Control Software*. Thereby they directly influencing the further behavior of the vehicle.

TEST SCENARIO AND RESULTS

The test environment initially consists of a small test basin with the extension of 2m * 3m * 1m [L, B, T]. The used test vehicle is a BlueROV2 from BlueRobotics Inc. Since the acoustic position determination was not fully implemented when the paper was finished, the position data has been emulated in GPS format. The orientation data are recorded by an inertial measurement unit (IMU) installed in the ROV. The control of the ROV was done manually via an XBox controller. All data are distributed in the network by the DUNE instance of the BlueRov2 and can thus be recorded by the visualization environment, which is also registered via another DUNE instance in the network. Furthermore we integrated up to ten more virtual ROVs and USVs to the simulation. Each vehicle was controlled by an own DUNE instance. By registration the single instances in the same network it is possible to send steering data to each vehicle as same as to send sensordata and messages from a vehicle to each other vehicle in the network. Each virtual vehicle was equipped with an simulated IMU and up to three simulated laser distance measurement systems. The framerate in the running visualization was stable with about 50fps by using an Intel Core i7-6820HQ CPU with 2.7GHz, 40GB RAM and a NVIDIA Quadro M2000M graphiccard. Another test scenario is planned and will be run in the near future. The test environment will be a basin with the size of 20m * 10m * 5m [L, B, H]. The integration of the acoustic modems into the DUNE framework will be done then, so the position data from the underwater vehicles does not need to be emulated anymore for the real vehicles.

Figure 4: Visualization of a real ROV; left: Simulation environment; right: Real environment

CONCLUSION AND OUTLOCK

The visualization environment presented here provides the basis implementation for a real-time simulation environment of underwater vehicles and surface vessels. It is able to visualize virtual and real ROVs and AUVs in a 3D environment. The application communicates bidirectionally with the control units of each vehicles in the network. The simulated vehicles can thus be moved via the control software like a real ROV. The design concept is to handle virtual vehicles in the simulation environment from the control software like real vehicles. So it is possible to use the control software equally for the simulation as well as for the real vehicles without further adjustments (Figure 5). Furthermore it is possible to use these software for other vehicletypes wich are using the DUNE framework for controlling.

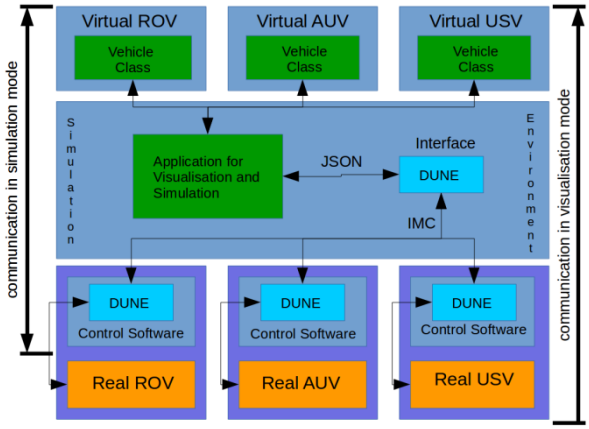


Figure 5: Principle of visualisation and simulation mode. The real vehicles (orange boxes) are only required in visualization mode

At present, the virtual vehicles receive only the coordinates and orientation data of the real vehicle. In the first implementation, the visualization thus always corresponds to the state of the real vehicles. In the further course, the vehicles in the simulation environment are to be expanded with the sensors of the real ROV. The sensors can then be used in the simulation environment e.g. to determine the distance to underwater obstacles and send these data to the control software via the communication interface. This data can than be incorporate into the control logic to generate new control commands for the vehicle. First simple virtual sensors, such as laser based distance sensors, implemented in a previous work, already integrated into the simulation

software presented in this paper. Until now these sensors are not based on the real vehicle sensors and could only deliver fictive data.

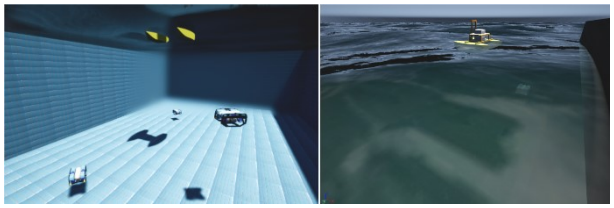


Figure 6: Different simulation scenes; left: Indoor basin; right: open water

An important step in the development of the simulation environment is an investigation into the extent to which water specific data, such as flow models, have an influence into the simulation environment for testing the control algorithms. It is conceivable that the control algorithms can initially also be evaluated by simple, randomly generated drift vectors. Furthermore, the simulation of underwater sensors within the simulation environment will be improved.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the received grants from the internal research and development funds from Jade University of Applied Sciences.
Furthermore, the work became possible by grants of Lower Saxony's Ministry for Science and Culture (Nds. Ministerium für Wissenschaft und Kultur) in the funding scheme VW-Vorab Science for sustainable development.

REFERENCES

Dias P. S., Paulo Sousa and Gomes, Rui M. F. and Pinto, José and Gonçalves, Gil Manuel and de Sousa, João Borges and Pereira, Fernando Lobo. 2006. "Mission Planning and Specification in the Neptus Framework". *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.,* 3220-3225

Dias P. S. and S. L. Fraga and R. M. F. Gomes and G. M. Goncalves and F. L. Pereira and J. Pinto and J. B. Sousa. 2005. "Neptus - a framework to support multiple vehicle operation". *Europe Oceans 2005* Vol. 2 963-968.

M. Prats and J. Pérez and J. J. Fernández and P. J. Sanz. 2012. "An open source tool for simulation and supervision of underwater intervention missions." *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2577-2582

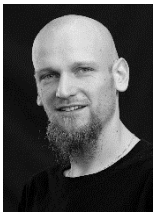
Quigley M. , B Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, A. Ng. 2009. "ROS: an open-source Robot Operating System.". *ICRA workshop on open source software*, Vol. 3 (May)

Theuerkauff T.; T. Werner; F. Wallhoff; T. Brinkhoff., 2017 „3D-Visualisierung von Über- und Unterwasserfahrzeugen zur Evaluation von Steuerungsalgorithmen mithilfe eine Game-Engine.". *Go-3D 2017 Mit 3D Richtung Maritim 4.0*, Fraunhoferverlag , 135-164

Tchakoua, Pierre, Wamkeue, René, Ouhrouche, Mohand, Slaoui-Hasnaoui, Fouad, Tameghe, Tommy Andy, Ekemb, Gabriel. 2014. "Wind Turbine Condition Monitoring: State-of-the-Art Review, New Trends, and Future Challenges." *Energies*, Vol 7, 2595-2630

Xiang X., B. Jouvencel, O. Parodi. 2010. "Coordinated Formation Control of Multiple Autonomous Underwater Vehicles for Pipeline Inspection".*International Journal of Advanced Robotic Systems,* SAGE Publications Vol. 7, No. 1, 75 – 84

AUTHOR BIOGRAPHIES



Tobias Theuerkauff was born in Leer, in Germany and lives in Oldenburg. He graduated in 2011 at the Jade University of Applied Sciences in Oldenburg as Master of Science in Geodesy and Geoinformatics. Since 2011 he has been working as a research assistant at the Jade University. Until 2011, he worked at the Institute of Photogrammetry and Geoinformatics (IAPG), where he specialized in the fields of virtual and augmented reality. Since 2017 he works at the Institute for Technical Assistance Systems (ITAS) at Jade University of Applied Sciences. There he focuses on the field of artificial intelligence in combination with 3d visualization. His e-mail address is: `tobias.theuerkauff@jade-hs.de` and his Web-page can be found at `https://www.jade-hs.de/team/tobias-theuerkauff`

Yves Wagner was born in Salzwedel, Germany and moved to Oldenburg. In 2014 he graduated as Bachelor of Science in the study course assistive technologies at Jade University for Applied Science in Oldenburg. After that he started as research assistant at the Fraunhofer Institute for Digital Media Technologies (IDMT). Since 2014 he is research assistant at the Institute for Technical Assistance Systems (ITAS) at the Jade University of Applied Science in Oldenburg. His e-mail address is: `yves.wagner@jade-hs.de`

Frank Wallhoff was born in Rheinhausen, now Duisburg in Germany and studied Electrical Engineering in Duisburg where he received his diploma degree in 2000. Hereafter he started as a research assistant at Duisburg University and changed to the Technical University of Munich in 2002 where he was promoted as Dr.-Ing. in 2006 in the area of face detection and recognition. In 2010 he became professor for Assistive Technologies at Jade University of Applied Sciences at the campus in Oldenburg. Since then he is the leader of the group Interactive Systems with focus on artificial intelligence at the Institute for Technical Assistance Systems (ITAS). He is also the director of a Fraunhofer Center of Transfer. His e-mail address is: `frank.wallhoff@jade-hs.de` and his Web-page ca be found at `https://www.jade-hs.de/team/frank-wallhoff`

MODEL CHECKING KNOWLEDGE AND COMMITMENTS IN MULTI-AGENT SYSTEMS USING ACTORS AND UPPAAL

Christian Nigro, Libero Nigro, Paolo F. Sciammarella

Software Engineering Laboratory
University of Calabria, DIMES - 87036 Rende (CS) – Italy
Email: christian.nigro21@gmail.com, l.nigro@unical.it, p.sciammarella@dimes.unical.it

KEYWORDS
Multi-agent systems, knowledge and commitments, NetBill protocol, actors, model checking, UPPAAL.

ABSTRACT

This paper proposes a method for modelling and analysis of knowledge and commitments in multi-agent systems. The approach is based on an actors model and its reduction onto UPPAAL. A key factor of the approach is the possibility of exploiting the same UPPAAL model for exhaustive verification or, when state explosion problems forbid model checking, for quantitative evaluation of system properties through statistical model checking. The article describes the method, shows its application to modelling the NetBill protocol, proposes a translation into the terms of the timed automata language of UPPAAL and demonstrates the analysis of the NetBill protocol together with some experimental results.

INTRODUCTION

Nowadays more and more distributed software systems are developed using the multi-agent systems (MAS) paradigm (Wooldridge, 2009). MAS enable the construction of highly modular and scalable systems, where the resultant behavior at the population level (macroscopic or society level) emerges (in a non-intuitive way) from the individual behavior of agents and their interactions (microscopic level). Fundamental abilities of agents contributing to the expressive power of MAS include autonomy, goal-directed mission, sociality (i.e., communications), pro-activity and reactivity.

As the applications of MAS continue to grow the need arises to support modelling and verification of an agent-based system in order to ensure correctness of a development.

This paper focuses on modelling and analysis of knowledge and commitments in MAS (Al-Saqqar et al., 2015)(Singh, 2000) which naturally occur, e.g., in business web-based protocols and applications. Commitments express the willingness for agents to do something (e.g., paying for an accepted offer of a good over the Internet). Commitments have a status which can switch from creation, to fulfillment, to discharge, to cancellation. Knowledge refers to the epistemic relation of agent awareness about the status of a commitment.

Reasoning on knowledge and commitments in MAS is often pursued through a particular temporal logic language like CTLKC$^+$ (Al-Saqqar et al., 2015)(Sultan, 2015) whose models can be verified through reduction to an existing model checking tool. In (Al-Saqqar et al., 2015) the use of the NuSMV model checker (Lomuscio et al., 2007) is demonstrated with the overall approach which favours model scalability.

The work described in this paper was triggered by the CTLKC$^+$ work, and the desire to offer timed automata modelling and particularly to simplifying the query language in order for it to become more easy to use and understand by final modellers. In the proposed approach agents are modelled by actors (Cicirelli & Nigro, 2016)(Nigro & Sciammarella, 2017a) and the analysis activities are carried out in the popular UPPAAL toolbox (Behrmann et al., 2004)(David et al., 2015). The presented reduction of actors to UPPAAL is novel and extends previous authors' work (Nigro & Sciammarella, 2017a-b)(Nigro et al., 2018). The new contribution refers to the fact that a reduced actor model in UPPAAL now enables both exhaustive model checking (through the construction of the model state-graph) and statistical model checking (based on simulations) which can be necessary when exhaustive verification is forbidden by state explosion problems.

The paper is structured as follows. First some background is provided concerning basic concepts of knowledge and commitments in MAS and of the adopted actors. Then the NetBill protocol (Sirbu, 1997)(Al-Saqqar et al., 2015) is modelled using actors. After that, the developed reduction of actors onto UPPAAL is detailed through the transformation of the NetBill model. Then some experimental work is reported using the achieved UPPAAL model. Finally, conclusions are presented with an indication of further work.

BACKGROUND CONCEPTS

Commitment issues

Knowledge and commitments (to do something) in MAS are issues that in the literature are often handled independently. This is unfortunate (Al-Saqqar et al., 2015) because, as occurs in business settings, agents need to reason on their social commitments *and* their knowledge, especially when they are engaged in conversations.

The following example, quoted from (Al-Saqqar et al., 2015), testifies the importance of expressing the interactions between knowledge and commitments:

"*Suppose that we asked a member from our team to buy a book for us last month. He made the online order and committed to pay. The credit card debit succeeded,*

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

meaning that the agent (our team member) knows that he fulfilled his commitment to pay. The publisher company committed to send the requested book to our address. Unfortunately, the book has never arrived. The publisher claimed they had send it out, but the shipping company they dealt with could not find it in their records. As a result, we asked them to send it again. However, knowing that the book is delivered (i.e., fulfilling the commitment of delivering the book) will help avoiding such situations."

Commitments are naturally tied to communications (Singh, 2000)(Sultan, 2015). Two main approaches can be distinguished. In the *mentalistic* approach communications are in terms of privately defined information of agents, that is agent mental states as beliefs, desires and intentions. In such approaches the assumption is made that "*each agent can read each other mind*". However (Sultan, 2015), mental based models do not favour verification. A different approach considers communications as based on publicly available information. In a social commitment, a *debtor* agent (sender) agrees with a *creditor* (receiver) agent about a debtor engagement to bringing a certain property. Social commitments provide an objective semantics to messages and naturally accommodate for agent heterogeneity.

A crucial issue when modelling social commitments is agent *uncertainty*. Uncertainty, that is non-deterministic behavior, concretely affects agent evolution and makes it challenging to analyze agent behavior. Non determinism specializes into probabilistic behavior when agent commitments are studied quantitatively. In this case, non deterministic courses of actions in the agents can be labelled by chosen probability values.

In (Al-Saqqar et al., 2015)(Sultan, 2015), social commitments between a debtor and a creditor are associated with the existence of *shared variables* of the debtor and creditor agents. Shared variables, in turn, mean suitable communication channels among the agents exist through which the value of a shared variable is updated following a social commitment operation.

Formally, in CTLCK$^+$, a social commitment is denoted by $C_{i\to j}\ \varphi$ whose meaning is: agent i (the debtor) commits toward agent j (the creditor) about φ which is the content of the commitment. Of course, during its lifecyle, a commitment has to be known to the involved agents. The relation $K_i\ \varphi$ formally represents the fact that agent i knows φ in some state. A commitment status should be known to both its involved agents as they proceed in a conversation.

In (Al-Saqqar et al., 2015) the CTLKC$^+$ temporal logic is mapped on an action-based logic (ARCTL) and then on to the modelling language of the NuSMV symbolic model checker (Lomuscio et al., 2007), which is based on Binary Decision Diagrams (BDD) and Boolean functions (Bf). The overall approach is proved to be efficient in space and time and favours model scalability.

The analysis of a probabilistic version of CTLKC$^+$ is described in (Sultan, 2015) along with a reduction to PRISM modelling language and toolbox (Hinton et al., 2006).

Actor issues

The contribution of this paper is to propose a different approach to modelling and verification of knowledge and social commitments in MAS, which rests on a lightweight actor model which can effectively be reduced to UPPAAL timed automata (Alur & Dill, 1994) for model checking. When the model size forbids exhaustive verification by state explosion problems, the UPPAAL model can be analyzed quantitatively, by inferring probability measures about required properties by using the UPPAAL Statistical Model Checker (SMC).

Adopted actors (see (Cicirelli & Nigro, 2016)(Nigro & Sciammarella, 2017a-b)(Nigro et al., 2018) for more information) encapsulate a *data status* and publish a particular *message interface*. Message-passing is asynchronous. Only through a received message, an actor can update its local variables. Actors are thread-less agents: they are at rest until a message arrives. Messages can be sent to known actors (*acquaintances*). For pro-activity, behavior an agent can also send a message to itself.

The programming/modelling style of actors is easily supported by Java. Each admitted message is processed in a *message-server* method (in Java the method is provided of the @Msgsrv annotation) having the same name, and which can have parameters. The basic send operation follows the syntax:

```
target_actor.send( message_name[, args] );
```

and exploits computational reflection for posting a corresponding (untimed) message object on an underlying message pending set (see below). A message can be timed by specifying an *after* clause (relative time) in the send operation:

```
target_actor.send( after, message_name[, args ] );
```

It is meant that the message has to be consigned to its target actor as soon as after time units are elapsed since the send time. When the after clause is missing, it evaluates to 0. Current absolute model time is returned by the now() primitive.

The message servers of an actor class can be programmed according to a finite state machine.

A collection of actors (*theatre*) runs on a computing node (JVM instance) and is (transparently) regulated by a *control machine* which governs the set of sent (timed and untimed) messages. The control machine is responsible of managing a time notion and follows a *control loop*. At each iteration, a message is chosen from the pending set and dispatched to its destination actor. A library of control machines was developed (see (Cicirelli & Nigro, 2016)).

A theatre system consists in general of a collection of theatres which coordinate each other, e.g., to guarantee a common time notion.

A fundamental semantic aspect is the *macro-step semantics of messages*. In a same theatre, actors are

activated by messages one at a time. It is not possible to dispatch the next message until the last dispatched one was completely processed (the corresponding message-server method is terminated). This way, a control machine induces a cooperative concurrency schema among local actors which depends on message interleaving. Message servers of distinct theatres can be executed in true parallelism.

As a final remark, it is worth noting that a concrete implementation of the actor model can guarantee that messages sent by an actor A to an actor B, at the same time, will be received in the sending order. Such an order does not exist among messages sent by different senders toward a same receiver. When messages have different timestamps, their dispatch follows the timestamp order. However, during modelling and analysis, for generality concerns, message delivery at a same time will be assumed not deterministic.

ACTOR-BASED NETBILL PROTOCOL

The NetBill protocol (Sirbu, 1997)(Al-Saqqar,2015) is played by a customer agent (*cus*) and a merchant agent (*mer*). The message conversation relates to buying and selling an encrypted software good over the Internet.

In the following, the NetBill protocol is modelled using actors by adapting the customer and merchant models reported in (Al-Saqqar et al., 2015). Two commitments are involved in the protocol: $C_{cus \to mer}\ Pay$ and $C_{mer \to cus}\ Deliver$. The Pay commitment is raised by the customer toward the merchant to testify the customer intention, after having accepted the offer from the merchant, to proceed with the payment of the good. The $Deliver$ commitment is played by the merchant toward the customer, to express its willingness to deliver to it the required (and paid!) good. However, uncertainty in the agent behaviors can change the course of expected actions following the initial intention to commit, as shown in the models of customer and merchant in the Figg. 1 and 2.

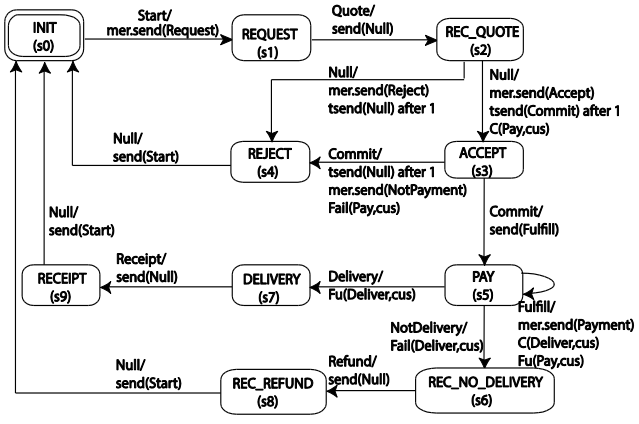


Fig. 1 - The Customer model

The customer starts by sending a request for a quote to the merchant. After receiving the quote the customer can, non deterministically, proceed by accepting the quote (state s3) or rejecting it (state s4) after which the model is restarted from the initial state (s0).

In the case the quote is accepted, the customer sends a message Accept to the merchant and a message Commit to itself, then it moves to state s3. Accepting the quote logically issues the Pay commitment.

In the work described in (Al-Saqqar et al., 2015) a social commitment implies a communication between the customer and the merchant in order to reflect the status of the commitment in the shared variables of the two agents. Such a communication can be naturally captured by an explicit message exchange in our actor modelling. However, the Accept message itself serves the purpose of transmitting to the merchant the customer intention to (possibly) proceed with the payment, then it plays also the role of communicating that the Pay commitment was issued. In addition, the customer in the state s2 sends to itself (pro-activity) the Commit message as a *reminder* to proceed with payment. On receiving the Commit message, the customer can, again non deterministically, choose to fulfil the Pay commitment or to send a NotPayment message to the merchant thus putting the Pay commitment into the failure state.

An important aspect of the actor model is concerned with message delivery which, as anticipated in the previous section, is supposed to be non deterministic. For example, the sequence of untimed messages Accept then NotPayment, without other provisions, can be actually received by the merchant in the order first NotPayment then Accept, thus incurring into a malfunction. To ensure the right order, the Commit message is sent as a timed message with a delay of 1 time unit. In the case the customer decides to proceed with the Pay fulfillment (state s5), it sends to itself the Fulfill message whose reception causes the customer to effectively send a Payment message to the merchant.

The status of a commitment c is assumed to be changed by the functions C(c,req), Fu(…) and Fail(…) which respectively set it, for the requestor agent req, to WillingToDo, Fulfilled and Invalid. Knowledge about commitments will be checked during model analysis.

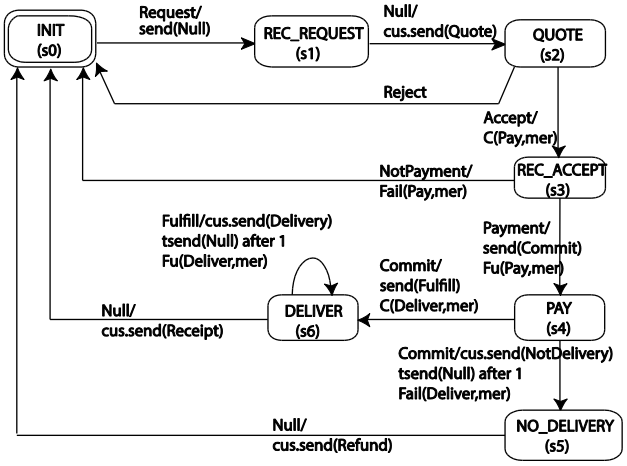


Fig. 2 - The Merchant model

By reciprocity, when the customer receives a Delivery message from the merchant, it calls Fu(Deliver,cus) to update its information about the *Deliver* commitment. Similarly, it calls Fail(Deliver,cus) when a NotDelivery is received, in which case a Refund is expected from the merchant. The Null message in the Fig. 1 is used pro-actively to ensure a proper state sequencing.

The merchant model in Fig. 2 follows the same design guidelines seen in Fig. 1. For example, on receiving an Accept message, it calls C(Pay,mer) to optimistically reflect the customer intention to proceed with payment. Similarly, the arrival of a Payment message in state s3, requires the merchant to update its status of the *Pay* commitment to Fulfilled. Other details should be self-explanatory.

A REDUCTION OF ACTORS TO UPPAAL

The rationale of proposed reduction of actors to UPPAAL is to express both actors and messages as template timed automata. The challenging is to correctly reproduce the asynchronous character of messages. Although UPPAAL SMC supports dynamic automata (Nigro & Sciammarella, 2017a-b), in this paper a certain number of statically created message instances are instead used and dynamically activated and then, after being dispatched, reset for them to be re-used. All of this opens to the possibility of making exhaustive model checking, provided the right number of message instances is prepared. The same model can also be used by simulations using the Statistical Model Checker of UPPAAL (David et al., 2015).

First, actors have to be uniquely identified:

```
const int CUS=2; //number of customers
const int MER=2; //number of merchants
const int N=CUS+MER; //number of total agents
typedef int[0,CUS-1] cus_id; //customer ids
typedef int[CUS,N-1] mer_id; //merchant ids
typedef int[0,N-1] aid; //type of agent ids
```

For simplicity, customers and merchants are supposed to operate in pairs. A customer id, e.g.,$cus \in [0, CUS-1]$, is paired with the merchant id: $mer = cus + CUS$.

Similarly, exchanged messages are classified:

```
const int Null=0, Start=1, Quote=2, Delivery=3, Receipt=4,
NotDelivery=5, Refund=6, Request=7, Reject=8, Accept=9,
Payment=10, NotPayment=11, Commit=12, Fulfill=13;
const int MSG=14; //number of different messages
typedef int[0,MSG-1] msg_id; //type of message ids
```

Then the array of broadcast channels msgsrv[][] is introduced which is used to dispatch a message to a given actor:

```
broadcast chan msgsrv[aid][msg_id];
```

Message automata ids corresponding to untimed and timed messages are dimensioned according to the needs of the models of customer and merchant (see Fig. 1 and Fig. 2):

```
const int UM=CUS; //number of untimed messages
typedef int[0,UM-1] umid;
const int TM=CUS; //number of timed messages
typedef int[0,TM-1] tmid;
bool freeUM[UM], freeTM[TM]; //initialized to all true
```

The following channel arrays serve to activate message instances:

```
broadcast chan send[umid][aid][msg_id],
              tsend[tmid][aid][msg_id];
```

Functions nM() and nTM() respectively return the id of the next free untimed or timed message instance. Obviously, when freeUM[.] or freeTM[.] are insufficiently dimensioned, an error will occur during model checking.

Fig. 3 and Fig. 4 show the automata for an untimed and a timed message. A sent message is first scheduled then dispatched.

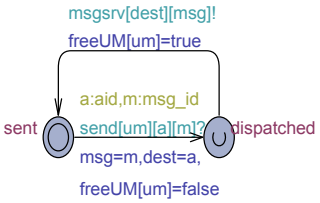


Fig. 3 - Message automaton

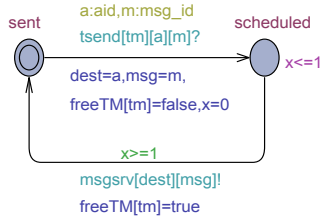


Fig. 4 - TimedMessage autom.

Message/TimedMessage automata have respectively the only parameter: const umid um, or const tmid tm.

The following global declarations introduce commitments and knowledge support:

```
const int PAY=0, DELIVER=1; //admitted commitments
typedef int[PAY,DELIVER] cid; //commitment ids
const int Invalid=0, WillingToDo=1, Fulfilled=2;
typedef int[Invalid,Fulfilled] status;
status k[cid][aid]; //agent knowledge of comm. states
```

Commitment states in agent knowledge are updated by the functions C(cid,aid), Fu(…) and Fail(…) (not shown for brevity) which receive the commitment id and the requestor agent. Function K(aid,cid) returns the status of a commitment as known to a given agent.

The Customer automaton shown in Fig. 5 has the only parameter: const cus_id cus. The associated mer id is implicitly estabilished as a local constant at initial time.

In the Receive location, the next arrived message is received through one msgsrv channel. In the Select location, the identity of the received message is checked and the corresponding response path selected, after which the Receive location is re-entered. It is worth noting that a processing path for responding to a message is achieved, in general, by committed locations, which have greater priority than urgent locations (see e.g. Fig. 3). All of this is a key to ensure the macro-step semantics of message processing. For the rest, the UPPAAL automaton in Fig. 5 closely corresponds to its design in Fig. 1. Fig. 6 shows the Merchant automaton.

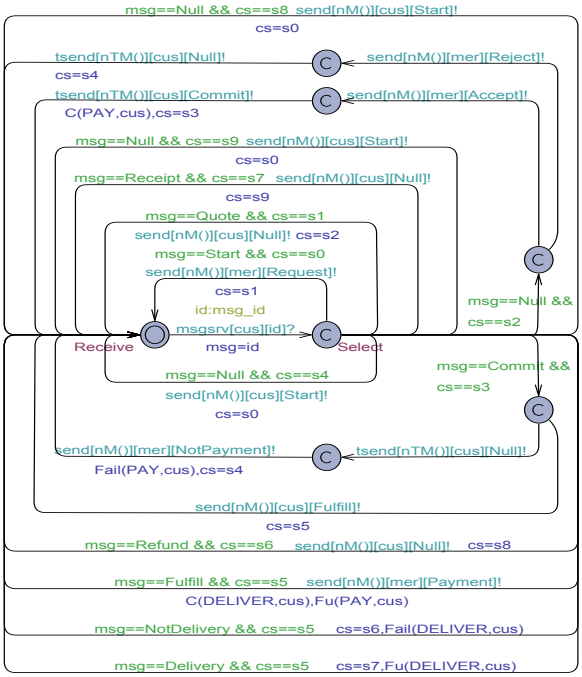


Fig. 5 - The Customer automaton

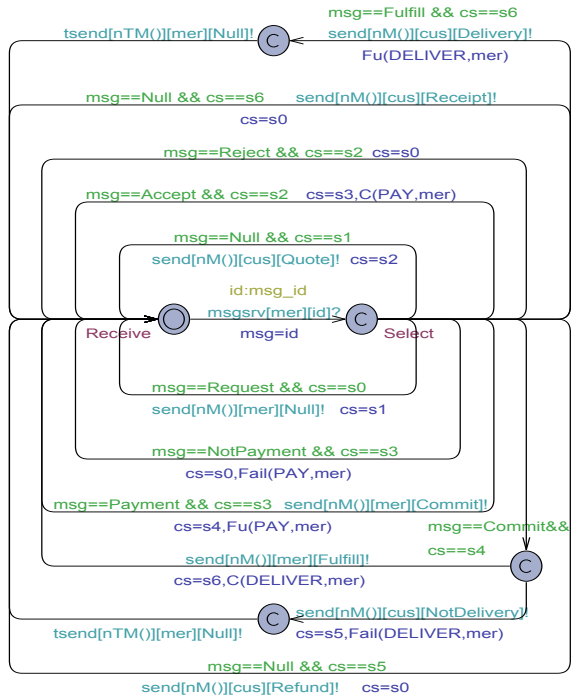


Fig. 6 - The Merchant automaton

The reduced NetBill protocol model is put into operation by a Main automaton (see Fig. 7) which initializes global data structures and sends a Start message to all the customers.

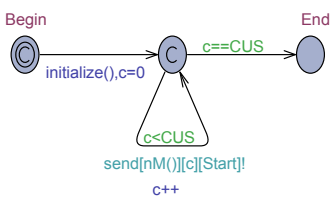


Fig. 7 - The bootstrapper Main automaton

Finally, the following is the system configuration line:

```
system Main,Customer,Merchant,Message,TimedMessage;
```

MODEL CHECKING THE NETBILL PROTOCOL

Both *safety* (a bad state is never reached) and *liveness* (a good state is eventually reached) properties were checked on the reduced actor-based NetBill protocol in UPPAAL.

Exhaustive model checking is centred on the TCTL-based query language offered by UPPAAL (Behrmann et al., 2004) which does not allow formulas to be nested, as instead is permited in CTL-like logics (see, e.g., (Al-Saqqar et al., 2015) and NuSMV queries). All of this contribute to simplicity and readability, without compromising correctness checking.

Qualitative analysis by model checking

Up to 4 agents (two pairs <cus,mer>) were easily model checked (also on a laptot with 4GB RAM). A greater number of agents causes state explosions. However, correct protocol behavior can be verified also on a single pair <cus,mer>.

Safety properties

The query

```
A[] ! deadlock
```

was found satisfied, thus guaranteeing there are no deadlock states in the model.

Another safety property is concerned with the situation: the customer fulfils its payment commitment, i.e., it sends the payment to merchant, but the merchant does not know that. In the following, the cus id 0 is considered, which works with mer id CUS. One way for checking the property is through the *leads-to* operator --> (cs is a local variable holding the current state of the customer):

```
Customer(0).cs==s5 -->
    K(0,PAY)==Fulfilled && K(CUS,PAY)==Fulfilled
```

which means: when the customer is in state s5, it inevitably will be followed by a successor state in which both the customer and the merchant know that the PAY commitment is fulfilled. The query was found satisfied. A similar query (satisfied) was prepared for checking the fulfillment of the DELIVER commitment (from mer to cus):

```
Merchant(CUS).cs==s6 --> K(CUS,DELIVER)==Fulfilled
    && K(0,DELIVER)==Fulfilled
```

Liveness properties

One such a property is the following: in the case the merchant commits to deliver the goods to the customer, it will eventually deliver them. The property was checked with the query:

```
K(CUS,DELIVER)==WillingToDo -->
    K(CUS,DELIVER)==Fulfilled
```

or the query:

Merchant(CUS).cs==s6 --> K(CUS,DELIVER)==Fulfilled

which are both satisfied. The following queries check that for both the customer and the merchant, it inevitably follows that the initial state s0 will be re-entered (now is a decoration clock measuring the global model time):

A<> now>0 && Customer(0).cs==s0
A<> now>0 && Merchant(CUS).cs==s0

Both queries are satisfied. As another example, would the customer or the merchant be in the state s1, it will definitely move to the initial state (queries are satisfied):

Customer(0).cs==s1 --> Customer(0).cs==s0
Merchant(CUS).cs==s1 --> Merchant(CUS).cs==s0

Quantitative analysis by statistical model checking

An important benefit of the developed reduction of actors to UPPAAL, is concerned with the possibility of exploiting the statistical model checker (SMC) for quantitative property checking. This, in turn, will help in the cases where state explosion problems do not allow for an exhaustive verification.

UPPAAL SMC (David et al., 2015)(Nigro et al., 2018) does not build the model state-graph. Instead it is based on simulations which imply a linear demand of memory. SMC predicts a number of simulation runs and infer from them (using Monte Carlo-like techniques or sequential hypothesis testing) a probability measure for a checked property. It comes equipped with a powerful query language for statistical analysis which is based on the Metric Interval Temporal Logic (MITL).

During the analysis using UPPAAL SMC of the NetBill protocol, non deterministic behavior is turned into a probabilistic one. For example, in state s3 of Fig. 1 where che customer can either choose to fulfil the payment toward the merchant, or send it the NotPayment message, the two alternatives have an equal probability to occur during SMC analysis. Would it be useful for the modelled system, the modeler can also label the two transitions with a probability weight. In the following, uncertainty by non determinism is implicitly replaced by probabilities without introducing specific probability weights. In addition, default statistical options of the toolbox are used (e.g., the uncertainty error of a confidence interval is $\varepsilon = 0.05$).

Functional behavior of the NetBill model was preliminarily checked with such queries like the following:

simulate 1 [<=10] { Merchant(CUS).cs }

which monitors, in 10 time units, the evolution of the local state variable cs of the Merchant model. An observed behavior is reported in Fig. 8 which confirm cs always regularly returns to the s0=0 value.

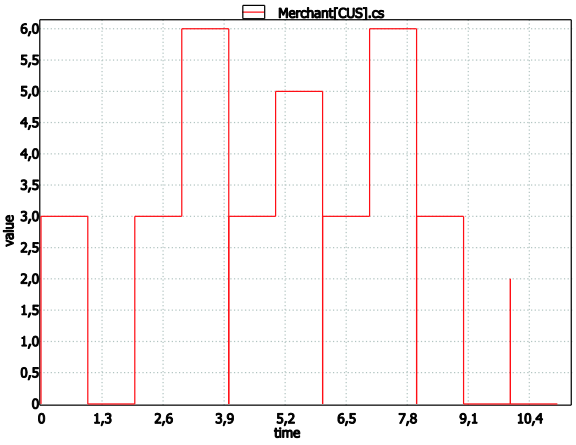


Fig. 8 - Evolution of Merchant(CUS).cs variable vs. time

In a different way, the liveness property, e.g., of Customer(0), that its local state variable cs will eventually assume again the s0 value, the following query estimates the probability of the event: "being the customer in a state different from s0 at an instant in [0,100], in at most 2 time units it will come back to its home state s0":

Pr(<>[0,100] (Customer(0).cs!=s0 &&
(<>[0,2] Customer(0).cs==s0)))

Uppaal SMC, using 738 runs, suggests a confidence interval of [0.95,1] with a confidence degree of 95%, to indicate that the event has a very high probability of occurrence.

The following query asks about the probability that, at any instant in [0,100], would the customer know that the PAY commitment is fulfilled, in 0 time the merchant will know it too.

Pr(<>[0,100] (K(0,PAY)==Fulfilled &&
(<>[0,0] K(CUS,PAY)==Fulfilled)))

Also in this case the probability has a confidence interval of [0.95,1] with confidence 95%.

The query:

Pr[<=1000] (<>Customer(0).cs==s5 &&
K(0,PAY)==Fulfilled)

asks to estimate the probability that when the customer is in the state s5, it knows the PAY commitment is fulfilled. Using 36 runs, Uppaal SMC says the probability has a confidence interval of [0.902606,1] with confidence 95%. The same event was monitored (see Fig. 9) with the query:

simulate 1 [<=100] {Customer(0).cs==s5 &&
K(0,PAY)==Fulfilled}

Obviously, as one can see from Fig. 9, not always the monitored condition is true because there are cases when the customer from s3 goes to state s4 rejecting the offer, instead to switching to s5 where it is guaranteed that the payment will be honored.

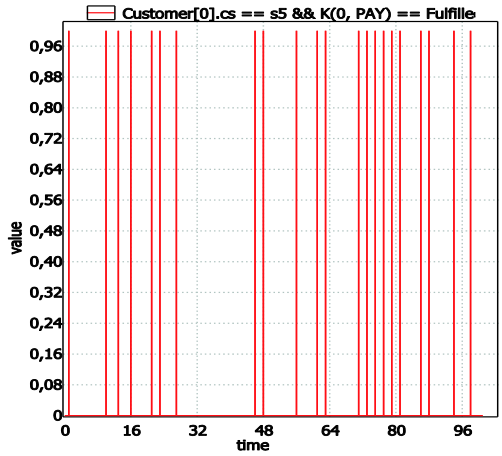


Fig. 9 – "Customer(0) in s5 knows PAY is fulfilled?" vs time

In the light of the reported qualitative/quantitative experimental results, the NetBill protocol was found to be correct.

Experiments were carried out on a Linux machine, Intel Xeon CPU E5-1603@2.80GHz, 32GB, using UPPAAL 4.1.19 64bit.

CONCLUSIONS

This paper proposes an approach to modelling and verification of knowledge and commitments in multi-agent systems (MAS) intended for business web-based protocols and applications.

The method is novel and it is based on lightweight actors (Cicirelli & Nigro, 2016)(Nigro & Sciammarella, 2017a-b)(Nigro et al., 2018) together with a reduction on top of UPPAAL timed automata (Behrmann et al., 2004)(David et al., 2015). This way a MAS model can be studied by exhaustive verification and/or through the statistical model checker which rests on simulations.

With respect to the inspiring work of (Al-Saqqar et al., 2015)(Sultan, 2015), the possibility of switching from model checking to statistical model checking on a same model, possibly enhanced by probabilistic weights at branch points, is a key contribution. In addition, the use of simpler and more readable queries than those CTL-like with nested formulas, should be pointed out.

Another benefit of the proposed approach stems from the adoption of an actor modelling language which has a clear link to implementation. Actually, the used actors are embedded in Java and can effectively be exploited for programming time-dependent distributed systems.

Prosecution of the research aims to:

- improving the proposed reduction of actors on to UPPAAL;
- continuing experimenting with modelling general MAS and analysing their properties;
- specializing the approach to modelling and qualitative/quantitative analysis of distributed, possibly probabilistic, timed actors (preliminary results are described in (Nigro & Sciammarella, 2017a)).

REFERENCES

Alur, R., Dill, D.L. (1994). A theory of timed automata. *Theoretical Computer Science,* **126**:183–235.

Al-Saqqar, F., Bentahar, J., Sultan, K., Wan, W., Asl, E.K. (2015). Model checking temporal knowledge and commitments in multi-agent systems using reduction. *Simulation Modeling Practice and Theory,* **51**, 45-68.

Behrmann, G., A. David, K.G. Larsen (2004). A tutorial on UPPAAL. In: *Formal Methods for the Design of Real-Time Systems*, Lecture Notes in Computer Science, Vol. 3185, Springer-Verlag, pp. 200-236.

Cicirelli, F., L. Nigro (2016). Control centric framework for model continuity in time-dependent multi-agent systems. *Concurrency and Computation: Practice and Experience*, **28**(12):3333-3356, Wiley.

David, A., K.G. Larsen, A. Legay, M. Mikucionis, D.B. Poulsen (2015). UPPAAL SMS Tutorial. *Int. J. on Software Tools for Technology Transfer*, Springer, **17**:1-19, 06.01.2015, DOI 10.1007/s10009-014-0361-y.

Hinton, A., Kwiatkowska, M.Z., Norman, G. and Parker, D. (2006). PRISM: a tool for automatic verification of probabilistic systems. In *Proc. of 12th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'06)*, Springer-Verlag LNCS Vol. 3920, pp.441–444.

Lomuscio, A., Pecheur, C., Raimondi, F. (2007). Automatic verification of knowledge and time with NuSMV. In *Proc. of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1384–1389.

Nigro, L., Sciammarella, P.F. (2017a). Statistical model checking of distributed real-time actor systems. In *Proc. of 21st IEEE/ACM Int. Symposium on Distributed Simulation and Real-Time Applications (DS-RT'17),* October 18-20.

Nigro, L., Sciammarella, P.F. (2017b). Modelling and analysis of distributed asynchronous actor systems using Theatre. *Advances in Intelligent Systems and Computing 661*, DOI 10.1007/978-3-319-67618-0_14, Springer.

Nigro, C., Nigro, L., Sciammarella, P.F. (2018). Modelling and analysis of multi-agent systems using Uppaal SMC. *Int. J. of Simulation and Process Modelling,* **13**(1):73-87.

Singh, M.P. (2000). A social semantics for agent communication languages. In *Issues in Agent Communication*, pages 31–45. Springer-Verlag.

Sirbu, M.A. (1997). Credits and debits on the Internet. *IEEE Spectr*. **34**(2):23-29.

Sultan, K.I. (2015). *Modeling and verifying probabilistic social commitments in multi-agent systems.* PhD thesis, Concordia University (CA), https://spectrum.library.concordia.ca/979616/

Wooldridge, M. (2009) *An Introduction to Multi-Agent Systems*, 2nd ed., John Wiley & Sons, Chichester, West Sussex, PO19 8SQ, UK.

PSEUDO NEURAL NETWORKS VIA ANALYTIC PROGRAMMING WITH DIRECT CODING OF CONSTANT ESTIMATION

Zuzana Kominkova Oplatkova, Adam Viktorin, Roman Senkerik

Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{oplatkova, aviktorin, senkerik}@.utb.cz

KEYWORDS
Pseudo neural networks, Analytic programming, Differential evolution.

ABSTRACT

This research deals with a novel approach to classification – pseudo neural networks (PNN). This technique was inspired in classical artificial neural networks (ANN), where a relation between inputs and outputs is based on the mathematical transfer functions and optimised numerical weights. Compared to ANN, the whole structure in PNN, i.e. the relation between inputs and output(s), is fully synthesised by evolutionary symbolic regression tool – analytic programming. Compared to previous synthesised models, the PNN in this paper were synthesised via a new approach to constant estimation inside the analytic programming – direct coding. Iris data was used for the experiments and PNN were used for the synthesis of a complex classifier for more classes. For experimentation, Differential Evolution (de/rand/1/bin) for optimisation in analytic programming (AP) was used.

INTRODUCTION

This paper follows and combines the previous research (Kominkova Oplatkova et al., 2014), (Kominkova Oplatkova et al. 2017a) and (Kominkova Oplatkova et al. 2017b). The new approaches for constant estimation (Kominkova Oplatkova et al. 2017b) inside the Analytic Programming (AP) (Zelinka et al., 2011) caused the interests in the behaviour of the synthesised pseudo neural networks (PNN). The aim is to observe the speed and the achieved quality of the classifier and compare it with the original meta-evolutionary approach.

The pseudo neural networks can be used in more or less the same tasks as artificial neural networks (ANN) like pattern recognition, prediction, control, signal filtering, approximation and others. All ANNs are based on some relation between inputs and output(s), which utilises mathematical transfer functions and optimised weights from training process. The setting-up of layers, number of neurons in layers, estimating of suitable values of weights is a demanding procedure. On account of this fact, pseudo neural networks, which represent the novelty approach using symbolic regression with evolutionary computation, namely Analytic Programming with the direct coding of constant estimation, is proposed in this paper.

Symbolic regression in the context of evolutionary computation means to build a complex formula from basic operators defined by users. The basic case represents a process in which the measured data is fitted and a suitable mathematical formula is obtained analytically. This process is widely known for mathematicians. They use this process when a need arises for a mathematical model of unknown data, i.e. the relation between input and output values. The symbolic regression can also be used for the design of electronic circuits or the optimal trajectory for robots and within other applications (Back et al., 1997), (Koza, 1998), (Koza, 1999), (O'Neill et al., 2003), (Zelinka et al., 2011), (Oplatkova, 2009), (Varacha et al., 2006). Everything depends on the user-defined set of operators. The proposed technique is similar to the synthesis of an analytical form of the mathematical model between input and output(s) in training set used in neural networks. Therefore the technique is called Pseudo Neural Networks.

Initially, John Koza proposed the idea of symbolic regression done by means of a computer in Genetic Programming (GP) (Back et al., 1997), (Koza, 1998), (Koza, 1999). The other approaches are e.g. Grammatical Evolution (GE) developed by Conor Ryan (O'Neill et al., 2003) and here described Analytic Programming (Zelinka et al., 2011), (Oplatkova, 2009), (Varacha et al., 2006), .

The above-described tools were recently commonly used for synthesis of artificial neural networks but in a different manner than is presented here. One possibility is the usage of evolutionary algorithms for optimization of weights to obtain the ANN training process with a small or no training error result. Some other approaches represent the special ways of encoding the structure of the ANN either into the individuals of evolutionary algorithms or into the tools like Genetic Programming. But all of these methods are still working with the classical terminology and separation of ANN to neurons and their transfer functions (Fekiac, 2011). In this paper, the proposed technique synthesizes the structure without a prior knowledge of transfer functions and inner potentials. It synthesizes the relation between inputs and output of training set items used in neural networks so

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

that the items of each group are correctly classified according to the rules for the cost function value.
In general, there are two options:

a) a continuous version of classification when just one node is enough even for more classes, alternatively said a multi-input-single-output (MISO) version (Kominkova Oplatkova et al., 2013), (Kominkova Oplatkova et al., 2016). The number of required classes determines the number of intervals which the range of output values will be divided into.

b) more output nodes are used, which means that AP is launched so many times the number of output nodes is required since the combination of output values predicts the final class. Alternatively, it is possible to call this approach as a multi-input-multi-output (MIMO) version (Kominkova Oplatkova et al., 2014). In this presented case three output nodes are used. Each node represents just one class, i.e. the activated node (output value =1) stands for the appropriate class, the rest of nodes should be deactivated (the output value = 0). The AP is performed for each particular node separately and independently.

The dataset used for training is Iris data set (Machine Learning Repository, Fisher 1936). It is a very known benchmark dataset for classification problem, which was introduced by Fisher for the first time.

As mentioned above, the paper deals with the comparison in the quality of the trained classifier via the original meta evolutionary approach of AP and the new one with the direct coding of the constant estimation.

Firstly, Analytic Programming as a symbolic regression tool with its used strategies is described. Subsequently, Differential Evolution used for the optimisation procedure within Analytic Programming is mentioned. Afterwards, the proposed experiment with differences and obtained results follow and a conclusion finishes the paper.

ANALYTIC PROGRAMMING

Basic principles of the AP were developed in 2001 (Zelinka et al., 2005), (Zelinka et al., 2008), (Zelinka et al., 2011).

The core of AP is based on a special set of mathematical objects and operations. The collection of mathematical objects is the set of functions, operators and terminals, which are usually constants or independent variables. Various functions and terminals can be mixed in this set. This set is called general functional set (GFS) due to its variability of the content. The structure of GFS is created by subsets of functions according to the number of their arguments. For example, GFSall is a set of all functions, operators and terminals, GFS3arg is a subset containing functions with only three arguments, GFS0arg represents only terminals. The subset structure presence in GFS is of vital importance for AP. It is used to avoid synthesis of pathological programs, i.e. programs containing functions without arguments and similar. The content of GFS is dependent only on the user (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova, 2009).

The second part of the AP core is a sequence of mathematical operations, which are used for the program synthesis. These operations are used to transform an individual of a population into a suitable program. Mathematically stated, it is a mapping from an individual domain into a program domain. This mapping consists of two main parts. The first part is called discrete set handling (DSH) (See Figure 1) (Zelinka et al., 2005), (Lampinen and Zelinka, 1999) and the second one stands for security procedures which do not allow synthesising pathological programs. The method of DSH, when used, allows handling arbitrary objects including nonnumerical objects like linguistic terms {hot, cold, dark…}, logic terms (True, False) or other user-defined functions. In the AP DSH is used to map an individual into GFS and together with security procedures creates the mapping mentioned above which transforms the arbitrary individual into a program.

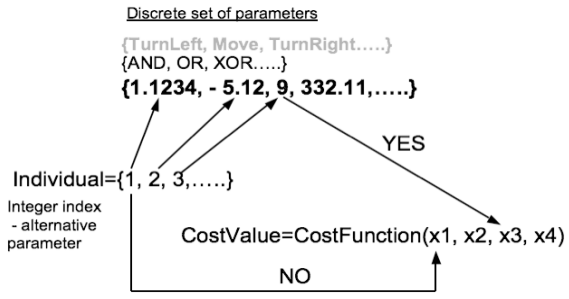


Figure 1: Discrete set handling

AP needs some evolutionary algorithm (Zelinka, 2004) that consists of a population of individuals for its run. Individuals in the population consist of integer parameters, i.e. an individual is an integer index pointing into GFS. The creation of the program can be schematically observed in Fig. 2.

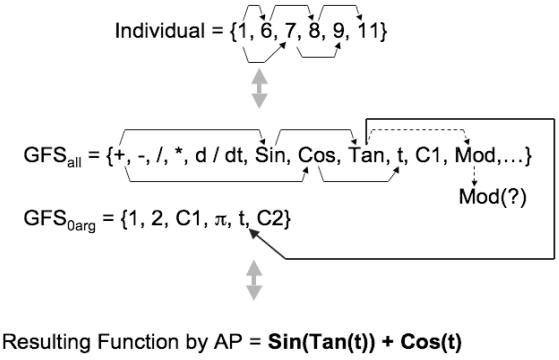


Figure 2: Main principles of AP

An example of the process of the final complex formula synthesis (according to the Fig. 2) follows.

The number 1 in the position of the first parameter means that the operator plus (+) from GFS_{all} is used (the end of the individual is far enough). Because the operator + must have at least two arguments, the next two index pointers 6 (sin from GFS) and 7 (cos from GFS) are dedicated to this operator as its arguments.

The two functions, sin and cos, are one-argument functions; therefore the next unused pointers 8 (tan from GFS) and 9 (*t* from GFS) are dedicated to the sin and cos functions. As an argument of cos, the variable *t* is used, and this part of the resulting function is closed (*t* has zero arguments) in its AP development. The one-argument function tan remains, and there is one unused pointer 11, which stands for Mod in GFS_{all}. The modulo operator needs two arguments but the individual in the example has no other indices (pointers, arguments). In this case, it is necessary to employ security procedures and jump to the subset of GFS_{0arg}. The function tan is mapped on *t* from GFS_{0arg} which is in the 11th position, cyclically from the beginning. The detailed description is represented in (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova et al., 2009).

ANALYTIC PROGRAMMING - VERSIONS

The above-described version is the basic one AP_{basic} (Zelinka et al., 2005) - without constant estimation. Except this, AP works with nonlinear fitting version AP_{nf}, meta-evolutionary strategy AP_{meta} and three novel direct approaches AP_{extend} (extended individuals) (Viktorin et al., 2016), $AP_{direct1}$ (Urbanek et al., 2016) and $AP_{direct2}$ (Kominkova Oplatkova et al. 2017a).
The following subsections are dedicated only to meta-evolutionary approach and one direct encoding $AP_{direct2}$ because the paper deals with their comparison in the case of PNN synthesis.
AP uses the general constant *K* (Zelinka et al., 2005) which is indexed during the evolution (1) - (3). The *K* is a terminal, i.e. GFS_{0arg}. So it is used as a standard terminal which is similar for instance to a variable *x* in the evolutionary process (1). When *K* is needed, a proper index is assigned – K_1, K_2, ... K_n (2). Numeric values of indexed *Ks* are estimated (3) via different techniques including AP_{meta},(meta-evolutionary approach with a second/slave evolutionary algorithm) (Zelinka et al., 2005, Oplatkova 2009) and $AP_{direct2}$.

$$\frac{x^2 + K}{\pi^K} \tag{1}$$

$$\frac{x^2 + K_1}{\pi^{K_2}} \tag{2}$$

$$\frac{x^2 + 3.156}{\pi^{90.78}} \tag{3}$$

AP_{meta} – meta-evolutionary approach

Generally, metaevolution means the evolution of evolution. Several directions, e.g. the usage of an evolutionary algorithm for tuning or controlling of another evolutionary technique or the evolutionary design of evolutionary algorithms, are discussed for instance in (Diosan, 2009), (Edmons, 2001), (Jones, 2002), (Oplatkova, 2009), (Kordik, 2010), Deugo, 2004), (Eiben, 2007).
In AP_{meta}, the metaevolution means that one evolutionary algorithm drives the main process of symbolic regression and the second is used for the constant estimation. This meta approach in AP is used when coefficients in the found model cannot be simply adjusted because of the character of the problem. It is not possible to interpolate the model to some "measured" values but the obtained solution is used further as a part of the complex technique with the aim to find the quality of the solution and cost value estimation.
AP_{meta} is a time-consuming process and the number of cost function evaluations, which is one of the comparable factors, is usually very high. This fact is given by two evolutionary procedures (Fig. 3).

$$EA_{master} \Rightarrow program \Rightarrow K_{indexing} \Rightarrow EA_{slave} \Rightarrow K_{estimation} \Rightarrow final \cdot solution$$

Figure 3: Schema of AP procedures

EA_{master} is the main evolutionary algorithm for AP, EA_{slave} is the second evolutionary algorithm inside AP. Thus, the number of cost function evaluation (CFE) is given by (4).

$$CFE = EA_{master} * EA_{slave} \tag{4}$$

The following $AP_{direct2}$ approach was developed to decrease the number of cost function evaluations to (5).

$$CFE = EA_{master} \tag{5}$$

$AP_{direct2}$ - direct encoding 2 of K in the individual

This version was developed after the analysis $AP_{direct1}$ behaviour. According to authors, the problematic issues were connected with the neighbourhood of arguments which are responsible for *K* estimation. Since *K* values are dependent only on the decimal part of the argument regardless the integer part of the value in $AP_{direct1}$ (Urbanek et al., 2016), two points placed on the opposite sides of the coordinate system can be neighbours from ind_K (6) point of view. It does not help the evolutionary optimisation process which expects that two points lie next to each other physically in the coordinate system for a successful performance. Surprisingly, (Kominkova Oplatkova et al. 2017b) has not proven such assumption and both variants work more or less in the same high-quality way. Despite this conclusion, authors will use the only $AP_{direct2}$ for further experiments.
$AP_{direct2}$ is based on the $AP_{direct1}$ (Urbanek et al., 2016). It works with a direct encoding in an individual and the difference is in the different computation of ind_K (9). No other slave evolutionary algorithm is necessary for *K* estimation as in AP_{meta}. The variable ind_K (6) works as

a proportional pointer which determines the value from the selected range of *K*.
It takes the value of the not rounded individual as the proportional part in respect of length of all components in GFS_{All}.

$$ind_K = \frac{ind}{Dim(GFS_{All})}, \quad (6)$$

where $ind = \{x_1, x_2, x_3 \ldots x_n\}$. Dim($GFS_{All}$) means the number of all non-terminals and terminals used in AP. For instance, if GFS_{All}={+, -, /, *, x, K}, the Dim(GFS_{All}) = 6 and the valid range for arguments in the individual is in the interval <1,6>.
The corresponding *K* is then computed easily from (7).

$$K = ind_K * \left| rangeK_{\max} - rangeK_{\min} \right| + rangeK_{\min} \quad (7)$$

The mapping is done in the standard procedure as in AP_{basic} and general approach of *K* indexation. When *K* is needed, the value in the corresponding position from (7) is directly used.

USED EVOLUTIONARY ALGORITHM - DIFFERENTIAL EVOLUTION

As mentioned above, the Analytic Programming needs an evolutionary algorithm for the optimisation - finding the best shape of the complex formula. This research used Differential Evolution (Price, 2005) in its canonical version DE/Rand/1/Bin. Future research expects to use some other strategies as DE/Best/1/Bin or SHADE which had a good performance in (Viktorin et al., 2016) and (Kominkova Oplatkova et al. 2017b).
DE is a population-based optimisation method that works on real-number-coded individuals (Price, 2005). For each individual $\vec{x}_{i,G}$ in the current generation G, DE generates a new trial individual $\vec{x}'_{i,G}$ by adding the weighted difference between two randomly selected individuals $\vec{x}_{r1,G}$ and $\vec{x}_{r2,G}$ to a randomly selected third individual $\vec{x}_{r3,G}$. The resulting individual $\vec{x}'_{i,G}$ is crossed-over with the original individual $\vec{x}_{i,G}$. The fitness of the resulting individual, referred to as a perturbed vector $\vec{u}_{i,G+1}$, is then compared with the fitness of $\vec{x}_{i,G}$. If the fitness of $\vec{u}_{i,G+1}$ is greater than the fitness of $\vec{x}_{i,G}$, then $\vec{x}_{i,G}$ is replaced with $\vec{u}_{i,G+1}$; otherwise, $\vec{x}_{i,G}$ remains in the population as $\vec{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimisation ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Description of used DERand1Bin mutation strategy is presented in (8). Please refer to (Price and Storn 2001, Price 2005) for the description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \bullet \left(x_{r2,G} - x_{r3,G} \right) \quad (8)$$

PROBLEM DESIGN

For this classification problem, iris data set was used (Machine Learning Repository, Fisher 1936). This set contains 150 instances. The half amount was used as training data and the second half was used as testing data. The dataset contains three classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are NOT linearly separable from each other. Each instance has four attributes (sepal length, sepal width, petal length and petal width (Fig. 3a)) and type of class – iris virginica (Fig. 3b), iris versicolor Fig. 3c) and iris setosa (Fig. 3d). The attributes contain real values.

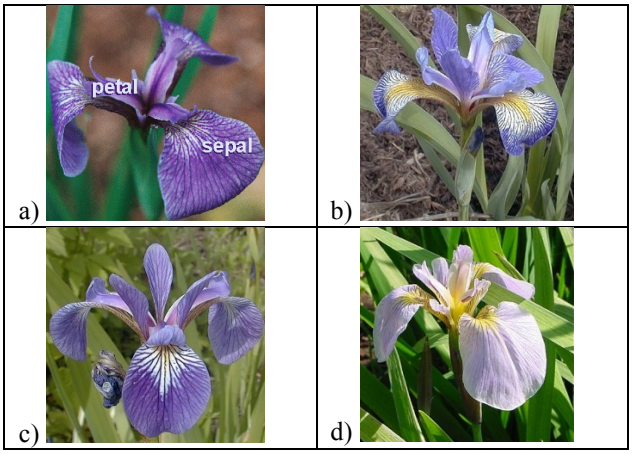


Figure 3: a) Iris - petal and sepal, b) Iris virginica, c) Iris versicolor, d) Iris setosa

Usually, the class is defined by three output nodes in classical artificial neural net and binary code. After the ANN training, an equation can be separated for each output node - relation between inputs, weights and the output node. The training is done in a parallel way - all weights are optimised and the final equations are produced in "one step". In this paper, MIMO approach was used. AP synthesised a pseudo neural net structure between inputs and three outputs independently. The procedure was carried out serially to reach the final solution. The output values were proposed similarly to ANN case, i.e. the combination of zeros and ones. The output vector for iris setosa was (1,0,0), for iris virginica (0,1,0) and for iris versicolor (0,0,1).
Since AP together with differential evolution are used for optimisation – finding of the best shape of the relationship between inputs and output – the cost function which measures the quality of the solution has to be designed. The cost function was selected based on the previous research and the natural character of the problem (9). Since it is necessary to synthesise three independent output equations, the cost function was given the same for all, i.e. if the cv is equal to zero, all training patterns are classified correctly.

$$cv = \sum_{i=1}^{n} \left| requiredOutput - currentOutput \right| \quad (9)$$

The training patterns were set up according to the following way. If the final classifier model synthesised by AP (one independent expression) determines a class iris setosa, the appropriate item in this group will be assigned to the output value 1 in the training set and the other two kinds of plants will be equal to zero. The same way, the training data were also prepared for the other two independent expressions created by AP.

RESULTS AND DISCUSSION

The paper will compare AP_{meta} and $AP_{direct2}$ strategies with differential evolution DE/Rand/1/Bin. The setting was based on some previous research in this field (Tab. 1 and Tab. 2 for AP_{meta} and Tab. 3. for $AP_{direct2}$). The total Max CFE for AP_{meta} was 6 000 000 according to (4) and for $AP_{direct2}$ only 8 000 as stated in Tab. 3. which is 750 times less. The real used CFEs were even much smaller which means the significant reduction of computation time.

Table 1: DE settings for the main process in AP_{meta}

PopSize	20
F	0.8
CR	0.8
Generations	50
Max. CF Evaluations (CFE)	1000

Table 2: DE settings for meta-evolution in AP_{meta}

PopSize	40
F	0.8
CR	0.8
Generations	150
Max. CF Evaluations (CFE)	6000

Table 3: DE settings in $AP_{direct2}$

PopSize	40
F	0.5
CR	0.8
Generations	2000
Max. CF Evaluations (CFE)	8 000

The set of elementary functions for AP was inspired in the items used in classical artificial neural nets. The components of input vector x contain values of each attribute (x_1, x_2, x_3, x_4).
Basic set of elementary functions for AP:
GFS2arg= +, -, /, *, ^, exp
GFS0arg= x_1, x_2, x_3, x_4, K

All simulations were performed 30 times out. Firstly, statistical results are shown for each class separately. Tab 4. deals with the class 1 which is linearly separable from the rest two classes. It corresponds with iris setosa was (1,0,0). The cost value for training converged to the value zero and also the testing was carried out correctly with the testing error equal to zero.

Tab. 5. works with the class 2 and tab. 6. with class 3 in a similar manner.

Table 4: Statistical results for cost function evaluations when cost value is equal to zero – class 1

	AP_{meta}	$AP_{direct2}$
Min	54 000.0	47.00
Max	288 000.0	123.00
Avg	102 400.0	71.77
Median	75 000.0	66.50
St.Dev.	52 589.2	18.14

Table 5: Statistical results for cost function evaluations when cost value reached only value two (training error is equal to two misclassified items) – class 2

	AP_{meta}	$AP_{direct2}$
Min	$2.868*10^6$	378.00
Max	$5.964*10^6$	987.00
Avg	$4.862*10^6$	758.97
Median	$4.971*10^6$	788.00
St.Dev.	797743.	159.93

Table 6: Statistical results for cost function evaluations when cost value reached only value one (training error is equal to one misclassified items) – class 3

	AP_{meta}	$AP_{direct2}$
Min	$2.268*10^6$	443.00
Max	$5.922*10^6$	959.00
Avg	$4.585*10^6$	740.93
Median	$4.728*10^6$	766.00
St.Dev.	957741.	143.44

From the tables mentioned above, it is visible that the training error was equal to 3 items (it means 4% error, 96% accuracy from all 75 training patterns). The testing error was equal to 7 in those cases (10.67% error, 89.33% accuracy).
The training error for the meta-evolutionary approach was equal to 2 misclassified items (it means 2.66% error, 97.34% success from all 75 training patterns) and testing error equal to 5 misclassified items (8% error, 92% accuracy).
The direct approach is slightly worse than the meta-evolutionary approach in this preliminary results but much faster. Both results suffer a bit from small overfitting but they are comparable with standard ANN approach. The future experiments with the standard validation techniques might help to make a better score.

From carried simulations, several notations of input dependency were obtained for both approaches. The advantage is that equations for each output node can be mixed. Therefore, the best shapes for each output node

can be selected and used together for correct behaviour. The following equations (10) - (12) are just examples of the found synthesised expressions of AP_{meta} and (13) - (15) for $AP_{direct2}$.

$$y1 = -692.512\,x3\exp\left(\frac{-x3\left(\left(x1^{438.154}-257.159\right)\left(x3-2.05754\right)-832.086\right)\left(\exp(\exp(x2))+\exp(x2)\right)^{x4}}{\exp(x2)-\exp(-0.0010047\,x1)}\right) \quad (10)$$

$$y2 = -5.73269\times10^{-48}-\exp\left(-\frac{0.0377017\,x1\left(-387.792^{320.911\,x4}+x4\right)}{x4}\right)+\frac{-713.828+700.061^{\frac{5.45627\times10^{221}}{953.763+\exp(x2)}}}{x1} \quad (11)$$

$$y3 = \exp\left(-\frac{950.429}{\exp(x1-1.\,x3\,x4)+\frac{\exp\left(\exp\left(-\exp(-238.261+x1)\right)^{\exp(\exp(57.0214-x4))}\right)^{5.0545765157800\times10^{-295}}+x2}{-16.4604+x3}}\right) \quad (12)$$

$$y1 = -3.24291+x4^{-\exp(x4)} \quad (13)$$

$$y2 = \frac{1}{\left(\left(-2.35294-\exp(x1)+x1\right)^{x2}\right)^{0.02353}} \quad (14)$$

$$y3 = -\frac{8.25994\exp\left(-2094.48\left(3.52941-x1-x2\right)\right)}{-5.46965+\exp(\exp(x1))-\exp(x3)+x4} \quad (15)$$

CONCLUSION

This paper deals with a novel approach to evolutionary synthesised Pseudo neural networks via Analytic programming and its two strategies – meta-evolutionary approach - AP_{meta} and direct coding of *K* estimation in the individual for speed up the process - $AP_{direct2}$. All simulations were performed with a DE/Rand/1/Bin strategy of differential evolution algorithm for the main optimisation process in AP and also the meta-evolutionary part (second slave algorithm).

The results showed that both approaches are comparable between themselves and also with classical artificial neural networks. However, $AP_{direct2}$ uses significantly less number of cost function evaluations than AP_{meta}, in our case even 750 times less. The speed of training is very important and each reduction of the training time is welcome.

Future plans include a comparison of other evolutionary techniques or their strategies, for instance SHADE since (Viktorin, 2016) presented better results with it in the case of data approximation or swarm algorithms – particle swarm optimisation, self.-organizing migrating algorithm and others.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

Back T., Fogel D. B., Michalewicz Z., *Handbook of evolutionary algorithms*, Oxford University Press, 1997, ISBN 0750303921

Deugo D., Ferguson D.: Evolution to the xtreme: Evolving evolutionary strategies using a meta-level approach, Proceedings of the 2004 IEEE congress on evolutionary computation, IEEE Press, Portland, Oregon, pp. 31–38, 2004

Dioşan, L., Oltean, M.: Evolutionary design of evolutionary algorithms. Genetic Programming and Evolvable Machines, Vol. 10, Issue 3, p. 263-306, 2009

Edmonds, B.: Meta-genetic programming: Co-evolving the operators of variation, Elektrik, Vol. 9, Issue 1, pp. 13-29, 2001

Eiben A.E., Michalewicz Z., Schoenauer M., Smith J.E.: Parameter control in evolutionary algorithms, pp. 19–46, Springer, 2007

Jones D.F., Mirrazavi S.K., Tamiz M.: Multi-objective meta-heuristics: An overview of the current state-of-the-art, European Journal of Operational Research, Volume 137, Issue 1, 16 February 2002, Pages 1-9, ISSN 0377-2217.

Kominkova Oplatkova Z., Senkerik R. (2014): MIMO Pseudo Neural Networks for Iris Data Classification. In Advances in Intelligent Systems and Computing. 285. Heidelberg : Springer-Verlag Berlin, 2014, p. 165-172. ISSN 2194-5357. ISBN 978-3-319-06739-1.

Kominkova Oplatkova Z., Senkerik R. (2016): Control Law and Pseudo Neural Networks Synthesized by Evolutionary Symbolic Regression Technique, in Al-Begain K., Bargiela A.: Seminal Contributions to Modelling and Simulation - Part of the series Simulation Foundations, Methods and Applications, pp 91-113, doi: 10.1007/978-3-319-33786-9_9, ISBN: 978-3-319-33785-2.

Kominkova Oplatkova Z., Viktorin A., Senkerik R. (2017a): Comparison of Three Novelty Approaches to Constants (Ks) Handling in Analytic Programming Powered by SHADE, Mendel, Springer Series, in print

Kominkova Oplatkova Z., Viktorin A., Senkerik R., Urbanek T. (2017b): Different Approaches For Constant Estimation In Analytic Programming, ECMS 2017, p. 326 – 334, doi: 10.7148/2017-0326, ISSN 2522-2414, ISBN: 978-0-9932440-4-9

Kordík P., Koutník J., Drchal J., Kovářík O., Čepek M., Šnorek M.: Meta-learning approach to neural network optimization, Neural Networks, Vol. 23, Issue 4, p. 568-582, 2010, ISSN 0893-6080.

Koza J. R. et al., *Genetic Programming III; Darwinian Invention and problem Solving*, Morgan Kaufmann Publisher, 1999, ISBN 1-55860-543-6

Koza J. R., *Genetic Programming,* MIT Press, 1998, ISBN 0-262-11189-6

Lampinen J., Zelinka I., 1999, "New Ideas in Optimization – Mechanical Engineering Design Optimization by Differential Evolution", Volume 1, London: McGraw-hill, 1999, 20 p., ISBN 007-709506-5.

O'Neill M., Ryan C., *Grammatical Evolution. Evolutionary Automatic Programming in an Arbitrary Language*, Kluwer Academic Publishers, 2003, ISBN 1402074441

Oplatkova Z.: Metaevolution: Synthesis of Optimization Algorithms by means of Symbolic Regression and Evolutionary Algorithms, Lambert Academic Publishing Saarbrücken, 2009, ISBN: 978-3-8383-1808-0

Price K., Storn R. M., Lampinen J. A., 2005, "Differential Evolution : A Practical Approach to Global Optimization", (Natural Computing Series), Springer; 1 edition.

Price, K. and Storn, R. (2001), *Differential evolution homepage*, [Online]: http://www.icsi.berkeley.edu/~storn/code.html, [Accessed 29/02/2012].

Urbanek T., Prokopova Z., Silhavy R., Kuncar A.: New Approach of Constant Resolving of Analytical Programming. In *30th European Conference on Modelling and Simulation*, 2016, p. 231-236. ISBN 978-0-9932440-2-5.

Viktorin A., Pluhacek M., Kominkova Oplatkova Z., Senkerik R.: Analytical Programming with Extended Individuals. In *30th European Conference on Modelling and Simulation*, 2016, p. 237-244. ISBN 978-0-9932440-2-5.

Volna, E., Kotyrba, M., & Jarusek, R. (2013). Multi-classifier based on Elliott wave's recognition. Computers & Mathematics with Applications, 66(2), 213-225.

Volna, E., Kotyrba, M., Kominkova Oplatkova Z., Senkerik R. (2018). Elliott waves classification by means of neural and pseudo neural networks. Soft computing, 2018, vol. 22, issue. 6., p. 1803-1813. ISSN 1432-7643 DOI: 10.1007/s00500-016-2097-y

Zelinka et al.: Analytical Programming - a Novel Approach for Evolutionary Synthesis of Symbolic Structures, in Kita E.: Evolutionary Algorithms, InTech 2011, ISBN: 978-953-307-171-8

Zelinka I., Varacha P., Oplatkova Z., *Evolutionary Synthesis of Neural Network*, Mendel 2006 – 12th International Conference on Softcomputing, Brno, Czech Republic, 31 May – 2 June 2006, pages 25 – 31, ISBN 80-214-3195-4

Zelinka I.,Oplatkova Z, Nolle L., 2005. *Boolean Symmetry Function Synthesis by Means of Arbitrary Evolutionary Algorithms-Comparative Study*, International Journal of Simulation Systems, Science and Technology, Volume 6, Number 9, August 2005, pages 44 - 56, ISSN: 1473-8031.

AUTHOR BIOGRAPHIES

ZUZANA KOMINKOVA OPLATKOVA

is an associate professor at Tomas Bata University in Zlin. Her research interests include artificial intelligence, soft computing, evolutionary techniques, symbolic regression, neural networks. She is an author of around 170 papers in journals, book chapters and conference proceedings. Her e-mail address is: oplatkova@utb.cz

ADAM VIKTORIN

was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the field of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@utb.cz

ROMAN SENKERIK

was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2004, Ph.D. degree in Technical Cybernetics in 2008 and Assoc. prof. in 2013 (Informatics). He is now an Assoc. prof. at the same university (research and courses in: Evolutionary Computation, Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). He is an author of around 290 papers in journals, book chapters and conference proceedings. His email address is: senkerik@utb.cz

STUDY ON VELOCITY CLAMPING IN PSO USING CEC`13 BENCHMARK

Michal Pluhacek, Roman Senkerik, Adam Viktorin and Tomas Kadavy
Tomas Bata University in Zlin , Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{pluhacek, senkerik, aviktorin, kadavy}@fai.utb.cz

KEYWORDS

Particle swarm optimization, PSO, Velocity clamping, Optimization

ABSTRACT

In this paper, we perform a new study of the importance of using clamping of the velocity of particles in the Particle Swarm Optimization algorithm. The velocity clamping is used to prevent the particles from rapid acceleration. We present results of testing different settings of maximal velocity on the extensive CEC´13 benchmark set and discuss the results, alongside the overall importance of the velocity clamping method.

INTRODUCTION

The Particle Swarm Optimization algorithm (PSO) (Kennedy and Eberhart 1995, Kennedy 1997) is one of the most prominent members of Swarm intelligence based category of evolutionary optimization algorithms (Volna and Kotyrba 2014). The PSO is very popular and widely applied. Many researchers focus on analyzing the inner principles of the algorithm in order to improve the understating of the inner dynamics and propose performance improvements (Shi and Eberhart 1998a, Van Den Bergh, and Engelbrecht 2006, Nickabadi et al. 2012). Like any other method, the PSO suffers from several drawbacks, and the continuous research focuses on addressing these problems.

Velocity clamping is a popular approach for dealing with one of the main drawbacks of PSO – the rapid particle acceleration. In this method, a maximal velocity value (v_{max}) is set. The popular choice for the setting of maximal velocity (Shi and Eberhart 1998b) is 20% of the search space range. However, it is not always favorable to use this value or to use velocity clamping at all.

In this study, we choose to investigate the impact of different settings of maximal velocity on the performance of PSO algorithm on the complex CEC`13 benchmark set that represents a large variety of fitness landscapes. Based on the result we choose to re-evaluate the significance of velocity clamping in PSO.

We provide evidence that the popular setting of the v_{max} = 20% of range might not be the best choice for many problems and that the velocity clamping might be omitted in some cases.

The paper is structured as follows: In the second section, a brief description of original PSO algorithm is given. In the next section, the experiment setup is detailed. Following is an extensive presentation of collected data, and the results are discussed in the following section.

PARTICLE SWARM OPTIMIZATION

The PSO algorithm is inspired by the natural swarm behavior of animals (such as birds and fish). It was firstly introduced by Eberhart and Kennedy in 1995 (Kennedy and Eberhart, 1995).

Each particle in the population represents a possible solution of the optimization problem, defined by the cost function (CF). In each iteration of the algorithm, a new location (combination of CF parameters) of the particle is calculated based on the previous location and velocity vector (velocity vector contains particle velocity for each dimension).

The velocity calculation formula is given in (1).

$$v_{ij}^{t+1} = w \cdot v_{ij}^{t} + c_1 \cdot Rand \cdot (pBest_{ij} - x_{ij}^{t}) + c_2 \cdot Rand \cdot (gBest_{j} - x_{ij}^{t}) \quad (1)$$

Where:

v_{ij}^{t+1} - New velocity of the ith particle in iteration t+1. (component j of the dimension D).

w – Inertia weight value.

v_{ij}^{t} - Current velocity of the ith particle in iteration t. (component j of the dimension D).

c_1, c_2 - Acceleration constants.

$pBest_{ij}$ – Personal best solution found by the ith particle. (component j of the dimension D).

$gBest_j$ - Best solution found in a population. (component j of the dimension D).

x_{ij}^{t} - Current position of the ith particle (component j of the dimension D) in iteration t.

Rand – Pseudo random number, interval (0, 1).

The new position of each particle is then given by (2), where x_i^{t+1} is the new particle position:

$$x_i^{t+1} = x_i^{t} + v_i^{t+1} \quad (2)$$

EXPERIMENT SETUP

In the experiment, 11 different settings of maximal velocity were used. The maximal velocity is typically set as a fractional multiplication of the search space range. This pattern was followed in this study with v_{max} set to values <0.1·Range ; Range> by step 0.1. In addition, the variant with no velocity clamping was tested (noted N/A). Therefore, 11 different settings of PSO were tested.

The performance of PSO algorithm with different settings of maximal allowed velocity was tested on the IEEE CEC 2013 benchmark set (Liang et al. 2013) for dimension setting (dim) = 10 and 30. According to the benchmark rules, 51 separate runs were performed for

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

each algorithm, and the maximum number of cost function evaluations (CFE) was set to 10000 · dim. The population size was set to 40. According to literature (Shi and Eberhart 1998, Shi and Eberhart 1999), the values of control parameters were set to popular values as follows: c_1, c_2 = 1.49618; w = 0.7298.

The results were tested for statistical significance using the Friedman rank test (α=0.05), followed by the Nemenyi post-hoc test.

RESULTS

In this section, the results overview is presented.

Firstly, the results for dim = 10 are presented. The Figures 1 – 8 depict selected examples of mean *gBest* history for various benchmark functions. Table 1 contains the median result values for all tested settings.

The ranking according to the Friedman rank test is given in Fig. 9. The critical distance from the lowest rank is displayed.

Further, the results for dim = 30 are presented in a similar way. The median result values are given in Table 2. Selected examples of mean *gBest* history are displayed in Figures 10 – 13. Finally, the Friedman ranking is presented in Fig. 14.

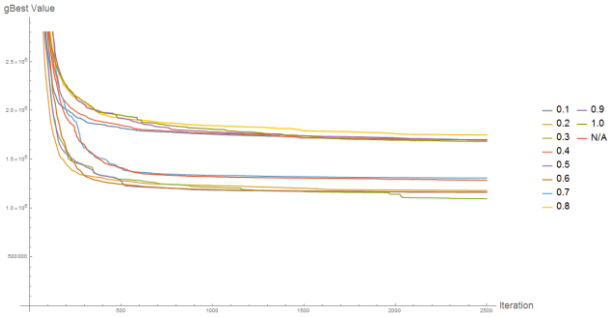


Figure 1. Mean gBest value history – f_2 – dim = 10

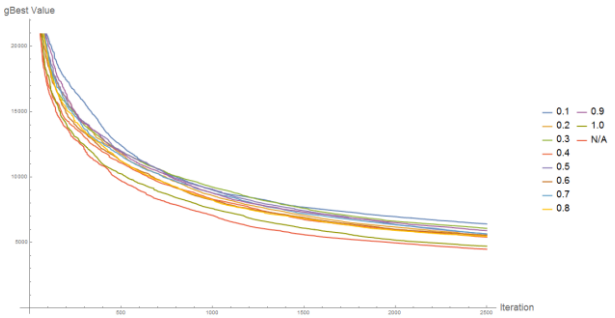


Figure 2. Mean gBest value history – f_4 – dim = 10

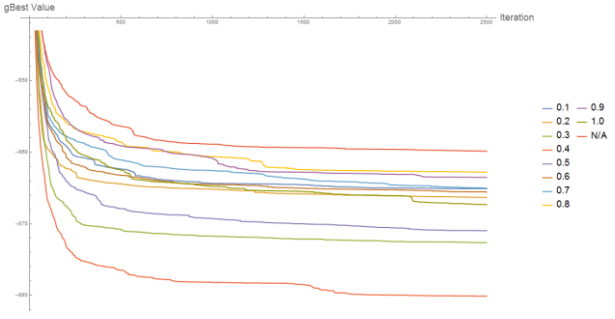


Figure 3. Mean gBest value history – f_6 – dim = 10

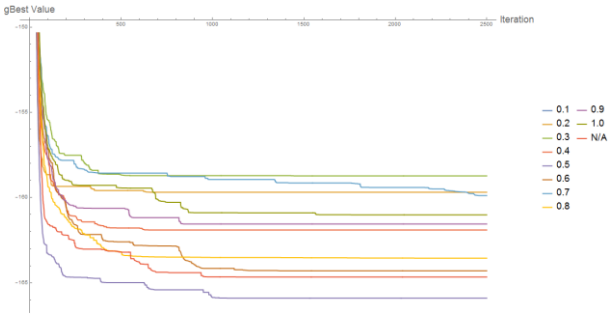


Figure 4. Mean gBest value history – f_{13} – dim = 10

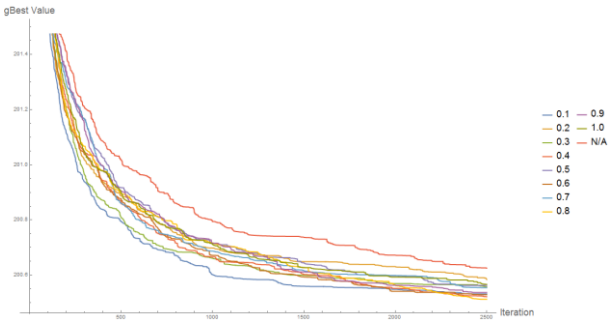


Figure 5. Mean gBest value history – f_{16} – dim = 10

Figure 6. Mean gBest value history – f_{21} – dim = 10

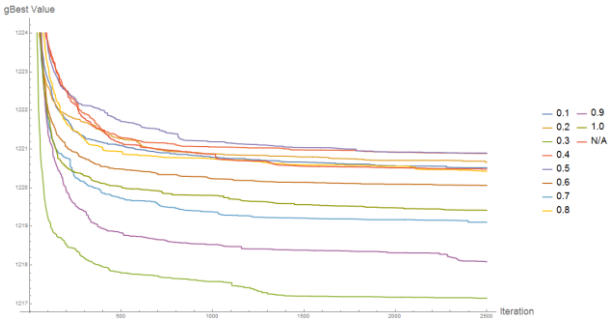


Figure 7. Mean gBest value history – f_{24} – dim = 10

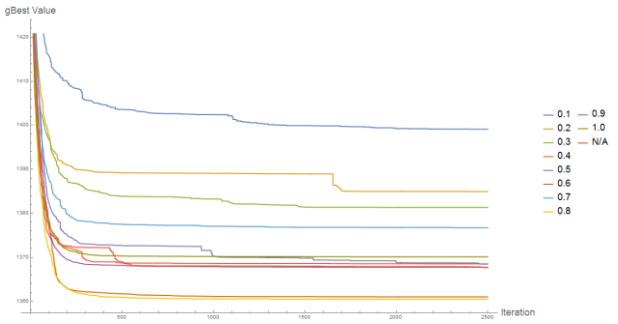


Figure 8. Mean gBest value history – f_{26} – dim = 10

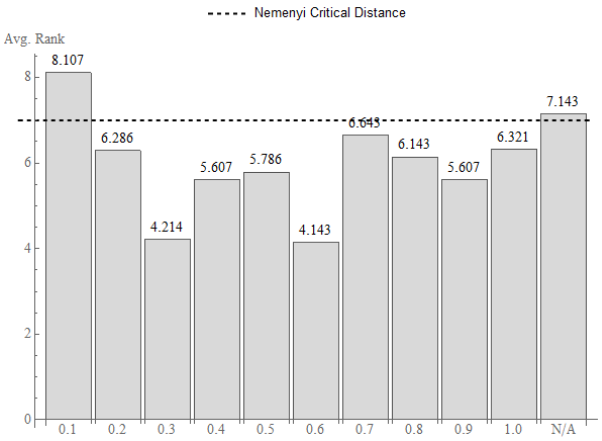


Figure 9. Friedman rank with Nemenyi critical distance dim = 10

TABLE I. MEDIAN OF RESULTS, DIM = 10;

v_{max}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	N/A
f_1	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03
f_2	1.11E+05	5.47E+04	1.47E+05	8.24E+05	4.10E+05	2.09E+05	4.12E+05	2.16E+05	4.55E+05	3.86E+05	4.22E+05
f_3	1.29E+07	5.90E+06	5.10E+06	8.69E+06	1.27E+07	1.22E+07	2.97E+07	4.36E+07	1.14E+07	2.11E+07	2.89E+07
f_4	5.69E+03	3.66E+03	4.36E+03	5.23E+03	4.74E+03	3.84E+03	5.46E+03	5.10E+03	3.22E+03	4.13E+03	4.58E+03
f_5	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03
f_6	-8.75E+02	-8.80E+02	-8.81E+02	-8.89E+02	-8.79E+02	-8.76E+02	-8.74E+02	-8.76E+02	-8.67E+02	-8.73E+02	-8.68E+02
f_7	-7.49E+02	-7.64E+02	-7.62E+02	-7.61E+02	-7.64E+02	-7.66E+02	-7.55E+02	-7.59E+02	-7.58E+02	-7.57E+02	-7.57E+02
f_8	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02	-6.80E+02
f_9	-5.94E+02	-5.94E+02	-5.95E+02	-5.95E+02	-5.95E+02	-5.94E+02	-5.94E+02	-5.94E+02	-5.95E+02	-5.94E+02	-5.94E+02
f_{10}	-5.00E+02	-5.00E+02	-5.00E+02	-4.99E+02	-4.99E+02	-4.99E+02	-4.99E+02	-4.99E+02	-4.99E+02	-4.99E+02	-4.99E+02
f_{11}	-3.79E+02	-3.88E+02	-3.90E+02	-3.88E+02	-3.90E+02	-3.88E+02	-3.88E+02	-3.86E+02	-3.90E+02	-3.88E+02	-3.88E+02
f_{12}	-2.67E+02	-2.77E+02	-2.77E+02	-2.79E+02	-2.81E+02	-2.82E+02	-2.76E+02	-2.77E+02	-2.77E+02	-2.79E+02	-2.78E+02
f_{13}	-1.50E+02	-1.59E+02	-1.62E+02	-1.62E+02	-1.68E+02	-1.67E+02	-1.61E+02	-1.67E+02	-1.61E+02	-1.60E+02	-1.67E+02
f_{14}	5.67E+02	3.92E+02	3.86E+02	3.27E+02	3.67E+02	3.45E+02	2.82E+02	3.60E+02	3.42E+02	2.76E+02	3.82E+02
f_{15}	8.37E+02	8.87E+02	9.37E+02	9.44E+02	9.80E+02	1.02E+03	9.20E+02	9.39E+02	9.51E+02	1.08E+03	9.96E+02
f_{16}	2.00E+02	2.01E+02	2.01E+02	2.01E+02	2.01E+02	2.00E+02	2.01E+02	2.01E+02	2.01E+02	2.01E+02	2.01E+02
f_{17}	3.25E+02	3.23E+02	3.22E+02	3.23E+02	3.23E+02	3.24E+02	3.24E+02	3.23E+02	3.22E+02	3.22E+02	3.21E+02
f_{18}	4.27E+02	4.28E+02	4.27E+02	4.25E+02	4.27E+02	4.25E+02	4.25E+02	4.27E+02	4.29E+02	4.28E+02	4.28E+02
f_{19}	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02	5.01E+02
f_{20}	6.03E+02	6.03E+02	6.04E+02	6.03E+02	6.03E+02	6.03E+02	6.03E+02	6.03E+02	6.04E+02	6.03E+02	6.03E+02
f_{21}	6.04E+02	6.03E+02	6.03E+02	6.03E+02	6.03E+02	6.03E+02	6.04E+02	6.03E+02	6.03E+02	6.03E+02	6.03E+02
f_{22}	1.70E+03	1.48E+03	1.47E+03	1.46E+03	1.45E+03	1.50E+03	1.41E+03	1.36E+03	1.47E+03	1.35E+03	1.30E+03
f_{23}	2.20E+03	2.20E+03	2.14E+03	2.11E+03	2.20E+03	2.11E+03	2.02E+03	2.15E+03	2.12E+03	2.14E+03	2.14E+03
f_{24}	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03	1.22E+03
f_{25}	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03	1.32E+03
f_{26}	1.40E+03	1.37E+03	1.37E+03	1.34E+03	1.36E+03	1.34E+03	1.36E+03	1.34E+03	1.34E+03	1.38E+03	1.36E+03
f_{27}	1.90E+03	1.81E+03	1.75E+03	1.74E+03	1.78E+03	1.72E+03	1.76E+03	1.75E+03	1.77E+03	1.73E+03	1.79E+03
f_{28}	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03	1.70E+03

TABLE II. MEDIAN OF RESULTS, DIM = 30;

v_{max}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	N/A
f_1	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03	-1.40E+03
f_2	1.49E+08	6.84E+07	6.38E+07	7.48E+07	6.27E+07	7.96E+07	8.24E+07	9.03E+07	8.53E+07	7.60E+07	9.20E+07
f_3	7.69E+10	2.71E+10	2.52E+10	2.97E+10	3.48E+10	2.74E+10	3.23E+10	2.87E+10	4.53E+10	4.41E+10	4.22E+10
f_4	3.82E+04	3.57E+04	3.45E+04	3.28E+04	3.16E+04	3.32E+04	3.23E+04	3.27E+04	3.34E+04	3.39E+04	2.82E+04
f_5	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03	-1.00E+03
f_6	-2.39E+02	-3.82E+02	-4.07E+02	-4.01E+02	-3.92E+02	-4.16E+02	-3.24E+02	-3.06E+02	-3.10E+02	-2.67E+02	-1.77E+02
f_7	-4.96E+02	-6.55E+02	-6.63E+02	-6.56E+02	-6.58E+02	-6.49E+02	-6.44E+02	-6.42E+02	-6.43E+02	-6.46E+02	-6.49E+02
f_8	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02	-6.79E+02
f_9	-5.67E+02	-5.68E+02	-5.67E+02	-5.67E+02	-5.67E+02	-5.67E+02	-5.68E+02	-5.67E+02	-5.67E+02	-5.66E+02	-5.67E+02
f_{10}	1.07E+03	-9.42E+01	3.35E+02	1.64E+01	1.81E+02	2.89E+02	1.52E+02	3.83E+02	5.26E+02	3.06E+02	2.43E+02
f_{11}	-1.49E+02	-2.26E+02	-2.45E+02	-2.48E+02	-2.48E+02	-2.52E+02	-2.59E+02	-2.65E+02	-2.47E+02	-2.55E+02	-2.42E+02
f_{12}	-1.84E+01	-9.70E+01	-1.32E+02	-1.31E+02	-1.14E+02	-1.24E+02	-9.70E+01	-1.04E+02	-1.04E+02	-9.50E+01	-9.03E+01
f_{13}	1.88E+02	8.74E+01	1.18E+02	9.26E+01	1.01E+02	1.06E+02	7.93E+01	9.54E+01	8.86E+01	8.46E+01	9.73E+01
f_{14}	3.25E+03	3.10E+03	2.85E+03	2.97E+03	2.73E+03	2.81E+03	2.92E+03	2.91E+03	2.73E+03	2.78E+03	2.76E+03
f_{15}	4.38E+03	4.38E+03	4.43E+03	4.45E+03	4.37E+03	4.32E+03	4.31E+03	4.32E+03	4.51E+03	4.42E+03	4.44E+03
f_{16}	2.01E+02	2.01E+02	2.01E+02	2.01E+02	2.01E+02	2.01E+02	2.02E+02	2.02E+02	2.01E+02	2.01E+02	2.02E+02
f_{17}	5.43E+02	5.01E+02	5.01E+02	5.12E+02	5.05E+02	5.06E+02	5.03E+02	5.01E+02	5.02E+02	5.05E+02	4.88E+02
f_{18}	6.64E+02	5.95E+02	6.14E+02	5.92E+02	6.09E+02	5.96E+02	6.12E+02	6.11E+02	6.12E+02	6.18E+02	6.07E+02
f_{19}	7.62E+02	6.36E+02	5.97E+02	5.87E+02	6.05E+02	6.39E+02	6.63E+02	6.29E+02	6.63E+02	6.70E+02	6.88E+02
f_{20}	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02	6.14E+02
f_{21}	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03	1.00E+03
f_{22}	5.48E+03	4.65E+03	4.56E+03	4.51E+03	4.66E+03	4.54E+03	4.61E+03	4.70E+03	4.54E+03	4.61E+03	4.47E+03
f_{23}	6.53E+03	6.49E+03	6.49E+03	6.43E+03	6.26E+03	6.59E+03	6.34E+03	6.14E+03	6.11E+03	6.19E+03	6.14E+03
f_{24}	1.32E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03	1.31E+03
f_{25}	1.47E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03	1.44E+03
f_{26}	1.59E+03	1.58E+03	1.58E+03	1.57E+03	1.58E+03	1.58E+03	1.58E+03	1.57E+03	1.57E+03	1.57E+03	1.58E+03
f_{27}	2.62E+03	2.54E+03	2.51E+03	2.51E+03	2.51E+03	2.48E+03	2.51E+03	2.51E+03	2.54E+03	2.51E+03	2.53E+03
f_{28}	3.95E+03	3.52E+03	3.65E+03	3.62E+03	3.53E+03	3.67E+03	3.60E+03	3.64E+03	3.62E+03	3.92E+03	3.79E+03

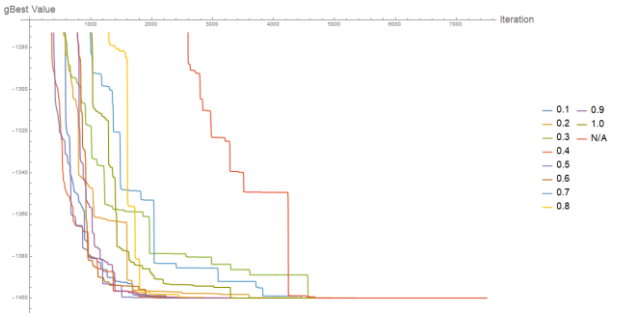


Figure 10. Mean gBest value history – f_1 – dim = 30

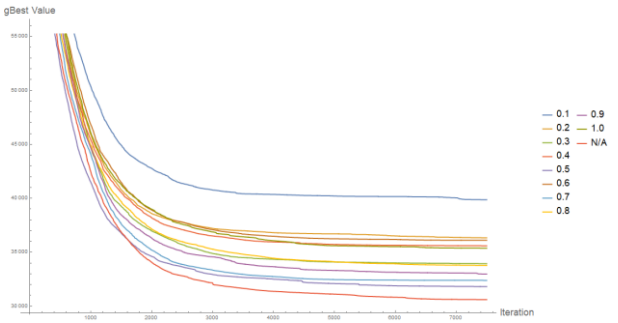


Figure 11. Mean gBest value history – f_4 – dim = 30

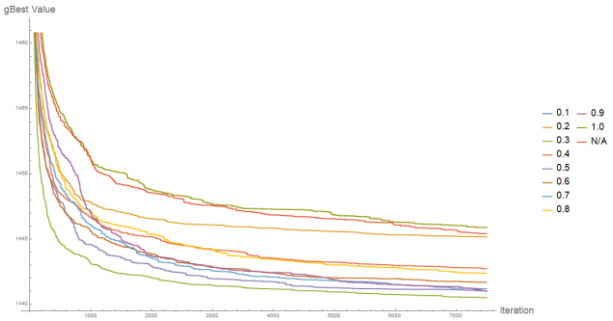


Figure 12. Mean gBest value history – f_{25} – dim = 30

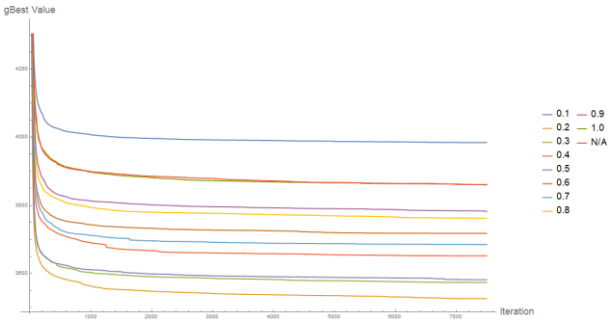


Figure 13. Mean gBest value history – f_{30} – dim = 30

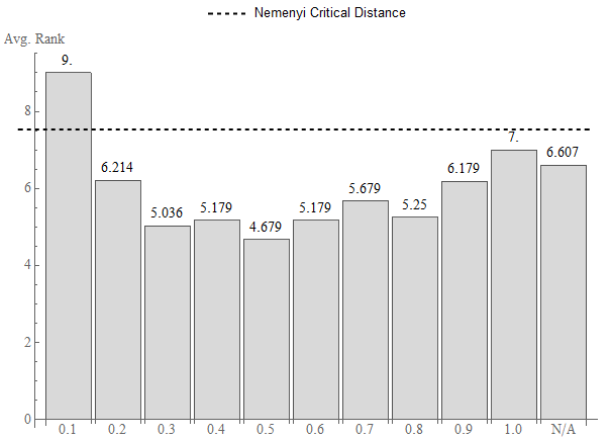


Figure 14. Friedman rank with Nemenyi critical distance dim = 30

RESULTS DISCUSSION

According to the statistical data presented in Fig. 9, the settings of v_{max} = 0.3 and 0.6 are best performing for dim = 10. The value of 0.1 does not perform well and so does the algorithm when velocity clamping is not used at all. Other settings do not perform significantly worse (according to statistical test).

As is displayed in Figures 1 – 8, the overall shape of the convergence history does not differ significantly regardless the velocity clamping value. A similar trend can be observed in Figures 10 – 13 for the dim = 30. This might also mean that the v_{max} setting does not directly affect the diversity of the swarm but a future study will be needed to support such conclusion.

According to statistics (Fig. 14) in higher dimensions (dim = 30), the value 0.1 is the only significantly outperformed setting by all other. However, the PSO without velocity clamping is no longer underperforming.

For researchers interested in comparing other approaches with our study the median result values are presented in Tables 1 and 2.

CONCLUSION

In this study, we tested the performance of standard PSO algorithm with different setting of maximal velocity and without velocity clamping on an extensive benchmark suite containing 28 different functions.

The observations can be summarized as follows:

- A very low value of v_{max} (0.1) is not advisable.
- In lower dimension, the velocity clamping must not be omitted. However, in higher dimensions (dim = 30), the performance of PSO with velocity clamping (v_{max} > 0.1) and PSO without the velocity clamping is comparable.
- Fine tuning of v_{max} value does not seem to be necessary as the performance does not differ with statistical significance for values over 0.1.
- The popular (according to literature) setting v_{max} = 0.2 does not seem to be favorable across this extensive benchmark suite
- The overall convergence behavior seems not affected by the maximal velocity value.
- It seems that the most favorable values for the maximum velocity are in the region from 0.3 to 0.6. However, more data is needed to support this claim.

Given the initial results and findings presented in this study, we will continue to study the mechanics of velocity clamping for a better understanding of this issue. Future study will focus on the relation between v_{max} setting and dimensionality of the problem. In addition, the relation between favorable v_{max} setting and fitness landscape parameters (such as ruggedness, etc.) will be closely investigated. In addition, the relation of v_{max} and other adjustable parameters of PSO will be taken into consideration.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

Kennedy J. and Eberhart R., “Particle swarm optimization,” in Proceedings of the IEEE International Conference on Neural Networks, 1995, pp. 1942–1948.

Kennedy J., “The particle swarm: social adaptation of knowledge,” in Proceedings of the IEEE International Conference on Evolutionary Computation, 1997, pp. 303–308.

Liang JJ, Qu B-Y., Suganthan PN, Hernández-Díaz AG (2013) Problem Definitions and Evaluation Criteria for the CEC 2013 Special Session and Competition on Real-Parameter Optimization , Technical Report 2012, Computational Intelligence Labora-tory, Zhengzhou University, Zhengzhou China and Technical Report, Nanyang Tech-nological University, Singapore.

Kennedy J. and Eberhart R., “Particle swarm optimization,” in Proceedings of the IEEE International Conference on Neural Networks, 1995, pp. 1942–1948.

Nickabadi A., Ebadzadeh M. M., Safabakhsh R., A novel particle swarm optimization algorithm with adaptive inertia weight, Applied Soft Computing, Volume 11, Issue 4, June 2011, Pages 3658-3670, ISSN 1568-4946

Shi Y. and Eberhart R. C., “A modified particle swarm optimizer,” in Proceedings of the IEEE International Conference on Evolutionary Computation (IEEE World Congress on Computational Intelligence), 1998a, pp. 69–73.I. S.

Shi Y., Eberhart R.C., Parameter selection in particle swarm optimization, in: Proceedings of the Seventh Annual Conference on Evolutionary Programming, New York, USA, 1998b, pp. 591–600.

Shi Y., Eberhart R.C., Empirical study of particle swarm optimization, in: Proceedings of the IEEE Congress on Evolutionary Computation, IEEE Press, 1999, pp. 1945–1950.

AUTHOR BIOGRAPHIES

MICHAL PLUHACEK was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011 and Ph.D. in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. He now works as a researcher at the same university. His email address is: pluhacek@utb.cz

ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2004, Ph.D. degree in Technical Cybernetics in 2008 and Assoc. prof. in 2013 (Informatics). He is now an Assoc. prof. at the same university (research and courses in: Evolutionary Computation, Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). His email address is: senkerik@utb.cz

ADAM VIKTORIN was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the field of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@utb.cz

TOMAS KADAVY was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2016. He is studying his Ph.D. at the same university and the fields of his studies are: Artificial intelligence and evolutionary algorithms. His email address is: kadavy@utb.cz

Van Den Bergh, F., and Engelbrecht, A. P. (2006). A study of particle swarm optimization particle trajectories. Information sciences, 176(8), 937-971.

Volna, E., and Kotyrba, M. (2014) A comparative study to evolutionary algorithms, In Proceedings 28th European Conference on Modelling and Simulation, ECMS 2014, Brescia, Italy, 2014, pp. 340-345.

TUNING OF THE BISON ALGORITHM CONTROL PARAMETERS

Anezka Kazikova, Michal Pluhacek and Roman Senkerik
Faculty of Applied Informatics
Tomas Bata University in Zlin
T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
E-mail: {kazikova, pluhacek, senkerik}@utb.cz

KEYWORDS

Bison Algorithm, parameter study, swarm algorithm, optimization, metaheuristic.

ABSTRACT

This paper studies the dependency of the Bison Algorithm performance on the control parameter configuration. The Bison Algorithm is a new swarm algorithm based on the protection mechanisms of bison herds. It operates with two groups: the exploiting swarming group and the exploring running group. Even though that adjusting the group size parameters affects both the time requirements and the performance of the algorithm, there was no investigation of the parameter settings carried out yet. This paper describes the Bison Algorithm and then investigates the control parameters for a better understanding of their meaning and influence on the overall optimization process.

INTRODUCTION

A recent trend in the modern optimization is to exploit known biological findings. This is done by simulating instances of typical nature optimization patterns: the Darwinian evolution (Bäck 1996), genomes (Goldberg 1989) or even animal behavior processes (Yang and Deb 2013).

The swarm algorithms model the swarm intelligence – a collective behavior of large animal groups, that can make intelligent decisions without an actual leadership. Based on partial information only, the swarms manage to optimize real-life problems like finding enough food supplies, shortening their travel distance, developing an ultimate hunting strategy, reproducing or escaping predators. The simulations such as the Bees Algorithm (Rajasekhar et al. 2017), Ant Colony Optimization (Dorigo and Stütle 2004; Duan and Ying 2009), Grey Wolf Optimizer (Mirjalili et al. 2014) and Cuckoo Search (Yang and Deb 2009) have been already successfully used in the optimization field.

The Bison Algorithm is a recent swarm optimization algorithm (Kazikova et al. 2017). It simulates two of the most typical bison behavior: protecting the weak by forming a circle around them and the advantages of a running herd.

Various extensions of the Bison Algorithm have been developed since (Kazikova et al. 2018). However, there has not been done any detailed parameter study, having all the previous papers on the subject relying on an early parameter test.

This paper studies the Bison Algorithm with various parameter configurations. It simulates the movement differences and the impact on solution quality. In Section 1 of the paper, the Bison Algorithm is outlined. Section 2 describes the methods used in the parameter study. The outcomes of the experiments are presented in Section 3 and discussed in Section 4. Finally, the meanings of the findings are considered in Section 5.

BISON ALGORITHM

The Bison Algorithm was inspired by the most typical behavior patterns of bison. Bison have two distinctive protective mechanisms: forming a circle around the weak ones and almost inexhaustible running manners (Berman 2008). The algorithm implements both, as described in pseudocode Algorithm 1.

Algorithm 1: Bison Algorithm Pseudocode

```
Initialization:
  Obj. function: f(x) = (x1,...,xd)
  Generate swarming group randomly
  Generate run. group around x_best
  Generate run direction vector Eq.(4)
For every iteration do
  Compute the swarming center Eq.(1,2)
  For every swarmer do
    Compute position candidate Eq.(3)
    if f(x_new) < f(x_old) then move to x_new
  End
  Adjust run direction vector r Eq.(5)
  For every runner do
    Move in run direction vector Eq.(6)
  End
  Copy success. runners to swarmers
  Sort swarming group by f(x) value
End for
```

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Definition of the Bison Algorithm

The algorithm is defined by two groups, each simulating different behavior. The first group models the swarming pattern. It starts by computing the center of the fittest swarming individuals (sorted by the objective function value). This paper operates with the ranked center computation. This approach sorts the fittest individuals, giving them corresponding weights according to their solution quality (Eq. 1) and then computes the center concerning their weights (Eq. 2). A new position candidate is computed (Eq. 3) and used only if it improves the objective function value of the swarmer.

$$weight = (10, 20, \dots, 10 * s) \quad (1)$$

$$c = \sum_{i=1}^{s} \frac{weight_i * \boldsymbol{x_i}}{\sum_{j=1}^{s} weight_i} \quad (2)$$

$$x_{new} = x_{old} + (c - x_{old}) \cdot rand(0, v) \quad (3)$$

Where s is the elite group size parameter, defining the number of the fittest bison to compute the center of, c is the center, v is the overstep parameter, x_i is the i[th] solution, x_{old} is the current solution and x_{new} is the new solution candidate.

The second group simulates the running behavior. During the initialization of the algorithm, a random run direction vector r is generated (Eq. 4) and then slightly altered after every iteration (Eq. 5). The running movement always happens in the run direction vector (Eq. 6), not concerning the quality of the solution. This means that unlike the swarmers the runners move even when it threatens their quality.

$$r = rand\left(\frac{ub-lb}{45}, \frac{ub-lb}{15}\right) \quad (4)$$

$$r = r \cdot rand(0.9, 1.1) \quad (5)$$

$$x_{new} = x_{old} + r \quad (6)$$

Where ub and lb are the upper and lower boundaries of the search space, r is the run direction vector and x_{old}, x_{new} are the original and the new solutions.

Since the Bison Algorithm considers the search space a hypersphere, the running movement can be based on the run direction vector throughout the whole optimization process. Whenever the individuals run over the boundaries, they appear on the other side of the exceeded dimensions, exploring the search space thoroughly.

The control parameters of the Bison Algorithm are described in Table 1. The recommended value of the overstep parameter 3.5 means that the swarmers can exceed the center 2.5 times in accordance with Eq. 3. The size of the running group is defined with the help of the swarm group size (Eq. 7).

$$run\ group\ size = NP - swarm\ size \quad (7)$$

Where NP is the population number and *swarm size* is the swarm group size parameter.

Table 1: Bison Algorithm Control Parameters

Parameter	Description
Population NP	Population size
Elite group size	Number of the fittest solutions used for center computation
Swarm group size	Number of the swarming group solutions
Overstep v	Maximum length of the swarming movement in relation to the center 0 = no movement 1 = to the center Recommended: 3.5 - 4.1

METHODS

Two parameter scenarios were examined. The first scenario called the *complete set* consists of 12 configurations described in Table 2 in the form, that first notes the swarm group size and then the elite group size. For example, S40E20 means a population of 40 swarmers and 20 elite individuals. The *40S set* examines only the configurations with 40 swarming individuals. Other parameters were set to: $NP = 50$, $v = 3.5$.

Table 2: Tested Sets of Parameter Configurations

Complete set	S20E1, S20E10, S20E20, S30E1, S30E10, S30E20, S30E30, S40E1, S40E10, S40E20, S40E30, S40E40
40S set	S40E1, S40E10, S40E20, S40E30, S40E40

This paper investigates the influence of the control parameter configuration on 1) movement patterns, 2) performance of the algorithm, 3) computation time.

The movement is presented by a 2D simulation of the population distribution on the Rastrigin's Function (Fig. 1). The included models are: S40E1 with one bison being the sole center, S40E40 with the center computed from all the swarmers and S20E10 with many runners.

For the performance experiments, we used the first 15 functions of IEEE CEC 2017 benchmark (Awad et al. 2016), on 30 independent runs, each consisting of $10\,000 \cdot dimension$ evaluations. The solutions were compared with the Friedman rank test ($p<0.05$) in Fig. 2 and 3 for the complete set and in Fig. 4, 5 for the 40S set. Table 3 presents the Friedman P-Values. Table 4 sum the results of the two most successful configurations Wilcoxon rank-sum tests ($\alpha=0.05$) comparing the in 10 and 30 dimensions. Table 5 shows the mean solution and standard deviation of the two approaches. The mean convergences of the 40S set are shown in Fig. 6, 7 and 8.

The time requirements were compared by the Friedman rank test in Fig. 9 and 10. Table 6 shows the mean time needed for solving 10-dimensional problems with the 40S testing set. The significantly better time results according to the Wilcoxon rank-sum test are bold.

RESULTS

Movement Patterns

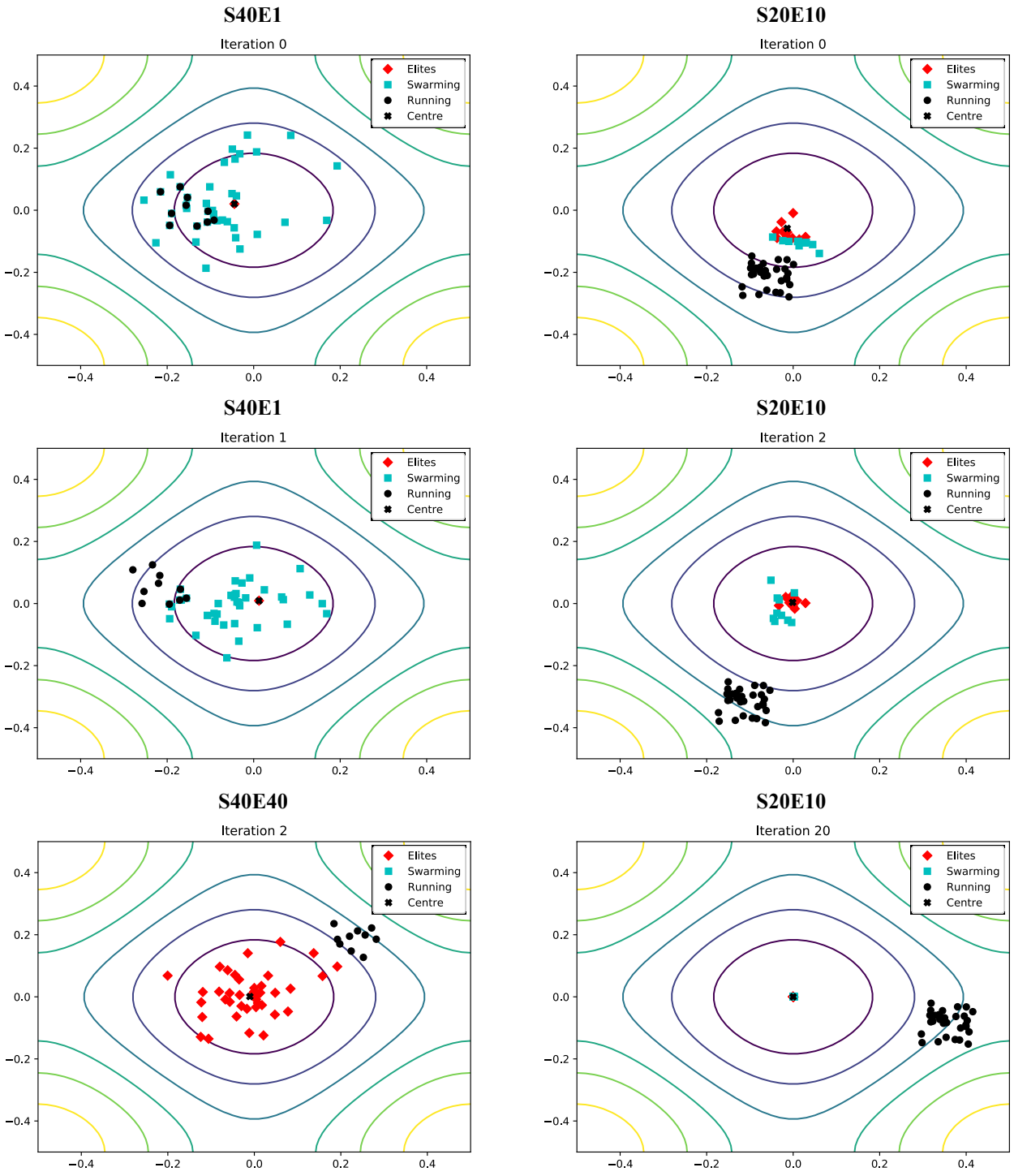


Figure 1: 2D Movement of Parameter Configurations S40E1, S40E40, S20E10

Performance Experiment Results

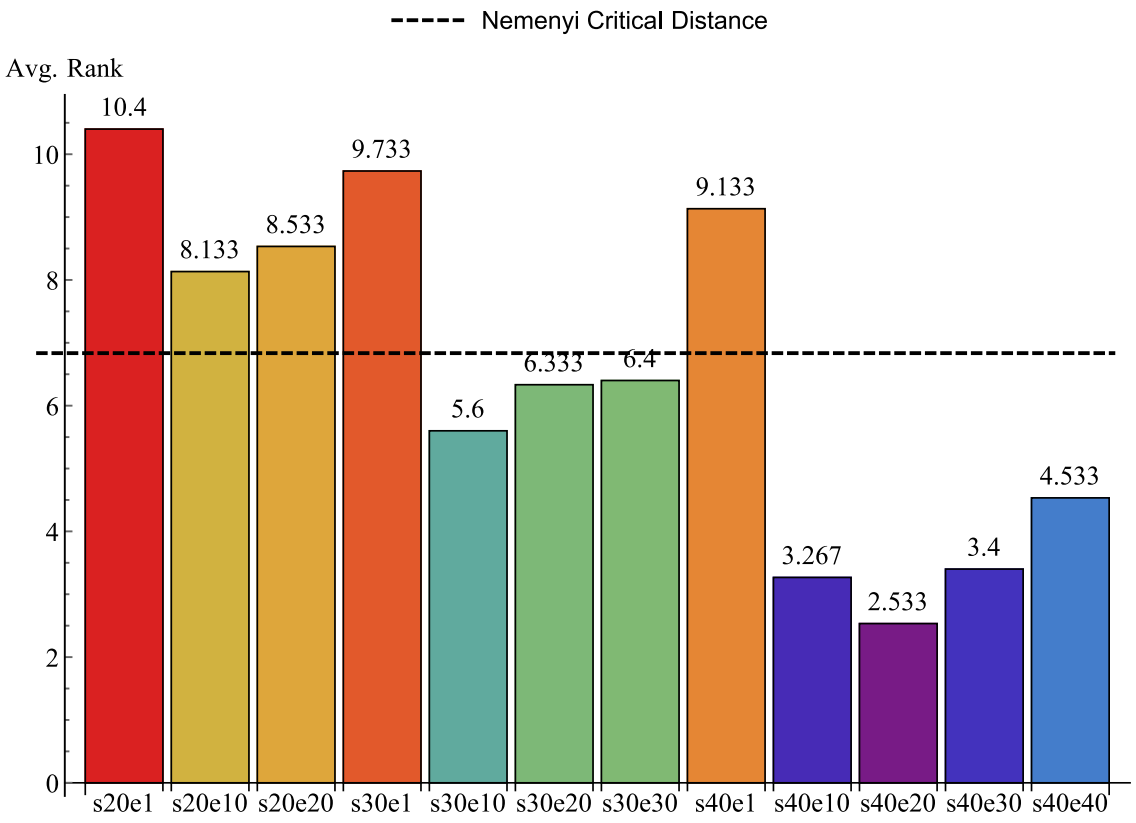


Figure 2: Friedman Rank Test on Complete Set of Parameter Configurations in 10D

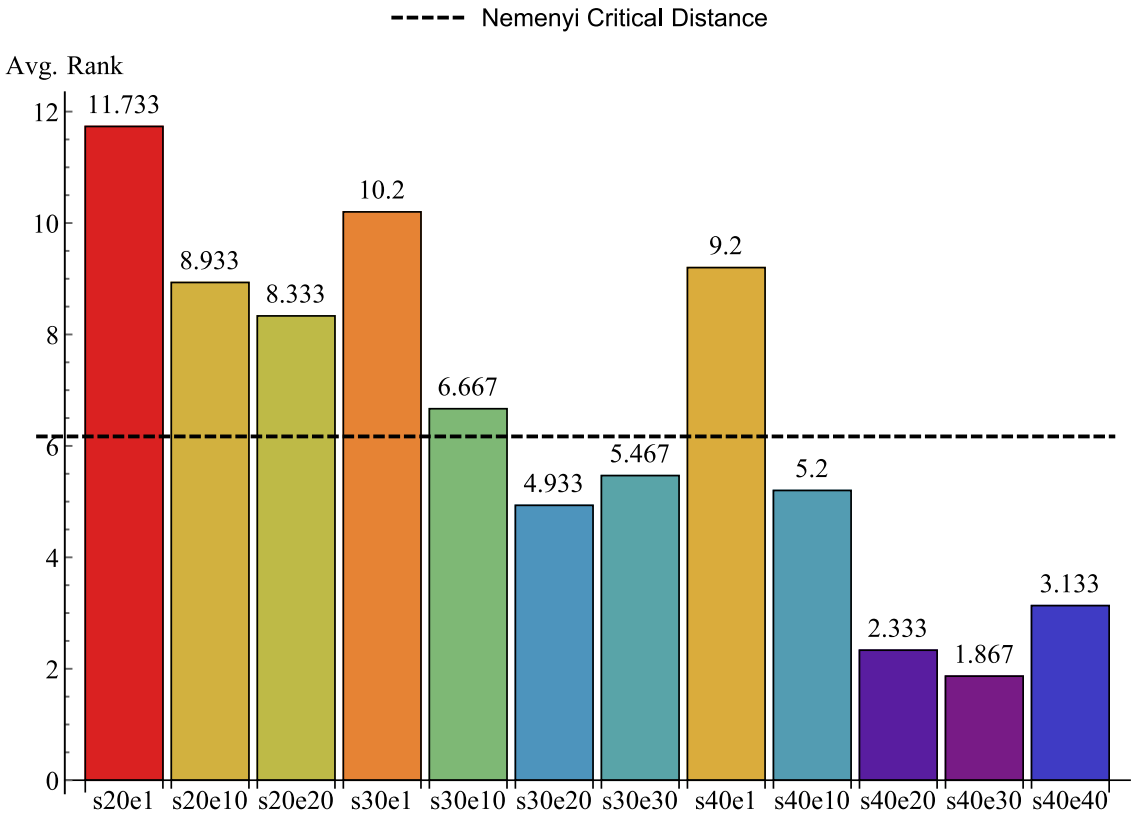


Figure 3: Friedman Rank Test on Complete Set of Parameter Configurations in 30D

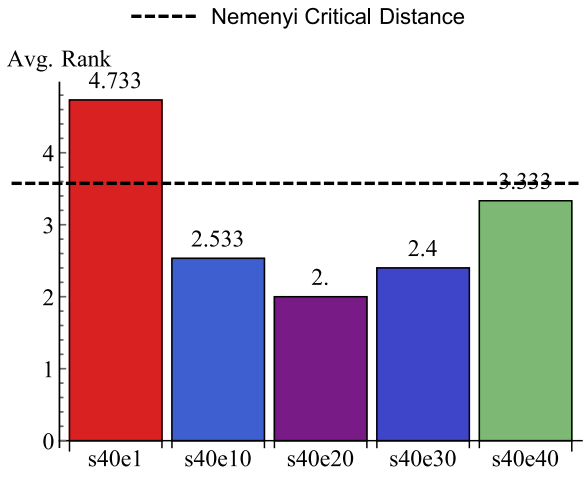


Figure 4: Friedman Rank Test on S40 Set 10D

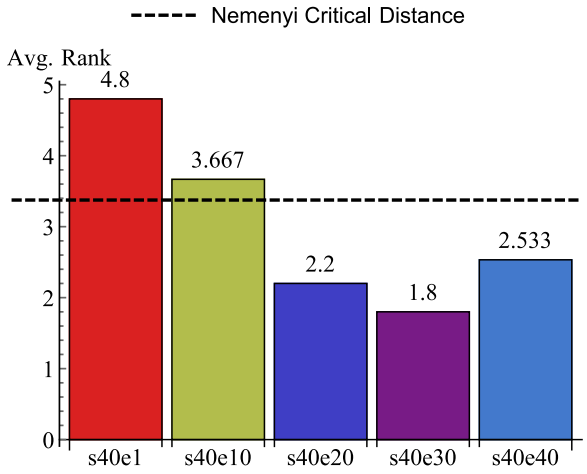


Figure 5: Friedman Rank Test on S40 Set 30D

Table 3: P-Values of the Friedman Rank Tests

	Complete set	S40 set
10 dimensions	2.26E-22	2.79E-07
30 dimensions	1.27E-46	1.45E-10
Time experiments in 10 dimensions	3.91E-74	1.87E-22

Table 4: Significant Wins of the 2 Most Successful Parameter Configurations S40E20 and S40E30 According to the Wilcoxon Rank-Sum Test (α=0.05)

Dimensionality	None	S40E20	S40E30
10 dimensions	11	3	1
30 dimensions	10	2	3

Table 5: Mean Solutions and Standard Deviations of S40E20 and S40E30 in 10 Dimensions

	S40E20 = set 1		S40E30 = set 2		Win
	avg	*std*	*avg*	*std*	*avg*
f_1	7.29E+02	1.12E+03	5.55E+02	8.29E+02	-
f_2	5.88E-02	2.38E-01	3.33E-02	1.83E-01	-
f_3	0.00E+00	0.00E+00	0.00E+00	0.00E+00	-
f_4	**2.70E-01**	2.13E-01	4.23E-01	9.93E-02	1
f_5	9.15E+00	7.86E+00	1.12E+01	8.81E+00	-
f_6	4.94E-05	2.65E-04	3.24E-06	1.23E-05	-
f_7	2.50E+01	8.27E+00	2.37E+01	8.02E+00	-
f_8	7.21E+00	6.10E+00	7.89E+00	7.63E+00	-
f_9	1.29E-01	4.53E-01	**0.00E+00**	0.00E+00	2
f_{10}	1.01E+03	4.32E+02	1.12E+03	2.52E+02	-
f_{11}	2.83E+00	2.57E+00	2.94E+00	2.01E+00	-
f_{12}	1.06E+04	1.02E+04	8.21E+03	7.32E+03	-
f_{13}	**3.21E+03**	3.06E+03	5.80E+03	4.24E+03	1
f_{14}	**3.41E+01**	6.54E+00	3.77E+01	5.47E+00	1
f_{15}	2.77E+01	1.54E+01	3.51E+01	2.04E+01	-

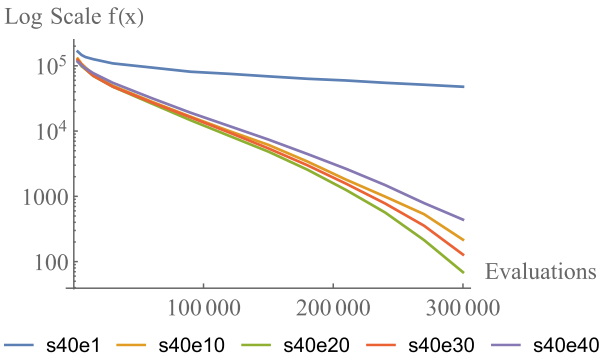


Figure 6: Mean Convergence of 40S Set on F3 in 30D

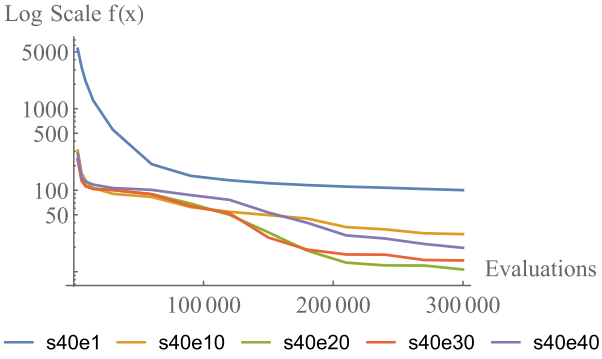


Figure 7: Mean Convergence of 40S Set on F4 in 30D

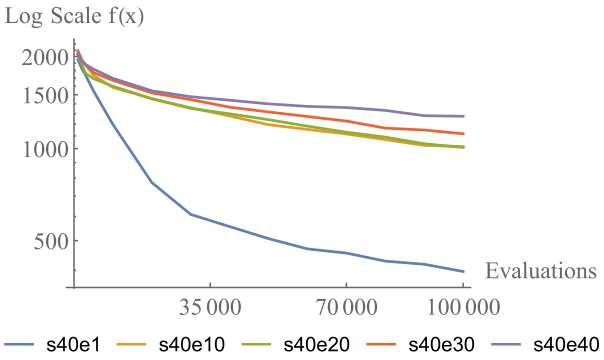


Figure 8: Mean Convergence of S40 Set on F10 in 10D

Computation Time Experiment Results

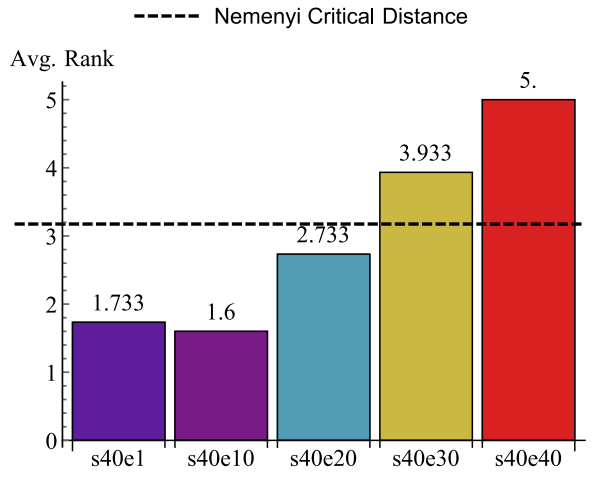


Figure 9: Friedman Rank Test for the Computation Time on S40 Set in 10 Dimensions

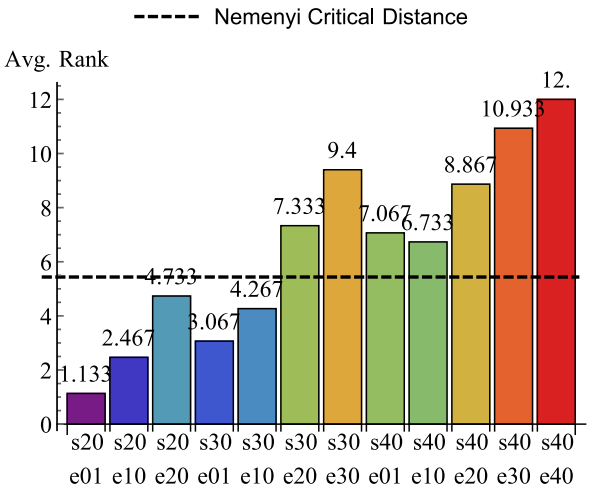


Figure 10: Friedman Rank Test for Computation Time on Complete Set of Parameter Configurations in 10 Dimensions

Table 6: Mean Time Needed for 100 000 Evaluations of 10D Functions in Seconds

	S40E1	**S40E10**	**S40E20**	**S40E30**	**S40E40**
f_1	5.07	5.02	5.25	5.59	5.75
f_2	5.50	5.42	5.55	5.68	5.83
f_3	5.23	5.10	5.33	5.57	5.81
f_4	5.30	5.13	5.33	5.57	5.79
f_5	5.25	**5.14**	5.38	5.59	5.82
f_6	5.21	5.27	5.53	5.80	6.04
f_7	**4.98**	5.18	5.39	5.64	5.89
f_8	**5.04**	5.18	5.37	5.61	6.11
f_9	**4.95**	5.16	5.37	5.62	5.97
f_{10}	5.75	5.64	5.77	6.02	6.28
f_{11}	**5.13**	5.20	5.39	5.61	5.82
f_{12}	5.20	5.21	5.42	5.63	5.89
f_{13}	5.63	**5.40**	5.51	5.76	5.97
f_{14}	5.84	**5.58**	5.70	5.93	6.08
f_{15}	5.66	5.78	5.58	5.78	5.99

Best times according to the Wilcoxon rank-sum test:

None	**S40E1**	**S40E10**	**S40E20**	**S40E30**	**S40E40**
8	4	3	0	0	0

DISCUSSION

The results of the performance experiments on the complete testing set indicated the superiority of the configurations with 40 swarming members except for the one with 1 elite bison. However, even the worst ranked S40E1 configuration demonstrated a promising convergence when solving F10 function (Fig. 8) in accordance with the No Free Lunch Theorem (Yang 2012).

A closer investigation of the S40 test set proved the efficiency of the S40E30 and S40E20. Comparing these two configurations proved, that the results were mostly comparable, S40E30 being slightly more successful in 30 dimensions, while S40E20 in 10 dimensions.

The time regarding experiments implied that the elite group size parameter might have a direct influence on the computation time as in most of the cases, the lower the elite group size parameter was, the better time was achieved. Based on these results, there seems to be a conflict between the time and performance requirements.

Even though the difference between the times shown in Table 6 might not seem very wide, it is important to remember, that this experiment included solving 10-dimensional problems only. In higher dimensions, the computation time tends to lengthen just as much as the difference between the time requirements of the parameter configurations.

CONCLUSION

This paper provided interesting findings regarding the parameter configuration of the Bison Algorithm. In most of the functions, the algorithm performed best with the parameter configuration of 40 swarming individuals and 20 or 30 elite members. However, the time experiments preferred the lower amount of both the swarming and the elite individuals.

Since in the real-time optimization are usually requirements for both the quality and the time of the optimization, the obtained results might be useful in the application of the Bison Algorithm on solving real-time problems. For a general optimization, we suggest the S40E20 configuration, as it provided faster evaluations while giving comparable results to the S40E30 configuration.

Since in the convergence analysis even the worst ranked configuration showed remarkable progress in comparison to the others, an adaptive parameter approach of the Bison Algorithm might be considered an exploitable extension of the algorithm and a possible subject of our future research.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

Awad, N. H., Ali, M. Z., Liang, J. J., Qu, B. Y. and P.N. Suganthan. 2016. "Problem Definitions and Evaluation Criteria for the CEC 2017 Special Session and Competition on Single Objective Bound Constrained Real-Parameter Numerical Optimization, Technical Report". Nanyang Technological University, Singapore.

Bäck, T. 1996. "Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms". Oxford University Press, Oxford, UK.

Berman, R. 2008. "American Bison (Nature Watch)". Lerner Publications, Minneapolis.

Dorigo. M. and T. Stütle. 2004. "Ant Colony Optimization". MIT Press, Cambridge, MA.

Duan, Y. and S. Ying. 2009. "A Particle Swarm Optimization Algorithm with Ant Search for Solving Traveling Salesman Problem". In: *International Conference on Computational Intelligence and Security*, Beijing, 137-141.

Goldberg, D. 1989. "Genetic Algorithms in Search, Optimization, and Machine Learning". Addison-Wesley, Reading, Mass, USA.

Kazikova, A., Pluhacek, M., Senkerik R. and A. Viktorin. 2017. "Proposal of a New Swarm Optimization Method Inspired in Bison Behavior". In: *Recent Advances in Soft Computing (Mendel 2017)*, Advances in Intelligent System and Computing, Springer, In press.

Kazikova, A., Pluhacek, M., Viktorin, A. and R. Senkerik. 2018. "New running Technique for the Bison Algorithm". In: *Lecture Notes in Artificial Intelligence (ICAISC 2018 Proceedings),* Springer, In press.

Mirjalili, S., Mirjalili, S.M. and A. Lewis. 2014. "*Grey Wolf Optimizer*". Adv. Eng. Softw. 69 (March 2014), 46-61.

Rajasekhar, A., Lynn, N., Das, S. and P.N. Suganthan. 2017. "Computing with the collective intelligence of honey bees–A survey". *Swarm and Evolutionary Computation, 32*, 25-48.

Yang, X.-S. and S. Deb. 2009. "Cuckoo Search via Levy Flights". In: *Proceeding Of World Congress on Nature Biologically Inspired Computing (NaBIC India Dec. 2009)*, IEEE Publications, USA, 210-214.

Yang, X.S. 2012. "Free Lunch or No Free Lunch: That Is Not Just a Question?". Int. J. Artif. Intell. Tools 21 (03)

Yang, X.S. and M. Karamanoglu. 2013. "Swarm Intelligence and Bio-Inspired Computation: An Overview. Swarm Intelligence and Bio-Inspired Computation". 3-23. 10.1016/B978-0-12-405163-8.00001-6.

AUTHOR BIOGRAPHIES

ANEZKA KAZIKOVA received her master's degree in Engineering Informatics from the Tomas Bata University in Zlin in 2015. She is now a Ph.D. student at the same university and researches the swarm algorithms and competitive behavior. Her e-mail is: kazikova@utb.cz. Web page of all the authors can be found at: www.ailab.fai.utb.cz.

MICHAL PLUHACEK received his Ph.D. degree in Information Technologies from the Tomas Bata University in Zlin in 2016. Currently works as a junior researcher at the Regional Research Centre CEBIA-Tech of Tomas Bata University in Zlin. His research focus includes swarm intelligence theory and applications and artificial intelligence in general. His e-mail is: pluhacek@utb.cz.

ROMAN SENKERIK received his Ph.D. degree in Technical Cybernetics from the Tomas Bata University in Zlin in 2008. He is currently an associated professor at the Tomas Bata University in Zlin, Faculty of Applied Informatics. His research interests include interdisciplinary, computational intelligence, optimization, cyber-security, theory of chaos and complexity. His e-mail is: senkerik@utb.cz.

COMPARATIVE STUDY OF THE DISTANCE/IMPROVEMENT BASED SHADE

Adam Viktorin
Roman Senkerik
Michal Pluhacek
Tomas Kadavy
Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{aviktorin, senkerik, pluhacek, kadavy}@utb.cz

KEYWORDS
Differential Evolution, SHADE, Db_SHADE, DIb_SHADE, Parameter Adaptation

ABSTRACT

In this paper, a comparative study of seven variants of the Success-History based Adaptive Differential Evolution (SHADE) algorithm is presented with the aim to show the influence of improvement and distance based parameter adaptation to its performance. Seven algorithms range from improvement only based SHADE through a balanced improvement and distance based DIb_SHADE to distance only based Db_SHADE. The algorithm set is compared on the CEC2015 benchmark set, results are statistically tested, commented and possible future research directions are provided.

INTRODUCTION

Differential Evolution (DE) is a tool for numerical optimization, which was developed by Storn and price in 1995 (Storn and Price 1995). Since then, it was studied and improved in numerous ways and a plethora of new versions appear every year. The research in the area of DE was summarized in (Neri and Tirronen 2010) and (Das et. al. 2016).

A recent trend in improving the performance of the DE is in self-adaptation of its control parameters – scaling factor F, crossover rate CR and population size NP. An algorithm that stands out from the self-adaptive DE crowd is Success-History based Adaptive Differential Evolution (SHADE) by Tanabe and Fukunaga (Tanabe and Fukunaga 2013). This algorithm is based on the JADE (Zhang and Sanderson 2009) and proposes a new way of storing of the successful scaling factor and crossover rate values in respective memories, which are periodically updated after each generation. SHADE is considered successful because it formed a basis for the last four CEC competition winners – CEC2014 L-SHADE (Tanabe and Fukunaga 2014), CEC2015 SPS-L-SHADE-EIG (Guo et. al. 2015), CEC2016 EpSin_LSHADE (Awad et. al. 2016) and CEC2017 jSO (Brest et. al. 2017).

One of the issues with SHADE algorithm is, that it suffers from premature convergence in higher dimensional spaces, and therefore a novel variant which addresses this issue has been proposed in (Viktorin et. al. 2017). The algorithm titled Db_SHADE uses the information about the distance between original and trial vectors to evaluate the weights for the parameter adaptation scheme and therefore promotes exploration of the decision space rather than its exploitation. It was shown, that this approach is beneficial for the performance of the algorithm on multi-modal complex functions.

In this paper, the weighted approach for distance and improvement based parameter adaptation is presented (DIb_SHADE) and seven SHADE algorithm variants with various preferences (improvement or distance) are compared on the CEC2015 benchmark set. The results provide an interesting insight into the performance of the algorithms and show future research direction possibilities.

The rest of the paper is structured as follows: The next section describes the basics of DE, SHADE, Db_SHADE and DIb_SHADE algorithms, the section that follows describes the experimental settings, second to the last section provides the results with discussion and the last section is devoted to concluding remarks.

FROM DE TO DIB_SHADE

In order to describe the Success-History based Adaptive Differential Evolution algorithm (SHADE) and its Distance based variants (Db_SHADE and DIb_SHADE), it is important to start from the canonical Differential Evolution (DE) by Storn and Price (Storn and Price 1995).

The canonical 1995 DE is based on the idea of evolution from a randomly generated set of solutions of the optimization task called population $\boldsymbol{P}$, which has a preset size of NP. Each individual (solution) in the population consists of a vector $\boldsymbol{x}$ of length D (each vector component corresponds to one attribute of the optimized task) and objective function value $f(\boldsymbol{x})$, which mirrors the quality of the solution. The number of

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

optimized attributes D is often referred to as the dimensionality of the problem and such generated population $\boldsymbol{P}$, represent the first generation of solutions.

The individuals in the population are combined in an evolutionary manner in order to create improved offspring for the next generation. This process is repeated until the stopping criterion is met (either the maximum number of generations, or the maximum number of objective function evaluations, or the population diversity lower limit, or overall computational time), creating a chain of subsequent generations, where each following generation consists of better solutions than those in previous generations – a phenomenon called elitism.

The combination of individuals in the population consists of three main steps: Mutation, crossover and selection.

In the mutation, attribute vectors of selected individuals are combined in simple vector operations to produce a mutated vector $\boldsymbol{v}$. This operation uses a control parameter – scaling factor F. In the crossover step, a trial vector $\boldsymbol{u}$ is created by selection of attributes either from mutated vector $\boldsymbol{v}$ or the original vector $\boldsymbol{x}$ based on the crossover probability given by a control parameter – crossover rate CR. And finally, in the selection, the quality $f(\boldsymbol{u})$ of a trial vector is evaluated by an objective function and compared to the quality $f(\boldsymbol{x})$ of the original vector and the better one is placed into the next generation.

From the basic description of the DE algorithm, it can be seen, that there are three control parameters, which have to be set by the user – population size NP, scaling factor F and crossover rate CR. It was shown in (Gämperle et. al. 2002) and (Liu and Lampinen 2002), that the setting of these parameters is crucial for the performance of DE. Fine-tuning of the control parameter values is a time-consuming task and therefore, many state-of-the-art DE variants use self-adaptation in order to avoid this cumbersome task. Which is also a case of SHADE algorithm proposed by (Tanabe and Fukunaga 2013) and since it is used in this paper, the algorithm is described in more detail in the next section along with the novel distance based parameter adaptation.

1.1 SHADE

As aforementioned, SHADE algorithm was proposed with a self-adaptive mechanism of some of its control parameters in order to avoid their fine-tuning. Control parameters in question are scaling factor F and crossover rate CR. It is fair to mention, that SHADE algorithm is based on Zhang and Sanderson's JADE (Zhang and Sanderson 2009) and shares a lot of its mechanisms. The main difference is in the historical memories $\boldsymbol{M}_F$ and $\boldsymbol{M}_{CR}$ for successful scaling factor and crossover rate values with their update mechanism.

Following subsections describe individual steps of the SHADE algorithm: Initialization, mutation, crossover, selection and historical memory update.

Initialization

The initial population $\boldsymbol{P}$ is generated randomly and for that matter, a Pseudo-Random Number Generator (PRNG) with uniform distribution is used. Solution vectors $\boldsymbol{x}$ are generated according to the limits of solution space – *lower* and *upper* bounds (1).

$$\boldsymbol{x}_{j,i} = U\left[lower_j, upper_j\right], \tag{1}$$

where i (i = 1, …, NP) is the individual index and j (j = 1, …, D) is the attribute index. The dimensionality of the problem is represented by D, and NP stands for the population size.

Historical memories are preset to contain only 0.5 values for both, scaling factor and crossover rate parameters (2).

$$M_{CR,i} = M_{F,i} = 0.5 \text{ for } i = 1, \dots, H, \tag{2}$$

where H is a user-defined size of historical memories.
Also, the external archive of inferior solutions $\boldsymbol{A}$ has to be initialized. Because of no previous inferior solutions, it is initialized empty, $\boldsymbol{A} = \emptyset$. And index k for historical memory updates is initialized to 1.

The following steps are repeated over the generations until the stopping criterion is met.

Mutation

Mutation strategy "current-to-$p$best/1" was introduced in (Zhang and Sanderson 2009) and it combines four mutually different vectors in a creation of the mutated vector $\boldsymbol{v}$. Therefore, $\boldsymbol{x}_{pbest} \neq \boldsymbol{x}_{r1} \neq \boldsymbol{x}_{r2} \neq \boldsymbol{x}_i$ (3).

$$\boldsymbol{v}_i = \boldsymbol{x}_i + F_i\left(\boldsymbol{x}_{pbest} - \boldsymbol{x}_i\right) + F_i(\boldsymbol{x}_{r1} - \boldsymbol{x}_{r2}), \tag{3}$$

where $\boldsymbol{x}_{pbest}$ is randomly selected individual from the best $NP \times p$ individuals in the current population. The p value is randomly generated for each mutation by PRNG with uniform distribution from the range [p_{min}, 0.2] and $p_{min} = 2/NP$. Vector $\boldsymbol{x}_{r1}$ is randomly selected from the current population $\boldsymbol{P}$. Vector $\boldsymbol{x}_{r2}$ is randomly selected from the union of the current population $\boldsymbol{P}$ and external archive $\boldsymbol{A}$. The scaling factor value F_i is given by (4).

$$F_i = C\left[M_{F,r}, 0.1\right], \tag{4}$$

where $M_{F,r}$ is a randomly selected value (index r is generated by PRNG from the range 1 to H) from $\boldsymbol{M}_F$ memory and C stands for Cauchy distribution.

Therefore the F_i value is generated from the Cauchy distribution with location parameter value $M_{F,r}$ and scale parameter value of 0.1. If the generated value F_i higher than 1, it is truncated to 1 and if it is F_i less or equal to 0, it is generated again by (4).

Crossover

In the crossover step, trial vector $\boldsymbol{u}$ is created from the mutated $\boldsymbol{v}$ and original $\boldsymbol{x}$ vectors. For each vector component, a PRNG with uniform distribution is used to generate a random value. If this random value is less or equal to given crossover rate value CR_i, current vector component will be taken from a trial vector, otherwise, it will be taken from the original vector (5). There is also a safety measure, which ensures, that at least one vector component will be taken from the trial vector. This is given by a randomly generated component index j_{rand}.

$$u_{j,i} = \begin{cases} v_{j,i} & \text{if } U[0,1] \leq CR_i \text{ or } j = j_{rand} \\ x_{j,i} & \text{otherwise} \end{cases}. \quad (5)$$

The crossover rate value CR_i is generated from a Gaussian distribution with a mean parameter value $M_{CR,r}$ selected from the crossover rate historical memory $\boldsymbol{M}_{CR}$ by the same index r as in the scaling factor case and standard deviation value of 0.1 (6).

$$CR_i = N\left[M_{CR,r}, 0.1\right]. \quad (6)$$

When the generated CR_i value is less than 0, it is replaced by 0 and when it is greater than 1, it is replaced by 1.

Selection

The selection step ensures, that the optimization will progress towards better solutions because it allows only individuals of better or at least equal objective function value to proceed into the next generation $G+1$ (7).

$$\boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} & \text{if } f\left(\boldsymbol{u}_{i,G}\right) \leq f\left(\boldsymbol{x}_{i,G}\right) \\ \boldsymbol{x}_{i,G} & \text{otherwise} \end{cases}, \quad (7)$$

where G is the index of the current generation.

Historical Memory Updates

Historical memories $\boldsymbol{M}_F$ and $\boldsymbol{M}_{CR}$ are initialized according to (2), but their components change during the evolution. These memories serve to hold successful values of F and CR used in mutation and crossover steps. Successful in terms of producing trial individual better than the original individual. During every single generation, these successful values are stored in their corresponding arrays $\boldsymbol{S}_F$ and $\boldsymbol{S}_{CR}$. After each generation, one cell of $\boldsymbol{M}_F$ and $\boldsymbol{M}_{CR}$ memories is updated. This cell is given by the index k, which starts at 1 and increases by 1 after each generation. When it overflows the memory size H, it is reset to 1. The new value of k-th cell for $\boldsymbol{M}_F$ is calculated by (8) and for $\boldsymbol{M}_{CR}$ by (9).

$$M_{F,k} = \begin{cases} \text{mean}_{WL}(\boldsymbol{S}_F) & \text{if } \boldsymbol{S}_F \neq \emptyset \\ M_{F,k} & \text{otherwise'} \end{cases} \quad (8)$$

$$M_{CR,k} = \begin{cases} \text{mean}_{WL}(\boldsymbol{S}_{CR}) & \text{if } \boldsymbol{S}_{CR} \neq \emptyset \\ M_{CR,k} & \text{otherwise'} \end{cases} \quad (9)$$

where $\text{mean}_{WL}()$ stands for weighted Lehmer (10) mean.

$$\text{mean}_{WL}(\boldsymbol{S}) = \frac{\sum_{k=1}^{|\boldsymbol{S}|} w_k \cdot S_k^2}{\sum_{k=1}^{|\boldsymbol{S}|} w_k \cdot S_k}, \quad (10)$$

where the weight vector $\boldsymbol{w}$ is given by (11) and is based on the improvement in objective function value between trial and original individuals in current generation G.

$$w_k = \frac{\text{abs}\left(f(\boldsymbol{u}_{k,G}) - f(\boldsymbol{x}_{k,G})\right)}{\sum_{m=1}^{|\boldsymbol{S}_{CR}|} \text{abs}\left(f(\boldsymbol{u}_{m,G}) - f(\boldsymbol{x}_{m,G})\right)}. \quad (11)$$

And since both arrays $\boldsymbol{S}_F$ and $\boldsymbol{S}_{CR}$ have the same size, it is arbitrary which size will be used for the upper boundary for m in (11).

The last equation (11) is the subject of change in the novel Db_SHADE algorithm, which is described in the next section.

1.2 Db_SHADE

The original adaptation mechanism for scaling factor and crossover rate values uses weighted forms of means (10), where weights are based on the improvement in objective function value (11). This approach promotes exploitation over exploration and therefore might lead to premature convergence, which could be a problem especially in higher dimensions.

Distance approach is based on the Euclidean distance between the trial and the original individual, which slightly increases the complexity of the algorithm by exchanging simple difference for Euclidean distance computation for the price of stronger exploration. In this case, scaling factor and crossover rate values connected with the individual that moved the furthest will have the highest weight (12).

$$w_k = \frac{\sqrt{\sum_{j=1}^{D} \left(u_{k,j,G} - x_{k,j,G}\right)^2}}{\sum_{m=1}^{|\boldsymbol{S}_{CR}|} \sqrt{\sum_{j=1}^{D} \left(u_{m,j,G} - x_{m,j,G}\right)^2}}. \quad (12)$$

Therefore, the exploration ability is rewarded and this should lead to avoidance of the premature convergence in higher dimensional objective spaces. Such approach might be also useful for constrained problems, where constrained areas could be overcome by increased changes of individual's components.

Below is the pseudo-code of the Db_SHADE algorithm for a clear overview.

Algorithm pseudo-code 2: Db_SHADE

Set NP, H and stopping criterion;
$G = 0$, $x_{best} = \{\}$, $k = 1$, $p_{min} = 2/NP$, $\boldsymbol{A} = \emptyset$;
Randomly initialize (1) population $\boldsymbol{P} = (\boldsymbol{x}_{1,G}, \ldots, \boldsymbol{x}_{NP,G})$;
Set $\boldsymbol{M}_F$ *and* $\boldsymbol{M}_{CR}$ according to (2);
$\boldsymbol{P}_{new} = \{\}$, $\boldsymbol{x}_{best}$ = best from population $\boldsymbol{P}$;
while stopping criterion not met
  $\boldsymbol{S}_F = \emptyset$, $\boldsymbol{S}_{CR} = \emptyset$;
  for $i = 1$ to NP **do**
    $\boldsymbol{x}_{i,G} = \boldsymbol{P}[i]$;
    $r = U[1, H]$, $p_i = U[p_{min}, 0.2]$;
    Set F_i by (4) and CR_i by (6);
    $\boldsymbol{v}_{i,G}$ by mutation (3);
    $\boldsymbol{u}_{i,G}$ by crossover (5);
    if $f(\boldsymbol{u}_{i,G}) < f(\boldsymbol{x}_{i,G})$ **then**
      $\boldsymbol{x}_{i,G+1} = \boldsymbol{u}_{i,G}$;
      $\boldsymbol{x}_{i,G} \rightarrow \boldsymbol{A}$;
      $F_i \rightarrow \boldsymbol{S}_F$, $CR_i \rightarrow \boldsymbol{S}_{CR}$;
    else
      $\boldsymbol{x}_{i,G+1} = \boldsymbol{x}_{i,G}$;
    end
    if $|\boldsymbol{A}| > NP$ **then** randomly delete individuals from $\boldsymbol{A}$ **end**;
    $\boldsymbol{x}_{i,G+1} \rightarrow \boldsymbol{P}_{new}$;
  end
  if $\boldsymbol{S}_F \neq \emptyset$ **and** $\boldsymbol{S}_{CR} \neq \emptyset$ **then**
    Update $\boldsymbol{M}_{F,k}$ (8) and $\boldsymbol{M}_{CR,k}$ (9) with distance based weights from (12), k++;
    if $k > H$ **then** $k = 1$, **end**;
  end
  $\boldsymbol{P} = \boldsymbol{P}_{new}$, $\boldsymbol{P}_{new} = \{\}$, $\boldsymbol{x}_{best}$ = best from population $\boldsymbol{P}$;
end
return $\boldsymbol{x}_{best}$ as the best found solution

1.3 DIb_SHADE

The DIb_SHADE algorithm is a trade-off between simple improvement based SHADE and distance based Db_SHADE. While Db_SHADE neglects the improvement in objective function value and works only with distance between solutions when it calculates parameter weights (12), DIb_SHADE combines both approaches. The resulting weight $\boldsymbol{w}$ for the parameter combination is computed as a weighted sum (13) of distance based weight $\boldsymbol{w}_d$ (11) and improvement based weight $\boldsymbol{w}_i$ (12).

$$\boldsymbol{w} = WD * \boldsymbol{w}_d + WI * \boldsymbol{w}_i, \tag{13}$$

where WD and WI are user-defined weights for distance and improvement parts of the calculation respectively. Therefore, this approach combines both explorative and exploitative weights in order to balance those two characteristics and can be tuned to optimally solve the given task.

EXPERIMENTAL SETTING

Algorithms ranging from improvement only based SHADE through a combination of distance and improvement based DIb_SHADE to distance only based Db_SHADE were tested on a basis of CEC2015 benchmark set of 15 test functions in 30D because, in lower dimensional spaces, the algorithm does not suffer the premature convergence. The weight pairs (WD, WI) for distance and improvement parts of the DIb_SHADE were as follows:

- (0, 1) – SHADE
- (1, 3) – DIb13_SHADE
- (1, 2) – DIb12_SHADE
- (1, 1) – DIb_SHADE
- (2, 1) – DIb21_SHADE
- (3, 1) – DIb31_SHADE
- (1, 0) – Db_SHADE

These seven versions were run with the same parameter settings:

- Population size $NP = 100$.
- Historical memory size $H = 10$.
- External archive size $|\boldsymbol{A}| = NP$.
- Stopping criterion according to CEC2015, maximum number of function evaluations $MAXFES = 10{,}000 \times D = 300{,}000$.
- Number of runs $R = 51$.

RESULTS

The resulting optimization values of the seven algorithm versions were subject to the Friedman rank statistical test to find out, whether there are significant differences. Friedman rank test yielded a p-value of 0.02, which confirmed the hypothesis, that there are significant differences and the resulting ranks for each algorithm variant are shown in Figure 1. Mean error values used for the Friedman rank test are provided in Table 1 and the best-obtained error value is highlighted in bold text.

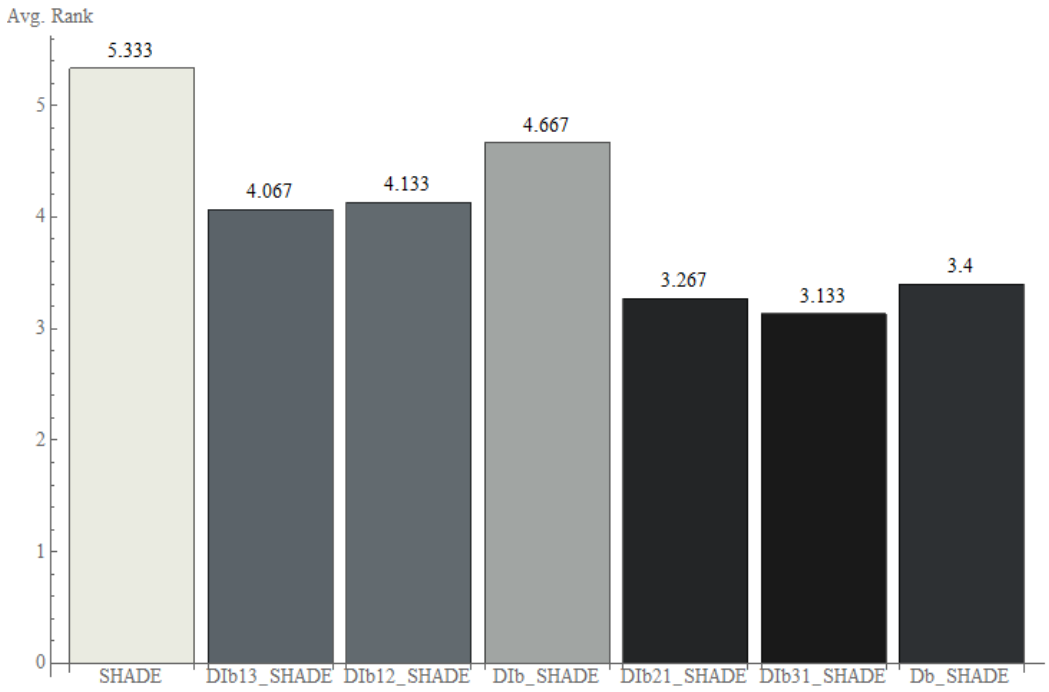


Figure 1: Friedman rank test results on CEC2015 in 30*D*.

Table 1: Mean results of 51 runs on the CEC2015 benchmark set.

f	SHADE	DIb13_SHADE	DIb12_SHADE	DIb_SHADE	DIb21_SHADE	DIb31_SHADE	Db_SHADE
1	2.62E+02	2.12E+02	1.72E+02	2.82E+02	1.94E+02	**1.37E+02**	2.42E+02
2	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**
3	2.01E+01	**2.01E+01**	2.01E+01	2.01E+01	2.01E+01	2.01E+01	2.01E+01
4	1.41E+01	1.36E+01	1.42E+01	1.38E+01	1.34E+01	1.34E+01	**1.31E+01**
5	1.50E+03	**1.46E+03**	1.49E+03	1.52E+03	1.51E+03	1.50E+03	1.52E+03
6	5.73E+02	6.22E+02	5.07E+02	4.80E+02	3.79E+02	4.70E+02	**3.48E+02**
7	7.26E+00	7.17E+00	6.91E+00	7.17E+00	6.83E+00	6.87E+00	**6.74E+00**
8	1.21E+02	1.01E+02	9.62E+01	1.05E+02	9.12E+01	7.84E+01	**7.38E+01**
9	1.03E+02	1.03E+02	1.03E+02	1.03E+02	1.03E+02	**1.03E+02**	1.03E+02
10	6.22E+02	5.96E+02	5.22E+02	5.72E+02	**5.18E+02**	5.87E+02	5.32E+02
11	4.50E+02	4.39E+02	4.36E+02	4.31E+02	4.27E+02	4.21E+02	**4.16E+02**
12	1.05E+02	**1.05E+02**	1.05E+02	1.05E+02	1.05E+02	1.05E+02	1.05E+02
13	9.50E+01	**9.40E+01**	9.59E+01	9.45E+01	9.42E+01	9.46E+01	9.50E+01
14	3.24E+04	3.24E+04	3.24E+04	3.23E+04	3.26E+04	**3.22E+04**	3.24E+04
15	**1.00E+02**	**1.00E+02**	**1.00E+02**	**1.00E+02**	**1.00E+02**	**1.00E+02**	**1.00E+02**

It can be seen, that the best rank was obtained by DIb31_SHADE variant closely followed by DIb21_SHADE and Db_SHADE. In the previous study (Viktorin et. al. 2017), Db_SHADE's improved performance was associated with the ability to maintain population diversity for longer optimization period and therefore, it would be only reasonable to assume that the ranks would be linearly decreasing towards distance only based SHADE variant (Db_SHADE). However, the results show that incorporating improvement factor into the weight calculation with lower importance can bring overall performance improvement (DIb31_SHADE and DIb21_SHADE). Additional Wilcoxon rank-sum tests were performed on the selected pairs of SHADE variants in order to obtain clearer answers to the performance question. These results are summarized in a form of wins/loses/draws triplets in Table 2. The significance level of the Wilcoxon rank-sum test was set to 5%.

From the Table 2, it is clear that Db_SHADE algorithm significantly outperforms the DIb31_SHADE on one test function and algorithms draw on the rest of the benchmark. This fact is in contradiction with their respective ranks from the Friedman test. It is also visible that introduction of the distance based parameter adaptation to the SHADE is beneficial in all cases. One of the reasons for surprising results in Friedman ranks is that the ranking is based on the mean error values of 51 runs on each test function. Therefore, even when the difference in obtained error values is not significant, the ranking system will prefer the smaller value and assign it with smaller rank.

Table 2: Selected pairwise comparisons.

Algorithm pair	wins/loses/draws
SHADE – DIb31_SHADE	0/4/11
SHADE - DIb_SHADE	0/2/13
SHADE - Db_SHADE	0/5/11
Dib31_SHADE - DIb_SHADE	2/0/13
Dib31_SHADE - Db_SHADE	0/1/14
DIb_SHADE - Db_SHADE	0/3/12

Another interesting aspect is that higher values of the distance weight factor *WD* appear to have a higher impact on the performance. This might be due to increased differences between distance weights by the given factor *WD* and therefore, originally significant weights will have an even higher influence to the weighted Lehmer mean computation in comparison with weights from the other side of the spectrum (10). This phenomenon is an attractive future research direction for the authors.

CONCLUSION

This paper presented a comparative study of the effect of improvement and distance based parameter adaptation to the overall performance of the SHADE algorithm. On the seven algorithm variants, ranging from improvement only based SHADE through balanced improvement and distance based DIb_SHADE to distance only based Db_SHADE, it was shown that introduction of the distance based parameter adaptation to the weight computation in historical memory update is a beneficial step. However, there is also a number of open questions in the weighting of both improvement and distance factors. These provide an interesting future research direction for the authors and will be addressed in their future work.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

Awad, N. H., Ali, M. Z., Suganthan, P. N., & Reynolds, R. G. (2016, July). An ensemble sinusoidal parameter adaptation incorporated with L-SHADE for solving CEC2014 benchmark problems. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 2958-2965). IEEE.

Brest, J., Maučec, M. S., & Bošković, B. (2017, June). Single objective real-parameter optimization: Algorithm jSO. In Evolutionary Computation (CEC), 2017 IEEE Congress on (pp. 1311-1318). IEEE.

Das, S., Mullick, S. S., & Suganthan, P. N. (2016). Recent advances in differential evolution–An updated survey. Swarm and Evolutionary Computation, 27, 1-30.

Gämperle, R., Müller, S. D., & Koumoutsakos, P. (2002). A parameter study for differential evolution. Advances in intelligent systems, fuzzy systems, evolutionary computation, 10, 293-298.

Guo, S. M., Tsai, J. S. H., Yang, C. C., & Hsu, P. H. (2015, May). A self-optimization approach for L-SHADE incorporated with eigenvector-based crossover and successful-parent-selecting framework on CEC 2015 benchmark set. In Evolutionary Computation (CEC), 2015 IEEE Congress on (pp. 1003-1010). IEEE.

Liu, J., & Lampinen, J. (2002). On setting the control parameter of the differential evolution method. In Proceedings of the 8th international conference on soft computing (MENDEL 2002) (pp. 11-18).

Neri, F., & Tirronen, V. (2010). Recent advances in differential evolution: a survey and experimental analysis. Artificial Intelligence Review, 33(1-2), 61-106.

Storn, R., & Price, K. (1995). Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous spaces (Vol. 3). Berkeley: ICSI.

Tanabe, R., & Fukunaga, A. (2013). Success-history based parameter adaptation for differential evolution. In Evolutionary Computation (CEC), 2013 IEEE Congress on (pp. 71-78). IEEE.

Tanabe, R., & Fukunaga, A. S. (2014). Improving the search performance of SHADE using linear population size reduction. In Evolutionary Computation (CEC), 2014 IEEE Congress on (pp. 1658-1665). IEEE.

Viktorin, A., Senkerik, R., Pluhacek, M., Kadavy, T. & Zamuda, A. (2017) Distance Based Parameter Adaptation for Differential Evolution. In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on (pp. 2612-2618) IEEE. In press.

Zhang, J., & Sanderson, A. C. (2009). JADE: adaptive differential evolution with optional external archive. Evolutionary Computation, IEEE Transactions on, 13(5), 945-958.

AUTHOR BIOGRAPHIES

ADAM VIKTORIN was born in the Czech Republic in 1989, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same University and the fields of his studies are: Artificial Intelligence, Data Mining and Evolutionary Computation. His most recent professional "hobby" is a development and analysis of self-adaptive strategies for Differential Evolution in the area of numerical optimization; and also application of such algorithms to real-world problems. His email address is: aviktorin@utb.cz

ROMAN SENKERIK received his Ph.D. degree in Technical Cybernetics from the Tomas Bata University in Zlin, Czech Republic in 2008. He is currently an associated professor at the Tomas Bata University in Zlin, Faculty of Applied Informatics. His research interests include interdisciplinary applications of evolutionary computation, modification and development of evolutionary and swarm based algorithms, computational intelligence, optimization, cyber-security, theory of chaos, emergence and complexity. His email address is: senkerik@utb.cz

MICHAL PLUHACEK received his Ph.D. degree in Information Technologies from the Tomas Bata University in Zlin, the Czech Republic in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. Currently works as a junior researcher at the Regional Research Centre CEBIA-Tech of Tomas Bata University in Zlin. His research focus includes swarm intelligence theory and applications and artificial intelligence in general. His email address is: pluhacek@utb.cz

TOMAS KADAVY was born in the Czech Republic in 1990, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlin, where he studied Information Technologies and obtained his MSc degree in 2016. He is studying his Ph.D. at the same University and the fields of his studies are: swarm based algorithms, computational intelligence and optimization. His email address is: kadavy@utb.cz

BOUNDARY STRATEGIES FOR FIREFLY ALGORITHM ANALYSED USING CEC`17 BENCHMARK

Tomas Kadavy
Michal Pluhacek
Adam Viktorin
Roman Senkerik
Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{kadavy, pluhacek, aviktorin, senkerik }@utb.cz

KEYWORDS
Firefly Algorithm, FA, FFPSO, Boundary, CEC17

ABSTRACT

In this paper, we are presenting a comparison of few selected boundary strategies for two popular optimization algorithms. The Firefly Algorithm (FA) and its hybridized modification, the Firefly Particle Swarm Optimization (FFPSO). The problem of boundary constrained optimization was already extensively studied for well-known heuristic optimization Particle Swarm Optimization (PSO). This suggesting importance for similar research for other swarm-based algorithms, like FA. The recent CEC´17 benchmark suite is used for the performance comparison of the methods and the results are compared and tested for statistical significance.

INTRODUCTION

Firefly Algorithm (FA) (Yang 2008; Yang 2009) is one of the modern and versatile optimization algorithms developed by Yang in 2008. Over the years, the FA proved its robustness in performance on several problems (Oosumi et al. 2016; Yang 2013). Another proof of the importance of this modern swarm-based algorithm is supported by a number of its modification. For example, Lévy flights (Yang 2010) and chaos-driven FA (Coelho, Mariani 2012) show the large potential for future research and possible application.

The definition of typical optimization task contains boundary limits for optimized parameters. Due to nature of the metaheuristic optimization algorithm, the trial particle (in this case firefly) can emerge outside of the area of the feasible solution. The paper focuses on the question what to do with particle if it tries to violate the defined boundaries. The research in the area of possible border strategies and their influence on performance was already extensively done for another well-known optimization technique the Particle Swarm Optimization (PSO). As the provided studies suggesting, it could be a truly difficult task (Helwig et al. 2013; Kadavy et al. 2017).

Since there is a lack of such studies for FA available in the literature, we have decided to perform and present this original experimental research. In this paper, four relatively common borders strategies (or rather methods) are implemented and compared on CEC'17 benchmark set (Awad et al. 2016). Also, one modern hybrid optimization technique was tested with the mentioned boundary strategies. The Firefly Particle Swarm Optimization (FFPSO) (Kora, Rama 2016) was particularly selected due to its similarity with the PSO. The previous studies on PSO show the importance of careful selection of border strategy. The main research questions can be then summarized as follows:

- Could the selection of border strategy influence the performance of canonical FA or the FFPSO?
- What is the most suitable border strategy in general?

The paper is structured as follows. The FA and the FFPSO are described in details within the next two sections. The implemented and tested border strategies follow afterwards. Last two sections discuss the experiment setting and results.

FIREFLY ALGORITHM

This optimization nature-based algorithm was developed and introduced by Yang in 2008 (Yang 2008). The fundamental principle of this algorithm lies in simulating the mating behavior of fireflies at night when fireflies emit light to attract a suitable partner. The main idea of Firefly Algorithm (FA) is that the objective function value that is optimized is associated with the flashing light of these fireflies. The author for simplicity set a couple of rules to describe the algorithm itself:

- The brightness of each firefly is based on the objective function value.
- The attractiveness of a firefly is proportional to its brightness. This means that the less bright firefly is lured towards, the brighter firefly. The brightness depends on the environment or the medium in which fireflies are moving and decreases with the distance between each of them.
- All fireflies are sexless, and it means that each firefly can attract or be lured by any of the remaining ones.

The movement of one firefly towards another one is then defined by equation (1). Where x_i' is a new position of a firefly i, x_i is the current position of firefly i and x_j is a selected brighter firefly (with better objective function

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

value). The α is a randomization parameter and *sign* simply provides random direction -1 or 1.

$$x_i' = x_i + \beta \cdot (x_j - x_i) + \alpha \cdot sign \quad (1)$$

The brightness *I* of a firefly is computed by the equation (2). This equation of brightness consists of three factors mentioned in the rules above. On the objective function value, the distance between two fireflies and the last factor is the absorption factor of a media in which fireflies are.

$$I = \frac{I_0}{1+\gamma r^m} \quad (2)$$

Where I_o is the objective function value, the γ stands for the light absorption parameter of a media in which fireflies are and the *m* is another user-defined coefficient and it should be set $m \geq 1$. The variable *r* is the Euclidian distance (3) between the two compared fireflies.

$$r_{ij} = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \quad (3)$$

Where r_{ij} is the Euclidian distance between fireflies x_i and x_j. The d is current dimension size of the optimized problem.

The attractiveness β (4) is proportional to brightness *I* as mentioned in rules above and so these equations are quite similar to each other. The β_0 is the initial attractiveness defined by the user, the γ is again the light absorption parameter and the *r* is once more the Euclidian distance. The *m* is also the same as in equation (2).

$$\beta = \frac{\beta_0}{1+\gamma r^m} \quad (4)$$

Finally, the pseudocode below shows the fundamentals of FA operations.

```
1.  FA initialization
2.  while(terminal condition not met)
3.    for i = 1 to all fireflies
4.      for j = 1 to all fireflies
5.        if(Ij < Ii) then
6.          move xi to xj
7.          evaluate xi
8.        end if
9.      end for j
10.   end for i
11.   record the best firefly
12. end while
```

FIREFLY PARTICLE SWARM OPTIMIZATION

The typical example of a hybrid of the FA and PSO algorithms, the FFPSO (Kora, Rama 2016) introduced in late 2016 by Padmavathi Kora and K. Sri Rama Krishna. The basic idea behind such an approach is that the new hybrid strategy can share advantages from both algorithms and hopefully eliminate their disadvantages. The main principle remains the same as in the standard FA, but the equation for firefly motion (1) is slightly changed according to PSO movement (Eberhart, Kennedy 1995) and is newly computed as (5).

$$x_i' = wx_i + c_1 e^{-r_{px}^2}(pBest_i - x_i) + c_2 e^{-r_{gx}^2}(gBest - x_i) + \alpha \cdot sign \quad (5)$$

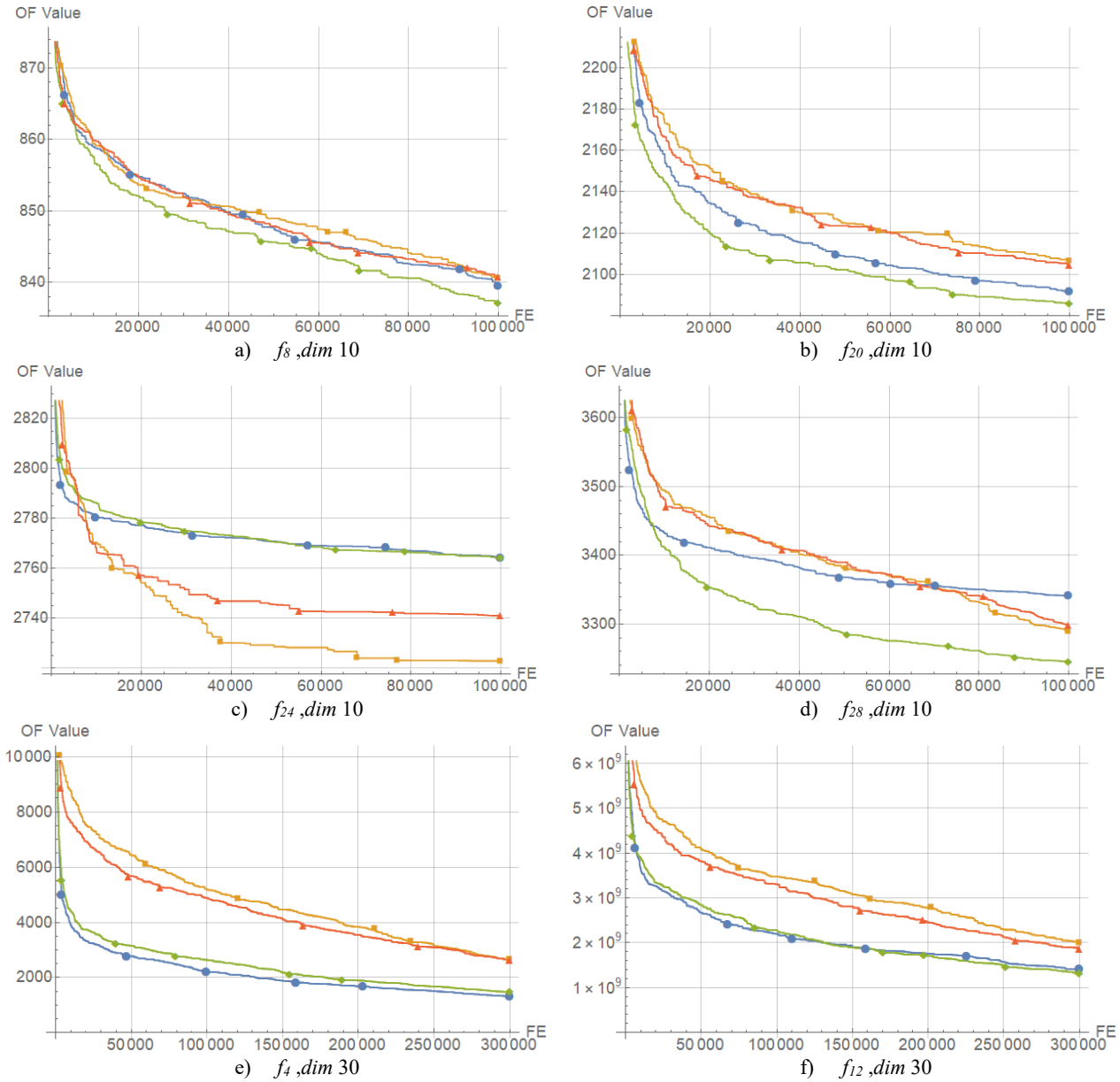
Where *w*, c_1, and c_2 are control parameters transferred from PSO and their values often depends on the user. Also, the *pBest* and *gBest* are variables originally belonging to PSO algorithm. They both represent the memory of the best position where *pBest* is best position of each particle and *gBest* is globally achieved best position so far. The remaining variables r_{px} (6) and r_{gx} (7) are distances between particle x_i and both $pBest_i$ and *gBest*.

$$r_{px} = \sqrt{\sum_{k=1}^{d} (pBest_{i,k} - x_{i,k})^2} \quad (6)$$

$$r_{gx} = \sqrt{\sum_{k=1}^{d} (gBest_k - x_{i,k})^2} \quad (7)$$

The pseudocode below shows the fundamentals of FFPSO operations. As it can be seen in given pseudocode, the main principle remains the same as in canonical FA.

```
1.  FFPSO initialization
2.  while(terminal condition not met)
3.    for i = 1 to all fireflies
4.      for j = 1 to all fireflies
5.        if(Ij < Ii) then
6.          calculate rpx and rgx
7.          move xi to xj
8.          evaluate xi
9.        end if
10.     end for j
11.   end for i
12.   record the best firefly
13. end while
```



a) f_8 ,*dim* 10 b) f_{20} ,*dim* 10

c) f_{24} ,*dim* 10 d) f_{28} ,*dim* 10

e) f_4 ,*dim* 30 f) f_{12} ,*dim* 30

Figure 1: Convergence graphs of canonical FA The clipping strategy is represented as blue line with circle marks. The random strategy is represented as orange line with rectangle marks. The reflection strategy is represented as green line with rhombus marks. Finally, the periodic strategy is represented as red line with triangle marks.

BORDER STRATEGIES

Every time, when a single objective function optimization problem has defined a range where the best value is being found by the metaheuristic algorithm, one of the many difficult tasks could arise to an operator or a user. After each step of an algorithm, in this case after position update of a firefly, the new position should be checked if it lies in the appropriate range or boundaries (inside space of feasible solution). In case that the new position of the particle is outside this allowed region a certain correction has to be made. Several possible correction methods or strategies could do the trick. However, select the most appropriate is not an easy task since each of them could have a very different effect on the algorithm ability to achieve a good solution (Helwig et al. 2013). For this paper, the most common ones were selected and compared together to show how they could affect the FA or FFPSO on different benchmark functions.

Clipping strategy

The first selected strategy is rather simple in principle. The particle (or in this case firefly) cannot cross the given boundaries in each dimension. This strategy is very simple to implement and is described as (8).

$$x_i' = \begin{cases} x_i = b^u, if\ x_i > b^u \\ x_i = b^l, if\ x_i < b^l \\ x_i, otherwise \end{cases} \quad (8)$$

Where x_i is the position of i firefly before boundary check, the x_i' is a newly updated position after the boundary check and the b^u and b^l are the upper and lower boundary given to each dimension.

Random strategy

If a firefly violates the boundary in any dimension, the new position for this firefly for a particular dimension is created between the lower and upper boundary (with a pseudo-random uniform distribution). Again this strategy is rather simple and very easy to implement, as the equation (9) shows.

$$x_i' = \begin{cases} x_i = U(b^l, b^u), if\ x_i > b^u\ OR\ x_i < b^l \\ x_i, otherwise \end{cases} \quad (9)$$

Where *U* stands for uniform distribution in range from b^l (lower boundary limit) to b^u (upper boundary limit).

Reflection strategy

The reflection strategy (Helwig et al. 2013) reflects the particle back to feasible space of solution if it tries to violate the defined borders. This strategy tries to emulate the reflection characteristic of for example a mirror. For violated dimension, the correction of a position of a particle is computed as (10). Where again the b^u and b^l are the upper boundary limit and lower boundary limit.

$$x_i' = \begin{cases} x_i' = b^u - (x_i - b^u), if\ x_i > b^u \\ x_i' = b^l + (b^l - x_i), if\ x_i < b^l \\ x_i, otherwise \end{cases} \quad (10)$$

Periodic strategy

The possible solution to prevent the infeasibility could lie in the method of infinite copies of the optimized hyperspace (Zhang et al. 2004). This strategy involves only mapping the particle back to the space of available solution using only the modulo function (11).

$$x_i' = b^l + \left(x_i\ MOD\ (b^u - b^l)\right) \quad (11)$$

EXPERIMENT SETTING

The experiments were performed on a set of well-known benchmark functions CEC'17 which are detailly described in (Awad et al. 2016). The tested dimensions were 10 and 30. The maximal number of function evaluation was set as 10 000 · *dim* (dimension size). The lower and upper boundary was as $b^l = -100$ and $b^u = 100$ according to CEC'17. The number of fireflies was set to 40 for both dimension sizes. Every test function was repeated for 51 independent runs and, the results were statistically evaluated. The benchmark itself includes 30 test functions in four categories: unimodal, multimodal, hybrid and composite types. The global minimum of each function is easy to determine as it is $100 \cdot f_i$ where *i* is an order of test function.

The parameters of FA were set as $\alpha = 0.5,\ \gamma = 1, \beta_0 = 0.2$ and $m = 1$ according to author (Yang 2008). The parameters of FFPSO were set the same as FA including the parameters borrowed from PSO ($c = 1.49445,\ w = 0.729$).

RESULTS

The results of performed experiments are presented in this section. Firstly, the examples of convergence behavior of the compared methods are given in Figure 1.

Furthermore, the results were tested for statistical significance using the Friedman Rank test (Demšar 2006). The null hypothesis that the mean is equal is rejected at the 5 percent level based on the Friedman Rank test. The corresponding p-values of Friedman Rank test are presented in Table 1. If the p-value is lower than 0.05, the further Friedman rankings are relevant (these values are given by bold numbers in Table 1). In Figure 2, the Friedman ranking for both algorithms (FA and FFPSO) on both dimension sizes is shown. The lower the rank is, the better is the performance of the strategy labeled strategy. Furthermore, the presented Friedman ranks are accompanied with critical distance evaluated according to the Nemenyi Critical Distance post-hoc test for multiple comparisons. The dashed line represents the critical distance from the best boundary method (the lowest mean rank).

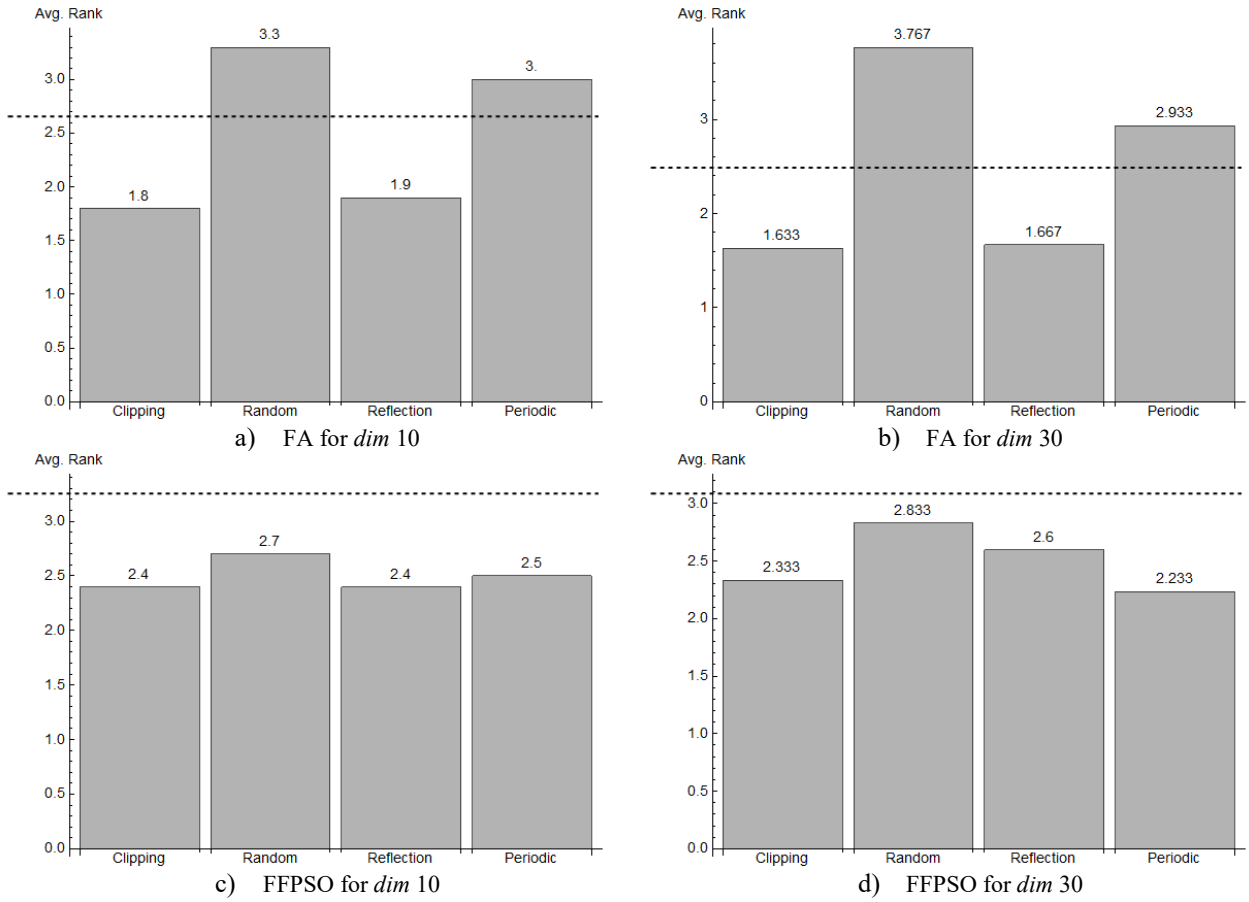
The critical distance (CD) value for this experiment has been calculated as 0.656757; according to the definition given in (12) and value q_a = 2.56892; using k = 4 boundary methods and a number of data sets N = 51 (51 repeated runs).

$$CD = q_a\sqrt{k(k+1)/(6N)} \quad (12)$$

Table 1: P-values of Friedman Rank tests

Algorithm	Dimension size	
	10	*30*
FA	**3.757E-08**	**1.208E-19**
FFPSO	0.788E-00	0.268E-00

According to this evaluation and the ranking shown in Figure 2, the significant impact of the selected border strategies is mostly observed on canonical FA. For the hybrid method FFPSO the results suggesting some sort of insignificance for the used method. For dimension sizes 10 and 30, the most favorable strategies seem to be clipping and reflection methods. Despite the fact that at first sight, the used strategies are unique to each other, some similarities are shared among them. The clipping and reflection strategies forcing the fireflies to move only in the feasible space of solution without loss of information of the previous position. On the contrary, the two remaining strategies have in common the loss of the previous position and their behavior resembling the random search optimization (Bergstra, Bengio 2012).



a) FA for *dim* 10 b) FA for *dim* 30

c) FFPSO for *dim* 10 d) FFPSO for *dim* 30

Figure 2: Friedman Rank tests for selected border strategies.

CONCLUSION

In this original study, the impact of various border strategies on the performance of the FA and FFPSO is tested. The topic is actual due to the increasing variety and complexity of optimization problems. As a benchmark for the performance comparisons, the CEC 2017 set was used. It represents the most recent collection of artificial optimization problems that vary in terms of modality and other characteristics of the fitness land scape.

It may be concluded, that according to statistical data, the clipping and reflection strategies seem to be favorable over the other two (random and periodic strategies). As the Friedman ranks showed, the slightly better performance is achieved by the clipping method. However, the same four strategies strangely seem to have no significant impact on the hybrid FFPSO algorithm. Moreover, all the observation suits for both dimension setting with almost the same results.

To answer the research question:

- Could the selection of border strategy influence the performance of canonical FA or the FFPSO?
- From the achieved results, the border strategy has a significant impact on the performance for only the canonical FA. For FFPSO, the influence is negligible.
- What is the most suitable border strategy in general?
- In general term of speaking, the most suitable strategies are two. Specifically, the clipping and reflection strategy.

Despite that, the results of this study are useful as an empirical study for researchers dealing with firefly algorithm. This research will continue in the future with exploring the performance of firefly algorithm with different boundary strategies on different fitness landscape models and real-world problems, especially with a focus on the algorithm setup to achieve the best performance.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied

Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

X. S. Yang, Nature-Inspired Metaheuristic Algorithms, Luniver Press, UK, (2008).

X. S. Yang, Firefly algorithms for multimodal optimization, Proc. 5th Symposium on Stochastic Algorithms, Foundations and Applications, Lecture Notes in Computer Science, 5792: 169-178 (2009).

Oosumi R., Tamura K., Yasuda K. Novel single-objective optimization problem and Firefly Algorithm-based optimization method. 9-12 Oct. 2016; IEEE; 2016.

Yang X. Multiobjective firefly algorithm for continuous optimization. Engineering with Computers. 2013; 29 (2):175-84.

X. S. Yang, Firefly algorithm, Lévy flights and global optimization, in: Research an Development in Intelligent Systems XXVI (Eds M. Bramer, R. Ellis, M. Petridis), Springer London, pp. 209-218 (2010).

Coelho LdS, Mariani VC. Firefly algorithm approach based on chaotic Tinkerbell map applied to multivariable PID controller tuning. Computers & Mathematics with Applications. 2012 October 1; 64(8):2371-82.

Helwig S, Branke J, Mostaghim S M. Experimental Analysis of Bound Handling Techniques in Particle Swarm Optimization. TEVC. 2013; 17(2):259-71.

Kadavy T, Pluhacek M, Viktorin A, Senkerik R. Comparing Border Strategies for Roaming Particles on Single and Multi-swarm PSO. In: Artificial Intelligence Trends in Intelligent Systems: Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol1. Cham: Springer International Publishing; 2017. p. 528-36.

N. H. Awad, et al. Problem Definitions and Evaluation Criteria for CEC 2017 Special Ses-sion and Competition on Single-Objective Real-Parameter Numerical Optimization, 2016.

Kora P, Rama Krishna KS. Hybrid firefly and Particle Swarm Optimization algorithm for the detection of Bundle Branch Block. International Journal of the Cardiovascular Academy. 2016 March 1;2(1):44-8.

Eberhart R, Kennedy J. A new optimizer using particle swarm theory. 1995.

Helwig, S., Branke, J., & Mostaghim, S. (2013). Experimental analysis of bound handling techniques in particle swarm optimization. IEEE Transactions on Evolutionary computation, 17(2), 259-271.

W.-J. Zhang, X.-F. Xie, and D.-C. Bi, "Handling boundary constraints for numerical optimization by particle swarm flying in periodic search space," in Proc. IEEE Congr. Evol. Comput., Jun. 2004, vol. 2, pp. 2307–2311.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), 1-30.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

AUTHOR BIOGRAPHIES

TOMAS KADAVY was born in the Czech Republic,

and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2016. He is studying his Ph.D. at the same university and the fields of his studies are: Artificial intelligence and evolutionary algorithms. His email address is: kadavy@utb.cz

MICHAL PLUHACEK was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011 and Ph.D. in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. He now works as a researcher at the same university. His email address is: pluhacek@utb.cz

ADAM VIKTORIN was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the fields of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@utb.cz

ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlín, where he studied Technical Cybernetics and obtained his MSc degree in 2004, Ph.D. degree in Technical Cybernetics in 2008 and Assoc. prof. in 2013 (Informatics). He is now an Assoc. prof. at the same university (research and courses in: Evolutionary Computation, Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). His email address is: senkerik@utb.cz

A REVIEW ON THE SIMULATION OF SOCIAL NETWORKS INSIDE HEURISTIC ALGORITHMS

Roman Senkerik, Michal Pluhacek, Adam Viktorin, Tomas Kadavy, Jakub Janostik and Zuzana Kominkova Oplatkova

Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{senkerik, oplatkova , pluhacek , aviktorin , kadavy}@fai.utb.cz

KEYWORDS
Social networks; Graphs; Analysis; Evolutionary algorithms; Swarm Algorithms.

ABSTRACT

This paper represents a comprehensive review of selected methods for visualization of the population dynamics of the swarm and evolutionary algorithms in the form of networks. The whole idea is based on the obvious similarity between interactions between individuals in a swarm and evolutionary algorithms and for example, users of social networks, society, etc. The analogy between individuals from the population used in an arbitrary evolutionary or swarm-based algorithm and vertices (nodes) of a network is discussed here, as well as between edges in a network and communication between individuals in a population. Simple experiments with four well-known heuristic algorithms are described here, giving an insight into different approaches to the building of the network during metaheuristic run.

INTRODUCTION

In this review paper, we have merged two different attractive areas of research: (complex) networks and evolutionary computation. Interactions in a swarm and evolutionary algorithms during the optimization process can be considered like user interactions in social networks or just people in society. It has been observed that networks generated by evolutionary dynamics show properties of complex networks in certain time frames and conditions (Skanderova et al. 2016).

Evolutionary computation is a sub-discipline of computer science belonging to the bio-inspired computing area. In recent decades, more robust and effective algorithms have been introduced. Like Differential Evolution (DE) (Das et al. 2016), Particle Swarm Optimization (PSO) (Engelbrecht 2010), Self Organizing Migrating Algorithm (SOMA) (Zelinka 2016), Artificial Bee Colony (Karaboga and Basturg 2007) or Firefly Algorithm (FA) (Fister et al. 2013).

Currently, the utilization of complex networks as a visualization tool for the analysis of population dynamics for evolutionary and swarm-based algorithms is becoming an interesting open research task. The population is visualized as an evolving complex network that exhibits non-trivial features – e.g., degree distribution, clustering, and centralities. These features offer a clear description of the population under evaluation and can be utilized for the adaptive population as well as parameter control during the metaheuristic run. The initial studies (Zelinka et al. 2014; Davendra et al. 2014) describing the possibilities of transforming population dynamics into complex networks were followed by the successful adaptation and control of the metaheuristic algorithm during the run through the given complex networks´ frameworks (Skanderova and Fabian 2015; Metlicka and Davendra 2015; Gajdos et al. 2015; Janostik et al. 2015).

This research paper reviews the complex network frameworks for DE, PSO, FA and Fireworks algorithm (FWA) (Tan and Zhu 2010). Currently, all algorithms above are known as powerful metaheuristic tools for solving optimization problems.

The organization of this paper is as follows: Firstly, the motivation, background of the heuristic algorithms and the concept of complex network framework for heuristics are briefly described; followed by the simple experiment designs, graphical visualizations, and conclusions.

MOTIVATION AND RELATED WORKS

This paper represents a comprehensive overview and continuation of the previous successful initial experiments, which are referred in particular sections. The motivation for the research presented herein can be summarized as follows:

- To show the different approaches in building complex networks to capture the dynamics either of evolutionary or swarm-based algorithms.
- To investigate the time development of the influence of either individual selection inside a DE or (density of) communication inside a swarm transferred into the complex network.
- To briefly discuss the possible utilization of complex network attributes that can be extracted from graph visualizations – e.g., adjacency graphs, centralities, clustering, etc for adaptive population and parameter control during the metaheuristic run.

HEURISTIC ALGORITHMS

This section contains the basic background for the metaheuristic algorithms DE, PSO, FA, and FWA that were used in this review paper. The algorithms workflows are limited to the minimum information required for the better understanding of social

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

interactions capturing to the network. More details about the algorithms can be found in the referred original literature sources.

Differential Evolution

DE is a population-based optimization method that works on real-number-coded individuals (Das et al. 2016). DE is quite robust, fast, and effective, with global optimization ability. There are essentially five inputs to the heuristic. *D* is the size of the problem, G_{max} is the maximum number of generations, *NP* is the total number of solutions, *F* is the scaling factor of the solution and *CR* is the factor for crossover. *F* and *CR* together make the internal tuning parameters for the heuristic.
In this research, we have used original DE "rand/1/bin" (1) mutation strategy and binomial crossover. The parent indices (vectors) are selected by standard PRNG with uniform distribution. Mutation strategy "rand/1" uses three random parent vectors with indexes *r1*, *r2,* and *r3*, where $r1 \neq r2 \neq r3$. Mutated vector $v_{i,G}$ is obtained from three different vectors $\boldsymbol{x}_{r1}, \boldsymbol{x}_{r2}, \boldsymbol{x}_{r3}$ from current generation *G* with the help of with the help of scaling factor F_i as follows (1):

$$\boldsymbol{v}_{i,G} = \boldsymbol{x}_{r1,G} + F_i\left(\boldsymbol{x}_{r2,G} - \boldsymbol{x}_{r3,G}\right) \quad (1)$$

After the mutation is done, the crossover procedure based on the defined and fixed *CR* value is performed between active individual from the population and newly created mutated vector. If both processes lead to the better solution, this one will be stored in the next generation *G*+1 replacing the active individual (solution).

PSO Algorithm

Original PSO algorithms take their inspiration from the behavior of fish and birds (Engelbrecht 2010). The knowledge of the global best-found solution (*gBest*) is shared among the particles in the swarm. Furthermore, each particle knows its own (personal) best-found solution (*pBest*). The last important part of the algorithm is the velocity of each particle, which is taken into account during the calculation of the particle´s movement. The new position of each particle is then given by (2), where $\boldsymbol{x}_i^{t+1}$ is the new particle position; $\boldsymbol{x}_i^t$ refers to the current particle position, and $\boldsymbol{v}_i^{t+1}$ is the new velocity of the particle. To calculate the new velocity, the distance from *pBest* and *gBest* is taken into account along with its current velocity (3), where c_1, c_2 represents the acceleration constants, and symbol *j* points to the *j*-th component of the dimension *D*..

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \boldsymbol{v}_i^{t+1} \quad (2)$$

$$v_{ij}^{t+1} = v_{ij}^t + c_1 \cdot Rand \cdot (pBest_{ij} - x_{ij}^t) + \\ + c_2 \cdot Rand \cdot (gBest_j - x_{ij}^t) \quad (3)$$

Firefly Algorithm

FA was firstly introduced by X. S. Yang (Yang 2010). This nature-based algorithm tries to simulate the mating behavior of fireflies at night. Every firefly emits flashing light to lure appropriate mating partner. For the formulation of the FA, the flashing light is associated with the objective function value that is optimized. For simplicity, the three following rules are used:

- All fireflies are sexless (each firefly can attract, or be attracted by, any of the remaining ones).
- The attractiveness of fireflies is proportional to their brightness. Thus the less bright firefly will move toward the brighter one. The brightness decreases with the distance between fireflies. If there is a no brighter firefly, the particular one will move randomly.
- The firefly brightness is based on the objective function value.

The brightness of firefly consists of three factors: the objective function value, the distance between two compared fireflies and absorption of media in which the fireflies are.

Fireworks Algorithm

The FWA is an algorithm that is inspired by fireworks explosion in a night sky. This algorithm is initialized with a random population of fireworks. The particular firework position is represented as coordinates in n-dimensional space of solutions. These coordinates are parameters of the optimized problem. The number of the fireworks is defined by the parameter NP. This algorithm consists of four parts: explosion operator, mutation operator, mapping rule and selection strategy. These parts and adjustable parameters are more explained in next sections. The realization of FWA is as follows:

- Randomly generate NP fireworks in the n-dimensional search space.
- Obtain fitness values of these generated fireworks by the fitness function.
- Calculate the number of generated sparks and their amplitude for each firework by explosion operator.
- Use Gaussian mutation to generate new random sparks by mutation operator.
- Apply mapping rule to all generated sparks.
- Calculate fitness values of sparks, and by applying selection, strategy pick the selected sparks as new fireworks.
- If the terminal conditions are met, stop the algorithm. Otherwise, continue the iteration process from the third step.

COMPLEX (SOCIAL) NETWORKS

A complex network (CN) is a graph which has unique properties - usually in the real-world graph domain. A complex network contains features which are unique to the assigned problem. These features are important markers for a population used in Evolutionary/Swarm

based algorithms (Davendra et al. 2016). The following two described non-trivial features are important for a quick analysis of the network thus created.

Degree Centrality is defined as the number of edges connected to a specific node. Degree Centrality is an important distribution hub in the network since it connects - and thereby, distributes most of the information flowing through the network. Using Degree Centrality, we could analyze if stagnation or premature convergence is occurring within the population. From the graphs, it can be seen either that the multiple nodes are increasing (emphasizing their prominence in the population – helping in generating of the better individuals) or the degree centrality values are stagnating (i.e., no improvements in population).

The second important feature is the *average clustering coefficient*, which for the entire network, is calculated from every single local clustering coefficient for each node. The clustering coefficient of a node shows how concentrated the neighborhood of that node is. It is possible to assume that such a feature can show the population diversity, its compactness or tendency to form heterogeneous subgroups (subpopulations).

SOCIAL NETWORKS INSIDE HEURISTICS

In this research, the complex network approach is utilized to show the linkage between different individuals in the population. Each individual in the population can be taken as a node in the complex network graph, where its links specify the successful exchange of information in the population.

Since the internal dynamics and principles are different for evolutionary (DE) and swarm-based algorithms (PSO, FA, and FWA), several different approaches for capturing the population dynamics have been developed and tested.

In the case of the DE algorithm, an Adjacency Graph was used. In each generation, the node is only active for the successful transfer of information, i.e., if the individual is successful in generating a new better individual who is accepted for the next generation of the population. If the trial vector created from three randomly selected individuals (DE/Rand/1/Bin) is better than the active individual, one establishes the connections between the newly created individual and the three sources; otherwise, no connections are recorded in the Adjacency Matrix.

Although the Firefly algorithm is a swarm type, the situation here is very similar to the evolutionary algorithms. To create a network, we decided to visualize every firefly as a node. The connection between nodes is plotted for every successful interaction between fireflies. Successful interaction is defined as such interaction where one of the individuals gets improved. In the case of Firefly algorithm, it is when firefly flies towards another and improves own brightness.

For the PSO algorithm, the main interest is in the communications that lead to population quality improvement. Therefore, only communication leading to improvement of the particles personal best (pBest) was tracked. Details and several different workflows are given later in this paper.

The last studied algorithm (FWA) is the original representative of *random search/local search engine* type algorithm. We have shown, that even for this type, it is possible to develop a scheme for capturing the communication in the form of a graph. An interesting phenomenon has been discovered. The network seems to have a lack of any other usable information, besides the ability to identify the surface type of optimized function.

EXPERIMENT DESIGN

A simple Schwefel's Test function was used in this experimental research for the generation of a complex network.

Experiments were performed, and the data were analyzed and visualized using the *Wolfram Mathematica SW* suite. Within the scope of this research, only one type of experiment was performed. It utilizes the maximum number of generations fixed at 100 with a population size of $NP = 50$ for DE and 30 for swarm algorithms, due to the better clarity of visualizations and differences between evolutionary and swarm systems. Other parameters were set up exactly as recommended by literature. Since only one run of heuristic algorithms was executed for each particular case, no statistical results and comparisons are given here.

VISUALIZATIONS FOF DE

The visualizations of complex networks are depicted in Figures 1 - 3 containing Adjacency Graphs for the selected case-studies - snapshots (beginning of the optimization process, middle part and the end of simulation). Figures 4 – 6 are depicting the corresponding community plots.

The *Degree Centrality* value is highlighted by the size of the node (red color). Analysis of CN from DE algorithm can be found in (Skanderova and Fabian 2015; Skanderova et al. 2017; Viktorin et al. 2016; Senkerik et al. 2016; Viktorin et al. 2017).

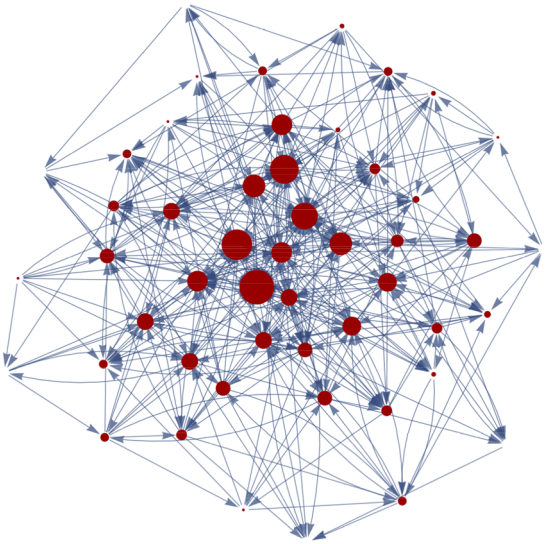


Fig. 1: CN for DE: the snapshot No.1. – the first 20 iterations.

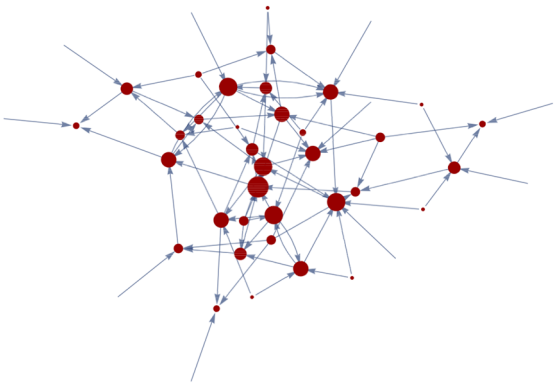


Fig. 2: CN for DE: the snapshot No.2. – the middle 20 iterations.

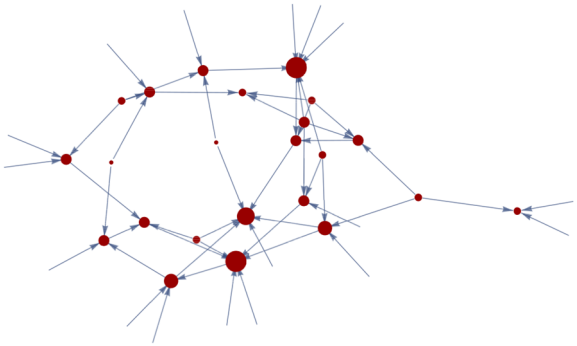


Fig. 3: CN for DE: the snapshot No.3. – the last 20 iterations.

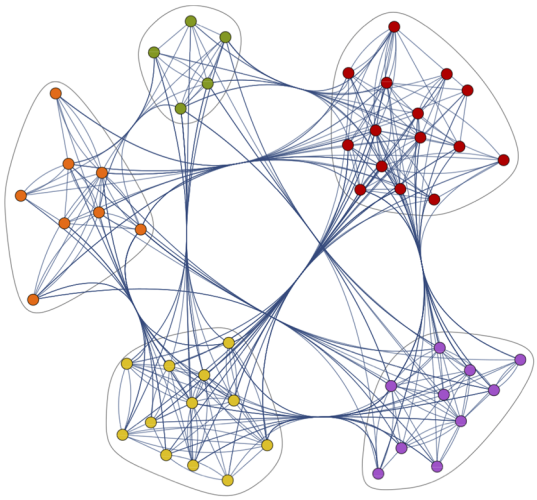


Fig. 4: Community plot for DE: the snapshot No.1. – the first 20 iterations.

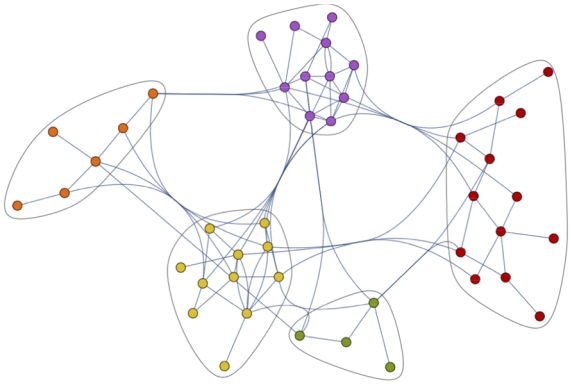


Fig. 5: Community plot for DE: the snapshot No.2. – the middle 20 iterations.

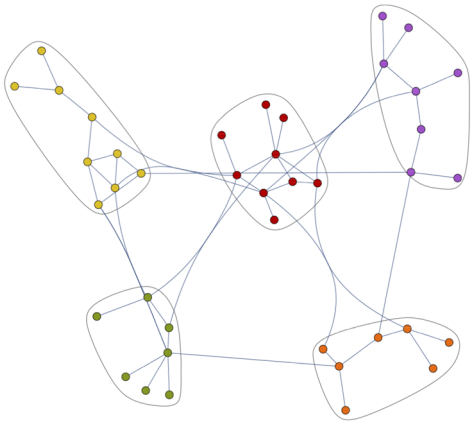


Fig. 6: Community plot for DE: the snapshot No.3. – the last 20 iterations.

VISUALIZATIONS FOR FA

In the case of FA, the connection is created, when firefly flies towards another and improves own brightness. This leads to network presented in Figure 7 (Janostik et al. 2016b). Duplicate connections were omitted.

Since across multiple iterations of the algorithm there may provide multiple connections between nodes we decided to improve upon the design by weighting the connections. If there is a connection between the firefly A and B, it starts with weight 1. If in another iteration there is another successful interaction between the firefly A and B, a new connection is not created, but the weight of the existing connection is incremented by 1. At the end of evolution, the weight is normalized. If the firefly gets improved by another in every iteration, at the end of the evolution their connection will have weight 1. If it never gets improved their connection will have weight 0. In Figure 8 we can see a network where connections have their weights visually distinguished. In the top left corner, we can see one dominant firefly which improved entire population more than 70% of iterations (blue lines). On the bottom right side we can observe few fireflies which took part in the improvement of the population only less than 30% of iterations (red lines). Also from the network, we can see that most of the fireflies improved one another only in between 30% and 70% of iterations.

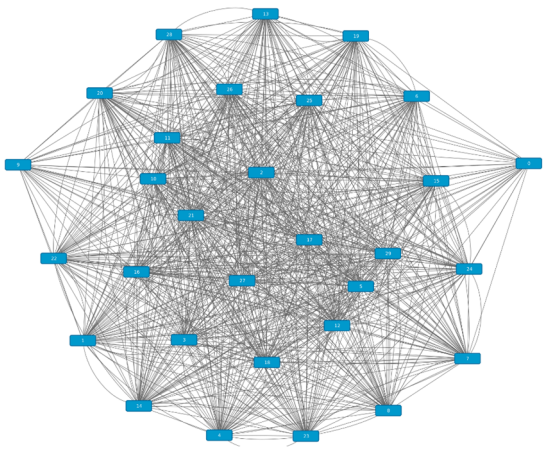


Fig. 7 Basic weighted oriented network for a population size 30 after 100 iterations.

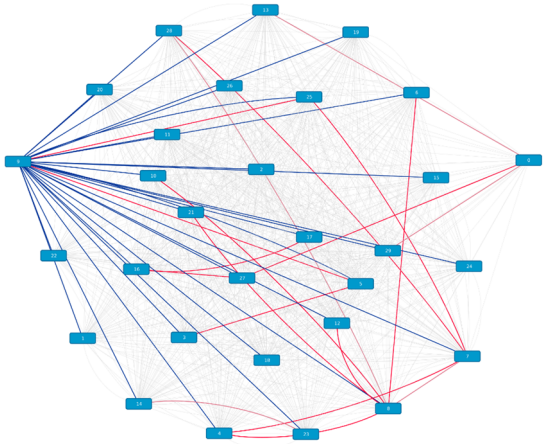


Fig. 8: a Basic weighted oriented network for a population of size 30 after 100 iterations with visually highlighted weights.

VISUALIZATIONS FOR PSO

The complex network for all iterations of the PSO algorithm that was created is depicted in Figure 9 (Pluhacek et al. 2016a). Nodes of a similar color represent particles with the same ID, and throughout different iterations. All links are from a particle that triggered the *gBest* update to a particle that has improved - based on that *gBest*. The nodes' code numbers represent a particle ID and its current iteration. This way, it is possible to precisely track the development of the network and the communication that occurs within the swarm. To be more precise, from a particular cluster, it can be observed that a single *gBest* update led to the improvement of multiple particles in different iterations. To capture the density of communication (Pluhacek et al. 2017a), the nodes in the network represent the particles in different time points (Particle ID with iteration code). This means that the theoretical maximal number of nodes in the network is the number of particles times the number of iterations. However, a new node in the network is created only when a particle manages to find a new personal best solution (*pBest*). When a node is created, two links are also created. The first link is between the newly created node and the previous node with the same particle ID (but different iteration code). This represents the information from *pBest*. Similarly the information from *gBest* represented by a link between the newly created node and a node that represents the last update of *gBest*. In the network visualizations (Figure 10) a color coding is used to differentiate the phases of the run as a percentage of the final number of cost functions evaluations (CFE). (The first 20% of CFE are represented by red color, magenta represents the 20-40% of CFE, green is the 40-60% CFE., 60-80% CFE is represented by yellow color and finally, the 80-100% CFE is represented as cyan). Such a representation can reveal the relations between the density of communication and convergence speed of the PSO.

Alternatively, it is possible to construct an Adjacency Graph and to benefit from its statistical features - as with the DE/FA case. The link is created between the particle that triggered the last *gBest* update and the particle that triggers a new *gBest* update. The self-loops (when a new *gBest* is found by the same particle as the previous *gBest*), are omitted. More studies aimed at PSO and CN framework are in (Pluhacek et al. 2016b; Pluhacek et al. 2017b).

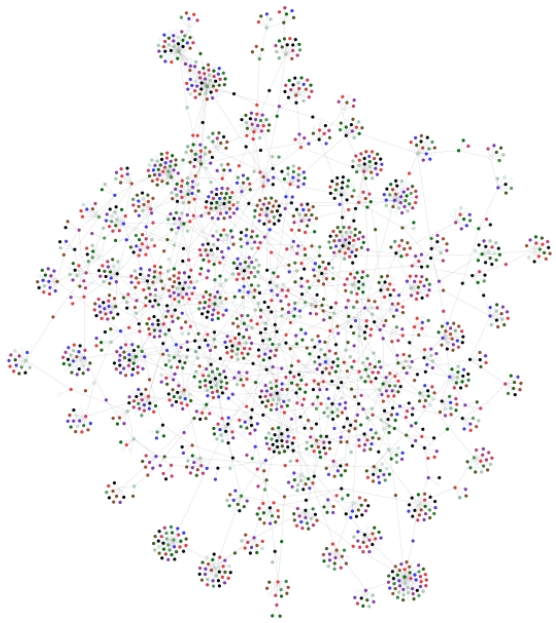


Fig. 9 PSO Dynamic as a Complex Network – Complete view (clusters).

VISUALIZATIONS FOR FWA

The network is created as a history of contributions. In each iteration, there are *NP* fireworks. These fireworks create *K* sparks. Some of these sparks are transferred into a new iteration as new fireworks. Fireworks are then represented as the nodes in the network. These nodes are labeled 1…*NP* for each iteration. The nodes (fireworks) are sorted by their fitness values before labeling so that the best node (smallest fitness value) gets number 1 and the worst node gets number *NP*. The edge between nodes represents spark that creates a new firework in next iteration. The initial node of the edge represents the firework from which the spark is created. The terminal node is the firework in the next iteration created by the spark. With that rule, the initial node from *t* iteration can have from 0 to *NP* edges, and the terminal node can only have one edge as input. The example is depicted in Figure 11 (Kadavy et al. 2017).

Fig. 10 PSO Dynamic as a Complex Network – Complete view (density of communication).

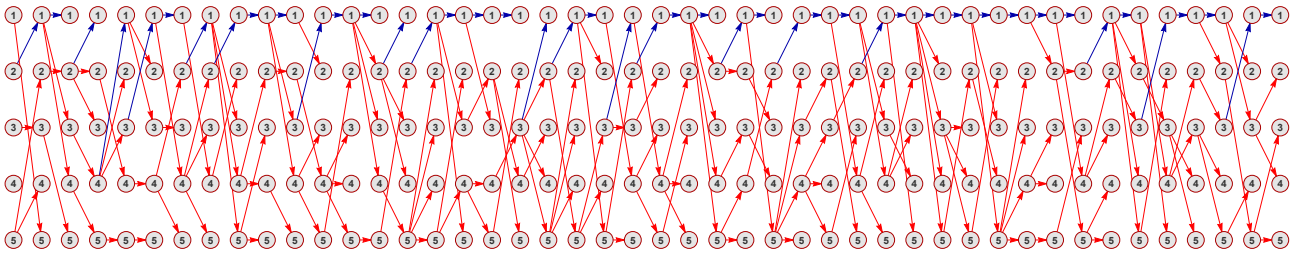


Fig. 11: FWA dynamics as a network - Blue edges indicate the spark with the best fitness function. The blue edge direction can only be towards the node number one. The first iteration is on the left side of the figure, and the last iteration is on the right side.

CONCLUSION

This work was aimed at the experimental investigation of the hybridization of a complex network framework using DE, PSO, FA and FWA algorithms. The population was visualized as an evolving complex network, which exhibits non-trivial features. These features provided a clear description of the population during evaluation.

The graphical and numerical data presented herein has fully manifested the influence of either time frame selection or type of construction to the features of the complex network. The findings can be summarized as follows:

The building of the Network: Since there is a direct link between parent solutions and offspring in the evolutionary algorithms, this information is used to build a complex network. In the case of swarm algorithms, the situation is a bit more difficult. It depends on the inner swarm mechanisms, but mostly, it is possible to capture the communications within the swarm during the updating of the information - based on the points of attraction. Several possible approaches are described herein, resulting in different graph visualizations and possible subsequent analyses.

Complex Network Features: A complex network created for evolutionary algorithms contains direct information about the selection of individuals and their success; therefore, many network features can be used for controlling a population during an EA run. At the beginning of the optimization process, intensive communication occurs (Figure 1). Later, hubs (centralities) and clusters are created (Figure 2), and it is possible to use such information either for the injection or replacement of individuals or to modify/alternate the evolutionary strategy. In the case of swarm algorithms, the communication dynamics are captured - thus the level of particle performance (usefulness) can be calculated, or some sub-clusters and centralities of such communication can also be identified - depending on the technique used for the transformation of swarm dynamics into the network.

Other Features – Fitness Landscape: Numerous previous experiments showed that there are no significant changes in complex network features for different test functions in the case of evolutionary algorithms. The capturing of communications (swarm dynamics) is sensitive to the fitness landscape. Thus, network features can be used for the raw estimation of a fitness landscape.

In this paper, we have reviewed several different approaches for visualizations, which can be, of course, hybridized and combined for any metaheuristic techniques according to user-defined requirements, which features are important to observe. Besides the presented approaches, more have been explored for a wider portfolio of algorithms (Tomaszek and Zelinka 2016; Kromer et al. 2015; Skanderova et al. 2014).

This novel topic has brought up many new open tasks, which will be resolved in future research. Another advantage is that this complex network framework can be used almost on any metaheuristic.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014), further by the European Regional Development Fund under the Project CEBIA-Tech no. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the Projects no. IGA/CebiaTech/2018/003. This work is also based upon support by COST (European Cooperation in Science & Technology) under Action CA15140, Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), and Action IC1406, High-Performance Modelling and Simulation for Big Data Applications (cHiPSet). The work was further supported by resources of A.I.Lab at the Faculty of Applied Informatics, Tomas Bata University in Zlin (ailab.fai.utb.cz).

REFERENCES

Skanderova, L., Fabian, T., Zelinka, I. (2016). Small-world hidden in differential evolution. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 3354-3361).

Das S., Mullick S.S., Suganthan P. (2016) Recent advances in differential evolution – An updated survey, Swarm and Evolutionary Computation, vol. 27, pp. 1–30.

Engelbrecht A (2010) Heterogeneous Particle Swarm Optimization. In: Dorigo M, Birattari M, Di Caro G et al. (eds) Swarm Intelligence, vol 6234. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 191-202.

Zelinka, I. (2016). SOMA—Self-organizing Migrating Algorithm. In Self-Organizing Migrating Algorithm (pp. 3-49). Springer International Publishing.

Karaboga, D. Basturk, B. (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, Journal of Global Optimization, 39 (3), pp. 459-471.

Fister I., Fister I. Jr., Yang X.S, Brest J., (2013) A comprehensive review of firefly algorithms, Swarm and Evolutionary Computation, Volume 13, Pages 34-46.

Zelinka I, Davendra D, Lampinen J, Senkerik R, Pluhacek M (2014) Evolutionary algorithms dynamics and its hidden complex network structures. In: Evolutionary Computation (CEC), 2014 IEEE Congress on, pp 3246-3251.

Davendra D, Zelinka I, Metlicka M, Senkerik R, Pluhacek M (2014) Complex network analysis of differential evolution algorithm applied to flowshop with no-wait problem. In: Differential Evolution (SDE), 2014 IEEE Symposium on, pp 1-8.

Skanderova, L., Fabian, T. (2015) Differential evolution dynamics analysis by complex networks. Soft Computing:1-15.

Metlicka, M., Davendra, D. (2015) Ensemble centralities based adaptive Artificial Bee algorithm. In: Evolutionary Computation (CEC), 2015 IEEE Congress on, pp 3370-3376.

Gajdos P, Kromer P, Zelinka I (2015) Network Visualization of Population Dynamics in the Differential Evolution. In: Computational Intelligence, 2015 IEEE Symposium Series on, pp 1522-1528.

Janostik J, Pluhacek M, Senkerik R, Zelinka I (2016a) Particle Swarm Optimizer with Diversity Measure Based on Swarm Representation in Complex Network. In: Abraham A, Wegrzyn-Wolska K, Hassanien EA, Snasel V, Alimi MA (eds) Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015. Springer International Publishing, Cham, pp 561-569.

Tan Y., Zhu Y. (2010) Fireworks Algorithm for Optimization. In: Tan Y., Shi Y., Tan K.C. (eds) Advances in Swarm Intelligence. ICSI 2010. Lecture Notes in Computer Science, vol 6145. Springer, Berlin, Heidelberg

Yang, X.S.: Nature-inspired metaheuristic algorithms. Luniver Press, Frome, U.K. (2010).

Skanderova, L., Fabian, T., Zelinka, I. (2017). Differential Evolution Dynamics Modeled by Longitudinal Social Network. Journal of Intelligent Systems, 26(3), 523-529.

Viktorin, A., Pluhacek, M., Senkerik, R. (2016). Network based linear population size reduction in SHADE. In Intelligent Networking and Collaborative Systems (INCoS), 2016 International Conference on (pp. 86-93).

Senkerik, R., Viktorin, A., Pluhacek, M., Janostik, J., Davendra, D. (2016). On the influence of different randomization and complex network analysis for differential evolution. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 3346-3353).

Viktorin, A., Senkerik, R., Pluhacek, M., Kadavy, T. (2017). Towards better population sizing for differential evolution through active population analysis with complex network. In Conference on Complex, Intelligent, and Software Intensive Systems (pp. 225-235).

Janostik, J., Pluhacek, M., Senkerik, R., Zelinka, I., Spacek, F. (2016b). Capturing inner dynamics of firefly algorithm in complex network—initial study. In Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 (pp. 571-577).

Pluhacek, M., Janostik, J., Senkerik, R., Zelinka, I., Davendra, D. (2016a). PSO as complex network—capturing the inner dynamics—initial study. In Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 (pp. 551-559).

Pluhacek, M., Senkerik, R., Janostik, A. V. J., Davendra, D. (2016b). Complex network analysis in PSO as an fitness landscape classifier. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 3332-3337).

Pluhacek, M., Senkerik, R., Viktorin, A., Kadavy, T. (2017a). Uncovering communication density in PSO using complex network. In Proceedings-31st European Conference on Modelling and Simulation, ECMS 2017. European Council for Modelling and Simulation.

Pluhacek, M., Viktorin, A., Senkerik, R., Kadavy, T., Zelinka, I. (2017b). PSO with Partial Population Restart Based on Complex Network Analysis. In International Conference on Hybrid Artificial Intelligence Systems (pp. 183-192).

Kadavy, T., Pluhacek, M., Viktorin, A., Senkerik, R. (2017). Firework algorithm dynamics simulated and analyzed with the aid of complex network. In Proceedings-31st European Conference on Modelling and Simulation, ECMS 2017. European Council for Modelling and Simulation.

Tomaszek, L., Zelinka, I. (2016). On performance improvement of the SOMA swarm based algorithm and its complex network duality. In Evolutionary Computation (CEC), IEEE Congress on (pp. 4494-4500).

Krömer, P., Gajdos, P., Zelinka, I. (2015). Towards a Network Interpretation of Agent Interaction in Ant Colony Optimization. In Computational Intelligence, 2015 IEEE Symposium Series on (pp. 1126-1132).

Skanderova, L., Zelinka, I., Saloun, P. (2014). Complex Network Construction Based on SOMA: Vertices In-Degree Reliance on Fitness Value Evolution. In ISCS 2013: Interdisciplinary Symposium on Complex Systems (pp. 291-297). Springer Berlin Heidelberg.

MAPPING OF ENCLOSED BUILDINGS USING MOBILE RADIO TOMOGRAPHY

Anastasia Ingacheva
Institute for Information
Transmission Problems RAS
B. Karetny per. 19, Moscow 127051,
Russia
National Research University Higher
School of Economics, Moscow 101000,
Russia
E-mail: ingacheva@gmail.com

Vladislav Kokhan
Institute for Information
Transmission Problems RAS
B. Karetny per. 19, Moscow 127051,
Russia
E-mail: vladkohan@list.ru

Dmitry Osipov
Institute for Information
Transmission Problems RAS
B. Karetny per. 19, Moscow 127051,
Russia
National Research University Higher
School of Economics, Moscow 101000,
Russia
E-mail: d_osipov@iitp.ru

KEYWORDS

Mobile radio tomography, convex optimization, regularization, simulation, robotics.

ABSTRACT

In this paper we consider the task of inner objects mapping for the building with a bunch of moving around it autonomous agents which use narrow beam of radio waves using WiFi frequency (2.4 GHz). Linear model of pixel-wise radio waves attenuation is considered. SIRT algorithm with TV and Tikhonov regularizations is used for the task of tomography reconstruction. Properties of the presented model are studied during simulation using synthetic data consisting of 8 buildings with inner object with different shapes. Dependency between mapping quality and transmission power is found. Simulation results confirm suggested approachs usability.

INTRODUCTION

In real life there are situations when military and special services need to determine positions of objects and people inside the building without possibility of entering. We would like to solve this problem with a bunch of mobile transmitters and receivers which locate around the building and send signals through it to each other. Method needs to be adequately protective of human health and the environment at the same time, therefore we cannot use the excessively hard and powerful radiation. The result of sounding and reconstruction of the building is a two-dimensional map of items locations (walls, objects, people) inside it, so this task is called the mapping an inaccessible building. Since the height of sounding sensors is fixed the resulting map is two-dimensional.

It is necessary to use a signal capable of passing through objects and spreading over long distances to build a solution of this problem. Methods based on radio waves are suitable for the task because it is important not to cause harm to environment and people. The narrow beam of high-frequency radio waves is preferable to use, since they do not slide around obstacles but travel through them being partially absorbed and reflected. This property of radio waves can be used for determining the characteristics of obstacles.

The most studied method of mapping is based on ground-penetrating radar. It uses principles of active radar properties, i.e. measurement of parameters (delay, phase, etc.) of the signal that is reflected from observed object. There are commercial systems based on this approach: PulsON (Time Domain Corporation 2016), ImpSAR (Eureka Aerospace 2009). An example of such systems is the ThruMapper (Tan et al. 2017). The ThruMapper system is a mobile platform which moves along the main corridor of the building and maps the building without entering rooms. The main disadvantage of this method is the exponential growth of the required signals power with increase in the thickness of walls and other charted objects.

Another group of methods for mapping is based on inverse tomography task. Diffraction and transmission radio tomography is used to solve the problem of mapping when it is possible to make a set of soundings from various positions around the building. Diffraction radio tomography is based on measurements of reflected electromagnetic waves parameters. Simulated experiments of diffraction tomography are presented in (Bardak and Saed 2014), possibility of mapping real data

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

is shown in (Sukhanov et al. 2015). Diffraction radio tomography requires to collect projection data frequently, which makes the method inapplicable in real conditions, despite the high detail of the information on the resulting maps. Also diffraction radio tomography has same problem as ground-penetrating radars (the exponential growth of the required signals power with increase in the thickness of walls and objects).

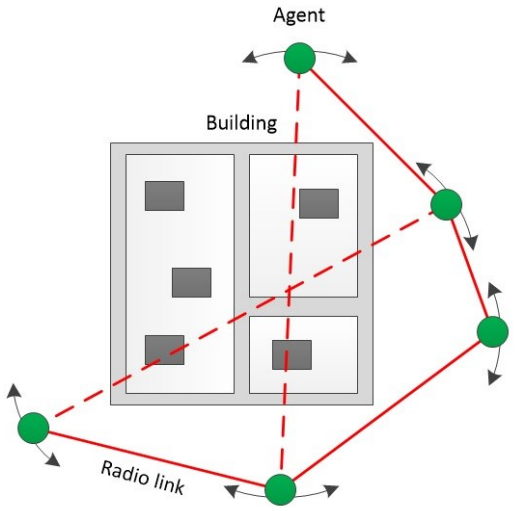
Radio tomography methods are based on absorption of radiation are divided into radio tomography and mobile radio tomography. In the radio tomography method network of stationary transceivers located around the observable area is used. With these transceivers the projection data is collected and then the inverse tomography problem is solved (Wilson and Patwari. 2010, Wilson and Patwari. 2011). The main disadvantage of this method is a fixed number of transceivers, which must be set up at pre-defined positions at the beginning of the measurements. Method of mobile radio tomography lacks this disadvantage, since it uses a network of mobile transceivers that allows to adjust the coverage of the observed zone if necessary (Van der Meij 2016, Helwerda 2016, Batenburg et al 2016). The usage of a mobile platform for radio tomography brings new challenges. For example, one needs to create the route of travel for transceivers, maintain precise positioning between them and keep the channel of data transfer up. However, method of mobile tomography makes it possible to obtain much more projection data, which makes it perspective for more accurate mapping of inaccessible building.

It would be feasible to use autonomous unmanned mobile tomographic robots, i.e. wheeled robots, if we considered the distinctive features of practical application (for special and military operations). At least two (transmitter and receiver) mobile agents for the purposes of transmission radio tomography must be used. It is better to use several pairs of mobile devices simultaneously for faster mapping which are capable to continue work even when a part of them is out of order. So, we are talking about a group of robots without centralized control, that perform remote mapping of inaccessible building together. The collaborative behaviour of a similar robots group for mapping the open area was considered in work (Shvets et al. 2015). Since the interaction of a group of mobile agents in practice is performed through WiFi (frequency 2.4GHz), it seems reasonable to choose this frequency for sounding, although we can use any other waves frequency of the microwave band that can travel through obstacles.

A distinctive feature of the application of mobile radio tomography to the considered task (especially using a group without centralized control) is non-fixed step of the sounding, which makes it impossible to use integral algorithms to find solution of the tomography problem. So we shall use algebraic methods then (Buzmakov et al. 2017, Ingacheva et al. 2017). A similar approach based on applying algebraic methods for mapping rooms with sonars is proposed in paper (Shvets et al. 2014). In this work the linear equations systems of large dimension are solved by continuous optimization methods.

SIMULATION MODEL

The model of mobile radio tomography we used in our experiments is shown in fig.1. It includes several mobile unmanned agents (tomography set-ups) which can emit and receive radio signals. They move around the building collecting projection data in certain positions and record the signal that has been partially attenuated along given direction (dashed line in Fig. 1). Solution of the inverse radio tomography problem can be found based on the collected projection data. The accuracy of resulting map of the building directly depends on the number of projections.



Figures 1: Radio tomography model.

It is necessary to accurately set the receiver and the transmitter, since it is required to know exact positions of both agents to solve inverse radio tomography problem. We consider that agents interact with each other using WiFi frequency (straight lines in Fig. 1) which is also used for sounding building.

Radio signal formation

For evaluation of received by agent radio signal power we use linear model of attenuation (Wilson and Patwari, 2010). This model relies only on dielectric properties of the objects the signal passes, but not considering the angle between the radio wave and object. Using this model, we can split entire mapping space into a separate pixels and work with them independently.

We can present power of registered signal P(l) for radio wave l as difference between the transmitted power P_t and summation of the losses in every pixel L(l), the losses of radio signal propagation L_m and noise η:

$$P(l) = P_t\text{-}L(l)\text{-}L_m\text{-}\eta\ . \qquad (1)$$

The losses of radio waves propagation L_m depend on the emitted signal power, coefficients of signal empowerment of transmitter, receiver and the distance between them. This distance is constant in our simulation model, thereby it is equal to the size of mapping space N. Total loss of radio wave can be written as a loss in every pixel multiplied by weight of the corresponding pixel:

$$L(l) = \sum_{i,j=1}^{N} w_{ij}(l) L_{ij}, \quad (2)$$

where $w_{ij}(l)$ is weight of pixel in coordinate (i, j) for radio wave l, L_{ij} is loss in corresponding pixel.
We can separate loss of radio wave into 2 types: loss during pass through dielectric materials and loss during pass of free space:

$$L_{ij} = \begin{cases} Lv_{ij}, & (i,j) \in Object; \\ 0, & otherwise. \end{cases} \quad (3)$$

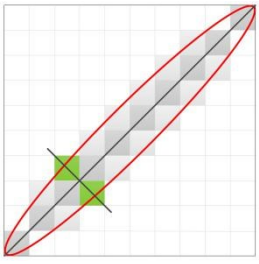
Loss in dielectric material can be written as:

$$Lv_{ij} = 10\lg(Pn_{ij}) - 10\lg(P_v(\Delta d)_{ij}), \quad (4)$$

where Pn_{ij} is radio wave power which crosses the border between different objects with different resistances in pixel (i, j), $P_v(\Delta d)_{ij}$ is radio wave power after pass through dielectric material of thickness Δd (in our case Δd is the size of a pixel and it equals 10 cm). Thereby loss during pass through pixel (i, j), can be calculated as:

$$P_v(\Delta d)_{ij} = (Pn_{ij} - Pr_{ij})A(\Delta d), \quad (5)$$

where Pr_{ij} is reflected part of the power, $A(\Delta d)$ – attenuation coefficient according to the interference with environment, which depends on Δd exponentially.
We use ellipsoidal model to determine weight of a pixel. Every matrix of weights defines only one radio wave. If we know positions of emitter and receiver we can describe every radio wave with this matrix. For example, on fig 2. we can see a matrix for a wave within ellipsoidal model.



Figures 2: Method of calculating radio wave weight.

This matrix can be also written as equation:

$$w_i = \begin{cases} \frac{1}{n}, i \in link; \\ 0, otherwise. \end{cases} \quad (5)$$

Where n is the number of pixels along the smallest eigenvector of the ellipse (green squares in Fig.2). Ellipsoidal model describes real life case better than a straight-line radio wave model since the wave forms an ellipsoid of revolution (spheroid) during propagation called Fresnel zone.

TOMOGRAPHIC RECONSTRUCTION

We consider the problem of mobile radio tomography as a system of linear equations

$$Wx = p. \quad (6)$$

where W is weight matrix in which the line corresponds to a radio wave, x is a column vector of an unknown region function of N × N pixels, p is a column vector of the obtained projection data. A solution of the system (6) is found using one of iterative algorithms of nonlinear optimization (Nocedal and Wright 1999). The standard way to solve system (6) is to minimize the follow expression:

$$\|Wx - p\|_2^2 \to \min_x. \quad (7)$$

The advantage of algebraic methods is its adaptability the particular task. For example, a priori knowledge about reconstruction function can be included in system (6). One way to modify the algorithm is to add a problem-specific regularization term.
TV-regularization (Total variation regularization) and Tikhonov regularization is used in this paper. TV-regularization provide the smoothness of an unknown function, so the minimal difference between neighbouring pixels should be achieved. Penalty for large value of functions gradient is imposed as ($\|\nabla x\|$) (Estrela et al. 2016). Tikhonov regularization is necessary to ensure that there will be no extremely large values in the unknown function. Penalty for L2 norm of unknown function is imposed as ($\|x\|^2$) (Gockenbach 2016). The final equation of the minimized functional with both regularizations can be written as:

$$\|Wx - p\|_2^2 + \alpha\|x\|_2^2 + \beta\|\nabla x\|_1 \to \min_x . \quad (8)$$

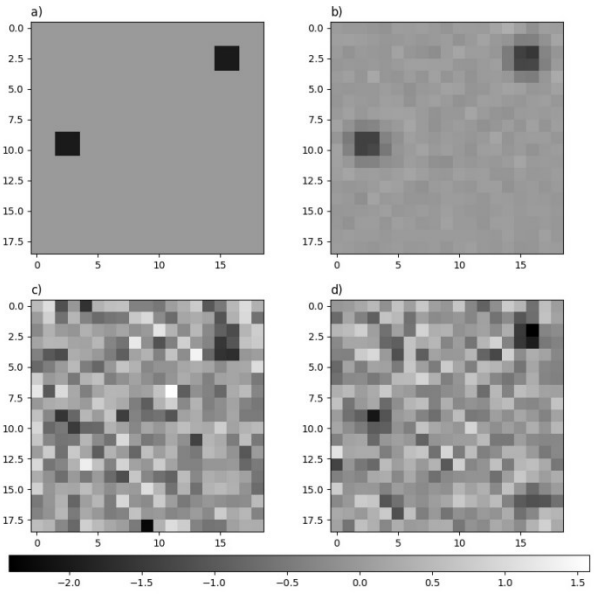
The numbers α and β are the configurable parameters of the algorithm. The steepest gradient method was chosen for minimization the functional (8).

NOISE MODEL

To add noise in our model we use the results of the research described in the paper (Ganesh and Pahlavan 1990). Author of this paper studied the approximation of noise distribution in the propagation of radio waves in buildings. They approximated the noise for radio wave by the log-normal distribution $lnN(\mu, \sigma)$, with mathematical expectation $\mu = 0$.
Since we use power in decibels in our model, we need to use a normal distribution.

COMPARISION WITH REAL DATA

To validate our simulation algorithm, we performed comparison of results from suggested approach with real data from paper (Batenburg et al 2016). In this work Dutch scientists performed a set of experiments with people in the building room without any obstacles. One of their experiment is illustrated in Fig. 3 a). The black squares in the image correspond to people, grey colour represents open area. The size of the grid 19x19 pixels, where one pixel is 16.75 cm2 square area. Using this simulation, we have calculated projection data with the same configuration of agents. Then we have select noise parameters and reconstruct obtained projection data with 100 iterations. The selected standart deviation $\sigma = 6.3$. We have not added any regularization to save real problem-specific noise in reconstructions. Examples of reconstruction data taken from the real experiment and our simulation using ellipsoidal weights models are shown in fig. 3.



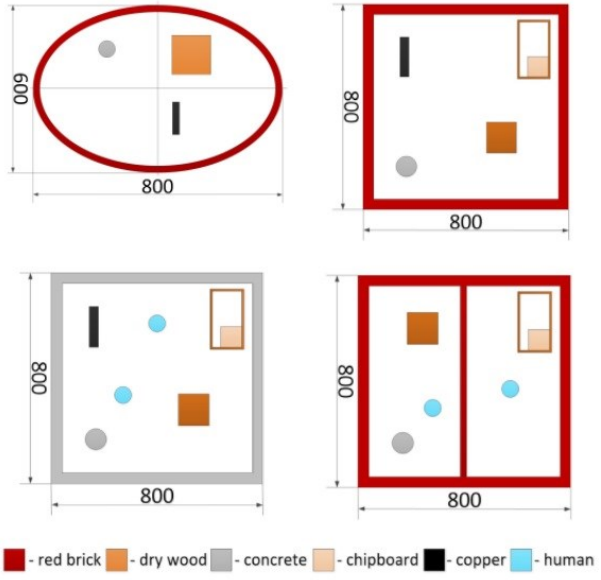
Figures 3: Real data comparison: a) model object, b) simulation data without noise reconstruction, c) real data reconstruction, d) simulation data with noise reconstruction.

To evaluate the quality or the similarity of experiment and simulation projection data we calculate mean square error (MSE). The MSE between real data and simulation without noise is 0.029, when for our noisy and not noisy data MSE is 0.024. From this comparison it is obvious that elliptical model with Fresnel zones and Log-normal noise distribution describes well conditions of real experiments enough.

EXPERIMENTAL SETUP

Dataset

For our synthetic data we have used five common materials in building: bricks, concrete, wood, chipboard and copper (any metal reflecting radio waves can be used). In some cases, we have added model of human body. Using these materials, we have created a dataset with 8 objects for studying possible applications of radio tomography methods in task of mapping enclosed buildings. Several images of objects from our dataset are shown in fig.4.



Figures 4: Examples of images from our dataset.

Mapping quality estimation

For quality measurement of reconstructed space, we calculated total error for every object from dataset using distance in L2 norm divided by N^2 - total number of pixels in target space:

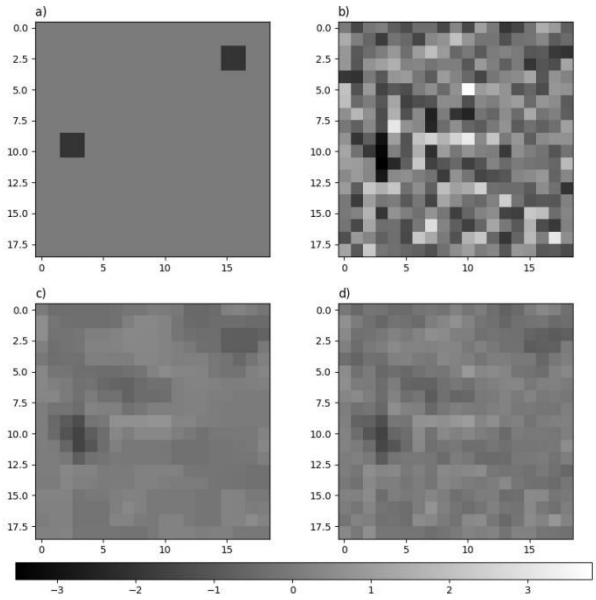
$$E = \sum_{z=1}^{8} \Big/ N^2 , \qquad (9)$$

where z is object number, M_{ij}^{z} is a pixel from synthetic object, R_{ij}^{z} is reconstructed pixel. Size of both simulated and reconstructed spaces is $\{N \times N\}$.

EXPERIMENTAL RESULT

According to the comparison of results against real data, it can be seen that existing WiFi devices for home usage are not suitable for mapping inaccessible buildings. The result of reconstruction is noisy and it is not possible to separate standalone objects from the background.

However, hardware-based and software-based approaches can be applied for noise reduction. The software approach is based on a modified spatial reconstruction algorithm, with addition of Tikhonov regularization and TV-regularization as noise suppressors. The experimental results show that Tikhonov regularization does not improve the quality of reconstructed map, because the noise in the projection data does not contain any sticking out values. The reconstruction results with different type of regularization are shown in Fig. 5. One can see structural similarity and different loss level in Fig. 5b) and Fig. 5d).



Figures 5: Reconstruction with different regularization type: a) model object, b) without regularization, c) TV regularization, d) Tikhonov regularization.

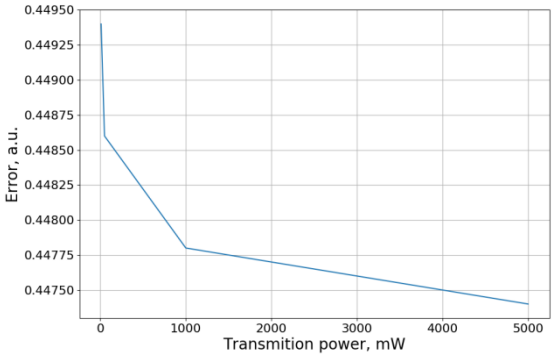
On the contrary, TV-regularization suppresses smoothly distributed noise very well (Fig. 5c)). We chose the regularization parameter for the visual noise on reconstruction to be as low as possible, while keeping the boundaries of objects visually distinguishable. Thus, for the dataset represented before (see Fig. 4) we took the parameter for TV-regularization equal to 5 and the parameter for Tikhonov regularization is 0.

The hardware approach of noise suppression is based on tuning the properties of receivers and transmitters. We have conducted experiments to demonstrate how increase in the transmitter power affects the reconstruction quality. For all model objects from our dataset we have calculated the projection data at 60 uniformly distributed rotation angles, the spatial step on each angle position has been chosen as minimal as possible (in each pixel). The reconstruction of projection data was done with 100 iterations in each experiment. Total error has been calculated by equation (9). The results of the experiment are shown in Table 1.

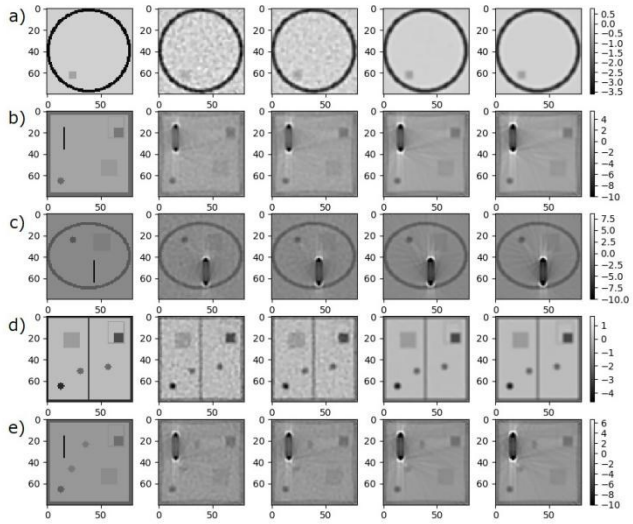
Table 1: Reconstruction and total error with different transmission power.

Object number \ Power	10 mW	50 mW	1W	5 W
1	0.0192	0.0147	0.0128	0.0122
2	0.0058	0.0056	0.0053	0.0052
3	0.7119	0.7114	0.7106	0.7102
4	0.7128	0.7124	0.7116	0.7112
5	0.7123	0.7119	0.7112	0.7108
6	0,7125	0,7122	0,7114	0,7110
7	0,0087	0,0085	0,0080	0,0078
8	0,7123	0,7119	0,7112	0,7109
Total	0,4494	0,4486	0,4478	0,4474

The curve of the dependence between reconstructed building map quality and the transmitter power is shown in Fig 6. Reconstructions of five model objects are shown in Fig. 7. Dependency between mapping quality and initial power is very similar for different values of regularization parameters in reconstruction algorithm.



Figures 6: Dependencies between initial power and reconstruction quality.



Figures 7: Different levels of power. Each line represents the same model object with different initial power.

One can see from the experimental graph that the quality of the mapped area is almost exponentially proportional to the power of the transmitter. It should be also noted that the quality of the reconstructed map is lower than building contains metal parts. The close objects to copper part are not selected on map (Fig. 7 b, c, e). Thus, the proposed radio tomographic algorithm is well suited for mapping of buildings that does not have metal elements.

CONCLUSION

In this paper we have demonstrated radio tomography technique usage possibility in the task of mapping enclosed buildings using simulated experiments. After the absorption model simulation accuracy of radio waves was compared with results from work (Batenburg et al 2016). The data acquirement process was considered as a set of consecutive steps, i.e. signals interference from different simultaneous projections was not counted. Therefore, there is also an interesting problem about possibility of receiving signals from more than one agents at the same time for faster mapping that was not consider in this paper.

Dependencies representing the reconstruction quality as a function of signal strength of the transmitter was obtained. The results of the experiment show that applying method of mobile radio tomography is well suited for mapping inaccessible building without metal items in their construction.

Experimental results also show that suggested approach is usable to map buildings in real life. Accuracy of inner objects map reconstruction is sufficient for usage in military or civilian operations.

ACKNOWLEDGMENT

The research was supported by the Russian Science Foundation grant (project No. 14-50-00150).

REFERENCES

Bardak C., Saed M. 2014. “Microwave imaging with a time-reversed finite-difference time-domain technique”. Journal of Electromagnetic Waves and Applications, V. 28, №. 12, 1455-1467.

Batenburg K. Joost, Helwerda L., Walter A. Kosters en Tim van der Meij. 2016. "Agents for Mobile Radio Tomography". In: Proceedings of the 28th Benelux Conference on Artificial Intelligence, 17-24.

Buzmakov Alexey, Anastasia Ingacheva, Victor Prun, Dmitry Nikolaev, Marina Chukalina, Claudio Ferrero and Victor Asadchikov. 2017. “Analysis of Computer Images in the Presence of Metals”. The 10th International Conference on Machine Vision, Vienna, Austria, 13-15 November 2017, http://icmv.org/, SPIE (In Press).

Estrela Vania V., Hermes Aguiar Magalhaes, Osamu Saotome. 2016. “Total Variation Applications in Computer Vision”. CoRR. V.abs/1603.09599.

Eureka Aerospace. 2009. “Impulse synthetic aperture radar: through-the-wall and underground imaging”. URL: http://www.eurekaaerospace.com/content/impulse-synthetic-aperture-radar-through-wall-and-underground-imaging (data of access: 31.10.2017).

Ganesh R., K. Pahlavan. 1990. “Effects of Traffic and Local Movements on Multipath Characteristics of an Indoor Radio Channel”. IEEE Electronics Letters, V. 26, № 12, 810-812.

Gockenbach Mark. 2016. “Linear Inverse Problems and Tikhonov Regularization”. The Mathematical Association of America, 333.

Helwerda L. 2016. “Mobile radio tomography: Autonomous vehicle planning for dynamic sensor positions”. Master’s thesis. LIACS, Universiteit Leiden.

Huang Y., Boyle K. 2008. “Antennas: from theory to practice”. John Wiley & Sons.

Ingacheva Anastasia, Marina Chukalina, Timur Khanipov, Dmitry Nikolaev. 2017. “Blur Kernel Estimation with Algebraic Tomography Technique and Intensity Profiles of Object Boundaries”. The 10th International Conference on Machine Vision, Vienna, Austria, 13-15 November 2017, http://icmv.org/, SPIE (In Press).

Nocedal Jorge, Wright Stephen J. 1999. “Numerical Optimization”. Springer. 634.

Shvets E. A., Nikolaev D. P. 2015. “Complex approach to long-term multi-agent mapping in low dynamic environments”. Proceedings SPIE. Eighth International Conference on Machine Vision (ICMV 2015), V. 9875, 98752A, 1-10.

Shvets E. A., Shepelev D., Nikolaev D. P. 2014. “Occupancy grid mapping with the use of a forward sonar model by gradient descent”. Journal of Communications Technology and Electronics, V. 61, №12, 1474-1480.

Sukhanov D. Ya.. Zavialova K. V. 2015. “Trekhmernaya radiotomografiya obyektov skrytykh za dielektricheski neodnorodnymi pregradami”. Zhurnal tekhnicheskoy fiziki. V. 85, №. 10, 115-120.

Tan B., Chetty K., Jamieson K. 2017. “ThruMapper: Through-Wall Building Tomography with a Single Mapping Robot”. Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications (ACM), 1-6.

Time Domain Corporation. 2016. “PulsON 440 (P440)”. URL: http://www.timedomain.com/products/pulson-440/ (date of access: 31.10.2017).

Van der Meij T. 2016. “Mobile radio tomography: Reconstruction and visualization of wireless sensor

networks with dynamically positioned sensors". Master's thesis, Leiden University.
Wilson J., Patwari N. 2010. "Radio tomographic imaging with wireless networks". IEEE Transactions on Mobile Computing, V. 9, №5, 621-632.
Wilson J., Patwari N. 2011. "See-through walls: Motion tracking using variance-based radio tomography networks". IEEE Transactions on Mobile Computing, V. 10, №. 5. – C. 612-621.

AUTHOR BIOGRAPHIES

ANASTASIA INGACHEVA was born in Kazan, Russia and went to Kazan (Volga region) Federal University, where she studied Economic Cybernetics and obtained her degree in 2012. She obtained master's degree of Higher School of Economics (National Research University) at computer science faculty in 2015. Now Anastasia is a PhD student in the same place. Since March 2013 Anastasia has been worked in the Vision Systems laboratory of the Institute for Information Transmission Problems RAS. Her research activities are in the areas of poly- and monochromatic X-Ray computed tomography and analysis of obtained CT data. Her e-mail address is: ingacheva@gmail.com and her Web-page can be found at http://tomo.smartengines.biz/peoples.en.html.

VLADISLAV KOKHAN was born in Moscow, Russia and went to the Moscow Aviation Institute (National Research University), where he studied radio engineering and obtained his bachelor degree in 2017. He is currently working toward the master degree in the Moscow Aviation Institute and working as a intern researcher in the Vision Systems laboratory of the Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute). His research interests include radio tomography, image processing and nonlinear optimization. His e-mail address is vladkohan@list.ru.

DMITRY OSIPOV was born in Moscow, USSR. He graduated from Bauman Moscow State Technical University in 2003 and obtained his Ph. D. from the Institute for Information Transmission Problems of the Russian Academy of Sciences in 2008. He is currently a senior researcher at the Institute for Information Transmission Problems of the Russian Academy of Sciences and Deputy Head of the Department at the National Research University Higher School of Economics. His research interests include: wireless communications, multiple access and channel models research and development. His e-mail address is d_osipov@iitp.ru.

ON A NOVEL SEARCH STRATEGY BASED ON A COMBINATION OF PARTICLE SWARM OPTIMISATION AND LEVY-FLIGHT

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

KEYWORDS
Particle Swarm Optimisation, Levy-flight, Submarine Groundwater Discharge, Swarm Intelligence.

ABSTRACT

The long term goal of this research is to develop a flexible, low-cost and autonomous platform for submarine exploration. Such a platform can be used for locating submarine points of interest. The search for submarine groundwater discharges (SGD) in coastal waters is one of the possible applications for such an observatory. The swarm should be guided by a search strategy. In this research a novel search algorithm based on Particle Swarm Optimisation and inertia Levy-flight is presented. It was demonstrated using a computer simulation that the novel algorithm is capable of improving the search performance compared to the performance of a swarm of homogenous particles or inertia Levy-flight to guide the search of the swarm to locate a source of submarine groundwater discharge.

INTRODUCTION

The aim of this project is to utilize a swarm of autonomous underwater vehicles (AUV) to develop a low cost and flexible environmental observatory. The search for submarine groundwater discharges (SGD) in coastal waters is one of the possible applications for such an observatory. Marine scientists are interested in locating and analysing these discharges because the nutrients discharged by SGD have a significant influence on the marine ecosystem (Dugan, et al., 2010; Moore W. , 2010; Nelson, et al., 2015).
AUVs can be used for the exploration of medium sized areas and measure some parameters, for example conductivity, temperature or nutrients to locate a SGD (Zielinski, et al., 2009). The interaction of the swarm must be managed by a search strategy. In this research, the behaviour of the swarm will be guided by a combination of particle swarm optimisation (Nolle, 2015) and Levy-flight (Tholen, et al., 2018).

Submarine Groundwater Discharge

Submarine groundwater discharge (SGD) consists of a flow of fresh groundwater and the recirculation of seawater from the sea floor to the coastal ocean (Moore W. , 2010). The fresh water and the sea water discharges commingle in the so-called mixing zone (Figure 1) (Evans & Wilson, 2016).

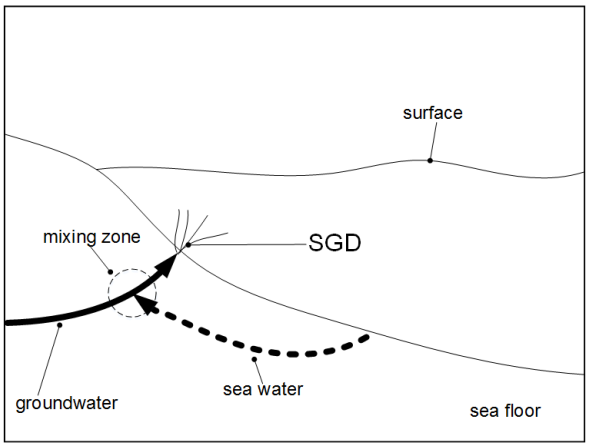


Figure 1: Submarine Groundwater Discharge of Fresh- and Recirculating-Water, modified after Evans and Wilson (2016)

Test Environment

The aim of the AUV is to localise a point of interest, for example SGDs. There is a constant input of different substances, i.e. nutrients or fluorescent dissolved organic matter (FDOM) (Nelson, et al., 2015) due to SGDs to the marine environment. While substances are discharged into the ocean, the concentration of these substances will be a function of the position and the time because of mixing processes in the water (Stedmon, et al., 2010). Dynamic behaviour could be simulated while using numerical models (Evans & Wilson, 2016) or a model based on cellular automata (Tholen, et al., 2017). However for fair comparison between the different algorithms, a static fitness function is used during this research. The conductivity in milli-Simens per centimeter (mS/cm) was chosen as tracer, to describe the distribution of the water mass inflow on a SGD in this work. The values of the conductivity are based on the measurements at the Black Point SGD in Maunalua Bay Hawaii (Richardson, et al., 2017).

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

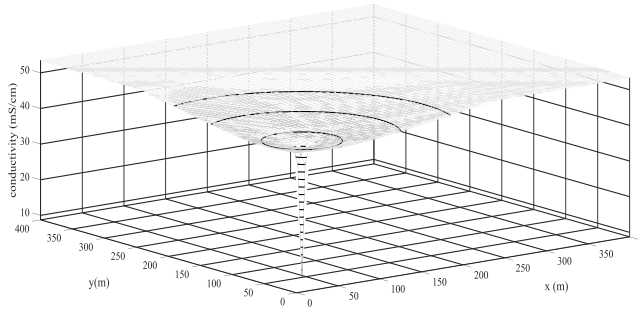


Figure 2: Fitness function

The value of the conductivity in the search space is simulated as follows:

$$f(x) = a * e^{b*x} + c * e^{d*x} \quad (1)$$

Where:
f(x): conductivity in mS/cm,
x: Euclidean distance between AUV and SGD,
a: scale parameter (45.62 mS/cm),
b: gradient parameter (0.00079),
c: scale parameter (-36.88 mS/cm),
d: gradient parameter (-0.3896),
max: maximum conductivity (53.42 mS/cm).

The search area is limited to 400 m x 400 m. The position of the SGD was set to position x_S = (50 m / 50 m). Figure 2 shows the shape of the described fitness function. As shown, gradient information are only sparsely available. This makes it a difficult test environment for direct search algorithms. During the search, the AUV moves through the search space and measure the conductivity after each second. To make the simulation more realistic, noise were added to the values of the fitness function (1), using a normal distribution with the measurement value as mean value and a standard deviation of 0.2 mS/cm. Furthermore, the measurement accuracy of the AUV is limited to a value of 0.01 mS/cm.

Particle Swarm Optimisation

PSO is modelled on the behaviour of collaborative real world entities (particles), for example fish schools or bird flocks, which works together to achieve a common goal (Kennedy & Eberhart, 1995; Bansal, et al., 2011). 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) (2):

$$\vec{v}_{i+1} = \vec{v}_i\omega + r_1c_1(\vec{p}_b - \vec{p}_i) + r_2c_2(\vec{g}_b - \vec{p}_i) \quad (2)$$

Where:

$\vec{v}_{i+1}$: new velocity of a particle,
$\vec{v}_i$: current velocity of a particle,
ω: inertia weight (2.0),
c_1: cognitive scaling factor (1.4),
c_2: social scaling factor (1.4),
r_1: random number from range [0,1],
r_2: random number from range [0,1],
$\vec{p}_i$: current position of a particle,
$\vec{p}_{best}$: best known position of a particle,
$\vec{g}_{best}$: 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:

$$\vec{p}_{i+1} = \vec{p}_i + \vec{v}_{i+1}\Delta t \quad (3)$$

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 (3) Δt, which always has the constant value of one unit, is multiplied to the velocity vector $\vec{v}_{i+1}$ in order to get consistency in the physical units (Nolle, 2015). In this research the control parameter values were chosen as follows (Tholen & Nolle, 2017):
ω = 2.0,
c_1 = 1.4,
c_2 = 1.4.

Inertia Levy-flight

Biologists have observed that animals, like sharks, bony fish, sea turtles and penguins, often move in patterns that can be approximated by Levy-flights (Reynolds, 2014) following the Levy-flight foraging hypotheses. This hypotheses states that natural selection should have led to adaptations for Levy-flight foraging, because Levy-flights can optimise search efficiencies (Viswanathan, et al., 2008). Since there is experimental, evidence for inherent Levy search behaviour in forging animals (Kölzsch, et al., 2015), Levy-flight has been selected as a search strategy for a single AUV. While using Levy-flight, the AUV has to choose a random direction as well as a random step length for each iteration. The direction is chosen from a uniformly distribution in a range of 0 to 360 degree. However the chosen step length is based on a power law cumulative distribution function:

$$s = r^{-\frac{1}{\alpha}} \quad (4)$$

Where
r: random number from the range [0,1],
α *:* parameter from the range of [1,2].

When using Levy-flight as a search algorithm, the value of α has to be chosen off-line by the user before the search. The value of α has direct impact on the step length *s* calculate in each iteration. Therefore, the search behaviour of the AUV depends heavily on the chosen

value of α. Instead of manually tuning of α, a self-adaptive scheme to tune this parameter α is used. The AUV in each step calculate a value of α, based on the information gained from the environment (Tholen, et al. , 2018) as follows:

$$\alpha = \frac{g_c - g_w}{g_b - g_w} + 1 \quad (5)$$

Where:

g_c: current fitness value,
g_w: worst score fitness found so far,
g_b: best score fitness found so far.

Furthermore, the AUV stores the fitness value of the previous iteration g_{c-1} and compare this value with the fitness value in the actual time step g_c. If there is an improvement, the AUV will keep its direction. Otherwise, it will choose a new direction randomly.

Combination of Particle Swarm Optimisation and inertia Levy-flight

Sometimes a scout vehicle can improve the success of a search in difficult search areas (Nolle, 2015). In this research, the inertia Levy-flight will be used as search strategy for the scout vehicle, to guide the search of the other particles using PSO as search strategy. During the search, the scout will move through the search space. In each step, it will measure the conductivity. The scout vehicle shares its best position with the swarm.

The movement of the scout vehicle is not affected by the search results of the other swarm particles. Hence, this allows an independent search for the scout vehicle, the scout will be able to reach new areas even if the other particles of the swarm will trapped into a local optima or stuck in an area without any gradient information.

EXPERIMENTS

To compare the performance of the proposed search algorithm, simulations were carried out, using the described test environment.
Three different algorithms with different configurations was evaluated using a swarm of four AUVs. In the first configuration, all particles were guided by a PSO. In the second configuration, the search of the AUVs were guided by the inertia Levy-flight. In this configuration the AUVs do not exchange any information about the environment. For the third configuration, three of the AUVs perform a PSO, while the fourth AUV perform an inertia Levy-flight.

For each configuration, 1,000 simulations were carried out. The search time is limited to 5,400 seconds, due to the limitations of the operation time of the hardware platform. In each run, the g_{best} value of the whole swarm is stored during the search.

RESULTS

The results of the 1,000 simulations were classified into three categories, to compare the performance of the different search algorithms. The minimal distance between the SGD and a particle during the search will be used for the classification. However, the g_{best} value of the swarm can be used for the classification process. Due to the specific nature of the search topology this distance information is transformed into conductivity values using equation (1). Table 1 shows the attributes of different classes in terms of the minimum values and the maximum values the distance to the SGD and conductivity values. Experiments with results that fall within class one can be named as successful runs, while others can be referred to as unsuccessful runs.

Table 1: Attributes for Classification

Class	Min-Value		Max-Value	
	x (m)	*f(x)* (mS/cm)	*x* (m)	*f(x)* (mS/cm)
1	0.0	8.74	0.4	14.08
2	>0.4	14.08	5	40.54
3	>5	40.55	495	53.42

Figure 3 shows the results of the simulations, using the classification described above. It can be observed from the figure, that the swarm guided by PSO was not able to find the SGD during specified search time. While using the inertia Levy-flight, the swarm is able to find the SGD in 595 runs. Furthermore using the combination of PSO and inertia Levy-flight, the swarm is able to find the SGD during 685 runs.

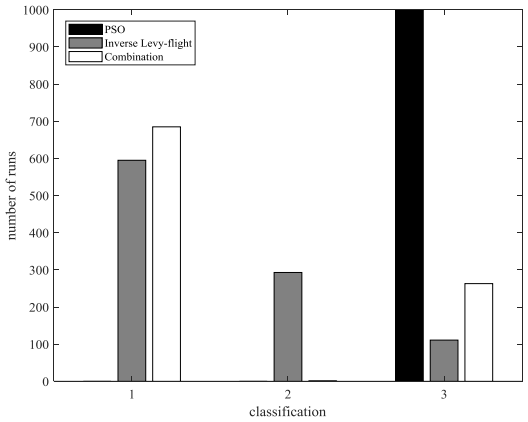


Figure 3: Frequency Diagram sorted by classes

Figure 4 shows the trajectories of the AUVs during the search for the three different algorithms. The graphs on the left hand side show the trajectories of the runs with the best performance. While the ones on the right hand side show the trajectories of the runs with the worst performance.

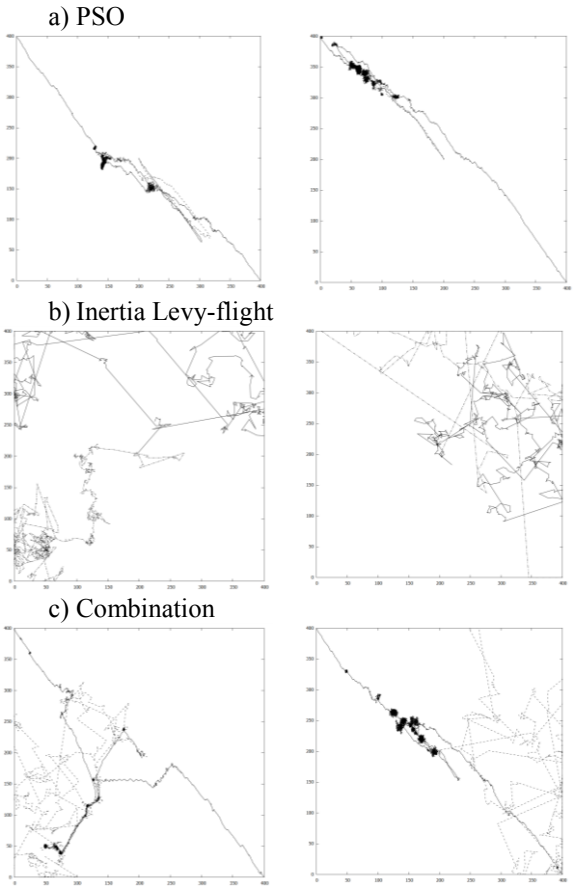


Figure 4: Trajectories of the AUVs during the search for the best run (left) and the worst run (right) for the three search algorithms

DISCUSSION

It can be observed from Figure 4-a that the PSO was unable to explore the search area. The particles of the swarm using inertia Levy-flight do not exchange any information and each particle explores the search area by itself. If one of the particles is able to find the SGD, the run will be successful (left). In the combination configuration, the particle using inertia Levy-flight adds extra exploring capabilities to the swarm. That has a great impact on the performance as depicted in figure 4-c. If this scouter particle moves near the SGD, it can redirect the swarm towards the SGD even if all the swarm is far from it. The combination can fail to find the SGD if both the swarm fails to approach the SGD or this particle fails to scout the SGD as in the case shown in figure-c right.

The experiments show that a search strategy based on PSO is not able to guide a swarm of four AUVs to a SGD due to the specific topology of the search space. While the performance of a swarm using the inertia Levy-flight is much better. However, by using a combination of both algorithms, the search performance of the swarm can be improved.

Table 2 summarise the results of the simulations. It can be depicted from the table, that the average performance of the inertia Levy-flight is better than for the combination. However the median value of the combination is better than the value of the inertia Levy-flight.

Table 2: Aggregation of Statistical Parameters

	PSO	Levy-flight	Combination
Min	51.2	8.70	8.38
Max	52.76	52.67	52.70
Average	52.42	17.74	19.76
Standard deviation	0.36	11.46	17.75
Median	52.61	13.15	9.25

CONCLUSION AND FUTURE WORK

A novel search algorithm based on PSO and inertia Levy-flight was developed to guide the search of a swarm of AUVs for a submarine point of interest, i.e. a SGD. It is shown, using a computer simulation, that this algorithm is capable to improve the search performance of a small swarm of AUVs during its search. With the sparsely availability of gradient information, a swarm of AUVs guided by PSO is not capable to reach the SGD during the limited search time. However, the inertia Levy-flight in general is able to help the swarm of AUVs to locate the SGD. When using the inertia Levy-flight, the AUVs do not exchange any information. The proposed configuration enable sharing knowledge about the environment without limiting the exploring capabilities of inertia Levy by allowing the scouter to provide information to the swarm without taken any commands from the swarm that can limit its exploring ability.
Further simulations will be done to evaluate the ability of the proposed algorithm to locate SGD in different search topologies and to fine tune the control parameters of the search algorithm. The algorithm will also be tested using the AUV platform that is currently under development.

REFERENCES

Bansal, J., Singh, P., Saraswat, M., Verma, A., Jadon, S., & Abraham, A. (2011). Inertia Weight Strategies in Particle Swarm Optimization. *Third World Congress on Nature and Biologically Inspired Computing*, (S. 633-640).

Dugan, J., Defeo, O., Jaramillo, E., Jones, A., Lastra, M., Nel, R., . . . Schoeman, D. (2010). Give Beach Ecosystems Their Day in the Sun. *Science, Vol. 329, ISSUE. 5996*, S. 1146.

El-Mihoub, T., Tholen, C., & Nolle, L. (2018). Blind Search Patterns for Off-Line Path Planning. *Proceedings of the 32nd European Conference on Modelling and Simulation ECMS 2018 (submitted).*

Evans, T., & Wilson, A. (2016). Groundwater transport and the freshwater–saltwater interface below

sandy beaches. *Journal of Hydrology, 538*, S. 563–573.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *IEEE International Conference on: Neural Networks*, (S. 1942-1948).

Kölzsch, A., Alzate, A., Bartumeus, F., de Jager, M., Weerman, E., Hengeveld, G., . . . van de Koppel, J. (2015). Experimental evidence for inherent Levy search behaviour in foraging animals. *Proceedings of the Royal Society B 282: 20150424*.

Moore, C., Barnard, A., Fietzek, P., Lewis, M., Sosik, H., White, S., & Zielinski, O. (2009). Optical tools for ocean monitoring and research. *Ocean Science, Vol. 5*, S. 661-684.

Moore, W. (2010). The effect of submarine groundwater discharge on the ocean. *Annual Review of Marine Science, Vol. 2*, S. 59-88.

Nelson, C., Donahue, M., Dulaiova, H., Goldberg, S., La Valle, F., Lubarsky, K., . . . Thomas, F. (2015). Fluorescent dissolved organic matter as a multivariate biogeochemical tracer of submarine groundwater discharge in coral reef ecosystems. *Marine Chemistry, 177*, S. 232–243.

Nolle, L. (2015). On a search strategy for collaborating autonomous underwater vehicles. *Proceedings of Mendel 2015, 21st International Conference on Soft Computing*, (S. 159-164). Brno, CZ.

Reynolds, A. (2014). Levy flight movment patterns in marine predators may derive from turbulance cues. *Proceedings of the Royal Society A 470 20140408*.

Richardson, C. M., Dulai, H., Popp, B. N., Ruttenberg, K., & Fackrell, J. K. (2017). Submarine groundwater discharge drives biogeochemistry in two Hawaiian reefs. *Limnol. Oceanogr., 62:*, S. 348–363.

Stedmon, C., Osburn, C., & Kragh, T. (2010). Tracing water mass mixing in the Baltic–North Sea transition zone using the optical properties of coloured dissolved organic matter. *Estuarine, Coastal and Shelf Science, 87(1)*, S. 156–162.

Tholen, C., & Nolle, L. (2017). Parameter Search for a Small Swarm of AUVs Using Particle Swarm Optimisation. *Proceedings of the 37th SGAI International Conference on Artificial Intelligence, AI 2017*, (S. 384-396). Cambridge, UK.

Tholen, C., El-Mihoub, T., Nolle, L., & Zielinski, O. (2018). On the robustness of self-adaptive Levy-flight. *Proceedings of the OCEANS'18 MTS (in press)*. Kobe (Japan): IEEE.

Tholen, C., Nolle, L., & Zielinski, O. (2017). On The Effect Of Neighborhood Schemes And Cell Shape On The Behaviour Of Cellular Automata Applied To The Simulation Of Submarine Groundwater Discharge. *31th European Conference on Modelling and Simulation ECMS 2017*, (S. 255-261).

Viswanathan, G., Raposo, E., & da Luz, M. (2008). Levy flights and superdiffusion in the context of biological encounters and random searches. *Phys. Life Rev. 5*, S. 133–150.

Zielinski, O., Busch, J., Cembella, A., Daly, K., Engelbrektsson, J., Hannides, A., & Schmidt, H. (2009). Detecting marine hazardous substances and organisms: sensors for pollutants, toxins and pathogens. *Ocean Science, Vol. 5*, S. 329-349.

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.

Predicting system level ESD performance

Guido Notermans, Sergej Bub, and Ayk Hilbrink
Nexperia Germany GmbH, Stresemannallee 101, 22529 Hamburg, Germany
E-mail: guido.notermans@nexperia.com

KEYWORDS
Verilog-A, Model, ESD, System Level, Gun Test, Protection, clamp, SEED.

ABSTRACT

This paper presents an ESD circuit model for a complete system, which allows accurate prediction of system pass levels during a system level ('ESD gun") test. The paper presents a Spice model for the ESD gun and a Verilog-A model for the protection device, both on-chip and on-board. Magnetic field scanning during an ESD discharge is used to optimize the model.

INTRODUCTION

Since electronic systems are used in electrostatically unprotected environments, such as an end user's home, it is important for system vendors to be able to predict the system-level ESD pass level, which is usually measured according to IEC 61000-4-2 (IEC 2008). Unfortunately, typical ESD conditions fall outside the range of small-signal parameters for which well-calibrated component models are readily available. This has given the field of ESD over the years a touch of 'black magic' among electronic engineers. In order to remedy this situation, the JEDEC organization has published two white papers (JEP161 2011, JEP162 2013) on system level testing, which describe a modeling approach which is called 'System-efficient ESD Design' (SEED).

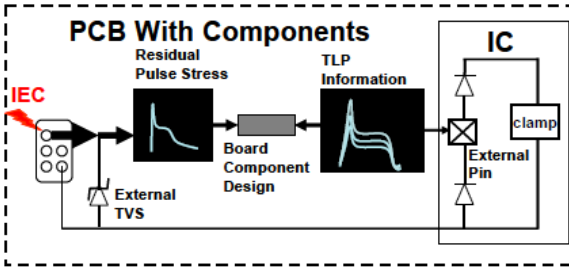


Figure 1 System-Efficient ESD Design (SEED) Concept (from JEP161 2011)

Figure 1 illustrates the basic concept. When a system-level pulse ('IEC' pulse) enters a system, by firing an ESD gun into a specified port, the main ESD current is supposed to flow into the external TVS, but inevitably a small residual current will flow into the IC, which may cause an over-voltage there. Either the residual current may cause thermal damage to the IC or the resulting over-voltage may damage sensitive gate oxides in the IC.
It is the goal of SEED to simulate the residual current and over-voltage with sufficient accuracy to allow an accurate prediction of the system failure level.
For a valid prediction, it is important to assess the impact of the parasitics in the system (Johnsson et al. 2012, Notermans et al. 2016), in particular of inductances. Furthermore, it is essential to have proper clamp models. This paper will focus on the development of proper clamp models, using Verilog-A, which allow a flexible implementation in a circuit simulation environment, such as ADS. Suitable models will be described in the modeling section. In the calibration section, the impact of parasitics will be explored and proper ways to extract the parameter values will be proposed. The application of the methodology on a USB3 interface board will be described. Finally, the use of an H-field scanning tool to optimize the system model is discussed.

MODELING INDIVIDUAL COMPONENTS

A full system model according to SEED needs at least three components: A gun model, a model for the external clamp, and a model for the parasitics on the system board and the IC receiving pin.
Note that the model of the IC pin does not need to include any (possibly proprietary) information on the layout or function of the IC I/O circuitry. The residual current is completely defined by the internal IC protection. For a SEED model, it is sufficient to characterize the I-V curve in the ESD timeframe, which will be described in the calibration section.

Gun model

The gun generator model (Figure 2) is derived from the work of (Wang et al. 2003) and (Caniggia et al. 2006).

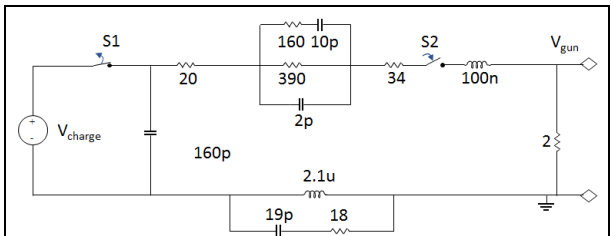


Figure 2 ESD Gun Model

Several workers have implemented similar gun models with small variations (Yang 2018). For our purpose, we optimized the model to generate a worst-case pulse with the minimum rise-time of about 0.6 ns and a high first peak of about 4 A. Furthermore, we tried to strike a balance between the waveforms measured with different guns. Figure 3 shows a simulated current waveform

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

compared to a measured waveforms from a NoiseKen ESS-S3011A and a Schloeder SESD 30000 gun. The relatively wide IEC 61000-4-2 specification is indicated as a hashed band. The simulated and Noiseken waveforms are in spec. The Schloeder guns is slightly out of spec in the second, slower peak. The simulated waveform is about midway between the two extremes in the second peak.

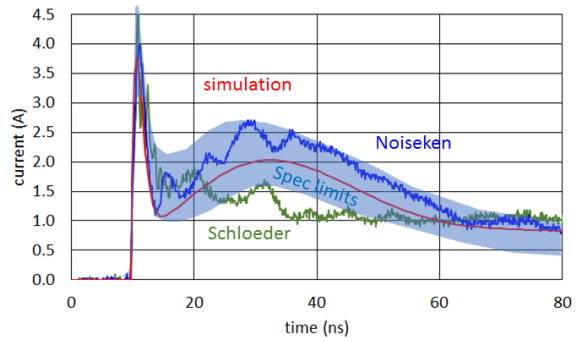


Figure 3 Simulated and Measured Waveforms

Clamp models

Verilog-A uses a reduced syntax from Verilog-AMS. It is integrated in Keysight's Advanced Design System (ADS®). Verilog-A offers a flexible and simple way to implement diodes and snapback devices, with and without hysteresis. The clamps are modeled using a quasi-static, piecewise linear (PWL) I-V curve. Because the derivative at the inflection points of such a curve is not continuous, convergence problems cannot be avoided completely.
An advantage of piecewise linear curves is the ease of calibration. All that is needed to calibrate the model is to enter the measured (V,I) values for each of the inflection points.

Diode model
The simplest possible form describes a non-ideal (avalanche) diode, defined by the inflection points Von, Ion, Ron for forward polarity and Vrev, Irev, Rrev for reverse polarity (Figure 4). This model is suited for clamps that do not exhibit snapback.

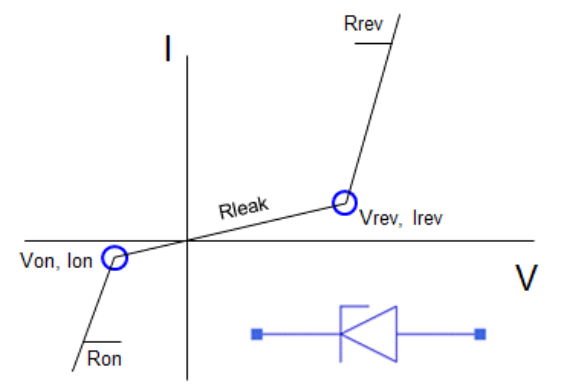


Figure 4 Piecewise Linear Model without Snapback

Snapback model
The clamp model can be extended in a straightforward manner to include snapback (Figure 5). Two additional inflection points are defined: [Vt1, It1] and [Vh, Ih]. The clamp triggers at [Vt1 It1] and for I>It1 it will enter its low-impedance state (given by Rrev). The minimum voltage and current for this low-impedance state are defined by [Vh,Ih]. Once I<Ih the clamp will return to its high-impedance state.

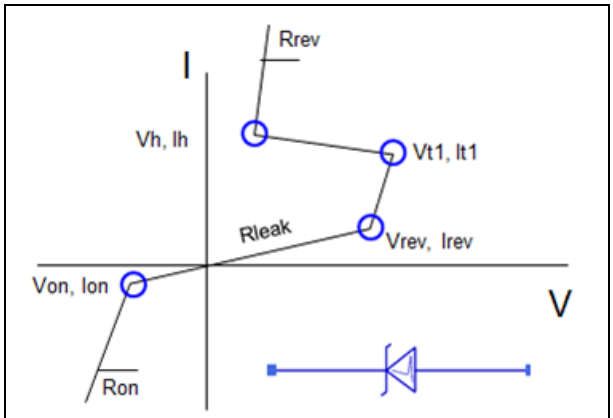


Figure 5 Piecewise Linear Model with Snapback

Note that the same curve is traversed when the current is increasing or decreasing (no hysteresis). In other words, the function is single-valued in current. In reality, Ih<It1 always, but the error introduced by the simplification Ih>It1 is small, because both Ih, It1 << It2.
Note further that Vrev and Vt1 are distinct. Vrev is the voltage at which the clamp leaves its high-impedance state and starts to conduct, typically when a trigger device kicks in. Snapback to the low-impedance state occurs at [Vt1, It1], when the main clamp triggers.
Forward (on) and reverse (rev) polarity are defined in accordance with the definitions for a zener diode.

Hysteresis
Depending on the timescale, the protection device may exhibit hysteresis. For instance, an SCR may need some conductivity modulation to enter its low-impedance state.

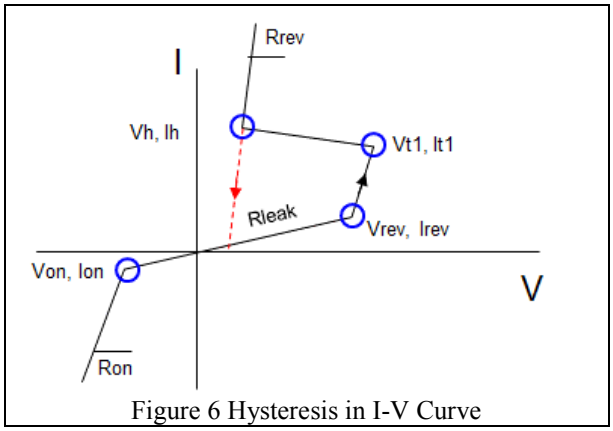


Figure 6 Hysteresis in I-V Curve

When switching off such a device, the charge takes some time to disappear and the SCR remains triggered for a

long time, typically microseconds. In the I-V curve, this effect manifests itself as a hysteresis (Figure 6).
When the current is increasing, the I-V curve follows the branch via Vrev, Irev and Vt1, and It1, which corresponds to the triggering of the protection. When the current is decreasing, the high-current curve with resistance Rrev is followed, i.e. the current just fades away'. without voltage increase (to Vt1). It is important to add this distinction to the model, since otherwise an unrealistic voltage overshoot would also appear at the end of pulse (Figure 7 red solid line), when the device is triggered by a square current pulse. With current hysteresis, the voltage waveform only shows an overshoot at the beginning of the pulse (Figure 7 blue dashed line), which is accurate at short (ESD) timescales.

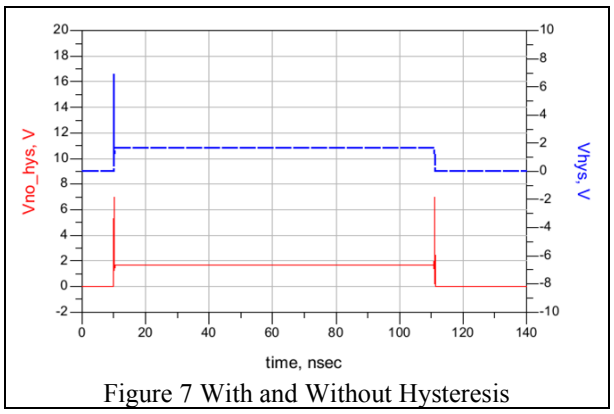


Figure 7 With and Without Hysteresis

Note that the hysteresis only occurs for timescales shorter than the time it takes to remove the injected charge of the device, which is around 10 µs. For longer timescales, notably during DC simulation, the hysteresis does not occur. This timescale dependence also needs to be taken into account in the model.

Dynamic overshoot
Dynamic effects due to metal inductances are implemented by adding a small inductance of about 2 nH in series to the Verilog-A model, which accounts for the L.dI/dt overshoot. The additional overshoot due to conductivity modulation can also be modeled by current-dependent resistor (Manouvrier et al. 2008), but this extension is out of scope for this paper and will be reported elsewhere (Notermans et al. 2018).

System model

Typical systems comprise a processor IC with its internal protection, an external protection, and a system PCB which may contain several parasitic components (Figure 8). The internal protection of the IC is modeled in the same way as an external clamp. The parasitic board components are represented as lumped elements. For high-speed applications, capacitors are typically very small (< 1 pF) and they may be left out of the SEED simulation.
Figure 9 shows a typical simulation result for the residual current Is (blue) into the IC at a total ESD current It (red) for a 4 kV gun discharge. A residual current Is ≈ 12 mA remains after the external protection for this particular configuration. In the same way, the residual voltage at the IC pin can be simulated. By comparing residual current and voltage with the known failure levels for the IC, the system robustness can be simulated. An example for a real application will be given below.

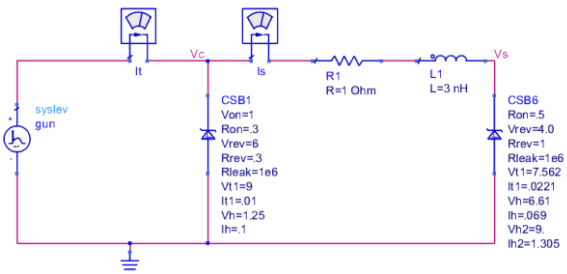


Figure 8 System-Level Simulation (SEED)

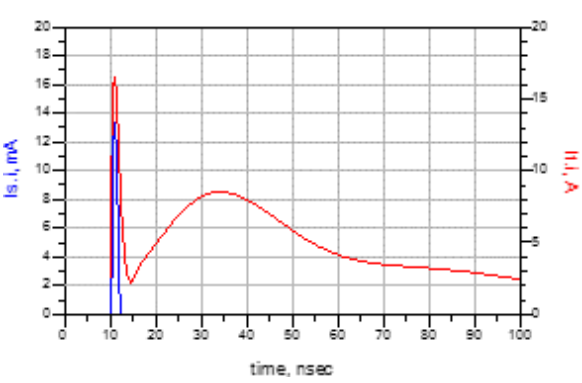


Figure 9 Simulated Total Current It (red) and Residual Current into the IC Is (blue).

CALIBRATING THE MODEL

Before the system robustness can be calculated, we need to calibrate the clamp parameters. The calibration is performed using Transmission Line Pulse (TLP) measurements (Maloney et al. 1985), which has the advantage that it provides a square current pulse with a flat 100 ns plateau (cf. Figure 7) as opposed to a gun pulse which has a double exponential waveform (cf. Figure 9). It has been shown (Notermans et al. 2012) that TLP measurements correlate well with gun discharge measurements, as far as thermal failure is concerned.

Clamps

Figure 10 shows a comparison between simulated and (TLP) measured I-V curves. The calibration is performed by entering the (V,I) points for the inflection points, as discussed in the previous section. It is easy to extend the model with additional inflection points, if required, e.g. to model thermal effects for high currents as well.
Figure 11 shows a zoom-in of the I-V curve of Figure 10 to highlight the fit for small currents. In the simulation a leakage resistor of 1 MΩ is used, which facilitates convergence. In reality, the leakage current maybe well below 1 nA, but for ESD simulation purposes the very low current behavior is not important.

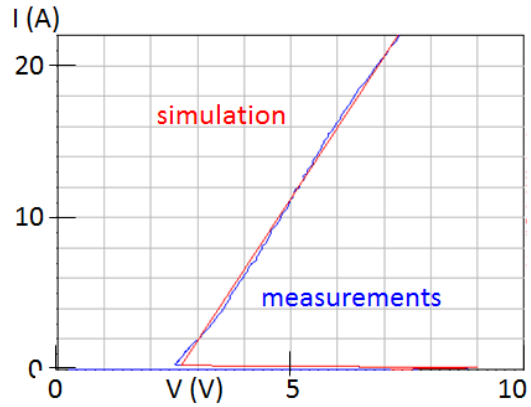


Figure 10 Comparison of Simulated and Measured (TLP) I-V Curve

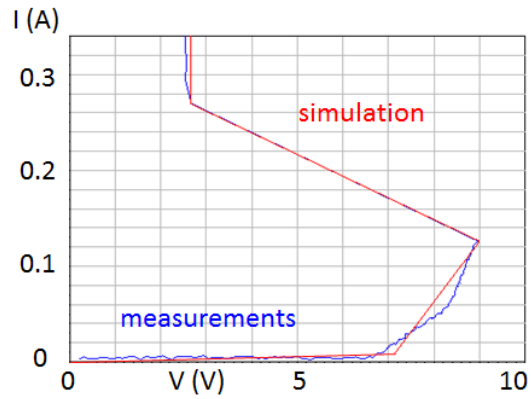


Figure 11 Zoom of I-V curve in Figure 10 for low Currents

Parasitics

For a complete system simulation, the parasitics of both (IC) internal and external protections, as well as the system board need to be determined.

Capacitance

The capacitance of the external protection is either measured directly using an Agilent E4980A precision LCR meter, at 1 MHz, or extracted from S-parameter measurements using a Rohde & Schwarz ZVA40 Vector Network Analyzer, up to 40 GHz.

For high-speed applications, such as USB3, typical capacitance is about 0.25 pF or less.

Inductance

Measuring the parasitic inductance is usually a bit more complicated. One possibility is to measure the S21 parameter and extract the capacitance at low frequency. From the resonance frequency and the known capacitance, the inductance can be derived. Typical inductances for external protections range from about 0.1 nH for CSP packages to about 1.5 nH for wire-bonded packages. Similar values are usually obtained for the IC, depending on the package type as well.

TLP Inductance measurement

In case RF extraction is not possible, e.g. due to too many reflections, the inductance may be estimated using TLP measurements and plotting the measured inductive over-voltage against dI/dt. If the relation is linear, the slope of the resulting line is the inductance L.

Figure 12 shows the peak voltage in a very fast vf-TLP pulse of 5 ns width in a protection device. The risetime of the current pulse is 0.6 ns. For high currents, the curve becomes linear and the slope corresponds to the inductance of the device, according to L = Vpeak/(dI/dt). For this example, an extracted L ≈ 0.75 nH results. The non-linear part below 10 A is caused by the conductivity modulation part, which is described elsewhere (Notermans et al. 2018).

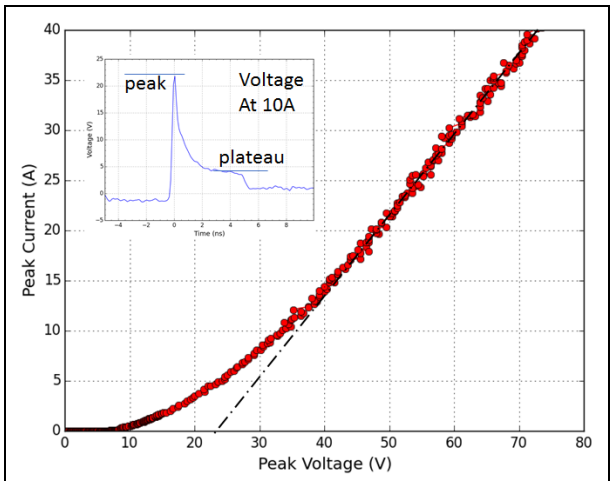


Figure 12 Peak (=Overshoot) Voltage versus Current

ESD CURRENT FLOW

It may not always be obvious which board components need to be included in a SEED simulation. It is, therefore, helpful to use a current spreading tool to determine the ESD current flow. We have used an Amber Precision SmartScan ESD-350 to show the ESD current on the USB3 board using a magnetic field probe. Figure 13 (left) shows a scan of the board with the original on-board protection set-up. The connector is at the bottom right. A first protection (prot1) is connected directly behind the connector, followed by a 1 Ω resistor and a second protection (prot2). The scan shows that the residual ESD current enters the IC at the RX1 pin. Part of the residual current is absorbed in the internal protection, but part of it exits the IC at a Vdd pin and flows via the decap to ground. The bottom graph shows that the maximum H-field is about 20 A/m.

The original protection scheme comprises two on-board protections and a 1 Ω resistor in a PI-configuration, presumably in an effort to improve the system protection. A second H-field scan, after the second protection was removed (Figure 13 right) reveals, however, that placing a second protection is actually counterproductive: more current is flowing into the IC (Figure 13 left).

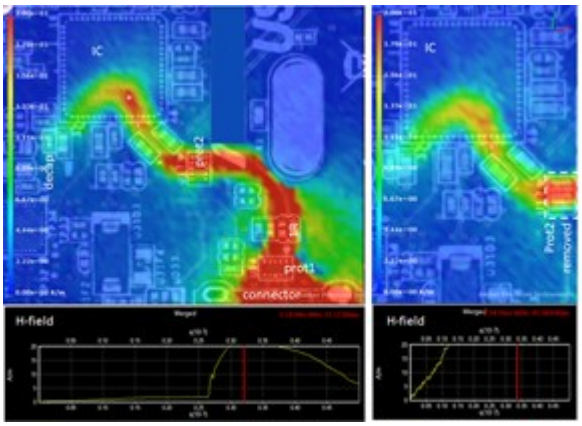


Figure 13 ESD Current on USB3 Interface with Double Protection (left) or Single Protection (right)

This example illustrates that the H-field scanner can be a valuable tool to assess the effectiveness of a proposed system protection scheme.
The residual ESD current is, in fact, strongly related to the inductance of the protection and, thus, to the overshoot (cf. Figure 12 inset) during an ESD (or TLP) pulse.

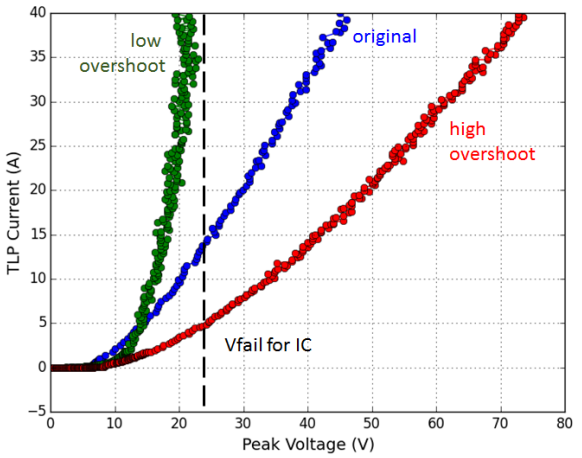


Figure 14 Peak Voltage against current for several Protections

Figure 14 shows the TLP peak voltage vs. TLP current for several on-board protections (alone). The original protection (blue) is compared with a protection with a lower overshoot voltage (green) and one with a higher overshoot voltage (red).
When the original on-board protection is replaced by a protection with a higher overshoot the residual current into the IC increases (Figure 15 left) compared to Figure 13 (right) and, conversely, using a protection with a lower overshoot reduces the residual current (right).
In fact, since the voltage for the low-overshoot device (PESD2V0Y1BSF) levels off around 22 V, which is lower than the failure voltage of the IC, the system ESD performance is in this case limited only by the ESD robustness of the protection (> 15 kV).

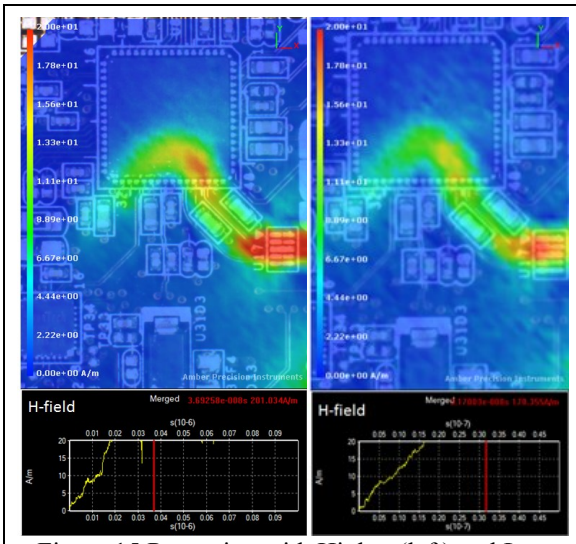


Figure 15 Protection with Higher (left) and Lower (right) Overshoot Voltage.

USING SEED ON A USB3 INTERFACE

The suitability of the SEED approach can be demonstrated on a USB3 PCIExpress board for a PC. Figure 16 shows the test set-up, which has been described in detail earlier (Notermans et al. 2016). The USB3 card is put into a PCIExpress slot of a PC and the gun is fired directly into one of the two receiver (RX) pins on the USB3 socket.

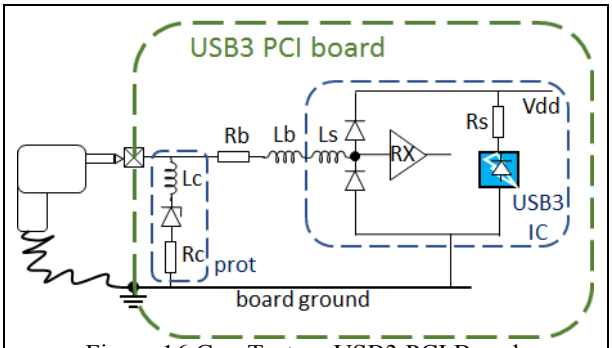


Figure 16 Gun Test on USB3 PCI Board.

The USB3 IC has a rail-based internal protection, with dynamic resistance Rs and parasitic inductance Ls. On the board there is an external protection with dynamic resistance Rc and inductance Lc, a resistor of Rb = 1 Ω and parasitic inductance Lb. The (parasitic) capacitances have not been included, because they are very small (typically < 0.25 pF) to allow a high bandwidth of > 10 GHz. Such small capacitances have negligible impact on the outcome of the SEED simulation. Of course, they would need to be included in a simulation under normal operating conditions.
It has been shown before (Notermans et al. 2016) that system performance is limited by the residual current into the IC during the fast first peak (cf. Figure 3) of the gun discharge. During the slower second peak the residual current is determined by the ratio of the dynamic resistances in the external vs. the internal protection:

Rc/(Rb+Rs).
During the first peak, however, the impedance of the inductances is much larger than the ohmic resistances and the residual current is determined by the ratio of the inductances Lc/(Lb+Ls).
This effect is illustrated in a measurement of a 1 kV gun discharge for three different inductance ratios (Figure 17).

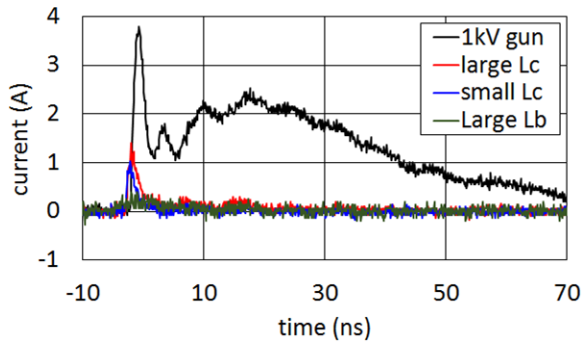


Figure 17 Residual Current Waveforms for three Inductance Ratios, compared to the Gun Waveforms

The current waveform of the gun is shown in black and the residual current for a relatively large Lc = 1.3 nH (red), Lc =0.7 nH (blue), and a very large LB = 35 nH (green) is compared. It is clear that the second peak is well suppressed in all cases, but a significant residual current remains during the first peak, which is highest for a large Lc (37%), followed by a small Lc (27%). The best solution is obtained when a large inductance Lb is added in the path to the IC (< 10%). The ESD performance scales accordingly (Notermans et al. 2016).
The effect of the inductive current distribution can be simulated using a SEED model (Figure 18). The correlation with measurements is excellent. The simulated residual current scales with the Lc/(Lb+Ls) ratio, as in the measurements.

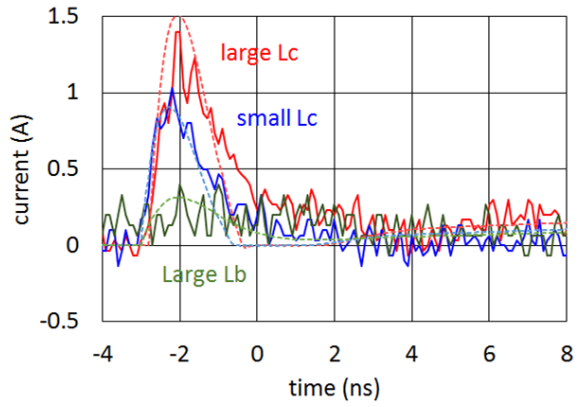


Figure 18 Measured (solid) and Simulated (dotted) Residual Current in the First Peak for several Lc/(Lb+Ls) Ratios

Note that the large inductance Lb = 35 nH is introduced into each of the two differential RX lines. The two air coils are coupled in such a way that the effective differential inductance is close to zero. Thus, the differential USB3 signal passes undamped. But an ESD signal couples to both lines as common mode signal and experiences the full inductance in each line (Werner et al. 2015 and 2016).
The measured and simulated peak currents in the first peak are summarized in Table 1. The correlation is very good.

Table 1 Comparison of Simulated and Measured First-Peak Current Amplitudes of Figure 17

1st peak (A)	measured	simulated
gun	3.76	3.64
large Lc	1.40	1.42
small Lc	1.03	0.86
large Lb	0.33	0.32

CONCLUSION

The paper has presented quasi-static Verilog-A models for the on-chip and on-board protections in a system. An H-field scanning tool may provide essential insight in where the residual current is flowing during a system level discharge, which facilitates setting up a complete system model.
The complete circuit model, incorporating these models allows accurate prediction of the system-level ESD performance.

REFERENCES

Caniggia S. and Maradei F. 2006. *Circuit and numerical modeling of electrostatic discharge generators*, IEEE Trans. Ind. Appl., vol. 42, no. 6, pp. 1350–1357.

International Electrotechnical Commission (IEC). 2008. *Electromagnetic Compatibility (EMC): Part 4-2: Testing and Measurement Techniques--Electrostatic Discharge Immunity Test*, IEC 61000-4-2, edition 2.

JEDEC 2011. JEP-161, System Level ESD Part 1: *Common Misconceptions and Recommended Basic Approaches*.

JEDEC 2013. JEP-162, System Level ESD Part 2: *Implementation of effective ESD robust designs*.

Johnsson, D. and Gossner, H. 2015. *Study of system co-design of a realistic mobile board*, EOS/ESD Symposium Proceedings.

Maloney, T. and Khurana, N. 1985. *Transmission Line Pulsing Techniques for circuit modeling of ESD phenomena*, EOS/ESD Symposium Proceedings.

Manouvrier J.R.; Fonteneau P.; Legrand C.A.; Beckrich-Ros H.; Richier C.; Nouet P.; Azais F. 2008. *A Physics-Based Compact Model for ESD Protection Diodes under Very Fast Transients*, EOS/ESD Symposium Proceedings.

Notermans, G.; Bychikhin S.; Pogany, D.; Johnsson, D.; and Maksimovic, D. 2012. *HMM–TLP correlation for system-efficient ESD design*, Microelec. Reliab. Journal, pp 1012-1019.

Notermans, G.; Ritter H.-M.; Seider S.; and Laue B.; 2016. *Gun tests of a USB3 host controller board*, EOS/ESD Symposium Proceedings.

Notermans, G.; Ritter H.-M.; Holland S.; and Pogany, D. 2018. *Modeling dynamic overshoot in ESD protections*, submitted

for publication at EOS/ESD Symposium.
Wang K.; Pommerenke D.; Chundru R.; Doren T.V.; Drewniak J.L.; and Shashindranath A. 2003. *Numerical modeling of electrostatic discharge generators*, IEEE Trans. Electromagn. Compat., vol. 45, no. 2, pp. 258-271.
Werner J.; Schuett J.; and Notermans, G. 2016. *Sub-Miniature Common Mode Filter With Integrated ESD Protection*, Proc. IEEE EMC, 2015.
Werner J.; Schuett J.; and Notermans, G. 2016. *Design And Simulation Of Integrated EMI Filter*, Proc. 30th ECMS, 270-276.
Yang, S.; Pommerenke D. 2018. *Effect of Different Load Impedances on ESD Generators and ESD Generator SPICE Models,* IEEE Trans. Electromagn. Compat.

GUIDO NOTERMANS is ESD fellow at Nexperia Germany in Hamburg. He graduated in Experimental Physics at Utrecht University in 1980 and received his PhD in Plasma Physics in 1984. He subsequently joined Philips Research Labs where he developed III-V semiconductor lasers until 1990. From 1995 he worked as senior ESD principal for Philips Semiconductors Nijmegen. In 1999 he moved to Berlin where he joined Infineon Fiber Optics as R&D director for electro-optical devices. In 2005 he joined Philips Semiconductors Zurich and returned to the field of ESD. In 2013, he moved to Nexperia Hamburg and is presently developing stand-alone ESD protections.

SERGEJ BUB is System Level ESD Expert at Nexperia Germany in Hamburg. He graduated as M.Sc. in Electrical Engineering at Technical University Hamburg with a focus on nanoelectronics and microsystem technic in 2017. His university study was finished by writing a master thesis in cooperation between TUHH and Nexperia concerning “Investigations of Secondary Breakdown behavior of power bipolar transistors for characterization of SOAR”.

Ayk Hilbrink was born in Stade Germany in 1992. He studied electrical engineering at Jade University of Applied Sciences, Wilhelmshaven and received the bachelor degree in 2016 and the Master of Science in 2018. His bachelor thesis “Generierung von Spice-Modellen in ADS” was written in cooperation with NXP Semiconductors and covers the field of device modelling for a high frequency range. His strong expertise in RF applications led him to his Master thesis at Nexperia about de-embedding. Since 2017 Ayk is working for Nexperia as characterization engineer with main focus on RF and EMI measurements.

Modelling, Simulation and Control of Technological Processes

MATLAB TOOLBOX FOR SELF-TUNING PREDICTIVE CONTROL OF TIME-DELAYED SYSTEMS

Radek Holiš, Vladimír Bobál
Department of Process Control
Faculty of Applied Informatics, Tomas Bata University in Zlin
Nad Stráněmi 4511, Zlin 76005, Czech Republic
E-mail: rholis@fai.utb.cz

KEYWORDS
Self-tuning Control, Model Predictive Control, MPC, Time-delay, MATLAB Toolbox

ABSTRACT

The designed MATLAB/SIMULINK Toolbox is dedicated to develop and design predictive Self-Tuning Control (STC) algorithm for the time-delayed systems. In practice, many processes can exhibit time-delay in their dynamic behavior, which is mainly caused by a time needed for transport of the energy, information, or mass. In lights of these facts, it is necessary to develop suitable algorithm and verify its correct dynamic behavior using simulation first, so this Toolbox can be used in advantage. This paper deals with the basic principles of Model Predictive Control (MPC), calculation of control law, design process of the predictive controller and recursive identification of control process using Recursive Least Squares Method (RLSM). There are also many cases when compensation of measurable disturbance is required, so this Toolbox allows compensating of this disturbance.

INTRODUCTION

Time-delayed systems appear in many processes in industry and other fields, including economical as well as biological systems (Camacho and Normey-Rico 2007). These processes are difficult to control using standard feedback controllers. When the relative time-delay is very large or a high performance of the control process is desired, we can choose MPC as the suitable algorithm for these types of processes. The predictive control strategy contains a model of the process in the structure of the controller. The first time-delay compensation algorithm was shown by (Smith 1957). This algorithm is known as the Smith Predictor (SP) and it contains a dynamic model of the time-delay process and it can be called as the first MPC algorithm. For more complex processes, containing time-delay and affected by a measurable disturbance, MPC strategy can be used (Maciejowski 2002).

The MPC is an attractive set of the control strategies widely used in the industry. The popularity of the MPC is mostly due to its leading to a safety operation of processes under all circumstances and ability to use constraints. The MPC is known as a control strategy where based on the measurements of plant's states at time, a mathematical model of the plant (often referred to as the prediction model) is being used for prediction of the evolution of the plant in the future.

The MPC with allowance to control of the time-delayed processes, ability of self-tuning, and possibility of the measurable disturbance compensation can be powerful and versatile algorithm for control of various processes.

This paper deals with the use of MPC for processes with time-delay with possibility of measuring disturbance compensation.

Strategy of MPC presents a series of advantages over other methods. The MPC can be used to control a great variety of processes, ranging from those with relatively simple dynamics to other more complex ones, including systems with long time-delay, unstable ones or non-minimum phase. The multivariable case can easily be dealt with. The additional advantage is that extension to the treatment of constraints is conceptually simple and these can be systematically included during the design process. This approach of control is a totally open methodology based on certain basic principles that is allowed for future extensions (Camacho and Normey-Rico 2007; Rossiter 2003; Haber et al. 2011).

The MPC has been deployed on slower processes in its early days (Kvasnica 2009).

It was caused by the large computational complexity of control algorithms and large time demands. Trends have expanded towards modifications of predictive control over the years. Nowadays, the MPC strategy can be used for controlling of very fast processes. These processes can have requirement for computation of control action in microseconds (e.g. explicit approach of MPC can be used). In practice, an excellent industrial survey reports many successful applications of the MPC in various industry areas (Qin and Badgewell 1997; Rawlings and Mayne 2009).

An extended version of the Generalized Predictive Control (GPC) algorithm is dedicated for design of the adaptive predictive controller in this paper.

This paper is arranged as follows. The extended GPC algorithm is described in the first section. The next section shows computation of the cost function for GPC and computation of control law. Brief description of the recursive identification procedure is introduced in the following section. The designed Toolbox is briefly described afterwards. The next section contains examples of the simulation control using designed Toolbox and the last section concludes this paper.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

EXTENDED VERSION OF THE GENERALIZED PREDICTIVE CONTROL ALGORITHM

The basic MPC structure with the extended GPC algorithm is schematically displayed in the Figure 1.

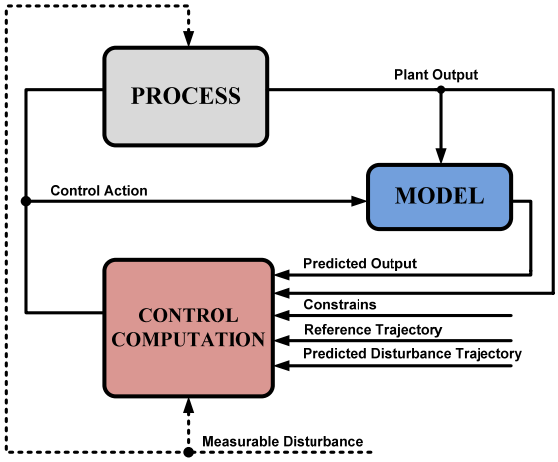


Figure 1 : Extended Structure of the MPC

The basic GPC algorithm minimizes a cost function that may be written as

$$J(N_x) = \sum_{i=N_1}^{N_2} \delta(i)\left[\hat{y}(k+i) - w(k+i)\right]^2 + \sum_{i=1}^{N_u} \lambda(i)\left[\Delta u(k+i-1)\right]^2 \quad (1)$$

where J is the function of the N_x, which represents N_1, N_2, and N_u. The N_1 and N_2 are the minimum and maximum horizons of cost function, and N_u is the control horizon of the cost function. This horizon should be chosen with regard to dynamics of controlled process to handle step response (Rossiter 2003; Moudgalya 2007). The $\hat{y}(k+i)$ is an optimum prediction of the system output. Coefficients $\delta(i)$ and $\lambda(i)$ are weighting coefficients and $w(k+i)$ is a vector of future reference sequence.

The goal of the predictive control is to calculate the future incremental control action of the $\Delta u(k)$, $\Delta u(k+1)$,... The cost function J (1) is minimized to obtain the final control law. This is realized by minimizing with respect to Δu, where the predictions $\hat{y}(k+i)$ are first expressed as a function of the past data and the future control actions $\Delta u(k+i-1)$. Thus, J can be considered as a function of the future control sequence. The control horizons as well as weighting factors are the tuning parameters, which can be changed to modify steepness and rapidity of the control course as required (Camacho and Normey-Rico 2007).

The horizons N_1 and N_2 are computed as $N_1 = d+1$ and $N_2 = N_u + d$, because of the time-delay characteristics of the process. In practice, N_1 and N_2 are hardcoded in the algorithm and N_u is the only changeable parameter.

The modified mathematical model of the Controlled Auto-Regressive Integrated Moving Average (2) (CARIMA) is used by the GPC to compute the predictions. It is the typical CARIMA model extended by the vector $v(k)$, which represents measurable disturbance sample

$$A(z^{-1})y(k) = z^{-d}B(z^{-1})u(k-1) + z^{-dv}D(z^{-1})v(k) + \\ + \frac{C(z^{-1})}{\Delta}e_s(k) \quad (2)$$

where d corresponds to number of steps of the time-delay for process, dv represents number of steps of the time-delay for disturbance, $e_S(k)$ is the white noise, and $\Delta = 1 - z^{-1}$. The polynomial $D(z^{-1})$ represents character of disturbance and the polynomial $C(z^{-1})$ describes character of the noise. This character is difficult to determine; therefore, polynomial $C(z^{-1})$ is chosen to be equal to one (Camacho and Bordons 2004; Fikar and Mikleš 2008; Clarke et al. 1987a).

Consider equation (2) multiplied by Δ. Then, predication model can be represented as follows, where output can be predicted as

$$\hat{y}(k+1) = \sum_{i=1}^{na+1} \tilde{a}_i y(k+1-i) + \sum_{i=1}^{nb} b_i \Delta u(k-d-i) + \\ + \sum_{i=1}^{nd} d_i \Delta v(k+1-dv-i) \quad (3)$$

where na, nb, and nd are degrees of polynomials $A(z^{-1})$, $B(z^{-1})$, and $D(z^{-1})$. The white noise $e_S(k)$ and its future values are considered to be equal to zero for the prediction of the future output values.

In case where time-delay is present, following equation should be used when equation (3) is applied recursively for i = 1, 2, ..., N_u (Clarke et al. 1987b).

$$\hat{\mathbf{y}} = \mathbf{G}\mathbf{u} + \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 + \mathbf{H}_{V1}\mathbf{v}_1 + \mathbf{H}_{V2}\mathbf{v}_2 \quad (4)$$

Matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{S}$ are constant matrices of dimensions $N_u \times N_u$, $N_u \times nb$, and $N_u \times (na+1)$, respectively. Matrices $\mathbf{H_{V1}}$ and $\mathbf{H_{V2}}$ are of dimensions $N_u \times (nd-1)$ and $N_u \times N_u$. Matrix $\mathbf{H}_{V1}$ can be used only in case when degree of polynomial $D(z^{-1})$ is equal to 2 or higher.

Following equation corresponds to the free response of the system that is the output that would be obtained if the control signal was kept constant.

$$\mathbf{f} = \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 + \mathbf{H}_{V1}\mathbf{v}_1 \quad (5)$$

Forced response of the system can be written as the next equation

$$\mathbf{f}_r = \mathbf{G}\mathbf{u} + \mathbf{H}_{V2}\mathbf{v}_2 \quad (6)$$

Vectors $\mathbf{u_1}$, $\mathbf{y_1}$, $\mathbf{v_1}$, and $\mathbf{v_2}$ are defined in the following equations. Sum of the free response (5) and the forced response (6) leads to overall response of the system defined by the following equation

$$\hat{\mathbf{y}} = \mathbf{f} + \mathbf{f}_r \quad (7)$$

Control Law Computation and Cost Function

The predicated output $\hat{y}$, expressed in the previous paragraph, is part of the equation (1). It is evident that *J* is the cost function of $\mathbf{y_1}$, $\mathbf{u}$ and $\mathbf{u_1}$. The individual elements of the summation of the cost function in the equation (1) can be written in a matrix form. This cost function can be defined as

$$J = (\mathbf{Gu} + \mathbf{Hu_1} + \mathbf{Sy_1} - \mathbf{w})^T \mathbf{Q_\delta} (\mathbf{Gu} + \mathbf{Hu_1} + \mathbf{Sy_1} - \mathbf{w}) + \\ + \mathbf{u}^T \mathbf{Q_\lambda} \mathbf{u} \quad (8)$$

where $\mathbf{Q_\delta}$ and $\mathbf{Q_\lambda}$ are the diagonal weighting matrices of size $N_u \times N_u$ with elements $\delta(j)$ and $\lambda(j)$, respectively. Although, in practice, the most common choice is to set $\delta(j)$ and $\lambda(j)$ constants on the horizon. In a fact, the values of these weighting factors must be normalized in order to obtain a correct weighting of the different errors and controller outputs.
After some manipulations *J* can be written as

$$J = \mathbf{u}^T (\mathbf{Q_\lambda} + \mathbf{G}^T \mathbf{Q_\delta} \mathbf{G}) \mathbf{u} + 2(\mathbf{Hu_1} + \mathbf{Sy_1} - \mathbf{w})^T \mathbf{Q_\delta} \mathbf{Gu} + \\ + (\mathbf{Hu_1} + \mathbf{Sy_1} - \mathbf{w})^T \mathbf{Q_\delta} (\mathbf{Hu_1} + \mathbf{Sy_1} - \mathbf{w}) \quad (9)$$

Minimizing *J* with respect to $\mathbf{u}$, it means $\frac{\partial J}{\partial \mathbf{u}} = 0$, leads to

$$\mathbf{Mu} = \mathbf{P}_0 \mathbf{y_1} + \mathbf{P_1} \mathbf{u_1} + \mathbf{P_2} \mathbf{w} \quad (10)$$

where $\mathbf{M} = \mathbf{G}^T \mathbf{Q}_\delta \mathbf{G} + \mathbf{Q}_\lambda$ is of dimension $N_u \times N_u$, $\mathbf{P}_0 = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{S}$ of dimension $N_u \times (na + 1)$, $\mathbf{P}_1 = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H}$ of dimension $N_u \times nb$ and $\mathbf{P}_2 = \mathbf{G}^T \mathbf{Q}_\delta$ of dimension $N_u \times N_u$.
In a receding horizon algorithm only the current value of the $\Delta u(k)$ is computed, so if $\mathbf{m}$ is the first row of matrix $\mathbf{M}^{-1}$, then $\Delta u(k)$ is given by

$$\Delta u(k) = \mathbf{mP_0} \mathbf{y_1} + \mathbf{mP_1} \mathbf{u_1} + \mathbf{mP_2} \mathbf{w} \quad (11)$$

Equations (8) – (10) deal with the case of no disturbance rejection. For the expression of the final control law containing compensation of the measurable disturbance, $\Delta u(k)$ is computed as the following control law form

$$\Delta u(k) = \mathbf{mP_0} \mathbf{y_1} + \mathbf{mP_1} \mathbf{u_1} + \mathbf{mP_2} \mathbf{w} + \mathbf{mP_{V1}} \mathbf{v_1} + \mathbf{mP_{V2}} \mathbf{v_2} \quad (12)$$

where $\mathbf{P_{V1}} = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H_{V1}}$ is of dimension $N_u \times (nd - 1)$ and $\mathbf{P_{V2}} = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H_{V2}}$ of dimension $N_u \times N_u$.
$\mathbf{H_{V1}}$ and $\mathbf{H_{V2}}$ are matrices including the coefficients of the system step response to the disturbance.
Future values of the disturbance can be determined only in certain cases, e.g. be measurement or generally in case, when it is related to the process load. In other cases, it can be predicted using means, trends, past data, other information, or by combination of specified items. If this is the case, the term corresponding to future deterministic disturbance can be computed (Schwarz et al. 2010). After introducing vectors $\mathbf{y_1}$, $\mathbf{u_1}$, $\mathbf{w}$, $\mathbf{v_1}$ and $\mathbf{v_2}$, final control law is defined as

$$\Delta u(k) = \mathbf{mP_0} \begin{bmatrix} \hat{y}(k+d) \\ \hat{y}(k+d-1) \\ \vdots \\ \hat{y}(k+d-na) \end{bmatrix} + \mathbf{mP_1} \begin{bmatrix} \Delta u(k-1) \\ \Delta u(k-2) \\ \vdots \\ \Delta u(k-nb) \end{bmatrix} + \\ + \mathbf{mP_2} \begin{bmatrix} w(k+d+1) \\ w(k+d+2) \\ \vdots \\ w(k+d+N_u) \end{bmatrix} + \mathbf{mP_{V1}} \begin{bmatrix} \Delta v(k-1) \\ \Delta v(k-2) \\ \vdots \\ \Delta v(k-nd+1) \end{bmatrix} + \\ + \mathbf{mP_{V2}} \begin{bmatrix} \Delta v(k) \\ \Delta v(k+1) \\ \vdots \\ \Delta v(k+N_u-1) \end{bmatrix} \quad (13)$$

If the future load disturbance is constant and equal to the last measured value (i.e. $\Delta v(k) = 0$), the last term of the equation (13) vanishes (Pawlowskaa et al. 2012).
It is evident that matrices $\mathbf{H_{V1}}$ and $\mathbf{H_{V2}}$ are dependent on the relative difference between number of steps of time-delay of input-output and disturbance-output which is defined as

$$\rho = d - dv \quad (14)$$

This leads to three types of structures for matrices $\mathbf{H_{V1}}$ and $\mathbf{H_{V2}}$ based on the value of ρ:

- $\rho < 0$ and $\rho = 0$

$$\mathbf{H_{VX}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 \\ h_1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & 0 & 0 & 0 & 0 \\ h_{Nu+\rho} & \dots & h_1 & 0 & 0 & 0 \end{bmatrix} \Big\} |\rho| \quad (15)$$

- $\rho > 0$

$$\mathbf{H_{VX}} = \begin{bmatrix} h_{\rho+1} & h_\rho & \dots & h_1 & 0 & 0 \\ h_{\rho+2} & h_{\rho+1} & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & h_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & h_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N+\rho} & h_{N+\rho-1} & \dots & \dots & h_\rho & h_{\rho+1} \end{bmatrix} \Big\} \rho \quad (16)$$

where h_i are the coefficients of $\mathbf{H_{V1}}$ and $\mathbf{H_{V2}}$ matrices obtained from the delay free disturbance response moved up/down according to value of ρ (Pawlowskaa et al. 2012).

RECURSIVE IDENTIFICATION

The identification of systems deals with the problem of creating mathematical models of dynamical systems based on data observed from the system. It is an

alternative procedure for obtaining a model in case, when it is not possible to determine a set of differential equations that describes the dynamic behavior of the system. The MPC requires an internal model of the system; therefore, really precise model of process is necessary for correct behavior of predictive algorithm.
Identification of control processes can be divided into two groups, which are used most often. The first group is Offline (one-time) Identification Methods (OfIM) and the second is Online (ongoing) Identification Methods (OnIM).
The OfIM type as well as the OnIM type can be used during the real-time control of processes. The estimated parameters obtained from the OfIM are usually selected as a starting point for the STC. They can be also chosen as the internal model throughout the control procedure when the process does not change its dynamic behavior much and adaptive control is not required. These STCs can utilize an auto-tuning or adaptive approach in many practical applications (Bitmead et al. 1990). The most known adaptive approach is to use OfIM recursively.

Offline Identification Methods

The well known OfIM is the MATLAB function *fminsearch*. It finds the minimum of the entered function without restricting conditions. The entered function can be single variable or multivariable type. This function uses the simplex search method for finding the minimum of a function. This is a direct search method that does not use numerical or analytic gradients. However, the most known method for the identification of the discrete transfer function model parameters is the Least Squares Method (LSM) based on the idea of linear regression. This identification algorithm can be carried out in a recursive manner as well in an order to use it for STC. The LSM is based on minimizing the sum of squared subtraction of measured and model output value.
The LSM is defined as the vector $\hat{\mathbf{\Theta}}$ that minimizes the quadratic error

$$\hat{\mathbf{\Theta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{y} \tag{17}$$

Note, that $\hat{\mathbf{\Theta}}$ is a vector of estimated model parameters, which has dimension $2n$, $\mathbf{F}$ is a matrix of dimension $N - n - d \times 2n$, $\boldsymbol{y}$ is a data vector of dimension $N - n - d$, where N is a number of measured data, n is an order of system, and d is a number of steps of time-delay. The $\mathbf{F}$ depends on past inputs and outputs and that this condition can be fulfilled if the input signal sequence is adequately chosen in such a way that the obtained vectors are linearly independent. (Camacho and Normey-Rico 2007).

Online Identification Methods

The OnIM are mainly used to adjust the estimates of the process parameters from initial estimates in the each sampling period. Since the approach, when calculation of estimated parameters is performed each sampling period, these methods are capable to react on changes in a dynamic behavior of system as well as they are able to compensate slightly non-liner behavior of the system.
One of the advantages of the process parameter estimation using the LSM is fact, that this algorithm can be used recursively. The parameter vector computed at step k can be computed as a function of the parameter vector estimated at step $k-1$.
The recursive least squares method (RLSM) is the most known recursive method and it uses the AutoRegressive eXogenous (ARX) model (Bobál et al. 2005).

$$y(k) = \mathbf{\Theta}^T(k)\mathbf{\Phi}(k) + e_s(k) \tag{18}$$

where $\mathbf{\Theta}$ is a vector of model parameters

$$\mathbf{\Theta}^T(k) = \begin{bmatrix} a_1 & a_2 \ldots a_n & b_1 & b_2 \ldots b_n \end{bmatrix} \tag{19}$$

and $\mathbf{\Phi}$ is a regression vector

$$\mathbf{\Phi}^T(k) = \begin{bmatrix} -y(k-1) & -y(k-2) \ldots -y(k-n) \\ u(k-d-1) & u(k-d-2) \ldots u(k-d-n) \end{bmatrix} \tag{20}$$

Final RLSM algorithm can be defined as

$$\hat{\mathbf{\Theta}}(k) = \hat{\mathbf{\Theta}}(k-1) - \frac{\mathbf{C}(k-1)\mathbf{\Phi}(k)}{1+\xi(k)}\hat{e}(k) \tag{21}$$

where $\mathbf{C}$ is covariance matrix and

$$\xi(k) = \mathbf{\Phi}^T(k)\mathbf{C}(k-1)\mathbf{\Phi}(k) \tag{22}$$

The RLSM can be modified by weighting of the past data and forgetting of them to always work with the most actual and relevant data. Application of the RLSM with exponential forgetting results in a more realistic situations. Parameters of the control law are being continuously adjusted in order to track time-varying properties of the controlled plant (Bobal et al. 2005; Skormin 2016). Final algorithm is defined as

$$\hat{\mathbf{\Theta}}(k) = \hat{\mathbf{\Theta}}(k-1) + \frac{\mathbf{C}(k-1)\mathbf{\Phi}(k)}{\varphi^2+\xi(k)}\hat{e}(k) \tag{23}$$

where covariance matrix is defined as follows

$$\mathbf{C}(k) = \frac{1}{\varphi^2}\left[\mathbf{C}(k-1) - \frac{\mathbf{C}(k-1)\mathbf{\Phi}(k)\mathbf{\Phi}^T(k)\mathbf{C}(k-1)}{\varphi^2+\xi(k)}\right] \tag{24}$$

The RLSM with adaptive directional forgetting eliminates disadvantages of the RLSM with exponential forgetting. It forgets old information only in the direction in which new data bring new information, which also helps to avoid the estimator windup effect.
The RLSM with exponential forgetting as well as with adaptive directional forgetting has been chosen as an algorithm for the STC algorithm used in the introduced Toolbox (Bobal et al. 2005; Skormin 2016).

TOOLBOX DESCRIPTION

The Toolbox for the STC GPC of time-delayed processes with measurable disturbance compensation is depicted in the Figure 2. This Toolbox was developed using the MATLAB R2014b. Basic setting of Toolbox

is possible in the *init.m* file, which is an initialization routine. This routine is executed automatically once the simulation is started using the MATLAB/SIMULINK.

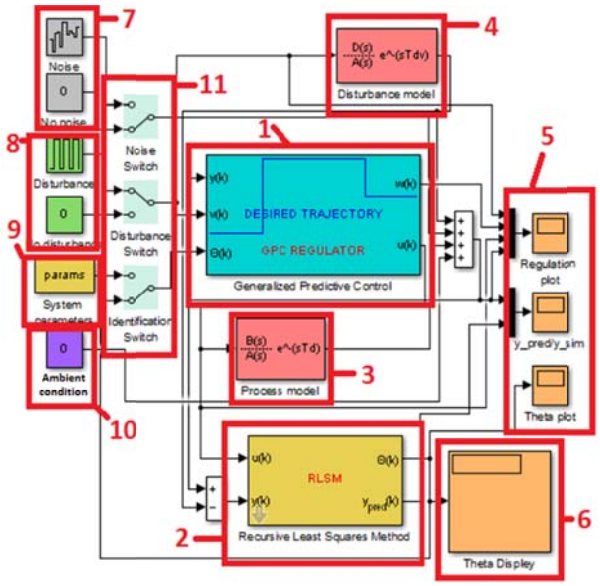


Figure 2 : STC GPC Toolbox MATLAB Scheme

The Toolbox consists of the following parts:
1 – GPC controller block
2 – RLSM identification block (used for STC)
3 – Controlled process model
4 – Disturbance model
5 – Resulting charts of control courses and predicates
6 – Estimated parameters of process from RLSM
7, 8 – Noise signal and disturbance signal
9 – System parameters setting
10 – Ambient condition setting signal
11 – Activation/deactivation switches of 9, 10, and 11

The GPC controller block 1 contains three tabs, see Figure 3. First tab is used for the setting of the Sample time, Dead times (time delays), Control horizon and Weighting parameters. Second tab is intended to design desired trajectory and last tab can set properties of disturbance compensation.

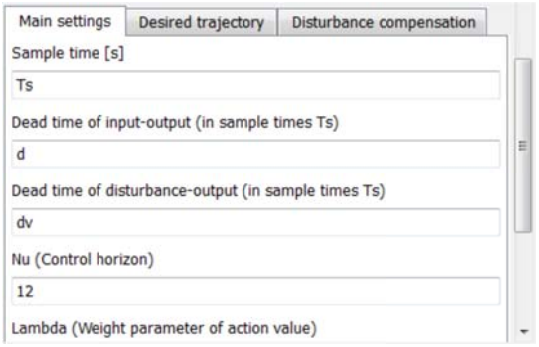


Figure 3 : Main Settings of GPC Window

The measurable disturbance compensation can be enabled/disabled based on the checkbox as it is visible on the Figure 4. If disabled, No Disturbance Compensation (NDC) approach is used. In case when disturbance compensation is enabled, user can choose one of the two possible ways of its compensation. First, the Normal Measurement Disturbance (NMD) means that disturbance is measured every sample interval and output is compensated based on the present and past data of disturbance. Second option is to use the Predicted Vector of Disturbance (PVD), where course of the disturbance over time is known during the whole simulated control process. Moreover, disturbance is measured for overcompensation of invalid data as well. For example, disturbance can be first measured and used for the control or disturbance vector can be statistically computed based on the past data, etc. Usage of the PVD significantly improves whole control process in terms of control quality.

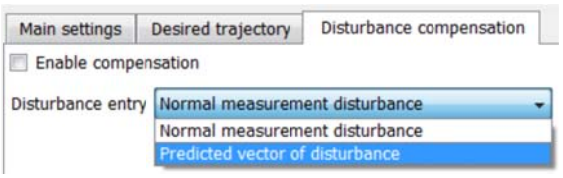


Figure 4 : GPC Disturbance Compensation Window

The RLSM identification block 2 is allowed for the self-tuning purposes by the Identification Switch 11. Otherwise, identified parameters of process are constants during the control process. The RLSM block allows to set Sample/dead time in the first tab. Second tab is designed for setting of the Type of identification (RLSM, RLSM with exponential forgetting, or RLSM with adaptive directional forgetting). Other boxes can modify the RLSM parameters.

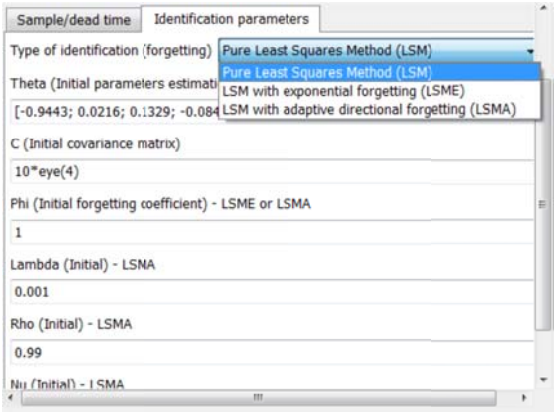


Figure 5 : RLSM Settings for the STC Window

To execute the simulation, set parameters of the controlled system in the *init.m* file first. Then, enable/disable Noise and external Disturbance using Switches 11. Choose STC with RLSM or non STC algorithm using the Identification Switch. Set Ambient temperature, if required. Set GPC controller block 2 and RLSM identification block according to description above. Run the simulation and display charts using 5.

SIMULATION VERIFICATION OF TOOLBOX

The designed STC GPC Toolbox was verified by simulation on three examples – GPC algorithm without STC, STC GPC algorithm, and GPC algorithm with disturbance compensation.

The Controlled process model (item 3 of the Figure 2) is represented as the following continuous transfer function, which was used for all simulations.

$$G_S(s) = \frac{B(s)}{A(s)} e^{-sT_d} = \frac{2}{20s^2 + 9s + 1} e^{-10s} \qquad (25)$$

The following second order linear discrete transfer function was used for the simulation purposes as a model for estimated parameters of controlled process for GPC and RLSM.

$$G_S(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} z^{-d} = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} z^{-d} \qquad (26)$$

GPC Algorithm without STC

First, the GPC algorithm without STC was verified. Very precise estimated model (27) of the controlled system was used.

$$G_S(z^{-1}) = \frac{0.1490 z^{-1} + 0.1104 z^{-2}}{1 - 1.2770 z^{-1} + 0.4066 z^{-2}} z^{-5}, \quad T_S = 2s \qquad (27)$$

Following parameters were used for all simulations:

$d = 5$, $\delta = 1$, $\lambda = 1$, $N_1 = 6$, $N_u = 10$, and $N_2 = 15$

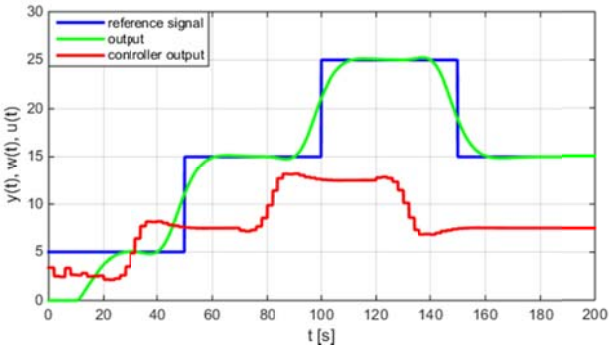


Figure 6 : GPC without STC – Exact Identification

From the Figure 6 it is obvious that the control quality is very good after start-up phase and overshoots are not significant. Next, the exact estimated parameters were changed subsequently

$$G_S(z^{-1}) = \frac{0.2000 z^{-1} + 0.1000 z^{-2}}{1 - 1.5000 z^{-1} + 0.5000 z^{-2}} z^{-5} \qquad (28)$$

where inaccurate identification is performed.

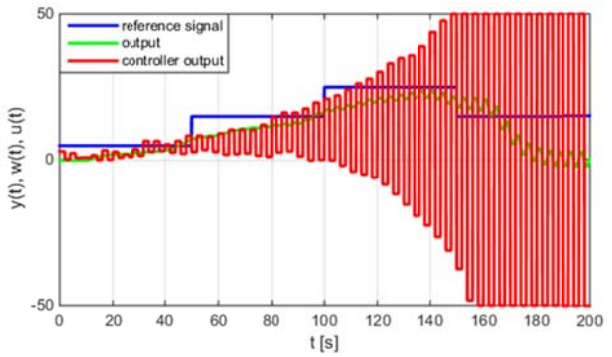


Figure 7 : GPC without STC – Inaccurate Identification

The Figure 7 shows that when identification process is underestimated and parameters are not accurate, whole control process becomes unstable and it cannot be effectively controlled using GPC.

STC GPC Algorithm

Disadvantages of inaccurate parameters estimation can be eliminated by STC GPC algorithm. The RLSM with adaptive directional forgetting and the controlled system model (28) was used for the simulation purposes.

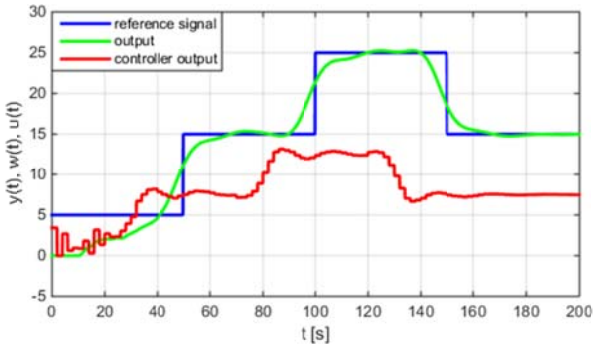


Figure 8 : STC GPC – Control Courses

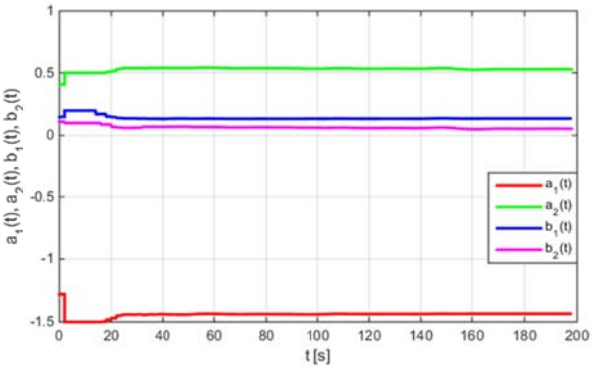


Figure 9 : STC GPC – Parametrs Evolution

The Figure 8 depicts that when GPC STC is used, even initial inaccurate system parameter estimation does not prevent high control quality. System parameters evolution is captured on the Figure 9.

GPC Algorithm – Disturbance Compensation

The controlled system model (27) was used to verify functionality of the GPC with NDC, NMD, and PVD. Control courses are depicted on the Figure 10.

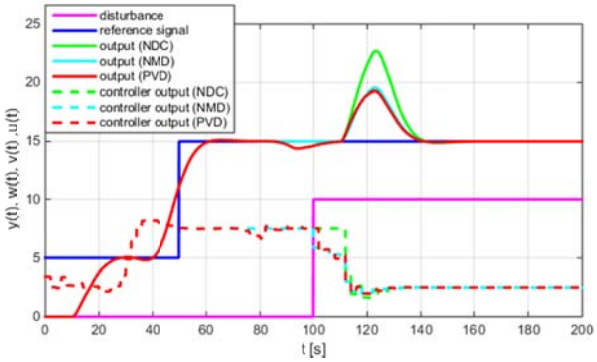


Figure 10 : GPC – Disturbance Compensation

CONCLUSION

This paper has presented an extended STC GPC Toolbox with possibility to compensate the measurable disturbance. The Toolbox has been created in the MATLAB/SIMULINK environment with a purpose to create a simulation suitable for the design and verification of adaptive control of time-delay systems with usage of the MPC strategy.

The GPC algorithm as itself without using the STC is able to control processes in a really good quality after start-up phase and overshoots are not significant. However, this is possible only in case when process parameters are estimated very precisely. Otherwise, in case when these parameters are not estimated with sufficient precision, control process becomes unstable. First option is to use suitable identification process for parameters estimation or use the STC algorithm. The STC algorithms have several advantages, e.g. initial parameter estimation can be only raw, slightly nonlinear process can be controlled using the STC, and influence of an unexpected conditions during the control process can be eliminated. The simulation shows that RLSM with adaptive directional forgetting can be suitable STC algorithm for the MPC, in general.

Incorporation of the disturbance compensation into the control law can have really positive effect on overall control processes. An overshoot caused by the disturbance can be eliminated when it is measured and predicted. The GPC with NMD and PVD improves the control quality and reduces the overshoot in comparison with the GPC with NMC.

Toolbox is maintained by the Department of Process Control, Faculty of Applied Informatics, Tomas Bata University in Zlin and for its downloading, feel free to contact authors or mentioned department.

ACKNOWLEDGEMENT

This work was supported by the Czech Republic Ministry of Education - grant IGA/CebiaTech/2018/002.

REFERENCES

Bitmead, R.R., Gevers, M. and V. Hertz. 1990.*Adaptive optimal control. The thinking man's GPC*, Prentice Hall, Englewood Cliffs, New Jersey.

Bobál, V., Böhm, J., Fessl, J. and J. Macháček. 2005. *Digital Self-tuning Controllers: Algorithms, Implementation and Applications.* Springer-Verlag, London.

Camacho E.F. and C. Bordons. 2004. *Model predictive control*, Springer Verlag, London.

Camacho E.F. and J.E. Normey-Rico. 2007.*Control of dead-time processes*, Springer-Verlag, London.

Clarke, D.W., Mohtadi, C. and P.S. Tuffs. 1987a. "Generalized predictive control, part I: the basic algorithm", Automatica 23, 137-148.

Clarke D.W.; C. Mohtadi and P.S. Tuffs. 1987b. "Generalized predictive control, part II: extensions and interpretations", Automatica 23,149-160.

Fikar, M. and J. Mikleš. 2008. *Process modelling, optimisation and control*, Springer-Verlag, Berlin.

Haber, R., Bars, R.and U. Schmitz. 2011. *Predictive control in process engineering: From the basics to the applications*. Weinhaim: Willey-VCH Verlag.

Kvasnica, M. 2009. *Real-Time Model Predictive Control via Multi-Parametric Programming: Theory and Tools*, VDM Verlag.

Maciejowski, J. M. 2002.*PredictiveControl, with Constraints*,Pearson Education.

Moudgalya, K.M. 2007.*Digital control*. Chichester: John Wiley.

Pawlowskaa A., Guzmána J. L., Normey-RicoJ. E. and M. Berenguela. 2012."Improving feedforward disturbance compensation capabilities in Generalized Predictive Control". Journal of Process Control 22, 527-539.

Qin, S. J. and T. A. Badgewell. 1997. *An overview of industrial model predictive control technology.* In: Chemical Process Control – V, volume 93, no. 316, 232-256. AIChe Symposium Series – American Institute of Chemical Engineers.

Rawlings, J. B. and D. Q. Mayne. 2009. *Model Predictive Control: Theory and Design*. Nob Hill Pub.

Rossiter, J.A. 2003.*Model based predictive control: a practical approach*, CRC Press.

Schwarz, M. H., Cox, C. S. and J. Börcsök. 2010."A Filtered Tuning Method for a GPC Controller". In: University of Kassel, Germany, 180-185.

Skormin, A. V. 2016. *Introduction to Process Control: Analysis, Mathematical Modeling, Control and Optimization,* Springer Verlag, London.

Smith, O. J. 1957. *Closed control of loops with dead time.* Chem. Eng. Progress 53,217-219.

AUTHOR BIOGRAPHIES

RADEK HOLIŠ studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master's degree in Automatic Control and Informatics in 2014. He now attends Ph.D. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interest focuses on modeling and simulation of discrete technological processes, adaptive control and model predictive control. He currently works at Honeywell HTS-CZ, Brno in Aerospace division as Software Design Engineer. His email address is rholis@fai.utb.cz.

VLADIMÍR BOBÁL graduated in 1966 from the Brno University of Technology, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are adaptive and predictive control, system identification and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.

NEW APPROACH TO MODELLING THE KINETICS OF THE FERMENTATION PROCESS IN CULTIVATION OF LACTIC ACID BACTERIA

Georgi Kostov*, Rositsa Denkova-Kostova**, Vesela Shopska*, Petar Nedyalkov*, 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; inj_petar_nedyalkov@abv.bg
Bogdan Goranov
LBLact, Plovdiv, Bulgaria, E-mail: goranov_chemistry@abv.bg
Vasil Iliev
Weissbiotech, Ascheberg, Germany, E-mail: illiev.vasil@gmail.com
Kristina Ivanova, Desislava Teneva
Food Research and Development Institute, E-mail: kriss_k@abv.bg; desi_gerinska@yahoo.com

KEYWORDS

Lactic acid bacteria, kinetic, modelling, semi-empirical models

ABSTRACT

The present paper reviews the methods for assessing the kinetics of the fermentation process for the cultivation of probiotic strains of lactic acid bacteria in a laboratory bioreactor with stirring. To describe the kinetics of the lactic acid fermentation process, an alternative approach with the use of semi-empirical degree laws that offer a new look into the biological parameters in the description models is proposed.

INTRODUCTION

Probiotic lactic acid bacteria

Lactobacilli belong to the natural microflora of human and animal organisms. They normally exist in the oral cavity, gastrointestinal tract, and vaginal microflora (Adams, 1999).

The most common lactobacilli species isolated from the gastrointestinal tract of humans are *Lactobacillus brevis, Lactobacillus casei, Lactobacillus acidophilus, Lactobacillus plantarum, Lactobacillus fermentum* and *Lactobacillus salivarius* (Slover, 2008).

The cellular components of certain lactic acid bacteria protect and modulate the immune system of the human body and improve its health status (Pirt, 1975; Salminen and von Wright, 1998a,b). *L. delbrueckii* ssp. *bulgaricus*, *L. acidophilus*, *L. casei, L. plantarum* are the major microorganisms that regulate the balance of the gastrointestinal microflora (Salminen and Wright, 1998b). Lactic acid bacteria provide the human organism with growth factors, amino acids, vitamins, organic acids, etc. through their metabolism. They have the ability to adhere to the intestinal mucosa, stop the formation of enterobacterial colonies, and prevent the colonization of the gut epithelium by bacteria coming from outside (Pirt, 1975). These regulatory functions are realized by the organic acids, bacteriocins, bacteriocin-like substances (BLIS), and other metabolites produced by lactic acid bacteria through which they inhibit enteropathogens (Fang, 2001).

Lactic acid bacteria and their metabolic products have beneficial effects on the digestive system and positive action during and after antibiotic treatment (O'Brien et al., 1999).

The mechanism of action of lactobacilli comprises: a) suppression of microbial putrefaction processes; b) prevention of constipation, colon cancer, etc.; c) prevention and treatment of antibiotic-associated diarrhea; d) stimulation of the immune system; e) suppression of toxic processes in the digestive tract (Pirt, 1975; Fuller, 1986; Salminen and Wright, 1998a,b). Lactic acid bacteria are also characterized by antimutagenic, anticancerogenic and antitumor activity (Hosono et al., 1990).

Lactic acid bacteria with proven probiotic properties are included in the composition of probiotic preparations, starter cultures, and in the production of dairy, meat and other products with functional properties. One of the requirements for a strain to be probiotic is that it would allow industrial processes to take place, including cultivation, and accumulation of high concentrations of viable probiotic cells in the cultivation process. One of the main stages in the production of probiotics and starter cultures is the cultivation of the selected strains in industrial bioreactors of different construction. The cultivation process defines mainly the qualitative and quantitative characteristics of the production process as a whole.

Methods of cultivation

A. Batch cultivation. In this method, the microbial population grows in a closed space without changing the medium volume; without the addition of any substrates, and with the addition of substances that correct some of the parameters only. Process parameters are a function of time, and the process is non-stationary. Batch cultivation of lactic acid bacteria is characterized by the following stages (Abdel-Rahmanq et al., 2013):

• Lag-phase: it occurs immediately after inoculation of the nutrient medium and aims at adapting the microbial population to the conditions of the medium. Cells undergo biochemical processes of synthesis of cellular structures needed for the binary fission. Additionally, if

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

the medium contains high molecular weight molecules, the cells release extracellular enzymes to break them down to low molecular weight compounds, thereby facilitating their intracellular transport.

• Exponential phase: the cells grow and divide intensively. They multiply at a maximum rate.

• Stationary phase: there is depletion of one or more substances from the growth medium, and the cells stop multiplying. During the stationary phase, the number of newly formed cells is equal to the number of dying cells, so no increase in biomass is observed, but the cells have a preserved metabolic activity.

• Death phase: during this phase, the number of cells that are dying is significantly greater than the number of newly formed cells.

B. Batch cultivation with pH correction.

The constantly increasing concentration of lactic acid during batch cultivation has an inhibitory effect on the growth of lactic acid bacteria. To remove it, pH adjustment is required during the batch fermentation process. The correction is accomplished by the addition of NaOH, KOH, $CaCO_3$, and ammonia water, and leads to a fuller absorption of the substrate and an increase in the amount of lactic acid produced and accumulated (Hetenyi et al., 2011, Abdel-Rahman et al., 2011a, b, Adsul et al., 2007; Tashiro et al., 2011). According to a number of authors, the optimum pH range for the growth of lactic acid bacteria is between 5 and 6 (Fu and Mathews 1999, Yuwono and Kokugan 2008).

C. Continuous fermentation.

Continuous cultivation systems provide increased productivity and reduce the inhibitory effect of lactic acid on cell growth. They are classified as open and closed continuous cultivation systems.

The most important factors for the production of probiotic products in continuous mode are the closed continuous cultivation systems. The targeted management of the lactic acid fermentation process is an important indicator for ensuring the quality of lactic acid foods and liquid probiotic preparations. The main point in the technological process is the provision of optimal conditions for the growth of the microbial cells, ensuring the accumulation of high active flora concentration, and creating conditions for obtaining standardized starters with constant properties and biochemical activity (Driessen et al., 1977).

Kinetic models for lactic acid fermentation description.

The description of the kinetics of microbial growth is done through a number of models. Currently, different types of dependencies are used in practice, many of which are based on Monod's classic equation, but there are other types of models that offer a new look at the fermentation process.

The Monod equation expressing the specific growth rate dependence on the concentration of the limiting substrate is analogous to the Michaelis-Menten equation (Bouguettoucha, et al., 2011; Abboud et al., 2010; Dey and Pal, 2013; Ghaly et al., 2005):

$$\mu = \mu_{\max} \frac{S}{K_S + S} \quad (1)$$

where: μ - specific growth rate, h-1; μ_{max} - maximum specific growth rate, h^{-1}; K_s - saturation constant, kg/m^3; S - substrate concentration kg/m^3. The saturation constant is equal to the substrate concentration at which the specific growth rate is half of its maximum value ($\mu = 0.5\mu_{max}$).

High concentrations of the substrate can lead to cell growth inhibition. This process should be taken into account when developing the mathematical model. The Andrews-Halden model is one of the most commonly used models for description of the inhibitory action of high substrate concentrations (Abboud et al., 2010, Dey and Pal, 2013; Bouguettoucha et al., 2011).

$$\mu = \mu_{\max} \frac{S}{K_S + S + \frac{S^2}{K_i}} \quad (2)$$

where: K_i - inhibition constant

Another model for describing the inhibitory action of high substrate concentrations is the Edward model (Abboud et al., 2010; Dey and Pal, 2013; Bouguettoucha et al., 2011):

$$\mu = \mu_{\max} \frac{S}{(K_S + S) + (1 + \frac{S^2}{K_i})} \quad (3)$$

The specific growth rate depends not only on the concentration of the limiting substrate but also on other factors, primarily the concentration of the metabolic products. Yerusalimski proves that this dependence is described by the non-competitive inhibition equation (Abboud et al., 2010; Dey and Pal, 2013; Bouguettoucha et al., 2011):

$$\mu = \mu_{\max} \frac{K_P}{K_P + P} \quad (4)$$

where: P - concentration of the inhibitory product, kg/m^3; K_p - constant, kg/m^3.

K_p is the concentration of the inhibitory product in which the specific growth rate is equal to half of its maximum value ($\mu=0.5\mu_{max}$).

All growth parameters in the previous models have no clear biological meaning. The biological requirement of the model parameters is satisfied by the Verhulst equation (Bouguettoucha et al., 2011):

$$\frac{dX}{d\tau} = \mu X - \beta X^2 = \mu X - \frac{\mu}{X_K} X^2 \quad (5)$$

where: X – biomass concentration, kg/m^3; X_k – final biomass concentration, kg/m^3; β - coefficient of internal population competition, kg/(kg.h)

SEMI-EMPIRICAL MODELS: A NEW APPROACH TO THE DESCRIPTION OF MICROBIAL KINETICS

Differential equations used in practice for describing the lactic acid process are rarely solved with high accuracy, largely due to the presence of too many ambiguous variables. This is particularly true of the more complex models describing the kinetics of the fermentation

process. Thus, a purely numerical solution can hardly be found, and in some cases the solution to a task is semi-empirical rather than analytical (Tishin and Fedorov, 2016).

In their works, Tishin and Fedorov, 2016; Tishin et al., 2015 suggest the use of a different principle for describing microbial kinetics. It is based on various assumptions about continuous cell division and the proportionality of biomass accumulation in time, its concentration in the culture medium and the cultivation time. In this case, cell multiplication can be described with the following dependence:

$$dX = kX^m \tau^n \quad (6)$$

where: ***k*** - a proportionality factor, ***m*** and ***n*** - degree indicators.

Depending on the values of the degree indicators, different mathematical models can be obtained for the same fermentation process.

For example, at $n = 0$ the step model (7) is obtained, and the coefficient k_1 is analogous to the specific growth rate μ (Tishin and Fedorov, 2016).

$$dX = k_1 X^m \quad (7)$$

In real form, after integration, equation (7) acquires the form:

$$X_B^{1-m} = 1 + \frac{(1-m)k1}{X_0^{1-m}} \tau \quad (8)$$

Equation (8) can be converted by entering parameter δ and assuming that:

$$m_1 = \frac{1}{1-m} \quad (9)$$

where:

$$X_B = (1 + \delta\tau)^{m_1} \quad (10)$$

According to Tishin et al., 2015, parameter δ is the average specific growth rate for the entire cultivation process interval.

In the second case, the differential equation can be solved if $m = 0$ and $n = 1$ are known:

$$dX = k_2 X^n \quad (11)$$

After integrating the obtained equation and dividing by x_0:

$$X_B = 1 + \frac{k_2}{(n_1+1)X_0} \tau^{(n_1+1)} \quad (12)$$

If the parameter γ, which is the ratio on the right side of equation (12), is introduced, the following is obtained:

$$X_B = 1 + (\gamma\tau)^n \quad (13)$$

γ can be considered an average specific growth rate, but it has a much clearer meaning since it can easily be shown that $1/\gamma$ represents the doubling time of the cell population (Tishin and Fedorov, 2016 Tishin et al., 2015).

Using Equations (8) and (13), equations for the substrate consumption during the fermentation process can also be derived. The detailed solution is presented in Tishin and Fedorov, 2016; Tishin et al., 2015, where dependencies of the following type are obtained:

$$S = \frac{1}{(1+\delta_s\tau)^{m_{1s}}} \quad (14)$$

$$S = \frac{1}{(1+\gamma_s\tau)^{n_s}} \quad (15)$$

Tishin and Fedorov, 2016; Tishin et al., 2015 show that equations (13) and (15) have a clearer biological sense of their parameters and are therefore preferable when describing the kinetics of the fermentation process.

On the basis of equations (6) to (15), other types of process equations are also proposed, which also include the knowledge of kinetics when using the classical equations (1) to (5) (Tishin and Fedorov, 2016):

- exponential degree model

$$X_B = a(d - e^{-\mu\tau}) \quad (16)$$

- modified empirical model of the logistic curve

$$X_B = \mu_m \frac{1}{1 + be^{-\mu\tau}} \quad (17)$$

- Weibull's equation for describing lactic acid biosynthesis

$$K_T = c_m - c_e^{-(q\tau)^\delta} \quad (18)$$

where X_B - biomass in a dimensionless form; μ - specific relative (average) growth rate of biomass at a time interval from $\tau = 0$ to $\tau = \tau_i$; μ_m - maximum specific growth rate, h^{-1}; ***a, b*** and ***d*** - empirical coefficients carrying certain biological meaning; c_m - maximum value of titratable acidity, °T; ***c*** - coefficient equal to the difference between the maximum and the initial titratable acidity, °T; ***q*** - specific rate of acid formation, °T/cfu.cm^3.h; δ - an indicator defining the change in the curve shape or the change in the rate of lactic acid accumulation over time; τ- time of cultivation, h.

The advantage of this type of model lies in the fact that it can easily be solved by simple methods and does not require complicated numerical solution programs, but also allows for a different kind of interpretation of the biological processes observed in the cultivation of microorganisms.

The aim of the present work was to apply these types of models to describe the process of cultivation of probiotic lactic acid bacteria under static and dynamic (cultivation in a bioreactor) conditions, and to present the possibilities for interpreting the obtained results and the biological meaning of the variables.

MATERIALS AND METHODS

- Microorganisms

The study was conducted with two strains of different lactobacilli species: *Lactobacillus delbrueckii* ssp. *bulgaricus* TAB2 isolated from spontaneously fermented dairy products, and *Lactobacillus plantarum* BZ3 isolated from spontaneously fermented vegetables.

Nutrient media (ISO 7889:2005)

• MRS - broth;
• MRS-agar;
• Saline solution.

Methods of analysis

• Determination of titratable acidity (ISO/TS 11869:2012);

• Number of viable lactobacilli cells (ISO 7889:2005).
Batch cultivation
• Under static conditions. Cultivation was carried out in flasks thermostated at 37±1°C;
• Cultivation in a bioreactor. Cultivation was carried out in the laboratory bioreactor shown in Fig. 1. The apparatus has a geometric volume of 2 dm^3 and a working volume of 1.5 dm^3 and is equipped with a Sartorius A2 control device, which includes all the measuring instruments for the fermentation process: temperature, pH, dissolved oxygen, etc. The fermentation process was carried out at a stirring speed of 150 rpm at 37±1°C.

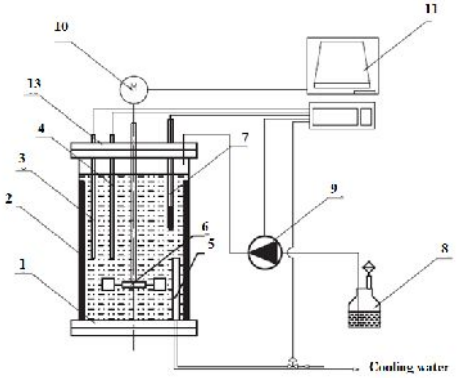


Figure 1: Laboratory bioreactor
1 - vessel with a geometric volume of 2 dm^3; 2 - baffles; 3 - temperature electrode (thermometer); 4 - cooling/heating device (water jacket); 5 - an additional cooling/heating device; 6 - turbine stirrer; 7 - pH/Eh electrode; 8 - fermentation medium/inoculum/pH adjustment medium; 9 - peristaltic pump; 10 - stirrer drive; 11 – Sartorius A2 control device;

Models for describing the kinetics of the fermentation process

The description of the kinetics was made using equations (16) to (18). The process parameters were defined using TableCurve2D and Excel. The software was also used for conducting statistical evaluation of the obtained models.

RESULTS AND DISCUSSION

The results of the studies on the dynamics of the fermentation process under static and dynamic conditions are presented in Fig. 2 and Fig. 3, and in Table 1 and Table 2. Biomass data are displayed in dimentionless form relative to the initial cellular concentration immediately after inoculation of the medium.

In the cultivation of *L.delbrueckii* ssp. *bulgaricus* TAB2, shortening of the lag-phase from 6 to 3 hours during dynamic cultivation was observed. At the same time, cultivation in a bioreactor led to the accumulation of one order higher concentration of viable lactobacilli cells, and as a result, about 10^{12} cfu/cm^3 were accumulated in the apparatus compared to static cultivation (Fig. 2.). The titratable acidity values were close and were in the range between 170-190°T (Fig. 2). The data on the static and dynamic cultivation of *L. planarum* BZ3 were similar (Fig. 3). About 10^{13}cfu/cm^3 of *L. plantarum* cells were grown in the bioreactor. At the end of the process, comparable acidity values for *L. planarum* BZ3 grown in a bioreactor and under static conditions were observed: 219°T and 214 °T, respectively.

The statistical analysis of the two models (part of the results are summarized in Table 1 and Table 2) showed that the semi-empirical models used described the kinetics of the fermentation process extremely accurately, and their accuracy was comparable to the classical kinetic models (data not presented). This was confirmed by the high correlation coefficient as well as by the low identification error of the model.

According to the exponential model for strain *L. delbrueckii* ssp. *bulgaricus* TAB2, a relatively higher specific growth rate was observed in its cultivation in the bioreactor (μ = 0.062 h^{-1}) compared to the same parameter during the static process (μ = 0.057 h^{-1}). Factor α is about 30% higher in bioreactor cultivation, indicating increased biochemical activity in the cells due to better stirring in the bioreactor compared to static cultivation. This was due to the presence of dissolved oxygen in the apparatus. It is well known that *L. delbrueckii* ssp. *bulgaricus* is the most sensitive lactic acid bacteria species in relation to oxygen, and in order to overcome the presence of oxygen in the apparatus, the oxid-peroxide system of the cells is activated (Table 1).

There was a similar trend in the cultivation of *L. plantarum* BZ3. Again, coefficient α was 22% higher, which also reflected the higher specific growth rate in the bioreactor cultivation (μ = 0.110 h^{-1}) compared to the value of the same parameter during static cultivation (μ = 0.103 h^{-1}). The reason for the observed difference was once again the dissolved oxygen, but unlike its impact on strains of the *L. delbrueckii* ssp. *bulgaricus* species, the presence of oxygen had another impact on *L. plantarum* strains. Although they do not contain the cytochromes involved in transferring electrons from the substrate to the oxygen, these lactic acid bacteria are microaerophilic, and thanks to their flavoprotein systems, the flavoprotein oxidases, in particular, they can oxidize different substrates in the presence of oxygen (Kwasnikov and Nesterenko 1975). Probably, through its flavoprotein oxidases, *L. plantarum* BZ3 manages to incorporate dissolved oxygen as the terminal acceptor of electrons in the oxidation of some substrates from the medium.

Coefficient d in the exponential model showed the influence of the culture conditions on the rate of the biochemical processes occurring in the cell. For both strains examined, higher values of this parameter were observed (2.263 for *L.delbrueckii* ssp. *bulgaricus* TAB2 and 2.462 for *L.plantarum* BZ3) compared to the values of the same parameter during static cultivation (1.994 and 2.218, respectively). The observed difference was a result of the delayed diffusion of nutrients to the surface of the cell; slow diffusion of secreted metabolic products from the cell into the culture medium; uneven temperature and pH distribution throughout the culture medium that put the microorganisms in the different microvolumes of the medium under different growth conditions. This was due to the lower kinetic parameter values for the studied strains cultivated under static conditions.

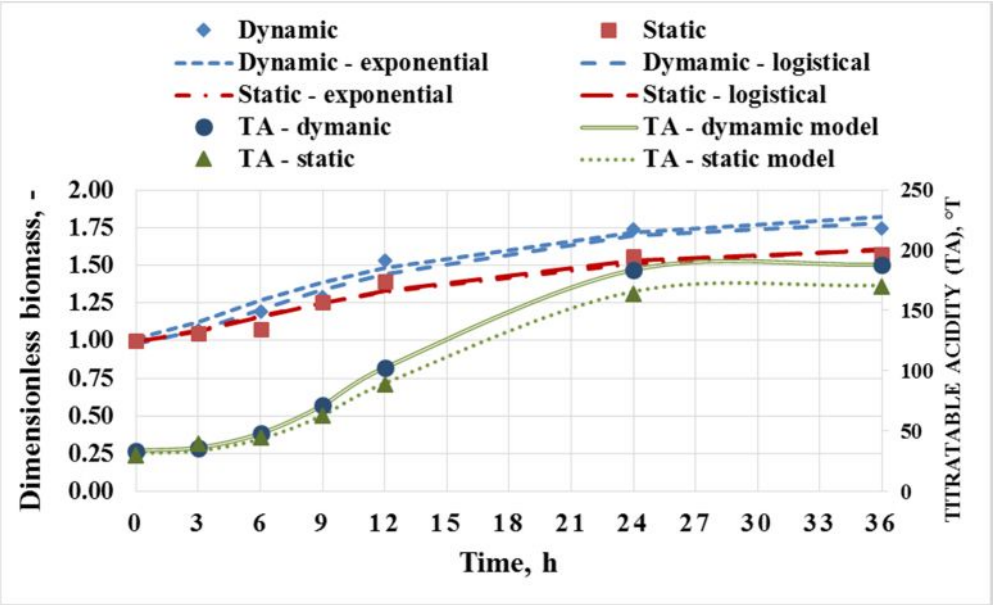


Figure 2: Dynamics of cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* TAB2 and comparison of experimental data with kinetic models.

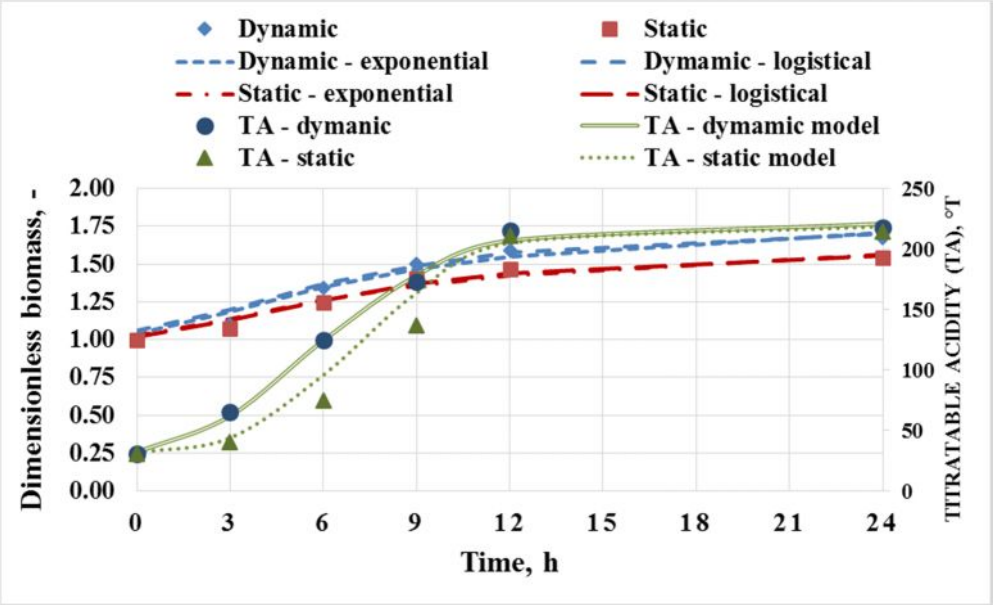


Figure 3: Dynamics of cultivation of *Lactobacillus plantarum* BZ3 and comparison of experimental data with kinetic models.

From the results presented in Table 1 for the logistic curve model, it can be seen that the relative and maximum specific growth rates for the two strains studied were higher in their cultivation in the bioreactor compared to static cultivation. For *L. delbrueckii* ssp. *bulgaricus* TAB2, $\mu = 0.108\ h^{-1}$ and $\mu_m = 1.818\ h^{-1}$ in dynamic cultivation, and $\mu = 0.096\ h^{-1}$ and $\mu_m = 1.637\ h^{-1}$ under static conditions. For *L. plantarum* BZ3, the values of these kinetic parameters when grown in a bioreactor with mechanical stirring and under static conditions were $\mu = 0.175\ h^{-1}$ and $\mu_m = 1.719\ h^{-1}$ and $\mu = 0.156\ h^{-1}$ and $\mu_m = 1.585\ h^{-1}$, respectively.

From the experimental data presented in Table 1, it appears that the exponential model yields lower relative growth rates for the two strains in their cultivation both in the bioreactor and in the static process compared to the logistic model.

An analogous trend was also observed for the coefficient ***b*** in the logistic model showing the difference in the intensity of the biochemical processes occurring in the cells. For both strains tested, the value of this parameter was higher in cultivation in a bioreactor than in static cultivation: 0.956 and 0.725, respectively, for *L. delbrueckii* ssp. *bulgaricus* TAB2, and 0.739 and 0.649, respectively, for *L. plantarum* BZ3 (Table 1).

The results on the lactic acid formation are also interesting. The data from Table 2 show that the type of cultivation did not affect the degree of acid formation of *L. delbrueckii* ssp. *bulgaricus* TAB2. In the cultivation of this strain under both dynamic and static conditions, almost the same acid formation rate was observed: q =

0.0009 °T/cfu.cm^3.h and q=0.0008°T/cfu.cm^3.h, respectively. The same trend was evident for parameter δ (Table 1). The value of δ indicates that the test strain produced high amounts of lactic acid.

Table 1: Kinetic parameters of the empirical mathematical models used

Strain	Empirical model	Cultivation in bioreactor					Static cultivation				
		μ, h^{-1}	*a*	*d (b*)*	R^2	error	μ, h^{-1}	*a*	*d (b*)*	R^2	error
L.delbrueckii ssp. *bulgaricus* TAB2	Exponential	0.062	0.953	1.994	0.9793	0.076	0.057	0.752	2.263	0.9701	0.070
	Logistic*	0.108	1.818	0.956	0.9706	0.065	0.096	1.637	0.725	0.9520	0.065
L.plantarum BZ3	Exponential	0.110	0.807	2.187	0.9828	0.065	0.103	0.658	2.462	0.9840	0.059
	Logistic*	0.175	1.719	0.739	0.9891	0.052	0.156	1.585	0.649	0.9840	0.059

** for logistic model*

Table 2: Kinetic parameters of the Weibull model for the lactic acid formation kinetics of *L. delbrueckii* ssp. *bulgaricus* TAB2 and *L. plantarum* BZ3

Cultivation conditions	Strain	c_m, °T	c, °T	q, °T/cfu.cm^3.h	δ	R^2	error
In the bioreactor	*L. delbrueckii* ssp. *bulgaricus* TAB2	188	154	0.0009	2.61	0.9986	0.341
Static		171	139	0.0008	2.62	0.9978	0.310
In the bioreactor	*L.plantarum* BZ3	221	189	0.020	1.96	0.9965	0.264
Static		219	187	0.037	2.64	0.9970	0.255

In *L. plantarum* BZ3, the type of cultivation method had an effect on the degree of acid formation. Upon cultivation of the strain in a bioreactor, a slower acidification rate of q = 0.020 °T/cfu.cm^3.h and a lower value of δ = 1.96 was observed compared to the values of these parameters in static cultivation: q = 0.037 °T/cfu.cm^3.h and δ = 2.64.

These results make it possible to conclude that the dissolved oxygen in the *L.plantarum* BZ3 cultivation in the bioreactor had greater influence on acid formation, while in the *L. delbrueckii* ssp. *bulgaricus* TAB2, cultivation the dissolved oxygen did not affect the acid formation. This was the reason for the equalization of the titratable acidity values at the end of the process during *L. plantarum* BZ3 static and dynamic cultivation. Almost the same amounts of lactic acid at the end of the two processes were accumulated as a result of the more intensive acid formation process under static conditions, although during the lag phase in static cultivation, a retention in the titratable acidity values followed by a more intensive process of acid formation was observed.

CONCLUSION

The present paper reviews a new approach to the description of the fermentation kinetics of lactic acid bacteria cultivation. For this purpose, semi-empirical dependencies were used that allow the differential equations of the fermentation process to be solved analytically and without the help of complex mathematical procedures for the identification of process parameters. The proposed model was applied to the description of the lactic acid process in the cultivation of representatives of the *Lactobacillus delbrueckii* ssp. *bulgaricus* and *Lactobacillus plantarum* species. Through the described process kinetics, it was shown that the presence of dissolved oxygen in the culture medium led to differences in the cultivation process and hence to substantial differences in kinetic parameters. The parameters of the proposed models also revealed differences in the biochemical intensity of the processes occurring in the lactic acid bacteria cells.

REFERENCES

Abboud M. M., Aljundi I. H., Khleifat K. M., Dmour S., 2010. "Biodegradation kinetics and modeling of whey lactose by bacterial hemoglobin VHb-expressing *Escherichia coli* strain." *Biochemical Engineering Journal*, 48, 166–172.

Abdel-Rahman M.A., Tashiro Y., Zendo T., Hanada K., Shibata K., Sonomoto K., 2011a. "Efficient homofermentative L-(+)-lactic acid production from xylose by a novel lactic acid bacterium, Enterococcusmundtii QU 25." *Appl Environ Microbiol*, 77, 1892–1895.

Abdel-Rahman M.A., Tashiro Y., Zendo T., Shibata K., Sonomoto K., 2011b. "Isolation and characterizationof lactic acid bacterium for effective fermentation of cellobiose intooptically pure homo L-(+)-lactic acid." *Appl Microbiol Biotechnol*, 89, 1039–1049.

Abdel-Rahman A. M., Tashiro Y. Sonomoto K., 2013. "Recent advances in lactic acid production by microbial fermentation processes." *Biotechnology Advances*, 31, 877–902.

Adams M. R., 1999. "Safety of industrial lactic acid bacteria." *J. Biotechnol*, 68, 171-178.

Adsul M.G, Varma A.J., Gokhale D.V., 2007. "Lactic acid production from waste sugarcane bagassederived cellulose." *Green Chem*., 9, 58–62.

Bouguettoucha, A., Balannec B., Amrane A., 2011. "Unstructured Models for Lactic Acid Fermentation –A Review." *Food Technol. Biotechnol*., 49(1), 3–12.

Dey, P., Pal P., 2013. "Modelling and simulation of continuous L(+) lactic acid production from sugarcane juice in membrane integrated hybrid-reactor system." *Biochemical Engineering*, 79, 15–24.

Driessen F. M., Ubbels S., Standhonders S., 1977. "Continuous manufacture of yoghurt. I Optimal condition and kinetic of the prefermentation process." *Biot. Bio*, 19, 821-839.

Fang H., 2001. "Adhesion of *Bifidobacterium* spp. to human intestinal mucus." *Microbiol Immunol*, 45(3), 259 - 262.

Fu W., Mathews A.P., 1999. "Lactic acid production from lactose by *Lactobacillus plantarum*: kinetic model and effects of pH, substrate, and oxygen." *Biochemical Engineering Journal*, 3 (8), 163-170.

Fuller R., 1986. "Probiotics." *J. Appl. Bacteriol.*, Symposium Supplement, 15-75.

Ghaly A. E., Kamal M., Correia L.R., 2005. "Kinetic modelling of continuous submerged fermentationof cheese whey for single cell protein production." *Bioresource Technology*, 96, 1143–1152.

Hetényi K., Nimeth Á., Sevella B., 2011. "Role of pH-regulation in lactic acid fermentation: Second steps in a processimprovement." *Chemical Engineering and Processing*, 50, 293–299.

Hosono A., Wardojo R., Otani H., 1990. "Binding of amino acid pyrolyzates by lactic acid bacteria isolated from `Dadih'." *Lebensmittel-Wissenschaft und-Technologie*, 23, 149-153

ISO 7889:2005.

ISO/TS 11869:2012.

Kwasnikov E. I., Nestsrenko O. A., 1975. "Lactic acid bacteria – A way ot use." *Nauka*, Moskow. (in Russian)

O'Brien J., Crittenden R., Ouwehand A. C., Salminen S., 1999. "Safety evaluation of probiotics", *Trends in Food Science and Technology*, 10, 418 - 424.

Pirt S. J., 1975. "Principles of microbe and cell cultivation", *Blackwell Sci. Publication*, London.

Salminen S., von Wright A., 1998a. "Lactic acid bacteria", *Marell Dekker INC*, New York.

Salminen S., von Wright A., 1998b. "Current Probiotics – Safety Assured", *Scandinavian University Press*, ISSN 0891-060X.

Slover C. M., 2008. "*Lactobacillus*: a Review", *Clinical Microbiology Newsletter*, 30, 4, 23-27.

Tashiro Y., Kaneko W., Sun Y., Shibata K., Inokuma K, Zendo T., Sonomoto K., 2011. "Continuous D-lactic acid production by a novel thermotolerant Lactobacillus delbrueckii subsp.lactis QU 41." *Appl Microbiol Biotechnol*, 89, 1741–50.

Tishin, V. B., Fedorov A. B., 2016. "Features of the search for mathematical models for the kinetics of cultivation of microorganisms.", *Scientific Journal of NIU ITMO, Series Process and Apparatus*, 4. (in Russian)

Tishin, V. B., Medelina T. V., Golovinskaya O.V., 2015. "About a choice of mathematical models of kinetics of cultivation of Saccharomyces cerevisiae yeast in the conditions of deficiency of oxygen." *Journal of ВГУИТ*, 3, 32-37. (in Russian)

Yuwono S. D., Kokugan T., 2008. "Study of the effects of temperature and pH on lactic acid production fromfresh cassava roots in tofu liquid waste by *Streptococcus bovis*." *Biochemical Engineering Journal*, 40, 175–183.

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is associated professor at the department "Technology of wine and beer" at University of Food Technologies, Plovdiv. He received his MSc in "Mechanical engineering" in 2007, PhD on "Mechanical engineering in food and flavor industry (Technological equipment in biotechnology industry)" in 2007 from University of Food Technologies, Plovdiv and DSc on "Intensification of fermentation processes with immobilized biocatalysts". His research interests are in the area of bioreactors construction, biotechnology, microbial population's investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, beer production.

ZAPRYANA DENKOVA is professor at the department "Microbiology" at University of Food Technologies, Plovdiv. She received her MSc in "Technology of microbial products" in 1982, PhD in „Technology of biologically active substances" in 1994 and DSc on "Production and application of probiotics" in 2006. Her research interests are in the area of selection of probiotic strains and development of starters for food production, genetics of microorganisms, and development of functional foods.

VESELA SHOPSKA is head assistant professor at the department "Technology of wine and beer" at University of Food Technologies, Plovdiv. She received her MSc in "Technology of wine and brewing" in 2006 at University of Food Technologies, Plovdiv. She received her PhD in "Technology of alcoholic and non-alcoholic beverages (Brewing technology)" in 2014. Her research interests are in the area of beer fermentation with free and immobilized cells; yeast and bacteria metabolism and fermentation activity.

ROSITSA DENKOVA-KOSTOVA is assistant professor at the department "Biochemistry and molecular biology" at University of Food Technologies, Plovdiv. She received her MSc in "Industrial biotechnologies" in 2011 and PhD in "Biotechnology (Technology of biologically active substances)" in 2014. Her research interests are in the area of isolation, identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is researcher in company "LBLact", Plovdiv. He received his PhD in 2015 from University of Food Technologies, Plovdiv. The theme of his thesis was "Production of lactic acid with free and immobilized lactic acid bacteria and its application in food industry". His research interests are in the area of bioreactors construction, biotechnology, microbial population's investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics.

VASIL ILIEV is a service manager in "Weissbiotech", Ascheberg, Germany. He received his PhD in 2016 from University of Food Technologies, Plovdiv. His research interests are in the area of bioreactors construction, biotechnology, microbial population's investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, beer and ethanol production.

DESISLAVA TENEVA is assistant professor at the department of "Food Technologies" at Food Research and Development Institute, Plovdiv. She received her MSc in "Analysis and Control of Food Products - University of Food Technologies – Plovdiv" in 2008 and PhD in "Microbiology (Biological sciences)" in 2017. Her research interests are in the area of isolation, identification and selection of probiotic strains and development of starters for functional foods.

PETAR NEDYALKOV is assistant professor at the department "Technology of wine and beer" at University of Food Technologies, Plovdiv. He received his PhD in "Technology of alcoholic and non-alcoholic beverages (Brewing technology)" in 2016. His research interests are in the area of beer fermentation, non-alcoholic beverages production and modeling of technological processes.

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.

A VARIABLE DETAIL MODEL SIMULATION METHODOLOGY FOR CYBER-PHYSICAL SYSTEMS

Ir. T.G. Broenink
Dr. Ir. J.F. Broenink
Robotics and Mechatronics
University of Twente
Enschede, Netherlands
Email: t.g.broenink@utwente.nl

KEYWORDS

cyber-physical systems, model driven design, port-based modeling, variable detail models

ABSTRACT

This publication is about simulating cyber-physical models with varying levels of detail, how to structure the models to make this possible and how to design the sub-models required. A structured method is posed to select which signals of the model to use for the structure and how to determine the different parts of the models to abstract away when varying the detail of sub-models. This method is applied to two examples based on a lab setup available. One example shows hard limits due to encoder processing, the other example shows the effects of simplified dynamic models. The method is effective in both examples and shows a clear trade-off between accuracy and performance.

INTRODUCTION

Design of cyber physical systems can benefit greatly from model based design (Broenink et al. 2016). Not only does model based design allow for solutions to common design conflicts (Ni and Broenink 2014), the current state of the art allows the partial synthesis of software and hardware description (Kuipers et al. 2016). However larger and especially hybrid systems are difficult to model and simulate (Derler et al. 2012), requiring complex simulations and more computations. This is why it is important to make models that are competent enough for the design challenges of cyber-physical systems (Lee 2008). Modeling and simulation of cyber-physical systems requires several decisions during the modeling process. Decisions on what to model, how detailed this should be modeled, and how to exactly implement it. Depending on the specific problem statement, different levels of detail are required to find solutions. However, more detailed models require more computations to solve, increasing the time required to simulate. Thus a model needs only as much details as required for the problem. It is beneficial for modeling and simulation to be able to select the level of detail at which a system is modeled and simulated, assuming that a method exists to simulate these different levels of detail in a model. This method can either be a versatile simulation package, or a way to co-simulate multiple different simulators. (Bastian et al. 2011; Nägele and Hooman 2017) This allows for a trade off between model accuracy and simulation speed. The same trade off that is made by taking a larger time-step when simulating a model. It also allows for a trade off between development time and accuracy. If it is easy to change the detail level of a sub-model it is possible to model a simpler model first and only creating the more complex model when necessary.

A model can be used to answer different questions about the final design. Depending on the question posed, some behavior of the model is more important then others. This means that some parts of the model need to be modeled and simulated in more detail than others. In order to do this, the level of detail of a model should be changeable per part. To be able to change the level of detail on different parts of the model, it is required that the structure of the model is able to accommodate this change, allowing parts of the model to be replaced. Thus it should be possible to create variable detail simulation with correctly structured modeled.

A method to structure these model is proposed here, a way to allow the varying of the implementation of the different parts of the model with a minimal change in structure. Furthermore a guideline is proposed for the sub-models used, a guideline on how to model these sub models and how to easily change the level of detail. This method is illustrated based on two different examples. The first example is based on a signal processing case. The second example is based on a dynamic modeling case.

METHODOLOGY

In order to simulate a system with varying amount of detail, two steps are needed. The first one is to structure the model in such a way that it is possible to change the detail of one element in such a way that the rest of the model is unaffected. The second step is to model the resulting sub models in different levels of detail, combined with analyzing the limits of certain approximations.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Model structure

The structure of a model is defined as the collection of sub-models composing the model, and the connections, or interfaces, defined between these sub-models. These interface consist of definitions of uni- and bi-directional signals. Included in this definition are not only direction, but also data-type and possibly update rate. When changing the details of one of the sub-models these interfaces are not allowed to change, thus all levels of detail will have to have the same interface. However it is possible to implement new sub-structures within an sub-model.

This requirement, to keep the interface unchanged, means that the interfaces used for the model structure will have to be based on signals that remain the same, irrelevant of the level of detail. As the encoding of information is very likely to change based on the different levels of detail used, signals that represent unencoded values should be used. These signals usually correspond to physical signals available in the system, instead of the IO signals of system parts. These signals can include:

1. Physical quantities, e.g. velocities, voltages, forces.
2. System states, e.g. started/stopped, switch pressed/released.
3. Representations of Physical quantities, e.g. set-points and sensor values.

This requires a split between input/output models and the actual core of the sub-model. All models should be transformations between physical, or detail-invariant signals. This is in contrast to conventional model boundaries, where physical system boundaries are often used, e.g. output pins, or communication interfaces.

Based on a simple electro-mechanical system, an example is given. Take the following system: A controller controlling an electromotor, applying a voltage to reach a desired angular position. Classically, this system can be split in a controller and a plant, giving the system of Fig.1. In the ideal case the controller can directly apply a voltage V to the motor and is able to directly read the angle ϕ. This is all that is required to determine the control laws and the plant dynamics. However the signals send between the controller and the motor are not directly the voltage V and the angle ϕ, even if these signals are supposed to represent those values. The controller needs some way to convert internal signals into an output voltage for the motor. The same goes for the angle of the motor, it has to be converted in some way to a signal that the controller understands. For the output voltage this is done by some form of motor driver, which has a signal input, and will supply a voltage based on this input, as seen in Fig 3. For the angle of the motor some sort of sensor is required to sense the physical angle ϕ and to turn it into a signal representing this angle in some way.

In the current configuration this is done either inside the Motor model or the Controller model. Depending on the actual encoding used for these signals this might be a complicated process.

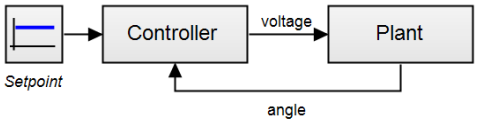


Figure 1: System for controlling a motor

When the detail invariant signals in this setup are identified, 5 signals are found:

1. The voltage on the motor (V)
2. The angle of the motor (ϕ)
3. The voltage signal as represented in the controller.
4. The angle signal as represented in the controller.
5. The set-point of the controller.

The conversion between these signals depends strongly on the implementation of the specific models. It can be as simple as assuming that the angle signal in the controller is the angle of the motor, or can be as complicated as a fully modeled PWM driver for the motor.

Based on these signals, a second model structure is created, using interfaces as defined by these 5 signals. This results in Fig.2. This structure allows the changing of the level of detail of one of the sub-models, without influencing the other sub-models.

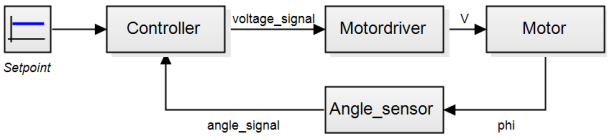


Figure 2: System for controlling a motor, with interfaces based on detail invariant signals

Variable detail

To model the sub-models with different levels of detail, it is important to analyze the impact of different parts of the sub-model on the performance. It is possible to reduce the detail the model using mathematical reduction techniques (Antoulas et al. 2001; Benner et al. 2013), but the focus of this method is on understanding the different parts of the model and deciding which to include in the final design, so that there are understandable effects that can be neglected or included. The interface of the sub-models was already determined in the previous step. Thus it is only

necessary to change the implementation of the relations between these signals. The sub-model should be modeled using a behavioral (Willems 2007) or port-based (Breedveld 2004) modeling technique, i.e. a technique where the different physical phenomena are visible as distinct elements of the model. This includes models like block-diagrams, bond graphs, or sub-structures. Representations in single equations e.g transfer functions or state-space representations are less suited for this purpose.

As the different physical effects are represented separately in the model, it should be possible to identify their individual effects. It is then possible to model what would happen if these effects were left out of the model. These effects can then be classified into three distinct categories:

1. Effects that are essential to a model's functioning.
2. Effects that are essential for a certain region of inputs/states (hard limits).
3. Effects that degrade overall model accuracy when left out (soft limits).

The first effects can never be abstracted away, as this would compromise model functionality completely. The second and third effect can be left out if the model is only used in certain regions. An example of these three effects can be made based on an electromotor, shown in Fig 3. The different effects of this model can be classified as:

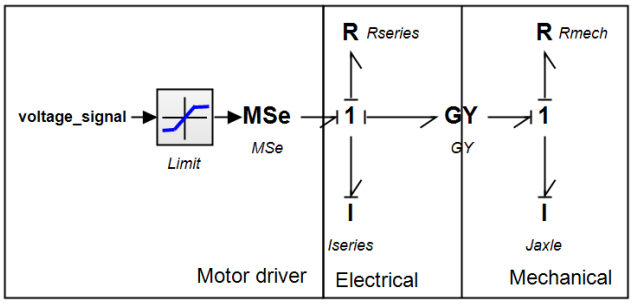


Figure 3: Port-based motor model using bond graphs

- Essential
  - The signal to electrical conversion of the MSe.
  - The transduction of the electromotor, represented as the gyrator (Gy).
- Hard limits
  - The limitation of the motor driver, this is can be left out as long as the input signal never reaches this limit.
- Soft limits
  - The Series inductance ($Iseries$) and moment of inertia ($Jaxle$), change the dynamic behavior of the motor, when left out.
  - The Electrical resistance ($Rseries$) changes the start-up behavior of the motor, and the maximum torque, when left out.
  - The Mechanical resistance ($Rmech$) reduces model accuracy at higher speeds when left out.

These classifications can then be used to create multiple models at different levels of detail. For example:

- A basic model only containing the essential effects.
- A model including the limit.
- A model for the detailed dynamic behavior including all the parasitic elements.

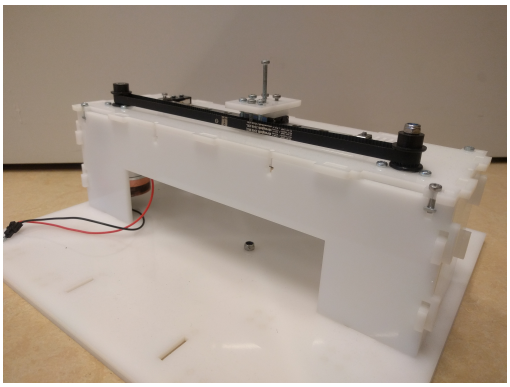
Which can all be used for different problem statements.

Method

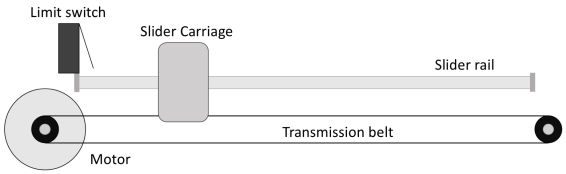
To summarize, to create a structure suitable for these detail variations, the following steps are performed.

1. Identify detail invariant signals, including but not limited to:
   (a) Physical quantities, e.g. velocities, voltages, forces.
   (b) System states, e.g. started/stopped, switch pressed/released.
   (c) Representations of Physical quantities, e.g. set-points and sensor values.
2. Define interface and model structure based on detail invariant signals
3. Model sub-models based on behavioral or port-based modeling principles
4. Identify region of validity of simplifications of model effects, classify these effects as:
   (a) Essential
   (b) Hard limit
   (c) Soft limit
5. Create models based on required regions of validity.

It is assumed that the created model(s) and sub-model(s) are validated and tested, as part of the creation of these models.



(a)



(b)

Figure 4: The test setup as inspiration for the examples (a), and a schematic overview of the structure (b)

TEST SETUP

To further demonstrate the use of this approach, two examples are provided. Both examples are based on a test setup available in our lab. A linear motion slider setup, shown in Fig 4. This test setup is patterned of a common motion used in mechatronic systems. A motor moves a linear slider via a transmission belt. The angle of the motor is encoded using a magnetic encoder, which gives the absolute motor angle. The motor is driven with a PWM signal. Identifying the detail invariant signals results in the same signals as mentioned en the methodology (Figure 2).

As the motor encoder gives an absolute motor angle from 0 to 2π rad, it is not possible to discern multiple rotations. However with some software processing, the actual rotation can be calculated. The dynamic model of the slider consists of three parts:

1. The dynamics of the electromotor.
2. The dynamics of the transmission belt.
3. The dynamics of the slider carriage.

Modeling this dynamic behavior gives the graph of Fig.8 The total model of this system can then be constructed equivalently to Fig.2.

ENCODER SIMULATION

As mentioned, the magnetic encoder present on the motor measures the absolute angle of the motor axis. The expression for the angle output from the sensor is:

$$\phi_{sensor} = \phi_{motor} \mod 2\pi \tag{1}$$

This value is then sampled by the controller to be processed back into the actual angle. In order to process this sensor angle back into the true angle of the motor the following algorithm is used, presented in pseudo-code:

Listing 1: Processing for angle sensor

```
angle_difference = angle - previous(angle)
if angle_difference < -pi then
        rotations +=1
if angle_difference > pi then
        rotations -=1
output = angle + rotations * 2 * pi
```

With this processing the true angle can be reconstructed, assuming that equation 2 holds for the angular velocity of the motor.

$$|\frac{\dot{\phi}}{f_{sample}}| < \pi \tag{2}$$

Whenever the angular velocity of the motor exceeds this limit, the processing interprets the motion as going in the other direction and with a different magnitude. The effect of this processing can be seen in the graph of Fig 5. It can be seen that after the maximum velocity ($\omega_{max} = 100\pi \mathrm{rad/s}$) is reached that the processing reports a velocity of $\omega = -\omega_{max}$. This effect can be likened to the aliasing that happens with digital sampling.

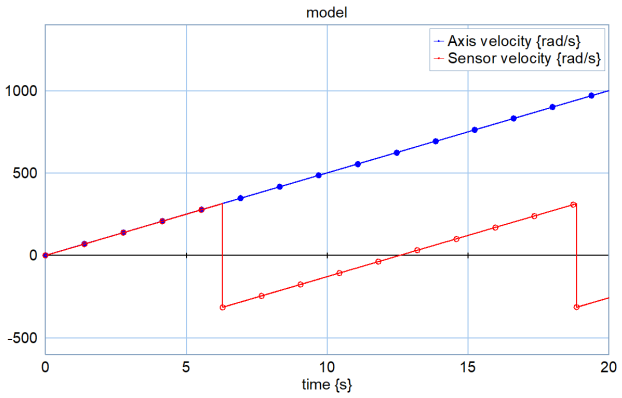
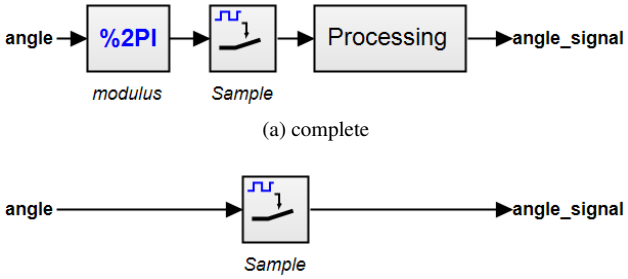


Figure 5: Effect of sensor aliasing when supplied with an ever increasing velocity

As the angle sensor is a sub-model with the actual angle as input and the angle signal as output, different implementations can be made. The first one is the full implementation. Using equation 1 as sensor model, sampling this, and processing it with the pseudo-code given. The second implementation is simplified up to the point of directly



(a) complete

(b) simple

Figure 6: A overview of both sensor models. The processing algorithm is shown in Listing 1

Model	time	speedup
simple	403ms	12%
complex	452ms	0%

Table 1: Simulation time and speed for a 1000 s simulation at 100 Hz and 100 rad/s

sampling the actual angle. An overview of both processing methods is given in Fig 6.

Both implementation are simulated using 20-sim. The system is given a constant velocity set-point of 100 rad/s and a controller-frequency of 100 Hz. The resulting simulations and speeds versus real time are shown in table 1. From Fig 5 and Table 1 it can be concluded that the simple implementation of the sensor can result in a speed increase when the model remains within the maximum angular velocity supported by the processing algorithm. When the model is used outside of this valid region, the simple model is no longer valid. To show the failure mode of the encoder aliasing, a simulation is done of a step to 50 rad. The result of this simulation is shown in Fig 7.

While it is expected that the angle_signal follows the angle exactly, the angular velocity becomes too large in this example, thus the angle processing cannot process correctly. The effect, which was predicted using Fig 5, is that the angle estimation goes in the opposite direction of the actual angle. This failure is especially catastrophic as the controller tries to correct for the deviation in processed angle by accelerating further, thus increasing the problem. This results in the actual angle overshooting the 50 rad.

This problem could be prevented using multiple methods. One of these methods would be to increase the controller frequency, or at least increase the sampling frequency of the angle sensor and sub-sample this signal. Another method of preventing this would be to design the system in such a way that this velocity is never required, either by limiting the controller, or changing the aggressiveness of the control.

Even when a system is not designed to be able to reach these maximum velocities, it is still important to keep in mind that this limit exists. An example of this limit unexpectedly applying was during a student project where a system with a motor loaded with a wheel could never reach these velocities. However, when the system was tested without the wheel to analyze the motor behavior, this limit was exceeded. This resulted in unexpected measurements for the students participating in this project.

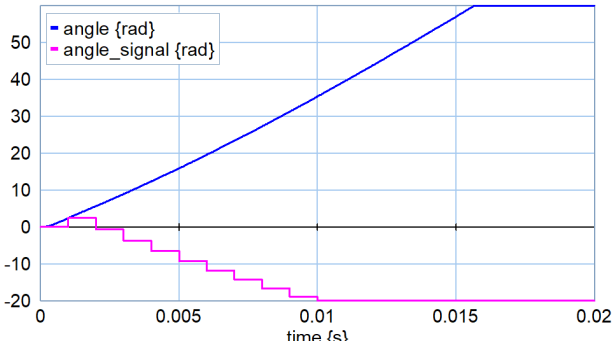


Figure 7: Breakdown of the angle sensor processing, the angle_signal goes negative, while the actual angle keeps increasing. Controlled at 100 Hz

DYNAMIC MODEL

The dynamic behavior of the plant can be modeled in multiple layers of detail. The dynamic behavior can be roughly divided into three segments:

- The motor
- The belt transmission
- The slider

A model containing these three parts is shown in Fig 8. This model is based on bond graphs, each half arrow representing the two conjugate power variables e.g. voltage and current, velocity and force, or torque and angular velocity. The

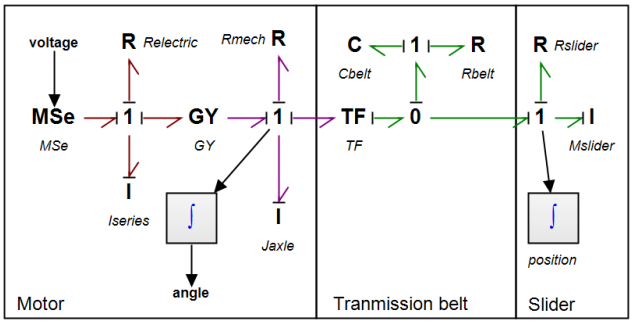


Figure 8: Full dynamic model of slider plant

model of the plant can be simplified. The elasticity of the belt introduces some oscillations during movement. Thus if the use of the model is to only estimate motion and position, instead of describing the full motion accurately it can be simplified by assuming the transmission belt to be ideal.

This is for example the case if a homing procedure of the controller has to be tested. When the transmission belt is assumed to be ideal it is also possible to combine both the rotational inertia of the motor with the mass of the slider, thus resulting in a system that is only second order, instead of fourth order. The resulting simplification is shown in Fig 9.

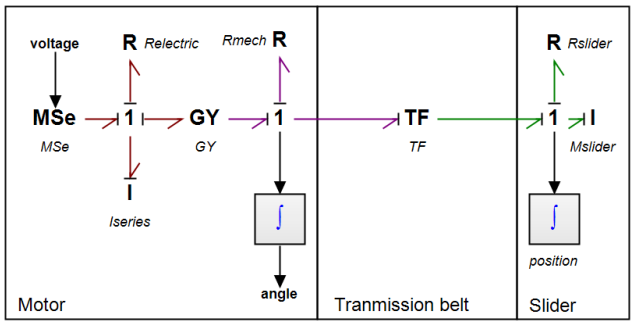


Figure 9: Simplified dynamic model of the slider

The results of these models can be compared. A step is applied to both models to move to 10 cm. The results of these motions are shown in Fig 10. It can be seen here that the final motion of both models is the same. However the oscillations at the beginning of the motion are not present in the simplified model. This same difference can be seen from the bode plots of both systems. These are shown in Fig 11. Here it can be seen that the higher frequencies of this model are not the same. Thus this simplification can only be used for low frequencies.

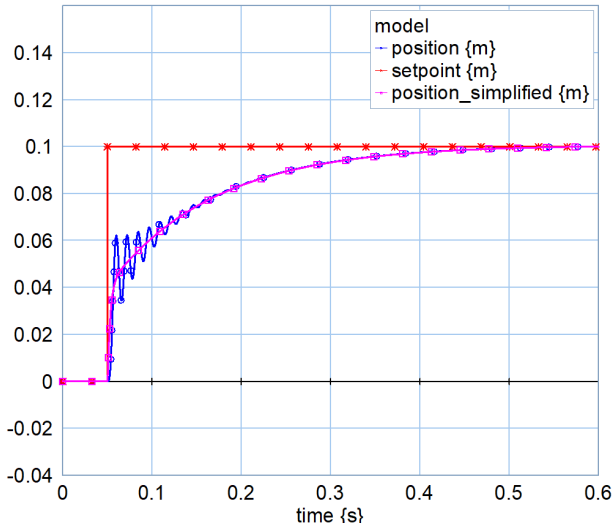


Figure 10: Simulation of (simplified)-plant model, controller frequency 100Hz

This lack of oscillation can be clearly seen in a speed comparison of both models. Both models are simulated for 100 of these steps. A controller frequency of 100Hz is used. The resulting simulations and speeds versus real time are found in table 2. Based on this data it is concluded that the simple model is suitable for either slow movements, or

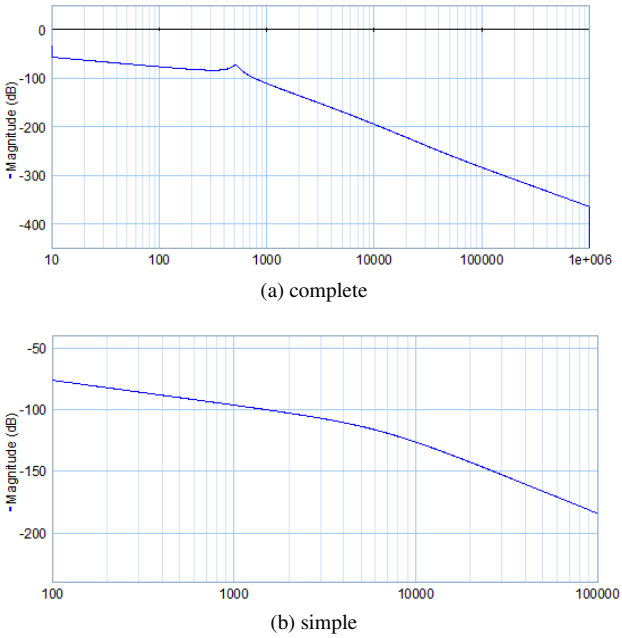


Figure 11: The bode plot of both plant implementations

Model	time	speedup
simple	4.733s	16%
complex	5.484s	0%

Table 2: Simulation time and speed for a 100 steps of 10 cm. Simulated at a 100 Hzcontroller frequency

situations in which the exact motion of the slider is less relevant. One example of this might be to test sequence control in the control system of the setup. This does not require exact motion, and should thus be modeled with the simpler plant model for increased performance. Faster motions, or situations where the exact position of the slider is relevant must be simulated with the more complex model. One example situation would be to evaluate the performance of the system when trying to reach a certain position.

CONCLUSION

Based on the previous examples it can be concluded that it is possible to do variable detail simulations on models, given that these models are structured in a correct manner. The methodology proposed in section 3 is successfully applied to two examples. This shows that the methodology is both applicable to hard-limits and soft-limits. The varying detail results in a speed increase for the simpler models. This effect could be much more pronounced in larger systems, where the differences in detail are bigger and in which more sub-models are present.

The next step in this research is to validate this methodology for larger and more complex systems. This will give better insight into the impact on larger systems. In order to maximize the efficiency of the multiple methods of detail, it is useful to automatically evaluate simulations based

on the limits of the sub-models. This allows switching to a more complex model if so required. This validation could be done after a simulation, but could also be implemented live, thus allowing an adaptable simulation.

REFERENCES

Antoulas, A. C., Sorensen, D. C., and Gugercin, S. 2001. "A survey of model reduction methods for large-scale systems". *Contemporary mathematics*, 280:193–220.

Bastian, J., Clau, C., Wolf, S., and Schneider, P. 2011. "Master for co-simulation using fmi". In *Proceedings of the 8th International Modelica Conference; March 20th-22nd; Technical Univeristy; Dresden; Germany*, number 63, pages 115–120. Linkping University Electronic Press; Linkpings universitet.

Benner, P., Gugercin, S., and Willcox, K. 2013. "A survey of model reduction methods for parametric systems".

Breedveld, P. C. 2004. "Port-based modeling of mechatronic systems". *Mathematics and Computers in Simulation*, 66(2-3):99–128.

Broenink, J. F., Vos, P.-J. D., Lu, Z., and Bezemer, M. M. 2016. "A co-design approach for embedded control software of cyber-physical systems". In *System of Systems Engineering Conference (SoSE), 2016 11th*, pages 1–5. IEEE.

Derler, P., Lee, E. A., and Vincentelli, A. S. 2012. "Modeling cyber–physical systems". *Proceedings of the IEEE*, 100(1):13–28.

Kuipers, F. P., Wester, R., Kuper, J., and Broenink, J. F. 2016. "Mapping CSP Models for Embedded Control Software to Hardware Using Clash". In *Communicating Process Architectures 2016, Copenhagen, DK*, Concurrent System Engineering Series, pages 133 – 150, Engeland. Open Channel Publishing Ltd. cpa bibtex: kuipers2016.

Lee, E. A. 2008. "Cyber Physical Systems: Design Challenges". pages 363–369. IEEE.

Nägele, T. and Hooman, J. 2017. "Co-simulation of cyber-physical systems using HLA". In *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*, pages 1–6. IEEE.

Ni, Y. and Broenink, J. F. 2014. "A co-modelling method for solving incompatibilities during co-design of mechatronic devices". *Advanced Engineering Informatics*, 28(3):232–240.

Willems, J. 2007. "The Behavioral Approach to Open and Interconnected Systems". *IEEE Control Systems Magazine*, 27(6):46–99.

AUTHOR BIOGRAPHIES

Tim Broenink (MSc 2016) is a PhD student at the Robotics and Mechatronics Lab of EE at the University of Twente. He is working on a project for a NWO-TTW perspective program regarding cyber physical systems. His track within this project relates to robust motion control for cyber-physical systems. His research interest includes behavior driven development of hardware and software, automated testing, and co-simulation.

Jan Broenink (MSc 1984; PhD 1990) is Associate Professor in Embedded Control Systems at the Robotics and Mechatronics Lab of EE at the University of Twente. His current research interests are on cyber-physical systems, embedded control systems (realization of control schemes on mostly networked computers) and software architectures for robotics. For that, he is interested in: model-driven design and meta-modeling of robot software architectures; designing software tools including (co)-simulation; concurrent and systems engineering. Since July 2017 he is chairman of the Robotics and Mechatronics group, together with prof. Stefano Stramigioli, who is Scientific Director.

BALL & PLATE MODEL FOR ROBOTIC SYSTEM

Lubos Spacek, Jiri Vojtesek, Frantisek Gazdos and Tomas Kadavy
Tomas Bata University in Zlin
Faculty of Applied Informatics
Nad Stranemi 4511, 760 05 Zlin, Czech Republic
E-mail: {spacek,vojtesek,kadavy,gazdos}@utb.cz

KEYWORDS

YuMi, robotics, LQ polynomial control, LQR state-space control, PD controller.

ABSTRACT

There are many solutions to control Ball & Plate model, ranging from hobby projects to more advanced control. This paper brings a new idea of control using robotic manipulator. This is quite challenging because industrial robots are not originally designed as a motion system for relatively fast and unstable system, which the Ball & Plate certainly is. This paper compares 3 controller designs to better comprehend the situation - a general LQR state-space control, LQ polynomial control and a basic PD controller. Results are also compared for a range of reference values to better understand advantages and disadvantages of chosen controllers, which will lead to future work and implementation for the real system. Data presented in this paper serve as a valuable background for next steps of the research and implementation.

INTRODUCTION

The Ball & Plate model is a well-known representative of fast and unstable systems ideal for designing and testing control algorithms. It is possible to find many designed Ball & Plate models on the Internet (mostly as hobby projects) with 2DoF separated control. There are of course more advanced structure with redundant degrees of freedom (Bruce et al. 2011) and even solution with 6DoF Stewart platform, which offers additional movement in the space.
Many control strategies are developed for Ball & Plate model starting at simple PID control (Jadlovska et al. 2009) and ending with fuzzy supervision (Moarref et al. 2008). This paper is aimed at the comparison of 2DoF LQ polynomial controller (Bobal et al. 2005), classic PD controller and LQR state-space controller. This comparison is also done for multiple reference values to better understand designed controllers. The quality of control is not the key aspect of this paper, but quality criteria for errors and controller effort are introduced in a table to better see the differences.
The paper is organized as follows. The first chapter deals with the theory and background behind Ball & Plate system and used robotic manipulator. It is divided into 3 subsections dealing with Ball & Plate, robot and identification separately. The next chapters describe 2DoF LQ polynomial controller and LQR state-space controller. The paper is closed with a chapter presenting obtained results and conclusion.

BALL & PLATE ROBOTIC SYSTEM

Ball & Plate

A great model for testing control algorithms for unstable systems is Ball & Plate model. This model has relatively fast dynamics and with its instability provides quite a challenging task for controller design. Its general simplified and linearized mathematical description can be described by (1) and (2). This was more closely described in the previous work of authors (Spacek et al. 2017).

$$\ddot{x} = K\alpha \quad \rightarrow \quad G_x(s) = \frac{K}{s^2} \tag{1}$$

$$\ddot{y} = K\beta \quad \rightarrow \quad G_y(s) = \frac{K}{s^2} \tag{2}$$

where x,y are ball coordinates from the center of the plate ($\ddot{x}$, $\ddot{y}$ are respective 2nd time derivatives), α, β are angles of the plate, K is constant dependent on the gravitational acceleration g and the momentum of the ball and $G_x(s)$, $G_y(s)$ are continuous-time transfer functions with complex variable s. It is obvious these functions are symmetric in nature and thus only one-dimensional solution (Fig. 1) will be presented in simulations in this paper.

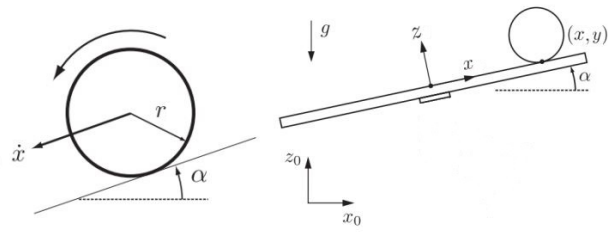


Fig. 1. Ball & Plate setup (Nokhbeh et al. 2011)

Collaborative Robot YuMi

The industrial robotic manipulator ABB IRB14000 YuMi (Fig. 2) was used as the motion structure for the simulations as a background to the real implementation. It has two manipulators with 7DoF and 0.02 mm repeatability each, which offers more possibilities and opportunities as the classic 2DoF solution with two servo motors or even 6DoF parallel manipulator commonly known as Stewart platform.
The YuMi is a collaborative robot, which means it can operate without external caging, which is safer and more efficient during implementation (Fryman and Matthias 2012). It would not be practical to have the

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

robot in a cage during control because it is possible to interact with the ball externally. The advantage is also that YuMi has two arms, which extends this application to the cooperative solution of more manipulators.

Fig. 2. ABB IRB 14000 YuMi

The development and simulation environment from ABB called RobotStudio is also a great advantage, as it excels in the deployment of similar applications and supports virtual sensors for measurements. The 7DoF manipulator has certainly complicated dynamics, but only last three joints actively influence the angle of the plate. The plate angle dynamics can be approximated by 2nd order transfer function, as can be seen from the change of the plate angle from 0 to 10 degrees in Fig. 3. This paper presents a solution, where the dynamics of the plate angle are approximated by 1st order transfer function. It is not the usual practice nowadays to simplify 2nd order dynamics to 1st order, but because Ball & Plate model consists of both the Ball & Plate dynamics and motor dynamics, it is the necessary step to reduce complexity. The identification chapter will show that it is not a very serious issue, as combined dynamics will negate this effect.

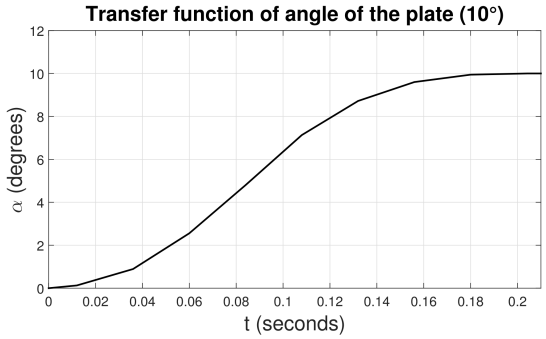


Fig. 3. Plate angle dynamics

Identification

The Ball & Plate model was identified for one dimension for both input-output (3) and state-space (4) models. Note that all measurements were in RobotStudio which is still the virtual environment, thus the identified system is still an ideal (semi-real) representation of the real system.

$$G(s) = \frac{K}{s^2\left(Ts+1\right)} = \frac{K}{Ts^3+s^2} \tag{3}$$

where $G(s)$ is continuous transfer function with complex variable s, K is velocity gain and T is the time constant of the system.

$$\begin{bmatrix}\dot{x}\\ \ddot{x}\\ \dot{\alpha}\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & K_s\\ 0 & 0 & T_s\end{bmatrix}\begin{bmatrix}x\\ \dot{x}\\ \alpha\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}\alpha$$

$$y = \begin{bmatrix}1 & 0 & 0\end{bmatrix}\begin{bmatrix}x\\ \dot{x}\\ \alpha\end{bmatrix} \tag{4}$$

where x is the position of the ball in one dimension, $\dot{x}$ is its time derivative, $\ddot{x}$ is its 2nd-time derivative, α is angle of the plate, y is the output of the system, K_s and T_s are identified constants of the state-space system.
Resulting identifications for different input angles showed a nonlinearity in the system, as shown in Fig. 4 which displays identification for the system described by (3). Both mathematical models were discretized with the period $T_0 = 0.05s$. Their discretization is necessary to design a discrete controller for future implementation of the real system.

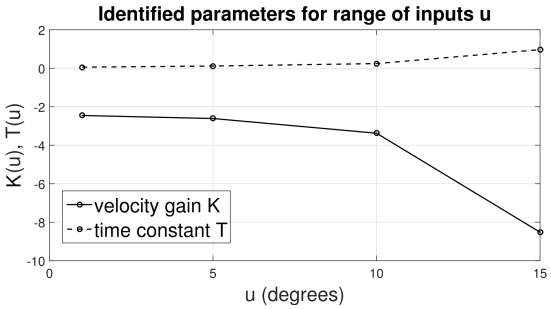


Fig. 4. Identified parameters

LQ POLYNOMIAL CONTROL

Discrete polynomial control is easy to implement because the discrete transfer function can be easily implemented in basic programming language and needs just previous values of input and output. Its disadvantage is a generally worse quality of control and limited possibilities compared to the state-space control, due to obvious reasons. The design of the 2DoF LQ digital controller is described in previous work of authors (Spacek et al. 2017).
The problem is solved by minimizing linear quadratic criterion in (5). This criterion can be solved in the polynomial form using spectral factorization (Bobal et al. 2005), instead of solving Riccati algebraic equation.

$$J = \sum_{k=0}^{\infty}\left\{\left[e(k)\right]^2 + q_u\left[u(k)\right]^2\right\} \tag{5}$$

where $e(k) = w(k) - y(k)$ is the error, $u(k)$ is controller output and q_u is a penalization constant, which

influences the controller output during the minimization process. Half of the solution (poles of characteristic polynomial in (6)) can be obtained from spectral factorization and other half is user-defined to adjust the behavior if needed (Spacek et al. 2017). This helps during the design of the controller, especially for the unstable system when poles of characteristic polynomial are harder to "guess".

$$D = AKP + BQ \tag{6}$$

where D is characteristic polynomial of a closed-loop system, B and A are numerator and denominator polynomials of the plant respectively, K is summation element, P and Q are denominator and numerator polynomials of the controller respectively. This equation is the same for continuous-time and discrete-time version, thus dependence notation of polynomials on complex variables s or z^{-1} is omitted.

The control law is designed for 2DoF controller structure (Fig. 5) because of its softer response to step changes. The 2DoF controller has 3 parts in this case: feedforward part C_f (7), feedback part C_b (8) and summation part $1/K(z^{-1}) = 1/(1 - z^{-1})$, which is extracted solely for practical purpose. Resulting implementation of the controller to programming environment is thus expressed in (9).

$$C_f(z^{-1}) = \frac{R}{P} = \frac{r_0}{1 + p_1 z^{-1} + p_2 z^{-2}} \tag{7}$$

$$C_b(z^{-1}) = \frac{Q}{P} = \frac{q_0 + q_1 z^{-1} + q_2 z^{-2} + q_3 z^{-3}}{1 + p_1 z^{-1} + p_2 z^{-2}} \tag{8}$$

$$\begin{aligned} u_k = (1 - p_1)u_{k-1} + (p_1 - p_2)u_{k-2} + p_2 u_{k-3} + \\ + r_0 w_k - q_0 y_k - q_1 y_{k-1} - q_2 y_{k-2} - q_3 y_{k-3} \end{aligned} \tag{9}$$

where q_i, p_i, and r_0 are controller parameters, u_{k-i} and y_{k-i} are outputs of the controller and plant respectively and w_k is desired (reference) value.

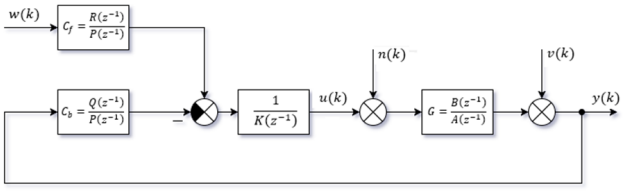


Fig. 5. Structure of 2DoF controller

LQR STATE-SPACE CONTROL

The state-space control is harder to implement, especially to the software of the robot because it not designed for such tasks in terms of speed and optimization. On the other hand, it has more options and may provide better results while using state observer, Kalman filter, MPC, and others. Because this paper deals with only simulation implementation it is not needed to use state observer

The state-space control (Fig. 6) is derived from the state-space description of the system either for continuous form in (10) or discrete form in (11).

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{10}$$

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k \end{aligned} \tag{11}$$

where x, u and y are state, input and output vectors respectively, A, B, and C are system, input, and output matrices respectively and D is zero matrix for most real systems.

As mentioned above, the discrete controller is needed for future implementation, thus it will be designed for the discrete state-space model in (11). The goal is to obtain the optimal gain matrix K for state feedback control in (12) by minimization of the quadratic criterion in (13). The gain matrix K is then computed in MATLAB from (14).

$$u_k = -Kx_k \tag{12}$$

$$J = \sum_{k=1}^{\infty} \{x_k^T Q x_k + u_k^T R u_k\} \tag{13}$$

$$K = (B^T S B + R)^{-1} B^T S A \tag{14}$$

where Q and R are state (x_k) and input (u_k) weight matrices respectively, A and B are matrices described in (11) and S is the solution of discrete-time Riccati equation. N in Fig. 6 is precompensator equal to the 1st term of the gain matrix K (in this case). The precompensator is required because the reference value is not directly compared to the output, thus it needs a compensation.

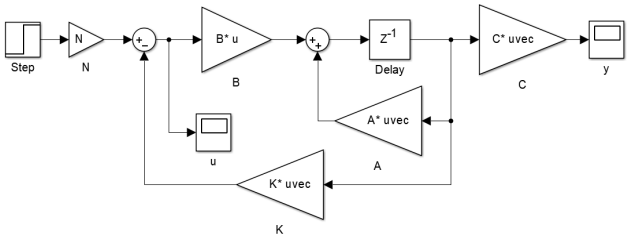


Fig. 6. State-space control in Simulink

As already mentioned, it is highly appropriate to use at least state observer to correctly estimate not measurable states, but in a simulation are all states "measurable", thus the more advanced control will be designed for the real system implementation only.

RESULTS

Designed controllers were implemented in MATLAB and Simulink to test and compare their results. Three types of controllers will be compared for this pilot study - adding a basic discrete PD controller to LQR state-space and LQ polynomial solutions. The PD controller

was designed by Naslin method (Balate 2003) and subsequently discretized, however its output was saturated to soften its output for large changes to interval $< -2; 2 >$. Results are shown for a step change (Fig. 7 and Fig. 8), sequence change (Fig. 9 and Fig. 10), ramp change (Fig. 11 and Fig. 12) and harmonic change (Fig. 13 and Fig. 14). Note that the plot is bounded for the step change, but the output of the PD controller reaches its saturation in $-2°$. The schematic of the Ball & Plate model virtualized in the RobotStudio for YuMi is shown in Fig. 15.

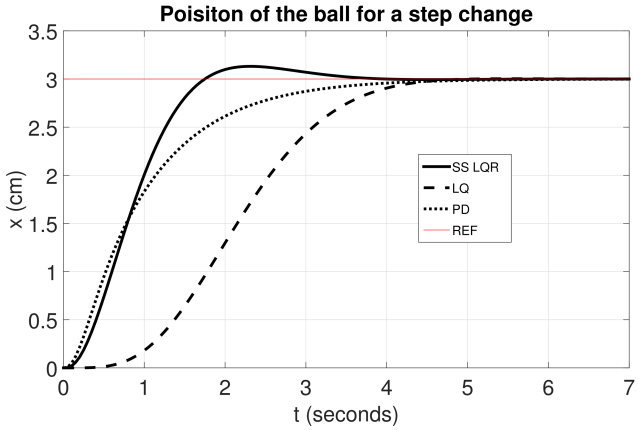


Fig. 7. Position of the ball - step change

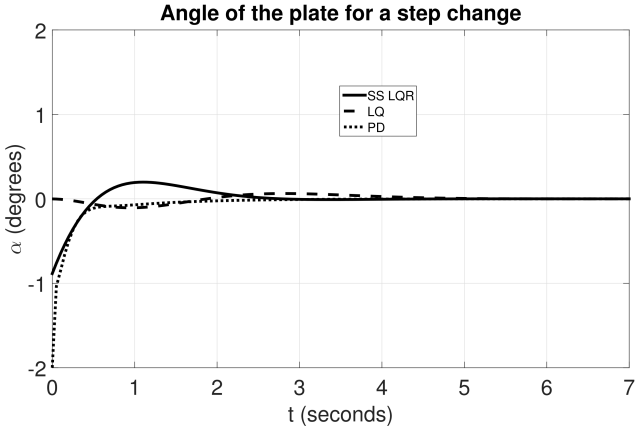


Fig. 8. Angle of the plate - step change

Quality criteria

The quality of control is shown in Tables 1-4. The criterion for the sum of squared errors is in (15) and for the sum of controller outputs in (16).

$$S_e = \frac{1}{N}\sum_{k=1}^{N} e^2(k) \tag{15}$$

$$S_u = \frac{1}{N}\sum_{k=1}^{N} u^2(k) \tag{16}$$

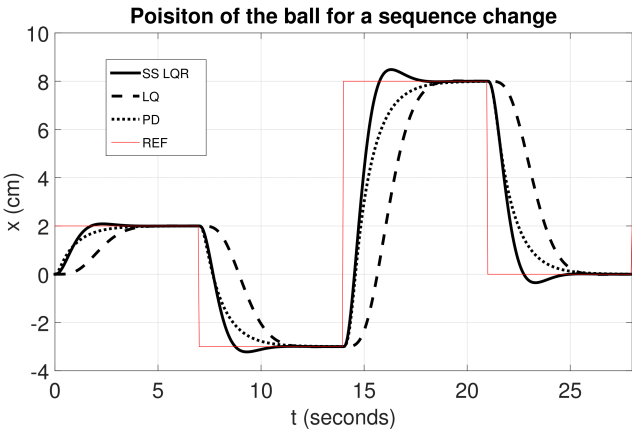


Fig. 9. Position of the ball - sequence change

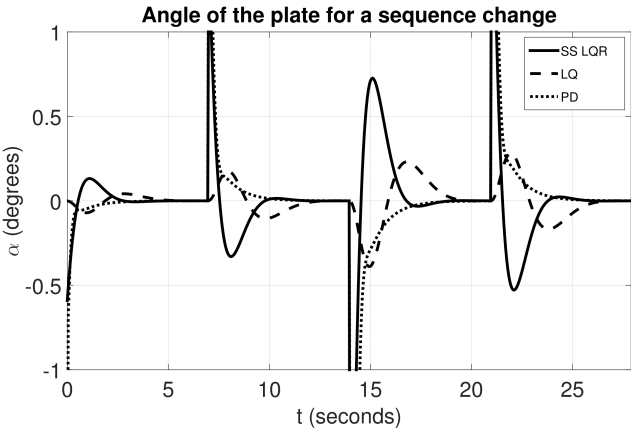


Fig. 10. Angle of the plate - sequence change

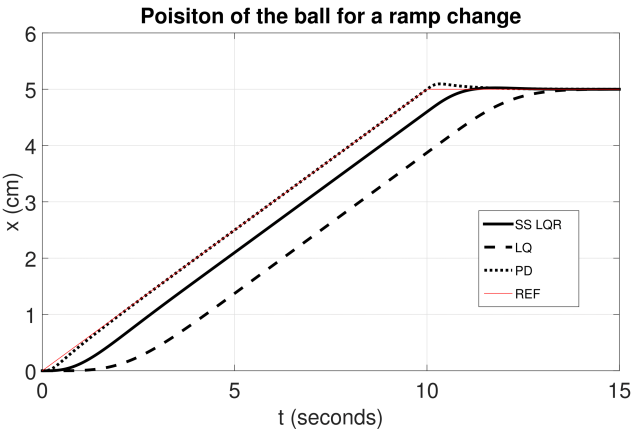


Fig. 11. Position of the ball - ramp change

Tab. 1. Quality control for the step change

	LQR	LQ	PD
S_e	0.0080	0.0229	0.0081
S_u	2.4203	0.2963	5.4281

Tab. 2. Quality control for the sequence change

	LQR	LQ	PD
S_e	0.0475	0.1348	0.0548
S_u	13.9880	1.4078	16.9862

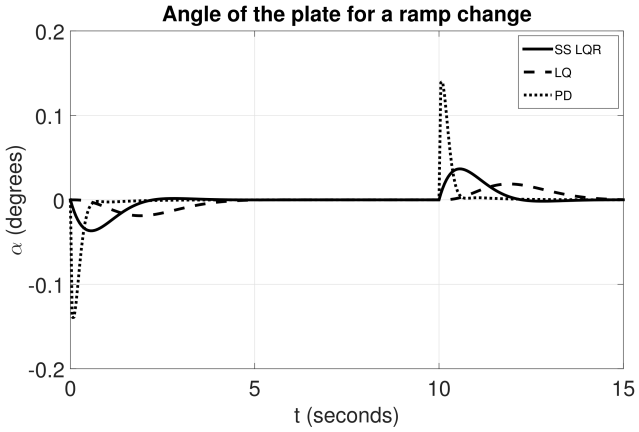


Fig. 12. Angle of the plate - ramp change

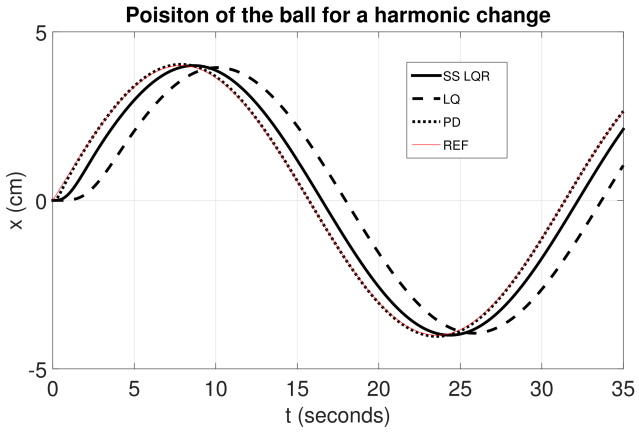


Fig. 13. Position of the ball - harmonic change

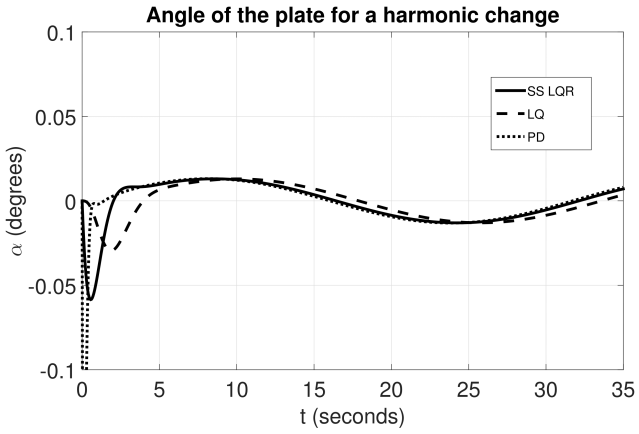


Fig. 14. Angle of the plate - harmonic change

Tab. 3. Quality control for the ramp change

	LQR	LQ	PD
S_e	0.0010	0.0076	~ 0
S_u	0.1292	0.1248	0.1637

Tab. 4. Quality control for the harmonic change

	LQR	LQ	PD
S_e	0.0022	0.0156	~ 0
S_u	0.0450	0.0490	0.0625

CONCLUSION

It is clearly visible from graphs that the PD controller shows a very good quality of control. However, from previous experiences with Ball & Plate system can be stated, that this type of controller would not work in the real system, because it has fast behavior and generates large changes in its output. Not to mention the noise from the environment and disturbances. It also nicely managed to follow a linear and sinusoidal trajectory. Two other controllers can be of course designed to be more competitive, but the goal was to show their true nature. For example, 2DoF LQ polynomial controller was quite slow and although it has similar settling time, it looks lazy. This is actually the point of using the 2DoF controller, which has a very subtle behavior well suited for the Ball & Plate model. State-space LQR controller could be also modified for better control, but its behavior is sufficient enough for this paper.

Tables clearly show that the LQ polynomial control excels at the quality of controller effort. This is caused by its 2DoF structure and desired, especially when the settling time is almost as small as for two other designed controllers. PD controller shows very good results for ramp and harmonic change because of their relatively slow rise time.

Results presented in this paper are helpful for future research on the topic and authors are keen to continue further. Next steps are to add state observer or Kalman filter to the state-space control, implement designed algorithms to the programming language of the robot and run tests in RobotStudio simulation environment. RobotStudio virtualizes the robot quite precisely, thus it is a good intermediate step in the implementation. The hardest part will be to implement state-space control in the programming language of the robot because it does not have appropriate tools and it is not optimized for this type of application. Using an external control unit is considered as the best option, which can simply send angles directly to the robot and do the "heavy" work.

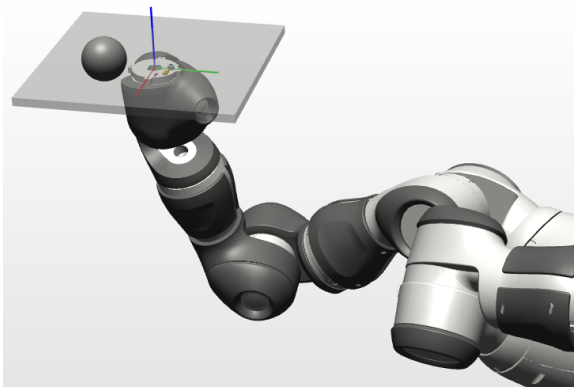


Fig. 15. Ball & Plate modeled for the YuMi in the RobotStudio

ACKNOWLEDGMENT

This article was created with support of the Ministry of Education of the Czech Republic under grant IGA reg. n. IGA/CebiaTech/2018/002.

REFERENCES

Bruce, J.; Keeling C. and R. RODRIGUEZ. "Four Degree of Freedom Control System Using a Ball on a Plate". Southern Polytechnic State University, 2011

Bobal, V.; J. Bohm; J. Fessl; and J. Machacek. 2005. *Digital Self-tuning Controllers*. Springer-Verlag, London, 2005.

Fryman J. and B. Matthias. "Safety of Industrial Robots: From Conventional to Collaborative Applications". 7th German Conference on Robotics, ROBOTIK 2012, pp.1-5, May 2012

Jadlovska, A.; S. Jajcisin; and R. Lonscak. 2009. "Modelling and PID Control Design of Nonlinear Educational Model Ball & Plate". In *Proceedings of the 17th International Conference on Process Control 09*. Strbske Pleso, Slovakia, 475-483.

Moarref, M.; M. Saadat; and G. Vossoughi. 2008. "Mechatronic design and position control of a novel ball and plate system". In *16th Mediterranean Conference on Control and Automation*. Ajaccio, France.

Nokhbeh, M.; Khashabi, D. and H.A., Talebi. "Modelling and Control of Ball-Plate System". Amirkabir University of Technology, 2011.

Spacek, L.; Bobal, V.; and J. Vojtesek. "LQ Digital Control of Ball & Plate System". *31st European Conference on Modelling and Simulation ECMS 2017*, pp. 427–432. European Council for Modelling and Simulation (2017)

Balate, J. 2003. *Automaticke rizeni*. BEN, Prague.

AUTHOR BIOGRAPHIES

LUBOŠ SPAČEK studied at the Tomas Bata University in Zlin, Czech Republic, where he obtained his masters degree in Automatic Control and Informatics in 2016. He currently attends PhD study at the Department of Process Control, Faculty of Applied Informatics. His e-mail address is `lspacek@utb.cz`.

JIŘÍ VOJTĚŠEK was born in Zlin, Czech Republic in 1979. He studied at Tomas Bata University in Zlin, Czech Republic, where he received his M.Sc. degree in Automation and control in 2002. In 2007 he obtained Ph.D. degree in Technical cybernetics at Tomas Bata University in Zlin. In the year 2015 he became associate professor. He now works at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlin, Czech Republic. His research interests are modeling and simulation of continuous-time chemical processes, polynomial methods, optimal, adaptive and nonlinear control. You can contact him on e-mail address `vojtesek@fai.utb.cz`.

TOMÁŠ KADAVÝ was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zln, where he studied Information Technologies and obtained his MSc degree in 2016. He is studying his Ph.D. at the same university and the fields of his studies are: Artificial intelligence and evolutionary algorithms. His email address is: `kadavy@utb.cz`.

FRANTIŠEK GAZDOŠ was born in Zlin, Czech Republic in 1976, and graduated from the Brno University of Technology in 1999 with MSc. degree in Automation. He then followed studies of Technical Cybernetics at Tomas Bata University in Zlin, obtaining Ph.D. degree in 2004. He became Associate Professor for Machine and Process Control in 2012 and now works as the Head of the Department of Process Control, Faculty of Applied Informatics of Tomas Bata University in Zlin. He is author or co-author of more than 80 journal contributions and conference papers giving lectures at foreign universities, such as University of Strathclyde Glasgow, Instituto Politcnico do Porto, Universita di Cagliari and others. His research covers the area of process modelling, simulation and control. His e-mail address is: `gazdos@utb.cz`.

MULTIMODEL APPROACH IN STATE-SPACE PREDICTIVE CONTROL

Lukáš Rušar and Vladimír Bobál
Department of process control
Faculty of applied informatics, Tomas Bata university in Zlin
Nad Stráněmi 4511, Zlin 76005, Czech Republic
E-mail: rusar@fai.utb.cz

KEYWORDS
predictive control, state-space, multimodel, inverted pendulum, predictor-corrector.

ABSTRACT

This paper presents a multimodel approach to control nonlinear systems. The system of the inverted pendulum which has one control input and two measured outputs was chosen as an exemplar process. This system is an example of the nonlinear process with a sampling period in order of milliseconds. The state-space predictive control of the system described by CARIMA model was chosen as a control method. This paper presents a description of the inverted pendulum nonlinear mathematical model, its linearization and the control signal calculation using predictor-corrector method. The results compare three methods of linearized model combination. All of the simulations were done in Matlab.

INTRODUCTION

Process control area offers a variety of different processes with different level of complexity. Even the sampling period can be very different. This paper focuses on nonlinear processes with fast sampling period. The complex and fast processes need a modern control method to control them effectively such as model predictive control (Bobál 2008).
This method predicts the output values on the chosen time interval based on the mathematical model of the controlled process. The model of the inverted pendulum is described by the state-space CARIMA mathematical model for the single-input multi-output (SIMO) system (Bars et al. 2011; Wang 2009).
The predictive control method uses a minimization of the cost function, which is usually in quadratic form, to calculate the control signal. The quadratic form of the cost function minimize the differences between the reference value and the output value and the control signal increments. The predictive control method offers a possibility to constrain the process variables then the chosen predictor-corrector minimization method can be used to minimize the cost function (Camacho and Bordons 2004; Maciejowski 2002; Rossiter 2003).
However, the chosen state-space predictive control method uses a linear CARIMA mathematical model for prediction of the output values but the chosen process of the inverted pendulum has nonlinear behaviour. That means we have to linearize the nonlinear model first. the nonlinear behaviour of the process can be described with a set of the linear models. The final linear model used to output values prediction is obtained by combination of two or more linear models out of the set of the linearized models according to the current output value (Albertos Peréz and Sala 20014; Hangos et al. 2004).
This paper is divided into the following sections. The model of the inverted pendulum is described in the first section. The predictive control and the calculation of the control signal are described next. The final sections shows the results of the research and the conclusion.

MATHEMATICAL MODEL OF THE CONTROLLED SYSTEM

The controlled system in this paper is represented by Amira PS600 inverted pendulum system which is shown at figure 1. The pendulum rod of this system is attached to the cart which is driven by a servo motor (Amira 2000; Chalupa and Bobál 2008).

Figure 1 : Amira PS600 Inverted Pendulum system

The servo motor produces the input force (input variable) that move with the cart. Position of the cart is the first measured output variable and the pendulum angle is the second measured output variable.. The figure 2 shows the analysis of the forces acting in the system (Amira 2000; Chalupa and Bobál 2008).

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

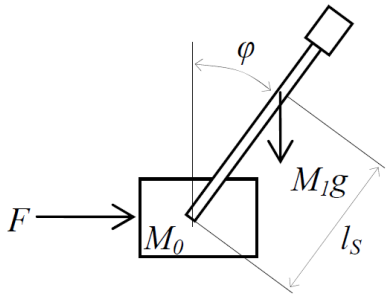


Figure 2 : Analysis of the inverted pendulum

The angle of the pendulum rod is expressed as φ, the input force produced by DC motor is symbol F, symbol ls means distance between pendulum gravity centre and rotation centre of the pendulum, weight of the cart is expressed as M_0, weight of the pendulum is expressed as M_1 and g is gravity acceleration constant.
The equations (1) and (2) describe the horizontal and the vertical forces that the pendulum causes on cart.

$$H = M_1 \frac{d^2\left(r + l_s \sin\varphi\right)}{dt^2} \tag{1}$$

$$V = M_1 \frac{d^2\left(l_s \cos\varphi\right)}{dt^2} \tag{2}$$

where r is the position of the cart.
The equation (3) describe a motion equation of the cart and the equation (4) represents the rotary motion of the pendulum rod about its centre.

$$M_0 \frac{d^2 r}{dt^2} = F - H - F_r \frac{dr}{dt} \tag{3}$$

$$\Theta_s \frac{d^2\varphi}{dt^2} = V l_s \sin\varphi - H l_s \cos\varphi - C\frac{d\varphi}{dt} \tag{4}$$

where F_r is the constant of a velocity proportional friction of the cart, Θ_s is the inertia moment of the pendulum rod with respect to the centre of gravity and C is the friction constant of the pendulum.
Substitution of the equations (1) and (2) into the equations (3) and (4) creates the nonlinear equations (5) and (6) describing the behaviour of the inverted pendulum system.

$$Mr'' + F_r r' + M_1 l_s \varphi'' \cos\varphi - M_1 l_s \left(\varphi'\right)^2 \sin\varphi = F \tag{5}$$

$$\Theta\varphi'' + C\varphi' - M_1 l_s g \sin\varphi + M_1 l_s r'' \cos\varphi = 0 \tag{6}$$

where following abbreviations were used:

$$\Theta = \Theta_s + M_1 l_s^2 \tag{7}$$

$$M = M_0 + M_1 \tag{8}$$

The position of the cart r, the angle of the pendulum φ and their derivations r' and φ' are chosen as state variables as is shown in the equation (9).

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} r \\ r' \\ \varphi \\ \varphi' \end{bmatrix} \tag{9}$$

And the output variables are the position of the cart r, the angle of the pendulum φ.

$$\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} r \\ \varphi \end{bmatrix} \tag{10}$$

The nonlinear state-space model is described in the equation (11).

$$\boldsymbol{x}' = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \\ x_4' \end{bmatrix} = \begin{bmatrix} x_2 \\ f_1(\boldsymbol{x}.u) \\ x_4 \\ f_2(\boldsymbol{x}.u) \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} \tag{11}$$

where the functions f_1 and f_2 are derived from the equations (5) and (6).

$$f_1(\boldsymbol{x}.u) = \frac{1}{M_1^2 l_s^2 \cos^2 x_3 - M\Theta}\Big[-CM_1 l_s x_4 \cos x_3 + \\ + M_1^2 l_s^2 g \sin x_3 \cos x_3 + \Theta F_r x_2 - \Theta M_1 l_s x_4^2 \sin x_3 - \Theta u\Big] \tag{12}$$

$$f_2(\boldsymbol{x}.u) = \frac{1}{M_1^2 l_s^2 \cos^2 x_3 - M\Theta}\Big[MCx_4 - MM_1 l_s g \sin x_3 - \\ - M_1 l_s F_r x_2 \cos x_3 + M_1^2 l_s^2 x_4^2 \sin x_3 \cos x_3 + M_1 l_s u \cos x_3\Big] \tag{13}$$

The described nonlinear model has to be linearized around an operating point. The linearization about the operating point is done with partial derivation of the nonlinear model which is shown in equations (14) and (15).

$$\begin{aligned} \dot{\boldsymbol{x}}_\delta &= \boldsymbol{A}\boldsymbol{x}_\delta + \boldsymbol{B}\boldsymbol{u}_\delta \\ \boldsymbol{y}_\delta &= \boldsymbol{C}\boldsymbol{x}_\delta \end{aligned} \tag{14}$$

where

$$\boldsymbol{A}=\begin{bmatrix} 0 & 1 & 0 & 0 \\ \dfrac{\partial f_1(\boldsymbol{x},u)}{\partial x_1} & \dfrac{\partial f_1(\boldsymbol{x},u)}{\partial x_2} & \dfrac{\partial f_1(\boldsymbol{x},u)}{\partial x_3} & \dfrac{\partial f_1(\boldsymbol{x},u)}{\partial x_4} \\ 0 & 0 & 0 & 1 \\ \dfrac{\partial f_2(\boldsymbol{x},u)}{\partial x_1} & \dfrac{\partial f_2(\boldsymbol{x},u)}{\partial x_2} & \dfrac{\partial f_2(\boldsymbol{x},u)}{\partial x_3} & \dfrac{\partial f_2(\boldsymbol{x},u)}{\partial x_4} \end{bmatrix}$$

$$\boldsymbol{B}=\begin{bmatrix} 0 \\ \dfrac{\partial f_1(\boldsymbol{x},u)}{\partial u} \\ 0 \\ \dfrac{\partial f_2(\boldsymbol{x},u)}{\partial u} \end{bmatrix}$$

$$\boldsymbol{C}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{15}$$

In case of using multimodel approach, this linearization is done within multiple operation points. Every linearization creates a new linear model in the set of linear models. The operating points of the linearization should be chosen generally from the most nonlinear area of the system behaviour.

However, this is still a continuous-time model and it needs to be transferred into a discrete-time model. The chosen predictive control method using a specific form of the state-space model for prediction of the output values. The transformation of the continuous-time model into the suitable discrete-time model is done by transferring the state-space model (14) into the input-output model

$$A(s)\,y(t)=B(s)\,u(t) \tag{16}$$

and then into its discrete representation

$$\tilde{A}(z^{-1})\,y(k)=B(z^{-1})\Delta u(k) \tag{17}$$

where the polynom $\tilde{A}(z^{-1})$ is

$$\tilde{A}(z^{-1})=(1-z^{-1})\,A(z^{-1}) \tag{18}$$

STATE-SPACE PREDICTIVE CONTROL

The output values are predicted using the state-space CARIMA (Controlled Auto-Regressive and Integrated Moving Average) model (19)

$$\begin{aligned} \boldsymbol{x}(k+1)&=\tilde{\boldsymbol{A}}\boldsymbol{x}(k)+\boldsymbol{B}\Delta u(k) \\ y(k)&=\boldsymbol{C}\boldsymbol{x}(k) \end{aligned} \tag{19}$$

where the vector of state variables has form

$$\begin{aligned} \boldsymbol{x}(k)=\big[&y(k),y(k-1),\cdots,y(k-na), \\ &\Delta u(k-1),\cdots,\Delta u(k-nb+1)\big]^T \end{aligned} \tag{20}$$

The matrices $\tilde{\boldsymbol{A}}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ from the model (19) can be expressed as

$$\tilde{\boldsymbol{A}}=\begin{bmatrix} -\tilde{a}_1 & \cdots & -\tilde{a}_{na} & -\tilde{a}_{na+1} & b_2 & \cdots & b_{nb-1} & b_{nb} \\ 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$\boldsymbol{B}=\begin{bmatrix} b_1 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \end{bmatrix}^T$$

$$\boldsymbol{C}=\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{21}$$

The values $-\tilde{a}_i$ for $i=1,\ldots,na+1$ and b_j for $j=1,\ldots,nb$, where *na* is order of the polynom $\tilde{A}(z^{-1})$ and *nb* is the order of the polynom $B(z^{-1})$, consist of the coefficients of the polynoms $\tilde{A}(z^{-1})$ and $B(z^{-1})$ from the equation (18) (Bars et al. 2011; Camacho and Bordons 2004).

In case of using multimodel approach, the state-space model used to output values prediction is calculated as a combination of two or more models from the set of the linearized models. The choice of the models is determined by current state.

The final model is calculated according to the equation (22)

$$\begin{aligned} \tilde{\boldsymbol{A}}&=w_m\cdot\tilde{\boldsymbol{A}}_n+(1-w_m)\cdot\tilde{\boldsymbol{A}}_{n+1} \\ \boldsymbol{B}&=w_m\cdot\boldsymbol{B}_n+(1-w_m)\cdot\boldsymbol{B}_{n+1} \\ \boldsymbol{C}&=w_m\cdot\boldsymbol{C}_n+(1-w_m)\cdot\boldsymbol{C}_{n+1} \end{aligned} \tag{22}$$

where the w_m is the weighting coefficient of the linear model from the set of the linear models and *n* is the model number.

This weighting coefficient can be calculated as a linear or nonlinear function or it can be equal 1 in case of an edge transition between models.

The equation for the output values prediction is obtained by recursive substitution of the state equation of the equation (19). The final matrix form of this prediction is

$$\hat{\boldsymbol{y}}=\boldsymbol{F}\boldsymbol{x}+\boldsymbol{H}_f\Delta\boldsymbol{u}_f \tag{23}$$

where $\hat{\boldsymbol{y}}$ is the vector of the predicted output values and $\Delta\boldsymbol{u}_f$ is the vector of the future control increments

$$\hat{\boldsymbol{y}}=\begin{bmatrix} \hat{y}(k+1) \\ \hat{y}(k+2) \\ \vdots \\ \hat{y}(k+N) \end{bmatrix}$$

$$\Delta\boldsymbol{u}_f=\begin{bmatrix} \Delta u(k) \\ \Delta u(k+1) \\ \vdots \\ \Delta u(k+N) \end{bmatrix} \tag{24}$$

where N is the chosen time horizon for prediction.

The aim of the predictive control is minimize the difference between the future reference values and the predicted output values and minimize the control signal demand. The quadratic cost function witch satisfied this requirement is the equation (25).

$$J = \left(\boldsymbol{w} - \hat{\boldsymbol{y}}\right)^T \boldsymbol{Q}_\delta \left(\boldsymbol{w} - \hat{\boldsymbol{y}}\right) + \Delta \boldsymbol{u}_f^T \boldsymbol{Q}_\lambda \Delta \boldsymbol{u}_f \tag{25}$$

where $\boldsymbol{w}$ is a vector of the future reference values, $\hat{\boldsymbol{y}}$ is the vector of the predicted outputs values, $\boldsymbol{Q}_\lambda$ and $\boldsymbol{Q}_\delta$ are the diagonal weighting matrices containing the weighting coefficients λ and δ. The vector $\Delta \boldsymbol{u}_f$ is unknown vector of the future control increments (Camacho and Bordons 2004; Fikar and Mikleš 2008).
Because of the optimization method calculates with the possibility of the process variables constraints, this cost function needs to be modified into the form of the equation (26).

$$J = \frac{1}{2}\boldsymbol{u}^T \boldsymbol{H}_c \boldsymbol{u} + \boldsymbol{g}^T \boldsymbol{u} \tag{26}$$

where

$$\begin{aligned} \boldsymbol{H}_c &= 2\left(\boldsymbol{Q}_\lambda + \boldsymbol{H}_f^T \boldsymbol{Q}_\delta \boldsymbol{H}_f\right) \\ \boldsymbol{g}^T &= 2\left(\boldsymbol{F}\boldsymbol{x} - \boldsymbol{w}\right)^T \boldsymbol{Q}_\delta \boldsymbol{H}_f \end{aligned} \tag{27}$$

PREDICTOR-CORRECTOR METHOD

The predictor-corrector method is using to solve the inequality constrained convex quadratic problems

$$\begin{aligned} f(x) &= \frac{1}{2}\boldsymbol{x}^T \boldsymbol{G}\boldsymbol{x} + \boldsymbol{g}^T \boldsymbol{x} \\ \boldsymbol{A}^T \boldsymbol{x} &\geq \boldsymbol{b} \end{aligned} \tag{28}$$

which is exactly the problem that the predictive control solves. The equation (28) represents the general formulation of the constrained quadratic problem. The aim is to find the unknown vector $\boldsymbol{x}$ with respect to the chosen constrains representing the future values of the control signal increments (Nocedal and Wright 2000; Wright 1997).
The predictor-corrector is an iterative method so the starting points of the unknown vector $\boldsymbol{x}_0$, the vector of the Lagrange multipliers $\boldsymbol{\lambda}_0$ and the slackvector $\boldsymbol{s}_0$ where $\boldsymbol{s} = \boldsymbol{A}^T \boldsymbol{x} - \boldsymbol{b}, \boldsymbol{s} \geq 0$ have to be set first. These starting points are used to calculate the initial residual vectors $\boldsymbol{r}_d$, $\boldsymbol{r}_s$ and $\boldsymbol{r}_{s\lambda}$

$$\begin{aligned} \boldsymbol{r}_d &= \boldsymbol{G}\boldsymbol{x}_0 + \boldsymbol{g} - \boldsymbol{A}\boldsymbol{\lambda}_0 \\ \boldsymbol{r}_p &= \boldsymbol{s}_0 - \boldsymbol{A}^T \boldsymbol{x}_0 + \boldsymbol{b} \\ \boldsymbol{r}_{s\lambda} &= \boldsymbol{S}_0 \boldsymbol{\Lambda}_0 \boldsymbol{e} \end{aligned} \tag{29}$$

where $\boldsymbol{S}_0$ and $\boldsymbol{\Lambda}_0$ are the diagonal matrices containing the elements of the $\boldsymbol{s}_0$ and $\boldsymbol{\lambda}_0$. The $\boldsymbol{e}$ is vector of ones (Nocedal and Wright 2000; Wright 1997).
The initial complementarity measure μ has to be calculated as next step which is needed for centering parameter σ

$$\mu = \frac{\boldsymbol{s}_0^T \boldsymbol{\lambda}_0}{m} \tag{30}$$

where m is the number of the inequality constraints.
The whole algorithm can be divided into two parts. The first is the calculation of the predictor step and the second is the calculation of the corrector step. The predictor step is calculated by applying the Newton's method around the current point on the equations (29).

$$\begin{bmatrix} \boldsymbol{G} & -\boldsymbol{A} & \boldsymbol{0} \\ -\boldsymbol{A}^T & \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{S} & \boldsymbol{\Lambda} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{x}^{aff} \\ \Delta \boldsymbol{\lambda}^{aff} \\ \Delta \boldsymbol{s}^{aff} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{r}_d \\ \boldsymbol{r}_p \\ \boldsymbol{r}_{s\lambda} \end{bmatrix} \tag{31}$$

Solving these equations will give us the affine scaling direction $\left(\Delta \boldsymbol{x}^{aff}, \Delta \boldsymbol{\lambda}^{aff}, \Delta \boldsymbol{s}^{aff}\right)$. Then the scaling parameter α^{aff} for the predictor step witch satisfy the conditions in the equations (32) is chosen.

$$\begin{aligned} \boldsymbol{\lambda} + \alpha_\lambda^{aff} \Delta \boldsymbol{\lambda}^{aff} &\geq 0 \\ \boldsymbol{s} + \alpha_s^{aff} \Delta \boldsymbol{s}^{aff} &\geq 0 \end{aligned} \tag{32}$$

The final scaling parameter is chosen as follows:

$$\begin{aligned} \alpha_\lambda^{aff} &= \min_{i:\Delta\lambda_i<0}\left(1, \min \frac{-\lambda_i}{\Delta \lambda_i^{aff}}\right) \\ \alpha_s^{aff} &= \min_{i:\Delta s_i<0}\left(1, \min \frac{-s_i}{\Delta s_i^{aff}}\right) \\ \alpha^{aff} &= \min\left(\alpha_\lambda^{aff}, \alpha_s^{aff}\right) \end{aligned} \tag{33}$$

The predictor step is finished with the calculation of the complementarity measure μ^{aff} and the centering parameter σ.

$$\mu^{aff} = \frac{\left(\boldsymbol{s} + \alpha^{aff} \Delta \boldsymbol{s}^{aff}\right)^T \left(\boldsymbol{\lambda} + \alpha^{aff} \Delta \boldsymbol{\lambda}^{aff}\right)}{m} \tag{34}$$

$$\sigma = \left(\frac{\mu^{aff}}{\mu}\right)^3 \tag{35}$$

The adjustment of the right hand side of the equation (31) by the computed affine scaling direction and the centering parameter will prepare the equation for the corrector step (36) (Nocedal and Wright 2000; Wright 1997).

$$\begin{bmatrix} \boldsymbol{G} & -\boldsymbol{A} & \boldsymbol{0} \\ -\boldsymbol{A}^T & \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{S} & \boldsymbol{\Lambda} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{x} \\ \Delta \boldsymbol{\lambda} \\ \Delta \boldsymbol{s} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{r}_d \\ \boldsymbol{r}_p \\ \boldsymbol{r}_{s\lambda} + \Delta \boldsymbol{S}^{aff} \Delta \boldsymbol{\Lambda}^{aff} \boldsymbol{e} - \sigma \mu \boldsymbol{e} \end{bmatrix} \tag{36}$$

Solving this system gives us the final scaling direction $\left(\Delta \boldsymbol{x}, \Delta \boldsymbol{\lambda}, \Delta \boldsymbol{s}\right)$. The step length is chosen in the same way it was in the predictor step calculation in the equations (33).

$$\begin{aligned} \boldsymbol{\lambda} + \alpha_\lambda \Delta \boldsymbol{\lambda} &\geq 0 \\ \boldsymbol{s} + \alpha_s \Delta \boldsymbol{s} &\geq 0 \end{aligned} \tag{37}$$

The last calculation of this method is the update of the unknown vector $\boldsymbol{x}$, the vector of the Lagrange multipliers $\boldsymbol{\lambda}$ and the slackvector $\boldsymbol{s}$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha\Delta\boldsymbol{x}$$
$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha\Delta\boldsymbol{\lambda}$$
$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k + \alpha\Delta\boldsymbol{s} \tag{38}$$

and the residuals vectors $\boldsymbol{r}_d$, $\boldsymbol{r}_s$ and $\boldsymbol{r}_{s\lambda}$ and the complementarity measure μ.

$$\boldsymbol{r}_d = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{g} - \boldsymbol{A}\boldsymbol{\lambda}$$
$$\boldsymbol{r}_p = \boldsymbol{s} - \boldsymbol{A}^T\boldsymbol{x} + \boldsymbol{b} \tag{39}$$
$$\boldsymbol{r}_{s\lambda} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{e}$$

$$\mu = \frac{\boldsymbol{s}^T\boldsymbol{\lambda}}{m} \tag{40}$$

RESULTS

This section shows the results of the process simulation of controlled fall of the inverted pendulum from up to down position. The controller parameters N = 20 steps, λ = 0.001, δ = 10 and the sampling period T_0 = 40 ms were set for all of the simulations. The presented simulation results are differ in calculation of the model weighting coefficient. The simulations were compared by two quadratic criterions for analysis of the control quality. The first criterion, described in equation (41), compares the control increments made in every step and the second criterion, described in equation (42), compares a difference between the reference value and the output value.

$$S_u = \frac{1}{N}\sum_{k=1}^{N}\Delta u^2(k) \tag{41}$$

$$S_e = \frac{1}{N}\sum_{k=1}^{N}\left[w(k) - y(k)\right]^2 \tag{42}$$

Table 1 shows the system parameters used for the mathematical model of the system.

Table 1 : System parameters

Symbol	Value	Meaning
M_0 [kg]	4	Cart weight
M_1 [kg]	0.36	Pendulum weight
l_s [m]	0.42	Pendulum length
Θ [kg.m^2]	0.08433	Pendulum inertia moment
F_r [kg/s]	6.5	Cart friction
C [Kg.m^2/s]	0.00652	Pendulum friction
k_a [N/V]	7.5	Servo amplifier gain
g [m/s^2]	9.81	Gravity constant

The figure 3 shows the pulse response (uncontrolled fall) of the system for the input pulse F = -2 N for 1s.

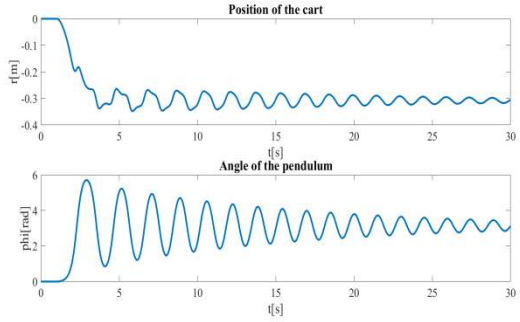


Figure 3 : Uncontrolled fall

The matrices $\tilde{\boldsymbol{A}}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ from the model (19) are calculated according the equation (43)

$$\tilde{\boldsymbol{A}} = w_m \cdot \tilde{\boldsymbol{A}}_n + (1 - w_m)\cdot\tilde{\boldsymbol{A}}_{n+1}$$
$$\boldsymbol{B} = w_m \cdot \boldsymbol{B}_n + (1 - w_m)\cdot\boldsymbol{B}_{n+1} \tag{43}$$
$$\boldsymbol{C} = w_m \cdot \boldsymbol{C}_n + (1 - w_m)\cdot\boldsymbol{C}_{n+1}$$

where w_m is the weighting coefficient and n=1...12 is the model number. The matrices $\tilde{\boldsymbol{A}}_n, \boldsymbol{B}_n$ and $\boldsymbol{C}_n$ are matrices $\tilde{\boldsymbol{A}}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ of the n-th model The nonlinear model was linearized in 12 operating points of the pendulum angle. These operating points have equidistant range from 0 to 2π rad. Therefore, the angle distance between operating points is $\frac{1}{6}\pi$ rad.

The figures 4 and 5 show the simulation results of edge transition between models where w_m=1.

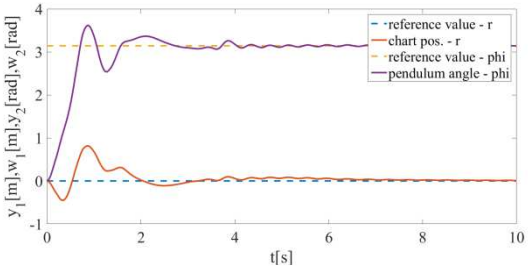


Figure 4 : System outputs - the edge transition

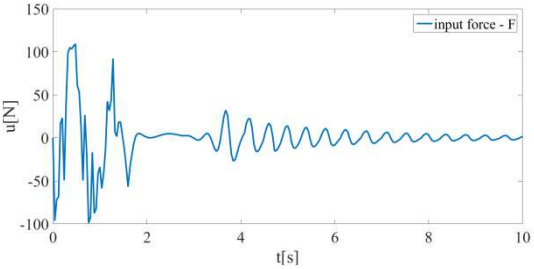


Figure 5 : System input - the edge transition

The figures 7 and 8 show the simulation results of the linear transition between models where w_m is calculated according to equation (44)

$$w_m = ax_3(k) + n \tag{44}$$

where $a = -\frac{6}{\pi}$ and n=1...12 is the model number. An example of this linear transition between models is shown in figure 6.

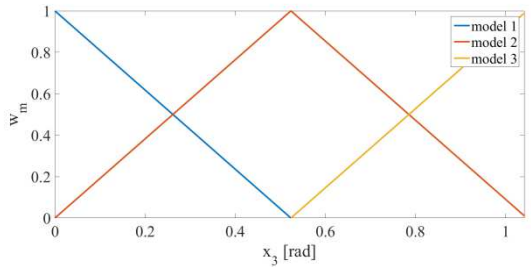


Figure 6 : Linear transition between models

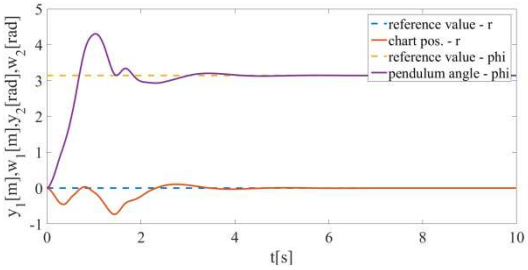


Figure 7 : System outputs - the linear transition

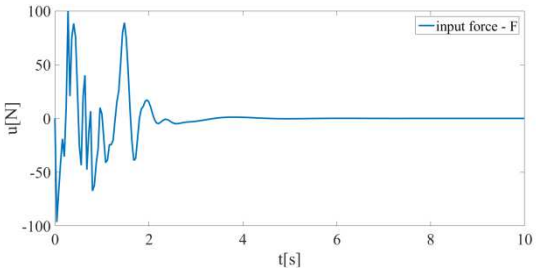


Figure 8 : System input - the linear transition

The figures 10 and 11 show the simulation results of the nonlinear transition between model where w_m is calculated according equation (45)

$$w_m = e^{-\frac{(x_3(k)-\mu)^2}{2\sigma^2}} \tag{45}$$

where $\mu = (n-1)\frac{1}{6}\pi; n = 1\ldots12$ and $\sigma = 0.2224$. An example of nonlinear transition between models is shown in figure 9.

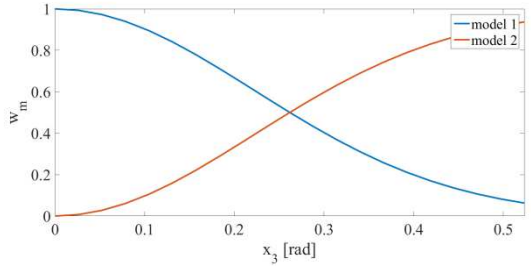


Figure 9 : Nonlinear transition between models

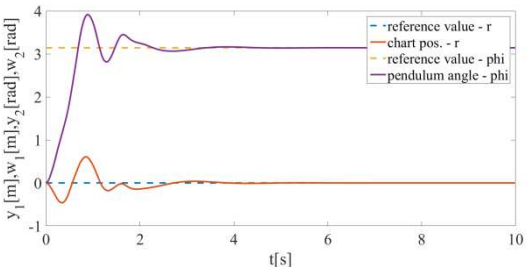


Figure 10 : System output - the nonlinear transition

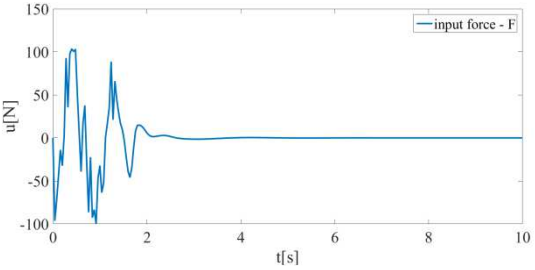


Figure 11 : System input - the nonlinear transition

The table 2 shows the results of the quadratic criterions.

Table 2 :Simulation results

	Edge transition	Linear transition	Nonlinear transition
S_{e1}[m^2]	0.034	0.029	0.017
S_{e2}[rad^2]	0.343	0.378	0.339
S_u[N^2]	361.06	262.40	301.37

CONCLUSION

In this paper, the multimodel approach of the predictive control based on the state-space CARIMA model was presented. The controller was tested on the inverted pendulum system which is an example of the nonlinear single-input two-output system. The goal of this multimodel approach in the predictive controller was the controlled fall of the pendulum from up to down position. The movement of the cart is acting like a disturbance in the system. The sampling period was chosen as T_0 = 40 ms. The mathematical model of the inverted pendulum was made according to the real laboratory model of the inverted pendulum Amira PS600. The aim of this paper is to present a possible approach in control of the nonlinear models using multimodel approach. That means using set of linear models to describe the behaviour of the nonlinear model. This set of models is created by linearization the nonlinear model in multiple operating points. The final linear model used to prediction of the output values is calculated as a combination of two linearized models according to the current state. The transition between linearized models is done using three methods. The first one is the edge transition, where one linearized model is using directly for the output values prediction. The second method is transition between linear models according to the linear function and the third method is transition between linear models according to the nonlinear function. The control signal is calculated by

the minimization of the cost function that minimize the differences between the output and the reference signals and the control signal increments. This minimization is achieved by predictor-corrector method. The result section compares the different methods of transition between linear models in multimodel approach. The examined criterions show that the nonlinear transition between models follows the reference signal as the best of the presented methods. The linear transition has the least control signal changes of the presented methods.

REFERENCES

Albertos Pérez P. and Sala A. 2004. *Multivariable Control Systems: an Engineering Approach*. Springer. London.

Amira. 2000. *PS600 Laboratory Experiment Inverted Pendulum.* Amira GmbH, Duisburg.

Bars R.; R. Haber and U. Schmitz. 2011. *Predictive control in process engineering: From the basics to the applications*. Weinhaim: Willey-VCH Verlag.

Bobál, V. 2008, *Adaptive and predictive control*. vol. 1. Zlin, Tomas Bata University in Zlin.

Camacho E.F. and C. Bordons. 2004. *Model predictive control*, Springer Verlag, London.

Chalupa P. and V. Bobál. 2008. "Modelling and Predictive Control of Inverted Pendulum". In: Proceedings 22nd European Conference on Modelling and Simulation. pp. 531-537.

Fikar M. and J. Mikleš. 2008. *Process modelling, optimisation and control*, Springer-Verlag, Berlin.

Hangos K.M.; Bokor J. and Szederkényi G. 2004. *Analysis and Control of Nonlinear Process Systems*. Springer. London.

Maciejowski J.M. 2002. *Predictive control with constraints*, Prentice Hall, London.

Nocedal J. and S. Wright. 2000. *Numerical optimisation second edition*. Springer, New York.

Rossiter J.A. 2003. *Model based predictive control: a practical approach*, CRC Press.

Wang L. 2009. *Model predictive control system design and implementation using MATLAB*, Springer Verlag, London.

Wright S. 1997 *Primal-dual interior point methods*. Philadelphia: Society for Industrial and Applied Mathematics.

ACKNOWLEDGMENT

This article was created with support of the Ministry of Education of the Czech Republic under grant IGA reg. n. IGA/CebiaTech/2018/002.

AUTHOR BIOGRAPHIES

LUKÁŠ RUŠAR studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2014. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interests focus on model predictive control. His e-mail address is rusar@fai.utb.cz.

VLADIMÍR BOBÁL graduated in 1966 from the Brno University of Technology, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are adaptive and predictive control, system identification, time-delay systems and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.

CONTROL OF TEMPERATURE INSIDE PLUG-FLOW TUBULAR CHEMICAL REACTOR USING 1DOF AND 2DOF ADAPTIVE CONTROLLERS

Jiri Vojtesek, Lubos Spacek and Frantisek Gazdos
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nam. TGM 5555, 760 01 Zlin, Czech Republic
Email: {vojtesek, lspacek, gazdos}@utb.cz

KEYWORDS

Simulation; Tubular chemical reactor; Steady-state analysis; Dynamic analysis; Adaptive control; Recursive identification; Pole-placement method; Polynomial synthesis

ABSTRACT

The goal of this contribution is to present simulation results of the adaptive control of the tubular chemical reactor as a typical equipment used in the industry. The mathematical model of the tubular reactor is described by a set of nonlinear partial differential equations. The numerical solution of such model is not trivial but the combination of the Finite difference and Rung-Kutta's method can be used for this task. The adaptivity of the controller is satisfied by the recursive identification of the external linear model as a linear representation of the originally nonlinear system. Control synthesis employs a polynomial approach together with the connection to the Pole-placement method and the spectral factorization satisfies basic control requirements and provides not only the structure of the controller but also computational relations. There are compared results for control systems with one degree-of-freedom (1DOF) and two degrees-of-freedom (2DOF).

INTRODUCTION

A tubular chemical reactor is an equipment that can be found in various industrial applications mainly in the chemical or biochemical industry (Ingham et al. 2000), (Russell and Denn 1972). Unfortunately, these systems are usually nonlinear and very complex for the mathematical description. The result of the modelling is often mathematical model in the form of a set of nonlinear partial differential equations (PDE) (Dostal et al. 1996). This set could be then numerically solved for example by the Finite differences method (Grossmann et al. 2007), (Evans 2010) that transform the set of PDE to the set of Ordinary Differential Equations (ODE) which are much better solvable using various numerical methods. The most common ones, Runge-Kutta's methods are very famous because they can be easily programmed or they are even build-in functions in the mathematical software like MathWork's Matlab (Mathews and Fink 2004), Wolfram Mathematica etc.
Adaptive control (Astrom and Wittenmark 1989) is not a new control technique but it is still used and developed in the practice. It's philosophy is taken from the nature where animals, plants or even human beings "adopts" they behaviour to the environment they are living in. Similarly, an adaptive controller changes its structure, parameters etc. according to requirements, settings, disturbances or generally changing conditions of the control (Bobal et al. 2005). This feature is a big advantage to the conventional methods that produces controllers with fixed parameters that could lead to nonoptimal or oven unwanted control results.
Use of the polynomial synthesis (Kucera 1993) in the adaptive control satisfies basic control requirements such as stability, reference signal tracking and disturbance attenuation. An advantage of this method is also in the fact, that it provides not only the structure of the controller but also relations for computing of controller's parameters.
All results presented in this contribution came from the simulation in the mathematical software MathWorks Matlab, version 7.1. Future work may aim to the verification of the controller on a real system or model of a real system.

PLUG-FLOW CHEMICAL REACTOR

A system which will be subjected to simulation experiments is a tubular chemical reactor (Dostal et al. 1996). We consider reaction inside has this simple scheme:

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \tag{1}$$

The simplified scheme of the tubular chemical reactor can be found in Fig. 1.

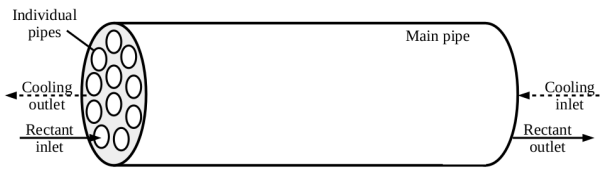


Fig. 1. Schematic representation of the Plug-flow Tubular Chemical Reactor

The mathematical model of this system is constructed with the help of material balances inside. Very important thing in the construction of the model is that state variables depend not only on the time variable, t, but also on the space variable, z. As a result, the mathematical model is described by the set of five Partial Differential Equations (PDE):

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

$$
\begin{aligned}
&\frac{\partial c_A}{\partial t} + v_r \frac{\partial c_A}{\partial z} = -k_1 \cdot c_A \\
&\frac{\partial c_B}{\partial t} + v_r \frac{\partial c_B}{\partial z} = k_1 \cdot c_A - k_2 \cdot c_B \\
&\frac{\partial T_r}{\partial t} + v_r \frac{\partial T_r}{\partial z} = \frac{h_r}{\rho_r c_{pr}} - \frac{4 \cdot U_1}{d_1 \rho_r c_{pr}} \cdot (T_r - T_w) \\
&\frac{\partial T_w}{\partial t} = \frac{4}{(d_2^2 - d_1^2)\, \rho_w c_{pw}} [d_1 \cdot U_1 \cdot (T_r - T_w) \\
&\qquad\qquad + d_2 \cdot U_2 \cdot (T_c - T_w)] \\
&\frac{\partial T_c}{\partial t} - v_c \cdot \frac{\partial T_c}{\partial z} = \frac{4 \cdot n_1 \cdot d_2 \cdot U_2}{(d_3^2 - n_1 \cdot d_2^2)\, \rho_c c_{pc}} (T_w - T_c)
\end{aligned}
\tag{2}
$$

Variables v_r and v_c represents fluid velocities of the reactant and cooling that are computed from

$$
v_r = \frac{q_r}{f_r};\ v_c = \frac{q_c}{f_c} \tag{3}
$$

where constants f_r and f_c are computed from

$$
f_r = n_1 \frac{\pi \cdot d_1^2}{4};\ f_c = \frac{\pi}{4} \left(d_3^2 - n_1 \cdot d_2^2\right) \tag{4}
$$

The nonlinearity of this system is mainly because of the reaction heat, h_r, and reaction velocities, k_i, that are computed using the so called Arrhenius law:

$$
\begin{aligned}
h_r &= h_1 \cdot k_1 \cdot c_A + h_2 \cdot k_2 \cdot c_B \\
k_i &= k_{0i} \cdot \exp\left(\frac{-E_i}{R \cdot T_r}\right) \quad \text{for } i = 1, 2
\end{aligned}
\tag{5}
$$

with k_{0i} as pre-exponential factors, E_j are activation energies, R represents universal gas constant and h_i as activation energies.

Fixed parameters of the reactor are shown in Table I (Dostal et al. 1996).

This system has five state variables - concentrations of compounds A and B $c_A(z,t)$, $c_B(z,t)$, temperature of the reactant $T_r(z,t)$, temperature of the metal wall $T_w(z,t)$ and the temperature of the coolant $T_c(z,t)$.

There is also wide range of input variables, for example input concentrations $c_{A0}(t)$ or $c_{B0}(t)$ or input temperatures of the reactant, $T_{r0}(t)$, or coolant at the end of the reactor as we consider counter-current cooling, $T_{cL}(t)$. However, the change of these input variables in control is very difficult which makes them useless from the practical point of view.

That is why the change of the flow rate of the coolant q_c was used in our work. This flow rate is used in the computation of the fluid velocity v_c as it is shown in (3). The controlled variable will be mean reactant temperature inside the reactor. Both, control ($u(t)$) and controlled ($y(t)$) variable are then

$$
\begin{aligned}
u(t) &= \frac{q_c(t) - q_c^s}{q_c^s} \cdot 100\ [\%] \\
y(t) &= T_{mean}(t) = \frac{\sum_{z=1}^{N} T_r(z,t)}{N}\ [K]
\end{aligned}
\tag{6}
$$

where q_c^s denotes initial (steady-state) volumetric flow rate of the coolant, z is space variable that corresponds to the discretization mentioned above where the length of the reactor L is divided into N equivalent parts.

Name and value of the parameter
Inner diameter of small pipe $d_1 = 0.02\ m$
outer diameter of the small pipe $d_2 = 0.024\ m$
diameter of the main pipe $d_3 = 1\ m$
number of pipes $n_1 = 1\ 200$
length of the reactor $L = 6\ m$
density of the reactant $\rho_r = 985\ kg \cdot m^{-3}$
density of the metal wall $\rho_w = 7\ 800\ kg \cdot m^{-3}$
density of the coolant $\rho_c = 998\ kg \cdot m^{-3}$
heat capac. of reactant $c_{pr} = 4.05\ kJ\ kg^{-1}K^{-1}$
heat capac. of metal wall $c_{pw} = 0.71\ kJ\ kg^{-1}K^{-1}$
heat capac. of the coolant $c_{pc} = 4.18\ kJ\ kg^{-1}K^{-1}$
heat transfer coef. 1 $U_1 = 2.8\ kJ\ m^{-2}\ K^{-1}\ s^{-1}$
heat transfer coef. 2 $U_2 = 2.56\ kJ\ m^{-2}\ K^{-1}\ s^{-1}$
pre-exponential factor 1 $k_{10} = 5.61 \cdot 10^{16}\ s^{-1}$
pre-exponential factor 2 $k_{20} = 1.128 \cdot 10^{16}\ s^{-1}$
activation energy 1 to R $E_1/R = 13\ 477\ K$
activation energy 2 to R $E_2/R = 15\ 290\ K$
enthalpy of reaction 1 $h_1 = 5.8 \cdot 10^4\ kJ \cdot kmol^{-1}$
enthalpy of reaction 2 $h_2 = 1.8 \cdot 10^4\ kJ \cdot kmol^{-1}$
input concentration of A $c_{A0}^s = 2.85\ kmol\ m^{-3}$
input temperature of the reactant $T_{r0}^s = 323\ K$
input temperature of the coolant $T_{c0}^s = 293\ K$

TABLE I: Fixed parameters of the tubular chemical reactor

Steady-state Analysis

Once we have the mathematical model of the system, we can move on to the simulation of the steady-state and dynamics which helps us with the choice of the optimal working point (from steady-state analysis) and External Linear Model (ELM) which was used later in the adaptive control (from dynamic analysis).

The steady-state analysis computes values of the state variables in time $time \to \infty$ which practically means that the partial derivations with respect to time variable in (2) are equal to zero, i.e. $\partial(\cdot)/\partial t = 0$ and the set of PDE is transformed to the set of ODE. The partial derivations with respect to space variable, $\partial(\cdot)/\partial z$ can be, on the other hand, solved using the Finite difference method (Grossmann et al. 2007), (Evans 2010) that replaces derivatives with respect to z by the first feedback difference

$$
\left.\frac{dx}{dz}\right|_{z=z_i} \approx \frac{x(i) - x(i-1)}{h_z}, \text{for } i = 1, 2, \ldots n \tag{7}
$$

for x as a general state variable, in our case c_A, c_B, T_r and T_w and with h_z representing discretization step. The coolant flows through the reactor in the different direction as we consider a counter-current cooling inside this reactor. Because of this, the first forward difference is used for the temperature of the coolant as a state variable

$$
\left.\frac{dT_c}{dz}\right|_{z=z_j} \approx \frac{T_c(i+1) - T_c(i)}{h_z}, \text{for } j = n, n-1, \ldots 0 \tag{8}
$$

As we consider the change of the input flow rate of the coolant, q_c as an input variable, the steady-state analysis was done for various values of $q_c =< 0.1, 0.35 > m^3 \cdot s^{-1}$.

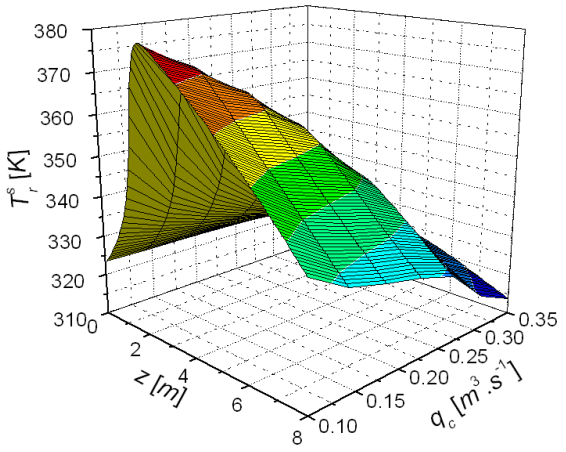


Fig. 2. Results of the steady-state analysis for various coolant flow rate q_c

Results of the steady-state analysis presented in Fig. 2 shows expected nonlinearity of the system.

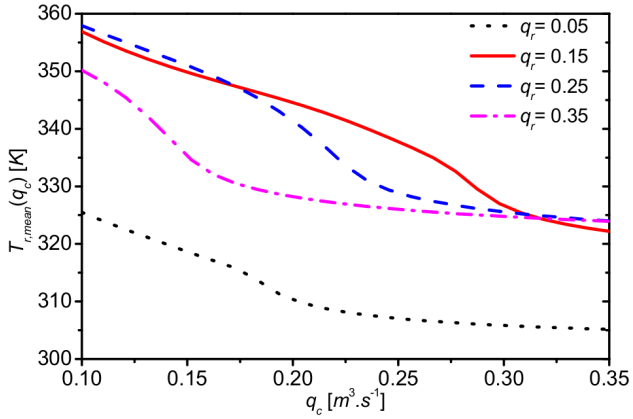


Fig. 3. Mean temperature of the reactant, $T_{r,mean}$, dependence on the flow rate of the coolant, q_c, for various values of the flow rate of the reactant, q_r

As the controlled output $y(t)$ will be the mean temperature of the reactant, the second steady-state analysis is focused on the observation of the $Tr, mean$ value for different values of flow rates of the reactant, q_r, and the cooling, q_c. Results are shown in Fig. 3 and we can confirm the nonlinearity of this output.

The working point should be then defined by flow rates of the reactant and the coolant $q_r^s = 0.15\ m^3 \cdot s^{-1}$, $q_c^s = 0.275\ m^3 \cdot s^{-1}$. The mean temperature of the reactant for this workint point in the steady-state is $T_{mean}^s = 333.12\ K$. Steady-state values of the state variables are used as initial conditions for the dynamic study.

Dynamic Analysis

The steady-state analysis is in fact solution of the set of nonlinear algebraic equations, the dynamic analysis is, on the other hand, pure numerical solution of the set of nonlinear PDE (2). We can use again Finite difference method described in the previous part that cope with the derivatives with respect to the space variable z. The set of PDE is then transformed to the set of Ordinary Differential Equations (ODE) that can be solved for example with the Runge-Kutta's methods (Johnson 1982). Advantage of these methods is that they are often build-in functions in the mathematical software and MathWorks Matlab (Mathews and Fink 2004) that is used in this work for the simulation is not an exception. As the reactor is discretized into $N = L/h_z$ steps using the first forward (7) or feedback (8) difference, the numerical computation of the dynamics was done in each step and the computed value of the state variable was used as a boundary condition for the numerical computation in the next step.

The dynamic analysis observes the behaviour of the system after the step change of the input variable, in our case the input variable $u(t)$ is a change of the flow rate of the coolant. Six various step changes were done and the resulted courses of the mean temperature of the coolant are shown in Fig. 4.

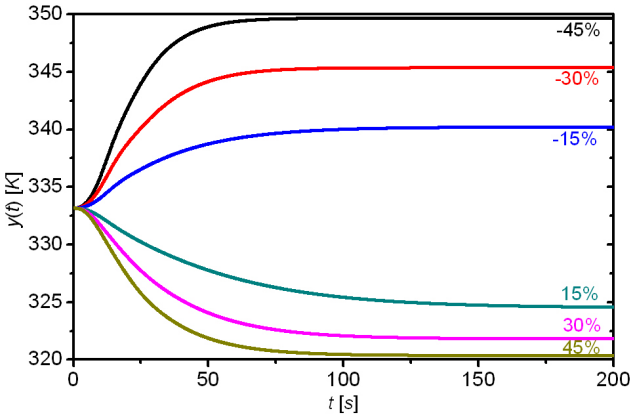


Fig. 4. Results of the dynamic analysis for various changes of $u(t)$

Step responses of the output variable in Fig. 4 present nonlinearity of the controlled system – final values of $y(t)$ after negative and positive steps are not equivalent. On the other hand, the course cold be described by the second order transfer function with the relative order one¿

$$G(s) = \frac{b_2 s + b_0}{s^2 + a_1 s + a_0} \tag{9}$$

ADAPTIVE CONTROL

The adaptive control (Astrom and Wittenmark 1989) approach here uses ELM in the form of (9) as a linear representation of the originally nonlinear system (2) parameters of which are estimated recursively during the control (Bobal et al. 2005). As parameters of the controller are computed with the use of identified

polynomials $a(s)$ and $b(s)$, the controller can react to a change of the state, appearance of a disturbance etc.

Control Synthesis

The result of the control synthesis is the structure of the control loop and also the structure of the controller and relations for computing its parameters. The polynomial approach (Kucera 1993) was used for this task.
There are various control configurations. This work will compare results for systems with one degree-of-freedom (often denoted as a 1DOF) and two degrees-of-freedom (2DOF) (Grimble 2006). The schematic representation of 1DOF control configuration is shown in Fig. 5.

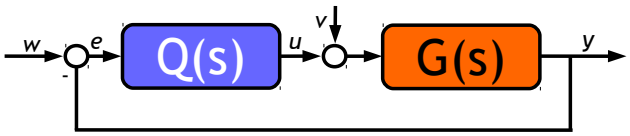


Fig. 5. One Degree-of-freedom (1DOF) Control Configuration

The block $G(s)$ represents transfer function (TF) of the ELM (9). The controller is represented as a block $Q(s)$ in the feedback part which is also a transfer function, generally

$$Q(s) = \frac{q(s)}{s \cdot \tilde{p}(s)} \tag{10}$$

with unknown polynomials $q(s)$ and $\tilde{p}(s)$ that can be computed from the Diophantine equation

$$a(s) \cdot s \cdot \tilde{p}(s) + b(s) \cdot q(s) = d(s) \tag{11}$$

with the Method of uncertain coefficients (Riley and Bence 2010). Polynomials $a(s)$ and $b(s)$ are known from the recursive identification and polynomial $d(s)$ on the right side of (11) is a stable optional polynomial.
Degrees of the polynomials $q(s)$, $\tilde{p}(s)$ and $d(s)$ are for the second order TF (9)¿

$$\begin{aligned} \deg q(s) &= \deg a(s) = 2 \\ \deg \tilde{p}(s) &= \deg a(s) - 1 = 1 \\ \deg d(s) &= \deg a(s) + \deg \tilde{p}(s) + 1 = 4 \end{aligned} \tag{12}$$

which means that controller $Q(s)$ in (10) has form

$$Q(s) = \frac{q(s)}{s \cdot \tilde{p}(s)} = \frac{q_2 s^2 + q_1 s + q_0}{s \cdot (\tilde{p}_1 s + \tilde{p}_0)} \tag{13}$$

The control configuration with two degrees-of-freedom (2DOF) (Grimble 2006) displayed in Fig. 6 has a controller divided into two parts – feedback $Q(s)$ and feedforward $R(s)$ part.
TF of the feedback and feedfoward part has again general form

$$Q(s) = \frac{q(s)}{s \cdot \tilde{p}(s)}; \quad R(s) = \frac{r(s)}{s \cdot \tilde{p}(s)}; \tag{14}$$

where parameters of the polynomials are computed from the set of Diophantine equations

$$\begin{aligned} a(s) \cdot s \cdot \tilde{p}(s) + b(s) \cdot q(s) &= d(s) \\ t(s) \cdot s + b(s) \cdot r(s) &= d(s) \end{aligned} \tag{15}$$

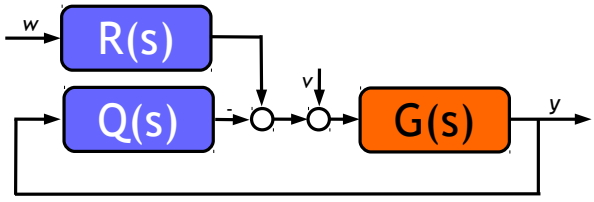


Fig. 6. Two Degrees-of-freedom (2DOF) Control Configuration

by the Method of uncertain coefficients. Polynomial $t(s)$ is an additional polynomial used for computation of (15). Degrees of polynomials $\tilde{p}(s), q(s), r(s), t(s)$ and $d(s)$ are again for the second order TF (9):

$$\begin{aligned} \deg q(s) &= \deg a(s) = 2 \\ \deg \tilde{p}(s) &= \deg a(s) - 1 = 1 \\ \deg r(s) &= 0; \deg t(s) = 0 \\ \deg d(s) &= 2 \cdot \deg a(s) = 4 \end{aligned} \tag{16}$$

and transfer functions of the controller (14) are

$$Q(s) = \frac{q_2 s^2 + q_1 s + q_0}{s \cdot (\tilde{p}_1 s + \tilde{p}_0)}; \; R(s) = \frac{r_0}{s \cdot (\tilde{p}_1 s + \tilde{p}_0)}; \tag{17}$$

Pole-placement Method

The polynomial synthesis presented in previous chapter shows the structure of the 1DOF a 2DOF controllers. Polynomials $a(s)$ and $b(s)$ are known from the recursive identification, polynomials $\tilde{p}(s), q(s)$ and $r(s)$ can be computed from the Diophantine equations (11) and (15) but there is still one unknown polynomial $d(s)$ on the right side of Diophantine equations. As it is an optional stable polynomial, we can choose a method for designing this polynomial. For example, simple Pole-placement method (Kucera 1993) can be used for this task.
In this method, the polynomial $d(s)$ has general form

$$d(s) = \prod_{i=1}^{\deg d(s)} (s + \alpha_i) \tag{18}$$

which means that we have 4 root α_i as $\deg d(s) = 4$.
Disadvantage of the Pole-placement method is in its generality. There is no sophisticated advice for the choice of the possition of roots, if it is better to use double, triple roots etc. Our previous experiments (Vojtesek and Dostal 2009) have shown that it is good to connect the choice of the polynomial with the controlled system somehow. This could be done for example with the use of spectral factorization of the polynomial $a(s)$ in the denominator of the identified ELM's TF (9) and obtain polynomial $n(s)$ that is, in fact, the stable mirror of the polynomial $a(s)$, i.e.

$$n^*(s) \cdot n(s) = a^*(s) \cdot a(s) \tag{19}$$

The stability is very important, because the whole polynomial $d(s)$ must be stable. It is clear, that $\deg n(s) = \deg a(s) = 2$ and the rest two degrees of the polynomial

$d(s)$ come from the Pole-placement method. Finally, polynomial $d(s)$ is then

$$d(s) = n(s) \cdot (s + \alpha)^2 \tag{20}$$

Advantage of this method is that optional position of the variable $\alpha > 0$ is also a tuning parameter that affect the course of the controlled output variable.

Recursive Identification of ELM

It was already mentioned, that the adaptivity is satisfied by the recursive identification of the ELM (9). We can discover, that both TF of the system $G(s)$ and controller $Q(s), R(s)$ are considered in a continuous-time (CT) that is more accurate than a discrete-time (DT) models. On the other hand, an online-line identification of the CT ELM is much complicated than DT ones.
The compromise between accuracy of the CT model and better applicability of the DT model can be found in the so called delta models that belong to the class of DT models but both input and output variables are recomputed to their δ value that reflects the sampling period T_v. It was proofed for example in (Stericker and Sinha 1993), that parameters of the delta model approach to corresponding parameters of the CT model. That is why we can call our method "hybrid" adaptive control because the controller is considered as CT but measurements of the input and the output variable are done in the discrete time intervals. We expect, that identified parameters of the delta ELM are similar to parameters of the CT ELM (9).
The CT model (9) could be then rewritten to the delta-form:

$$a'(\delta) \cdot y(t') = b'(\delta) \cdot u(t') \tag{21}$$

for DT polynomials $a'(\delta), b'(\delta)$ parameters of which are expected to be close to the parameters of CT polynomials $a(s), b(s)$ in (9) and t' is a discrete time. The relation to the sampling period T_v could be found in the new complex variable γ that is alternative to the complex variable s in the CT model or z in DT model. This variable could be computed for example (Mukhopadhyay et al. 1992) from this general equation

$$\gamma = \frac{z - 1}{\beta \cdot T_v \cdot z + (1 - \beta) \cdot T_v} \tag{22}$$

with β as an optional parameter. It is obvious, that we could have various delta models depending on the choice of the β. For example, $\beta = 0$ produces so called "forward delta model" with $\gamma = (z - 1)/T_v$.
The recursive identification then estimates the vector of parameters, $\boldsymbol{\theta}_\delta$, with known values in the data vector, $\boldsymbol{\phi}_\delta$, from the ARX model in vector form:

$$y_\delta(k) = \boldsymbol{\theta}_\delta^T(k) \cdot \boldsymbol{\phi}_\delta(k-1) + e(k) \tag{23}$$

where $e(k)$ is a general random immeasurable component.
The data vector $\boldsymbol{\phi}_\delta$ and the estimated vector of parameters $\boldsymbol{\theta}_\delta$ in (23) are for this concrete second order ELM (9)

$$\begin{aligned} \boldsymbol{\phi}_\delta(k-1) = [&-y_\delta(k-1), -y_\delta(k-2), \dots \\ &\dots u_\delta(k-1), u_\delta(k-2)]^T \\ \boldsymbol{\theta}_\delta(k) = &[a'_1, a'_0, b'_1, b'_0]^T \end{aligned} \tag{24}$$

where input and output variables are recomputed to the sampling period T_v

$$\begin{aligned} y_\delta(k) &= \frac{y(k) - 2y(k-1) + y(k-2)}{T_v^2} \\ y_\delta(k-1) &= \frac{y(k-1) - y(k-2)}{T_v} \\ y_\delta(k-1) &= y(k-2) \\ u_\delta(k-1) &= \frac{u(k-1) - u(k-2)}{T_v} \\ u_\delta(k-1) &= u(k-2) \end{aligned} \tag{25}$$

The task of the on-line identification is to estimate the vector of parameters $\boldsymbol{\theta}_\delta(k)$ from the ARX model (23). The Recursive Least-Squares (RLS) method (Mikles and Fikar 2007) can be used for this job. Advantage can be found in easy programmability of this method where the accuracy could be increased with the use of some kind of forgetting - for example exponential, directional etc.

SIMULATION RESULTS

Several simulation experiments were done on the mathematical model (2). The goal of these experiments was to compare 1DOF and 2DOF control configuration on this concrete type of systems.
As results must be comparable, the initial conditions and simulation parameters must be the same. The time of the simulation is 18 000 s (5 hours) and 4 different changes of the reference signal were done during this time. The sampling period was $T_v = 1$ s and initial vector of parameters for all simulations was $\boldsymbol{\theta}_\delta = [0.1, 0.1, 0.1, 0.1]^T$. The proposed adaptive controller has one tuning parameter - the value of parameter α and we will observe results for more values of α.

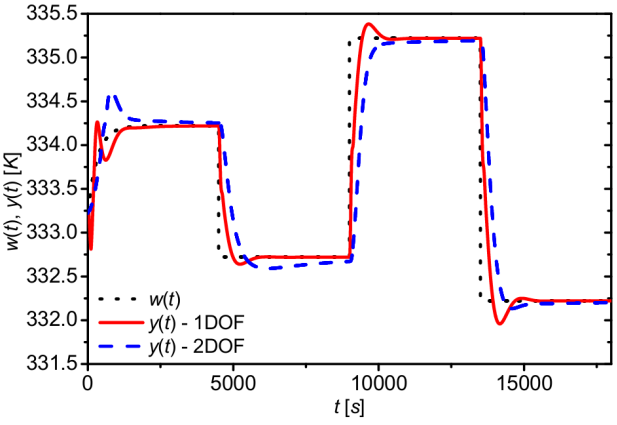


Fig. 7. The course of the reference signal, $w(t)$, and the output variable, $y(t)$, for various control configurations and $\alpha = 0.015$

The first simulation analysis for $\alpha = 0.15$ shows similar courses of the output variable $y(t)$ – see Fig. 7.

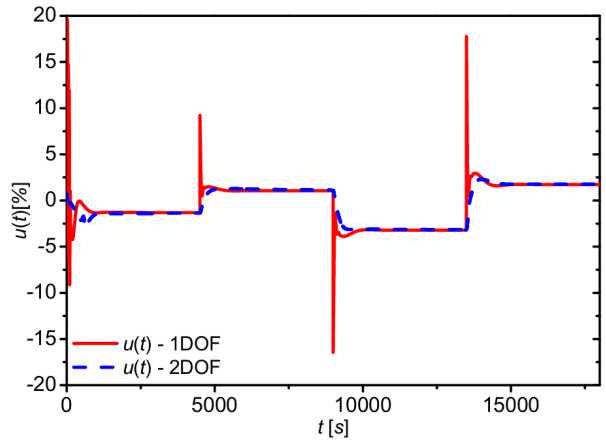


Fig. 8. The course of the input variable, $u(t)$, for various control configurations and $\alpha = 0.015$

2DOF controller has smoother courses of $y(t)$ after initial phase which can be seen also in Fig. 8 that represents courses of the input variable $u(t)$. This gentle course of the input variable could be also important from the practical point of view. As the input variable is the change of the flow rate of the coolant, the real mean of this change is the twist of the valve on the input pipe. It is clear, that the quick shocking changes of the input variable could affect the service time of the valve.

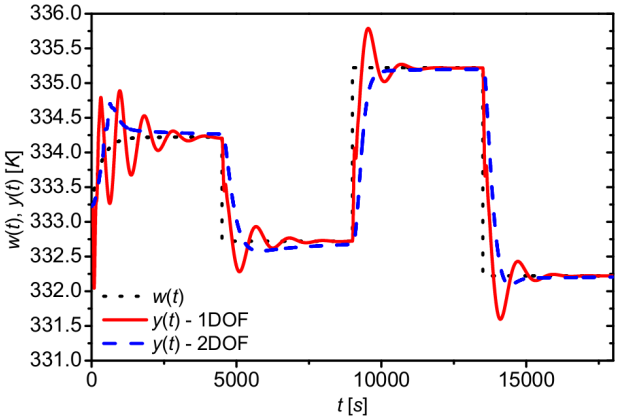


Fig. 9. The course of the reference signal, $w(t)$, and the output variable, $y(t)$, for various control configurations and $\alpha = 0.02$

Results of the second analysis for $\alpha = 0.02$ show that 1DOF controller could have problems with the control for some values of α and produce oscillating output. On the other hand, 2DOF controller has smooth course without oscillations for this setting.
In order to qualify results not only from the visual side, we can introduce the quality criteria S_u and S_y

$$\begin{aligned} S_u &= \sum_{i=2}^{N} \left(u(i) - u(i-1)\right)^2 \ [-] \\ S_y &= \sum_{i=1}^{N} \left(w(i) - y(i)\right)^2 = \sum_{i=1}^{N} e(i)^2 \ [K^2] \end{aligned} \tag{26}$$

where $N = T_f/T_v$ and $T_f = 18\,000\ s$ as a final time.

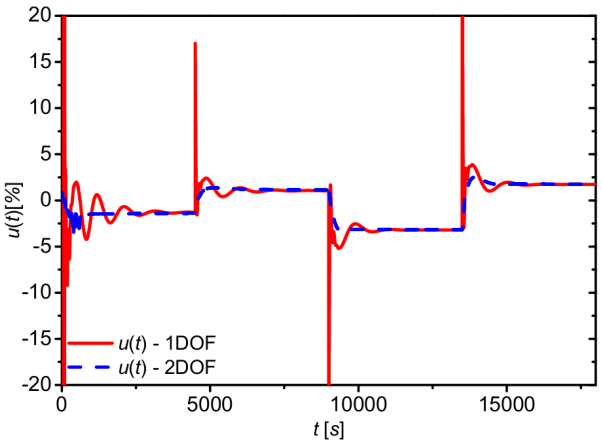


Fig. 10. The course of the input variable, $u(t)$, for various control configurations and $\alpha = 0.012$

Values of criteria S_u and S_y for previous studies are shown in Table II. Obtained results confirm our previous statements with better results of the 2DOF controller mainly from the input point of view. On the other hand 2DOF controller produces a bit worse values of the criterion S_y that is in fact sum of the control error $e = w - y$.

	1DOF		2DOF	
	S_u [-]	S_y [K^2]	S_u [-]	S_y [K^2]
$\alpha = 0.015$	1 870	1 585	0.71	4 653
$\alpha = 0.02$	30 063	2 134	2.82	4 201

TABLE II: Result of control quality criteria S_u and S_y

It was mentioned, that the adaptive approach here is based on the recursive identification and the Recursive Least-Squares method was used here for this task. The last graph in Fig. 11 shows example course of the identified parameters a'_0 and a'_1 for 2DOF controller and value of the parameter $\alpha = 0.02$.

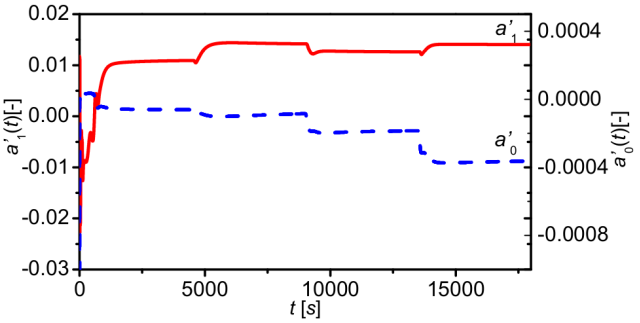


Fig. 11. The course of identified parameters $a'_1(t)$ and $a'_0(t)$ for 2DOF control configuration and $\alpha = 0.02$

We can see, that used RLS on-line identification with exponential forgetting does not have problems with the identification, parameters are recomputed after each change of the reference signal relatively quickly.

CONCLUSIONS

The paper deals with the hybrid adaptive control of the tubular chemical reactor as a typical member of the group of systems with continuously distributed parameters. The mathematical model of such a system is described by the set of partial differential equations. In this case, the mathematical model of the tubular chemical reactor with plug-flow of the reactant and the coolant consists of the set of five nonlinear partial differential equations. The steady-state and dynamic analyses use the Finite differences method for transformation of this set to the set of nonlinear ordinary differential equations that can be numerically solved for example by the Runge-Kutta's methods in Matlab.

Proposed control loop with 1DOF and 2DOF configurations were tested for the control of the mean reactant temperature inside. Both configurations produce good control results and they can be used for this task. An adaptation was satisfied by the online recursive identification of the external linear model of the controlled system. This external model has been chosen according to the results obtained from the dynamic analysis. The course of the output response could be influenced by the tuning parameter α whose negative value represents the choice of the position of the root from the Pole-placement method. 2DOF controller provides better control results than 1DOF controller mainly from the input variable point of view - the course of this computed variable is much smoother than for 1DOF which is important from the practical point of view.

REFERENCES

Astrom, K.J.; Wittenmark, B. 1989. *"Adaptive Control"*. Addison Wesley. Reading. MA, 1989, ISBN 0-201-09720-6.

Bobal, V.; Bhm, J.; Fessl, J.; Machacek, J. 2005 *"Digital Self-tuning Controllers: Algorithms. Implementation and Applications. Advanced Textbooks in Control and Signal Processing."* Springer-Verlag London Limited. 2005, ISBN 1-85233-980-2.

Dostal, P.; Prokop, R.; Prokopova, Z.; Fikar, M. 1996 "Control design analysis of tubular chemical reactors". *Chemical Papers*, 50, 195-198.

Evans, L. C. 2010 *"Partial differential equations"*. 2nd ed. Providence, R.I.: American Mathematical Society. Graduate studies in mathematics, v. 19. ISBN 978-0821849743.

Grimble, M. J. 2006 *"Robust industrial control systems: optimal design approach for polynomial systems"*. Hoboken, NJ: Wiley. ISBN 978-0-470-02073-9.

Grossmann, Ch.; Roos, H.-G.; Stynes, M. 2007 *"Numerical Treatment of Partial Differential Equations"*. Springer Science & Business Media. p. 23. ISBN 978-3-540-71584-9.

Ingham, J.; Dunn, I. J.; Heinzle, E.; Prenosil, J. E. 2000 *"Chemical Engineering Dynamics. An Introduction to Modeling and Computer Simulation."* Second, Completely Revised Edition. VCH Verlagsgesellshaft. Weinheim, 2000. ISBN 3-527-29776-6.

Johnston, R. L. 1982 *"Numerical Methods"*. John Wiley & Sons. 1982.

Kucera, V. 1993. Diophantine equations in control A survey. *Automatica*. 29, 1993, p. 1361-1375.

Mathews, J. H., Fink, K. K. 2004 *"Numerical Methods Using Matlab"*. Prentice-Hall 2004. ISBN 978-0130652485

Mikles, J.; Fikar, M. 2007 *"Process modelling, identification, and control"*. Berlin: Springer, ISBN 978-3-540-71970-0.

Mukhopadhyay, S.; Patra, A. G.; Rao, G. P. 1992 New class of discrete-time models for continuos-time systems. *International Journal of Control*, vol.55, 1992, 1161-1187.

Riley, K. F.; Bence, S. J. 2010 *"Mathematical Methods for Physics and Engineering"*. Cambridge University Press. ISBN 978-0-521-86153-3.

Russell, T.; Denn, M. M. 1972 *Introduction to chemical engineering analysis*. New York: Wiley, 1972, xviii, 502 p. ISBN 04-717-4545-6.

Stericker, D. L.; Sinha, N. K. 1993 Identification of continuous-time systems from samples of input-output data using the δ-operator. *Control-Theory and Advanced Technology*. vol. 9, 1993, 113-125.

Vojtesek, J; Dostal, P. 2009 "Adaptive Control of the Tubular Reactor with Co- and Counter-current Cooling in the Jacket." In *23rd European Conference on Modelling and Simulation*. Madrid, p. 544-550. ISBN 978-0-9553018-8-9.

JIRI VOJTESEK was born in Zlin, Czech. He studied at Tomas Bata University in Zlin, Czech Republic, where he received his M.Sc. degree in Automation and control in 2002. In 2007 he obtained Ph.D. degree in Technical cybernetics at Tomas Bata University in Zlin. In the year 2015 he became associate professor. His research interests are modeling and simulation of continuous-time chemical processes, polynomial methods, optimal, adaptive and nonlinear control. You can contact him on e-mail address vojtesek@utb.cz.

LUBOS SPACEK studied at the Tomas Bata University in Zln, Czech Republic, where he obtained his masters degree in Automatic Control and Informatics in 2016. He currently attends PhD study at the Department of Process Control. His e-mail address is lspacek@utb.cz.

FRANTISEK GAZDOS was born in Zln, Czech Republic in 1976, and graduated from the Brno University of Technology in 1999 with MSc. degree in Automation. He then followed studies of Technical Cybernetics at Tomas Bata University in Zlin, obtaining Ph.D. degree in 2004. He became Associate Professor for Machine and Process Control in 2012 and now works as the Head of the Department of Process Control, Faculty of Applied Informatics of Tomas Bata University in Zlin. His research covers the area of process modelling, simulation and control. His e-mail address is: gazdos@fai.utb.cz.

A MATLAB-BASED SIMULATION TOOL FOR THE ANALYSIS OF UNSYMMETRICAL POWER SYSTEM TRANSIENTS IN LARGE NETWORKS

Michael Kyesswa
Hüseyin K. Çakmak
Uwe Kühnapfel
Veit Hagenmeyer
Institute for Automation and Applied Informatics (IAI)
Karlsruhe Institute of Technology (KIT)
P.O. Box 3640, 76021 Karlsruhe, Germany
E-mail: {michael.kyesswa, hueseyin.cakmak, uwe.kuehnapfel, veit.hagenmeyer}@kit.edu

KEYWORDS
Modeling and simulation, power system dynamics, symmetrical components, unsymmetrical transients.

ABSTRACT

This paper presents an extendable Matlab-based phasor-time domain toolbox for modeling, simulation and analysis of unsymmetrical power system transients in large networks. Unlike most of the existing transient stability simulators which represent the transmission network on a per phase positive sequence basis, the new simulation function introduced in this paper is based on the symmetrical component technique which employs the three sequence networks. This representation allows consideration of network imbalances in order to include a wide range of disturbances during transient stability studies. The main aim of this paper is to describe the model details of the power system components required for unsymmetrical transients analysis and the solution methodology in the introduced simulation function. The performance of the simulation function is tested using standard IEEE test network models and the promising results are positively compared to respective results in DIgSILENT PowerFactory in terms of accuracy.

INTRODUCTION

Transient stability programs used for studying large complex networks usually assume balanced three-phase operating conditions (Kundur 1994) and any discrepancies that exist between the three phases are considered to be small in magnitude. This simplifies the power system model complexity by considering only the positive sequence network. Thereby, in the general large power system, the errors caused by neglecting the differences in the magnitudes of the voltage and the phase difference between the three phases are also considered to be small. However, there are many cases where the system imbalance cannot be ignored, especially due to unbalanced loads and unsymmetrical faults which form the majority of fault types in a real power system. This has become more pronounced in the current power system analysis problem with an increase in dimensions and complexity due to the growth in electricity demand, integration of renewable energy sources as well as expansion in power grids in form of large interconnected networks (IEEE/CIGRE Joint Task Force 2004). Therefore, there is a need to reconsider the methods used in system analysis to account for the changes in complexity of large scale systems and transients due to unbalanced network operation.

Unbalanced conditions can be accurately modeled and analyzed using detailed electromagnetic transient simulations packages (Dommel 1986). These tools are however limited to very small network sections due to the high computational burden involved in analyzing dynamics in large scale power systems. To make the computation as time-efficient as possible, simplifying assumptions in component modeling are considered in validated commercial software packages extensively used to study system wide transient behavior in large networks (Kaberere, et al. 2004). Such tools have proved to be computationally efficient and reasonably user-friendly but have a closed architecture. This implies that it is not possible for users to access the source code in order to modify system models and extend the functionality of the tools and therefore cannot be effectively explored in research to consider the changing network requirements.

In order to address the need for continuous development and improvement of power system analysis methods, several research and educational grade simulation tools have been developed with the advantage of providing easy access to source code and supporting modeling flexibility. A comparison of different open source software packages with their different levels of complexity is reported in (Milano and Vanfretti 2009). Among the listed simulation tools, the available open source simulation tools with a wide range of analysis features are mainly limited to power flow and transient analysis of balanced networks. In the present paper, analysis of unbalanced network transients using the symmetrical component technique is presented. Unlike in the traditional balanced methods where only the positive sequence network is considered, this approach takes the three sequence networks into consideration.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Earlier research efforts to extend power system analysis to unbalanced transients focused on modeling transformers and synchronous machines (Chen et al. 1991), (Halpin et al. 1993), (Tamura et al. 1997) as individual components using symmetrical component or individual phase modeling techniques. To include analysis of the entire power system, a modeling framework using dynamic phasor technique was introduced in (Stankovic´ and Aydin 2000) for analyzing unbalanced faults in poly-phasor systems. Further research in the field focused on developing models and methods for unbalanced system studies addressing the following issues in general power system:

- Three phase power flow calculations (Abdel-Akher et al. 2005), (Kamh and Iravani 2010), (Demirok et al. 2012),
- Assessment of small signal stability (Salim and Ramos 2012), and
- Analysis of dynamic transients in transmission networks (Saha and Aldeen, 2015) or distribution level systems (Elizondo et al. 2016).

The research highlighted above mainly focused on component modeling for three-phase unbalanced dynamic analysis and validating the developed models using commercial simulation tools. The main contribution of the present paper is the development of a Matlab algorithm for the analysis of unsymmetrical power system transients based on MatDyn simulation toolbox (Cole and Belmans 2011). The component models applied in the introduced algorithm are derived using the symmetrical component technique as presented in the various literature (Chen et al. 1991) - (Elizondo et al. 2016). In the present paper, the analyzed transients are limited to unbalanced network faults.

SYMMETRICAL COMPONENTS OVERVIEW

The system analysis introduced in this paper is based on quasi-stationary symmetrical components method. In this method, the three-phase time varying phasors are transformed into the positive-, negative- and zero-sequence components using the symmetrical components transformation matrix (Ts)

$$V^{abc} = T_s V^{012} \tag{1}$$

where $T_s = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix}$ and $a = 1\angle 120^o$

$V^{abc} = \left[V^a\ V^b\ V^c\right]^T$ is a vector of a, b, and c phase voltages, and $V^{012} = \left[V^0\ V^1\ V^2\right]^T$ is a vector of zero, positive and negative sequence voltages (Saadat 2010). A similar relation exists for currents.

The simplifying assumptions considered during component modeling include:

- Changes in network voltages and currents are instantaneous and therefore the transmission system can be represented using the lumped line model.
- System voltages and currents are represented using fundamental phasor quantities.
- The network frequency is considered to remain nearly constant (Jalili-Marandi et al. 2009).

The full power system dynamic model is generically represented by a set of differential algebraic equations

$$\dot{x} = f(x,y,u) \tag{2}$$

$$0 = g(x,y,u) \tag{3}$$

where x are dynamic state variables, y are algebraic variables, and u are system parameters. The differential equation is a set of uncoupled subsets representing all machines in the system and their controls coupled to each other through the network. The algebraic equation comprises the stator equations of each machine coupled to the equations of the network and loads.

The equations of the generator and connected controllers are developed in the rotor d/q-reference frame whereas the network is modeled using the common system reference frame. Fig. 1 shows the relationship between the d/q and sequence network coordinates. The d/q-axis is fixed on the field magnetic axis of each machine. The D1/Q1 axis is a network reference frame rotating synchronously in the positive sequence direction while D2/Q2 axis is rotating synchronously in the negative sequence direction. The angle between the D1-axis and the stationary frame is equivalent to ωt, where ω is the synchronous angular velocity. The displacement angle θ_m represents the angular position of the rotor and δ is the machine torque angle.

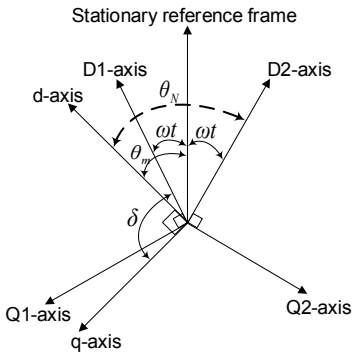


Figure 1: Reference frame transformation

The transformation between machine variables (v_d, v_q) and the respective sequence network variables (U_D, U_Q) as derived from Fig. 1 are

$$\begin{bmatrix} v_{d1} \\ v_{q1} \end{bmatrix} = \begin{bmatrix} \sin\delta & -\cos\delta \\ \cos\delta & \sin\delta \end{bmatrix} \begin{bmatrix} U_{D1} \\ U_{Q1} \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} v_{d2} \\ v_{q2} \end{bmatrix} = \begin{bmatrix} \cos\theta_N & -\sin\theta_N \\ -\sin\theta_N & -\cos\theta_N \end{bmatrix} \begin{bmatrix} U_{D2} \\ U_{Q2} \end{bmatrix} \tag{5}$$

where $\theta_N = 2\omega t + \delta - \pi/2$ is the displacement angle between the D2-axis and the d-axis, and ω is the synchronous rotational speed in the negative sequence

direction (Tamura et al. 1997). The relations in (4) and (5) hold for both currents and voltages. It has been shown in (Saha and Aldeen 2015) that the zero sequence stator quantities vanish in the d/q-coordinates.

POWER SYSTEM COMPONENT MODELS

Synchronous Machine Model

The synchronous machine equations comprise rotor electrical and mechanical, excitation system, and turbine-governor equations. The synchronous generator is represented by the fourth order model (Machowski et al. 2008), which is a simplified model generally assumed to accurately represent synchronous generators for studying the electromechanical dynamic behavior. The effects of the rotor damper windings are neglected in this model and results into a generator represented by transient emfs ($E_d^{'}$ and $E_q^{'}$) behind transient reactances ($X_d^{'}$ and $X_q^{'}$). The generator stator transients are also neglected and the stator becomes represented as a simple impedance with reactance components in the d- and q-axes. The stator equations are therefore algebraic equations.
The machine dynamics are divided into the respective sequence components which are isolated from each other. The positive sequence circuit is represented using the traditional synchronous generator models since the models used in balanced transient stability studies are developed considering only the positive sequence operation of the generator. The voltages and currents in the negative and zero sequence circuits are a result of unbalanced operations and therefore no power generation occurs in these circuits. A braking torque is included in the mechanical equation of the generator to account for sizeable negative sequence currents. The zero sequence currents do not produce an effective torque in the machine and are not included in the mechanical equation.

Positive Sequence Circuit
The differential and algebraic equations of the positive sequence circuit are similar to the traditional transient analysis studies. The difference is that the voltage and currents only refer to the positive sequence of the machine, and not the total machine quantities as in the balanced mode simulation. The differential equations describing the change in flux are given as

$$T_{do}^{'}\dot{E}_q^{'} = E_f - E_q^{'} + (X_d - X_d^{'})I_d \tag{6}$$

$$T_{qo}^{'}\dot{E}_d^{'} = - E_d^{'} - (X_q - X_q^{'})I_q \tag{7}$$

and the resulting algebraic equations of the positive sequence circuit that show the relation between the voltages and current components are given by

$$\begin{bmatrix} V_d \\ V_q \end{bmatrix} = \begin{bmatrix} E_d^{'} \\ E_q^{'} \end{bmatrix} - \begin{bmatrix} R & X_q^{'} \\ -X_d^{'} & R \end{bmatrix} \begin{bmatrix} I_d \\ I_q \end{bmatrix} \tag{8}$$

The parameters and variables (6) – (8) are described in (Machowski et al. 2008). For given initial conditions, the differential equations describing the change in flux can be solved, from which the stator and rotor positive sequence currents can be obtained using (8). The positive sequence generator current in network reference frame is obtained from (4).

Negative Sequence Circuit
The negative sequence circuit has no current source and is therefore represented by a pure impedance connected between the machine bus and ground. The resulting algebraic equation of the negative sequence stator voltage-current relationship is

$$\begin{aligned} 0 &= V_2 + Z_2 I_2 \\ Z_2 &= R_s + jX_2 \end{aligned} \tag{9}$$

where I_2 is the complex negative sequence current and V_2 is the complex voltage at the generator bus. Impedance Z_2 is the negative sequence impedance, and R_s, X_2 denote the stator resistance and negative sequence reactance. The value of X_2 can be approximated by $(X_d^{''} + X_q^{''})/2$, where $X_d^{''}$, $X_q^{''}$ are the direct axis and quadrature axis subtransient reactances. However, If transient saliency is neglected, then $X_2 = X_d^{''} = X_q^{''}$. The transformation matrix between the negative sequence components in network reference (U_{D2}, U_{Q2}) and the rotor d-q reference coordinates (v_{d2}, v_{q2}) is given in (5).

Zero Sequence Circuit
The zero sequence circuit is also represented by a pure impedance connected between the machine bus and ground. The resulting algebraic equation of the zero sequence stator voltage-current relationship is

$$\begin{aligned} 0 &= V_0 + Z_0 I_0 \\ Z_0 &= R_s + jX_0 \end{aligned} \tag{10}$$

where Z_0 is the zero sequence impedance, and R_s, X_0 denote the stator resistance and zero sequence reactance. I_0 is the complex zero sequence current and V_0 is the complex zero voltage at the generator bus. However, no zero sequence stator component exist in the d-q coordinate. These dynamics are expressed in the real and imaginary coordinates as

$$\begin{aligned} v_0(t) &= V_0(t)\cos(\omega_s t + \theta_0) \\ v_0(t) &= V_R(t)\cos(\omega_s t) - V_I(t)\sin(\omega_s t) \end{aligned} \tag{11}$$

where V_0 is the magnitude and θ_0 is the phase angle of the zero sequence quantity. $V_R(t) = V_0 \cos\theta_0$ and $V_I(t) = V_0 \sin\theta_0$ are the real and imaginary components of the phasor representing the zero sequence stator quantity, respectively (Saha and Aldeen 2015).

Mechanical Equations
The mechanical equation differs from the traditional transient stability equation in that it is modified to include the effect of unbalanced system operation. The positive sequence current represents the main electrical torque

and the negative sequence current produces an opposing torque referred to as the braking torque. The resulting mechanical part of the dynamic equations is given by

$$\dot{\delta} = \omega_b(\omega - \omega_0) \quad (12)$$

$$\dot{\omega} = \frac{1}{2H}\left[P_m - D(\omega - \omega_0) - P_e - (R_2 - R_s)I_2^2\right] \quad (13)$$

where R_2 is the negative sequence resistance and R_s is the stator resistance (Elizondo et al. 2016). The air gap power of the generator is only due to the positive sequence component since there are no voltage sources in the negative and zero sequence circuits. The power is given by

$$P_e = \left(E_d'I_d + E_q'I_q\right) + \left(X_q' - X_q'\right)I_dI_q \quad (14)$$

The changes in mechanical power P_m and emf E_f are computed as described in the turbine-governor and excitation system model definitions. The parameters and variables in (12) – (14) are described in (Machowski et al. 2008).

Synchronous Machine Controllers

Excitation System

The excitation system is used to control the terminal voltage by modifying the generator field voltage. The dynamic behavior of the excitation system implemented in the presented tool is according to the model of the type DC1A excitation system described in (IEEE Std 421.5-2005 2006).

Turbine-Governor

The turbine-governor model is developed using a proportional-integral (PI) controller. The dynamics of the system are given by (15) - (19).

$$y_1 = P_{ref} - K_R(\omega - \omega_{ref}) - P_m \quad (15)$$

$$\dot{y}_2 = \frac{1}{T_p}y_1 \quad (16)$$

$$y_3 = K_p y_1 + y_2 \quad (17)$$

$$\dot{y} = \frac{1}{T_m}\left(y_3 - y\right) \quad (18)$$

$$\dot{P}_m = \frac{1}{T_k}(y - P_m) \quad (19)$$

where y is the valve position, P_{ref} is reference power, P_m is mechanical power, ω_{ref} is reference speed, ω rotor speed, K_R=1/R, R is the droop constant, K_p and T_p are the gain and time constant of the PI controller. The variables y_1 and y_3 are the integrator input and output signals, respectively, y_2 is the servo motor output signal, T_m and T_k are the servo motor and turbine time constants.

Network Model

The modeling of the transmission network is based on Matpower (Zimmerman et al. 2011) representation as applied in steady state power flow solutions. This is an acceptable simplification due to the assumption that during transient events, changes in network voltages and currents are instantaneous as compared to the machine dynamics. However, unlike in the balanced transient stability studies, the symmetrical components technique is used to model the network in the present paper. Positive, negative, and zero sequence networks are constructed using equivalent admittances and the resulting nodal network equation is represented in complex current balance form as

$$I_{012}^{inj} = Y_{012}^{bus}V_{012}^{bus} \quad (20)$$

where the matrix Y_{012}^{bus} represents the nodal admittance matrix for the corresponding sequence network, and vectors V and I contain the sequence voltage and current phasors at fundamental frequency.

The nodal admittance matrices for the corresponding sequence networks are modified by inserting the generator and load admittances as diagonal shunt admittances.

Generator Admittance

The equivalent admittances of the generator model (Y_{g1}, Y_{g2} and Y_{g0}) are given by

$$\begin{aligned} Y_{g1} &= 1/Z_1 \\ Y_{g2} &= 1/Z_2 \\ Y_{g0} &= 1/Z_0 \end{aligned} \quad (21)$$

Load Admittance

Unlike in power flow analysis where loads are treated as constant power, static loads are treated as constant admittances in dynamic studies. This simplifies the analysis by inserting the load admittances directly into the admittance matrix and ignoring the load current injections. The elements of the load admittance diagonal matrix Y_{abc}^l at a bus are given in terms of the initial load power $P_{abc}^l + jQ_{abc}^l$ and the load bus voltage V_{abc}^i

$$Y_{abc}^l = \frac{P_{abc}^l - jQ_{abc}^l}{\left|V_{abc}^i\right|^2} \quad (22)$$

$$Y_{012}^l = [T_s]^{-1}\left[Y_{abc}^l\right][T_s] \quad (23)$$

In (23), the admittance Y_{012}^l in terms of sequence components is calculated using the symmetrical components transformation matrix T_s (Saadat 2010). In the simulation tool, the load implementation assumes balanced load models on the three phases.

The network equations (20) – (23) and the equations of the dynamic components (6) – (19) complete the model of unbalanced power systems for a phasor-time domain simulation. The system components are modelled separately and connected together using well-known analysis techniques depending on the fault type (Saadat 2010).

SOLUTION METHODOLOGY

Solution Methods

The system of differential and algebraic equations (6) – (23) form the mathematical description of the full power system model. The solution of this system starts from steady state values obtained from a power flow calculation. In this simulation, Matpower is used for the power flow computation to obtain the system voltages, angles, active and reactive power generation at time t = 0. These values are used to compute the initial dynamic state variables x in order to progress the system solution from a steady state operation.

Numerical methods are applied to solve the system of differential algebraic equations for the dynamic state variables x and algebraic variables y at each time step. In the presented simulation tool, the partitioned approach is used to interface the differential equations in (2) and algebraic equations in (3). This implies that the two equations are solved alternately (Soman et al. 2002). Explicit integration methods are used to numerically integrate the differential equations.

The solution of the algebraic equation is simplified by the following assumptions: 1) The system loads are represented as constant admittance loads. This eliminates the nonlinear behavior of the overall network equations which is caused by current injections at load buses if the loads are represented by constant power; 2) Transient saliency of the generator is neglected whereby $X_d^{'} = X_q^{'}$ and this simplifies the stator algebraic equation in (8). This results into linear network algebraic equations which are solved using the sparsity-oriented triangular factorization direct linear solver (Crow 2009).

Transient Analysis Procedure

A power flow calculation is initially carried out to determine the internal operating states of the connected generators and their respective controllers. The system is assumed to start from a balanced state and therefore the Matpower power flow calculation based on a single phase positive sequence network is applied. As a result, system bus voltages are calculated in terms of magnitude and angle as well as the active and reactive power generation at the buses. The initial generator current is calculated using

$$i_{gen,i} = \frac{P_{gen,i} - jQ_{gen,i}}{V_i^*} \tag{24}$$

where $P_{gen,i} + jQ_{gen,i}$ is the generated power obtained from the power flow solution and V_i is the i^{th} generator terminal voltage. It is important to note that this current only represents the positive sequence value. The negative and zero sequence current values are set to zero due to the physical symmetry of the generator (Chen et al. 1991).

The initial steady state internal emf and rotor angle for each generator in the system is calculated using;

$$\begin{aligned} E_{0,i} &= V_i + (R_s + jx_{q,i}^{'})i_{gen,i} \\ \delta_{0,i} &= \text{angle}\,(E_{0,i}) \end{aligned} \tag{25}$$

and the initial transient emfs $E_{q0}^{'}$, $E_{d0}^{'}$ are calculated using (6) and (7), respectively.

The original node admittance matrix is augmented by adding the generator internal admittance and the equivalent load admittance to the diagonal components. This is carried out at every occurrence of a system transient to re-compute the admittance matrix. For the pre-transient and post-transient condition, only the positive sequence network is considered corresponding to balanced system operation. During the unbalanced transient period, the system components are connected together using the well-known analysis techniques depending on the fault type.

System stability state is analyzed from the plots of the system variables. The important variables analyzed include generator power or torque, speed, angle, and bus voltages for a period of up to several seconds after the transient is cleared. The system is unstable if the responses diverge from one another or exhibit growing oscillations.

SIMULATION AND ANALYSIS

The main aim of the introduced simulation function is to analyze different kinds of unsymmetrical network transients. The simulated transients include single line-to-ground fault, line-to-line fault, and double line-to-ground fault. Other transients such as three phase-to-ground fault and load variations can also be simulated. In the present paper, simulation results for a single line-to-ground fault and a three-phase fault are presented. The results are compared to those obtained using the commercial software package DIgSILENT PowerFactory (DIgSILENT 2016) to assess the level of accuracy in capturing transient responses, as well as validate the component models used in the simulation tool.

Simulation Case1: 9-Bus Network

The standard IEEE-9 bus test feeder is considered for the simulation results presented in this section. The network consists of three generators, nine buses, and three loads. Excitation and turbine-governor systems are connected to the synchronous machines at bus 1, 2 and 3. Similar components are selected for the example network in DIgSILENT PowerFactory which are; a synchronous generator model with IEEET1 exciter, and TGOV1 type governor. The model parameters are modified to match the parameters in the Matlab-based simulation function. The network structure and parameters of the test feeder are given in (Anderson and Fouad 2003).

A single line-to-ground fault on phase-A with zero fault impedance is applied on bus 6 at 5s and cleared at 15s. The response of the phase voltages, sequence voltages, generator rotational speed and mechanical power at the generator buses (bus 1, 2 and 3) and the faulty bus are

shown in Fig. 2 – Fig. 9 for the simulated fault scenario in the new simulation function and DIgSILENT PowerFactory.

Analysis Of Bus Voltages

The responses of the phase and sequence voltages at the buses are shown in Fig. 2 – Fig. 7 for the simulations in the introduced function and PowerFactory. It can be observed that the introduced simulation function captures the transient behavior of the bus voltages following the fault condition. Fig. 2 – Fig. 4 show the responses of the bus phase voltages with voltage transients recorded at the switching instant and during the fault period. Phase-A voltage experiences the largest sag, as expected, compared to Phase-B and C. The voltage sag at the faulty bus 6 is up to zero, reflecting the zero fault impedance applied in this simulation scenario. Voltage transients are also captured in sequence voltages (Fig. 5 – Fig. 7). It is observed that the voltages in the negative and zero sequence networks are a result of the network imbalance due to the applied unsymmetrical fault. The positive sequence bus voltages experience a sag which is largest at the faulty bus.

The transient response of the bus voltage is due to the action of the excitation systems and can be explained as follows: The voltages are initially in steady state until the fault occurs resulting into a change of voltage levels at the buses. The change in voltage at the generator terminals initiates the operation of the excitation systems connected to the generation units which adjust the internal field voltages of the respective machines and attain new steady state operating voltage levels at the respective terminals. When the fault is cleared, the initial network structure is re-established and the action of the excitation system recovers the initial operating voltage levels as seen from the captured responses. Similar system response can be observed in the PowerFactory simulation results. However, there is a slight difference in voltage levels reached during the transient period.

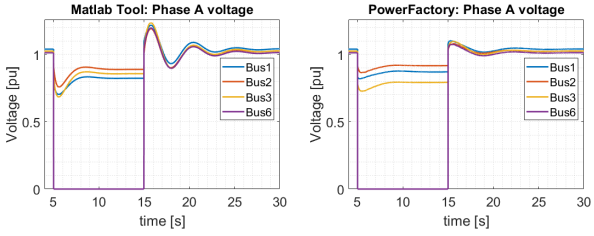


Figure 2: Phase-A voltage response

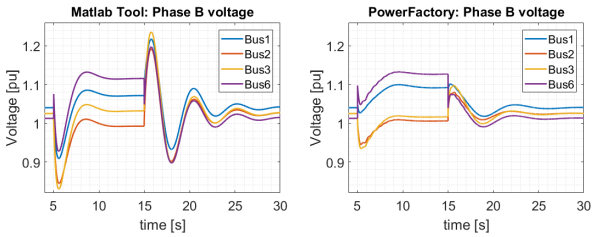


Figure 3: Phase-B voltage response

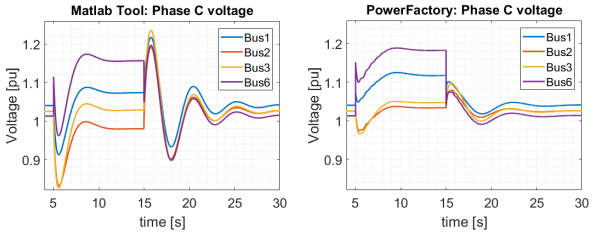


Figure 4: Phase-C voltage response

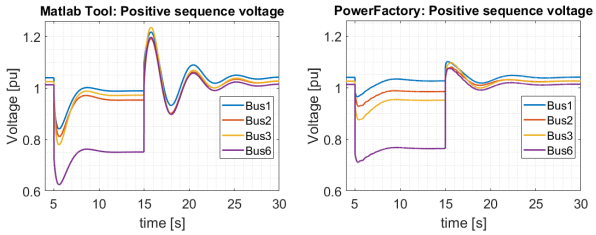


Figure 5: Positive-sequence voltage response

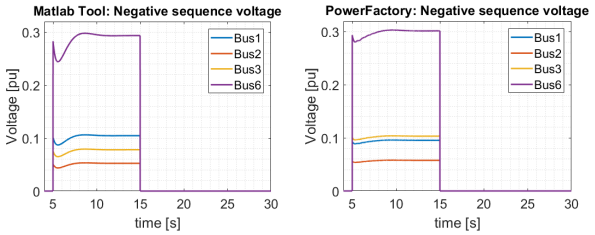


Figure 6: Negative-sequence voltage response

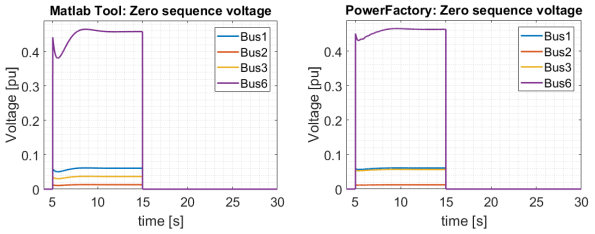


Figure 7: Zero-sequence voltage response

Analysis of Rotational Speed and Mechanical Power

The responses of the generator rotational speed and turbine mechanical power are shown in Fig. 8 and Fig. 9, respectively. It can be observed that the change in network structure due to the fault occurrence causes a change in accelerating power of the generation units. In this simulation case, there is a resulting acceleration of the units which initiates the action of the turbine–governor system of each unit to adjust the mechanical power contribution as shown in Fig. 9 during the fault period. The action of the governor controllers reduces the speed deviation until synchronous operation of the interconnected machines is achieved at a new operating point as shown in Fig. 8 during the fault period.

Clearing the fault results into restoration of the network structure to the initial state. In this case, the generation units experience a decrease in rotational speed as can be observed in Fig. 8. The speed deviation triggers further action of the governor systems to increase the mechanical

power as shown in Fig. 9 during the post-fault period until the speeds of the interconnected machines synchronize. The final operating point is observed to be similar to the initial operating point, as expected, since the initial structure is re-established. The results of the rotational speed and turbine-power responses of the introduced simulation function are seen to closely match the responses in DIgSILENT PowerFactory.

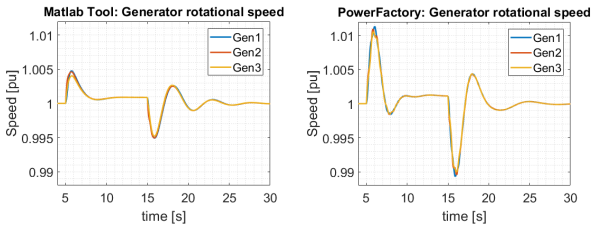


Figure 8: Generator rotational speed response

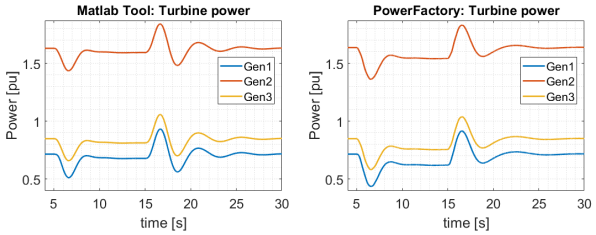


Figure 9: Mechanical power response

Simulation Case II: 9241-Bus Network

The simulation function is tested in a larger network representing the size and complexity of the European high voltage network. The test feeder (Case9241pegase) consists of 9,241 buses, 1,445 generators, and 16,049 branches operating at 750, 400, 380, 330, 220, 154, 150, 120, and 110 kV (Josz, et al. 2016).

In this simulation case, a three-phase fault is applied on bus 28 at time t = 3.0s and on bus 143 at time t = 5.0, each for a duration of 100ms. The results of bus voltage and generator rotational speed response obtained using the Matlab tool are shown in Fig. 10 and Fig. 11. The bus voltage response is presented for a single phase.

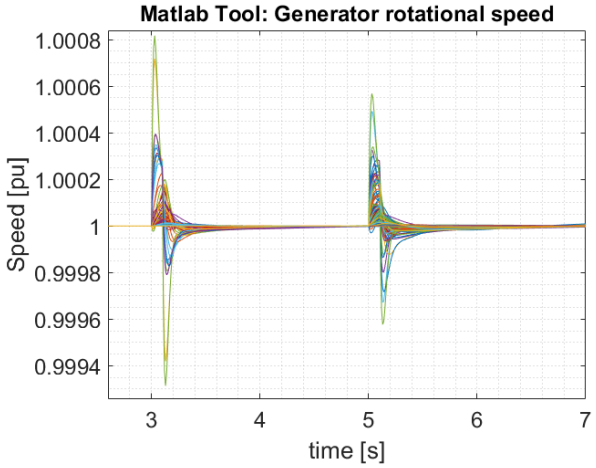


Figure 10: Generator rotational speed response

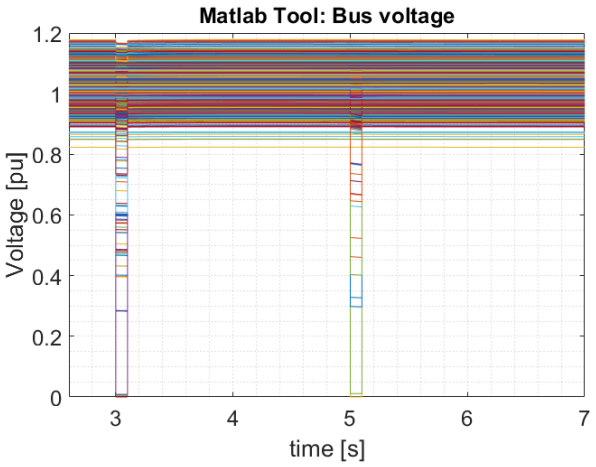


Figure 11: Bus voltage response

The results obtained from the simulation of the 9241-bus network reflect the expected behavior of the bus voltage and generator rotational speed in response to a network fault. This shows the ability of the tool to analyze network transients in complex networks as well.

Discussion of Results

It can be observed that the introduced Matlab-based function is able to capture the network transient responses. Qualitative analysis of the simulation results shows a close match with the DIgSILENT PowerFactory results. Similar steady state operating values are reached in both tools in the post-transient period. However, there is a noticeable difference in the response overshoot at the instant of change in the network structure, as well as in the steady state operating point attained during the transient period. Part of the ongoing work is to address the cause of the differences in simulation results and validation of the results through experimental setups.

CONCLUSION

The present paper has presented a new Matlab-based simulation function for analysis of unsymmetrical power system transients using the symmetrical component technique. The introduced approach provides the advantage of representing the power system using the three sequence networks which allows network imbalances to be taken into consideration during system analysis. The presented simulation function is an extension of MatDyn, a Matlab-based toolbox, which is only limited to analysis of balanced transients. This extension can facilitate the analysis of a wider range of transients especially due to unsymmetrical faults which form the majority of fault types in a real power system. The accuracy of the simulation algorithm and the developed models is verified against the commercial software package DIgSILENT PowerFactory for a single line-to-ground fault and the results are observed to closely match. The algorithm can be applied to analyze different types of faults in large networks. Part of the ongoing work is focused on including analysis of system imbalances due to unbalanced static or dynamic loads

and modeling renewable energy sources to analyze their effect on system stability. After including the above mentioned features and experimental validation of the simulation tool functionality, the toolbox extension will be made freely accessible to the research community.

ACKNOWLEDGMENT

This work is part of the "Energy System 2050" initiative of the Helmholtz Association.

REFERENCES

Abdel-Akher, M., Nor, K. and Rashid, A., 2005. "Improved three-phase power-flow methods using sequence components". *IEEE Trans on Power Systems,* 20(3), 1389 - 1397.

Anderson, P. M. and Fouad, A.-A. A., 2003. *Power System Control and Stability.* Piscataway, NJ: IEEE Press; Hoboken, NJ; Wiley-Interscience.

Chen, T. -H. et al., 1991. "Three-phase cogenerator and transformer models for distribution system analysis". *IEEE Transactions on Power Delivery,* 6(4), 1671 - 1681.

Cole, S. and Belmans, R., 2011. "MatDyn, A new Matlab-based toolbox for power system dynamic simulation". *IEEE Transactions on Power Systems,* 26(3), 1129 - 1136.

Crow, M. L., 2009. *Computational methods for electric power systems.* CRC Press.

Demirok, E., Kjær, S. B., Sera, D. and Teodorescu, R., 2012. "Three-phase unbalanced load flow tool for distribution networks". *Int Workshop on Integration of Solar Power*

DIgSILENT GmbH, March 2016. *DIgSILENT PowerFactory User manual Version 2016,* Gomaringen, Germany.

Dommel, H. W., 1986. *Electromagnetic Transient Program (EMTP) Theory Book.* Portland, Oregon: Bonneville Power Administration.

Elizondo, M. A., Tuffner, F. K. and Schneider, K. P., 2016. "Three-phase unbalanced transient dynamics and powerflow for modeling distribution systems with synchronous machines". *IEEE Transactions on Power Systems,* 31(1), 105 - 115.

Halpin, S., Gross, C. and Grigsby, L., 1993. "An improved method of including detailed synchronous machine representations in large power system models for fault analysis". *IEEE Transactions on Energy Conversion,* 8(4), 719 - 725.

IEEE Std 421.5-2005, 2006. *IEEE recommended practice for excitation system models for power system stability studies,* New York: IEEE.

IEEE/CIGRE Joint Task Force on Stability Terms and Definitions, 2004. "Definition and classification of power system stability". *IEEE Trans on Power Systems,* 19(2), 1387 - 1401.

Jalili-Marandi, V. et al., 2009. "Interfacing techniques for transient stability and electromagnetic transient programs". *IEEE Trans on Power Delivery,* 24(4), 2385 - 2395.

Josz, C., Fliscounakis, S., Maeght, J. and Panciatici, P., 2016. *AC Power Flow Data in MATPOWER and QCQP Format: iTesla, RTE Snapshots, and PEGASE.* [Online] Available at: http://arxiv.org/abs/1603.01533

Kaberere, K. K., Folly, K. A. and Petroianu, A. I., 2004. "Assessment of commercially available software tools for transient stability: Experience gained in an academic environment". *IEEE AFRICON,* 711 - 716.

Kamh, M. Z. and Iravani, R., 2010. "Unbalanced model and power-flow analysis of microgrids and active distribution systems". *IEEE Transactions on Power Delivery,* 25(4), 2851 - 2858.

Kundur, P., 1994. *Power System Stability and Control.* New York: McGraw-Hill.

Machowski, J.; Bialek, J. W.; and Bumby, J. R., 2008. *Power System Dynamics, Stability and control.* John Wiley and Sons, Ltd.

Milano, F. and Vanfretti, L., 2009. "State of the art and future of OSS for power systems". Calgary, AB, Canada.

Saadat, H., 2010. *Power system analysis.*PSA Publishing.

Saha, S. and Aldeen, M., 2015. "Dynamic modeling of power systems experiencing faults in transmission/distribution networks". *IEEE Transactions on Power Systems,* 30(5), 2349 - 2363.

Salim, R. H. and Ramos, R. A., 2012. "A Model-based approach for small-signal stability assessment of unbalanced power systems". *IEEE Transactions on Power Systems,* 27(4), 2006 - 2014.

Soman, S., Khaparde, S. and Pandit, S., 2002. *Computational Methods For Large Sparse Power Systems Analysis; An Object Oriented Approach.* Kluwer Academic Publishers.

Stankovic´, A. M. and Aydin, T., 2000. "Analysis of asymmetrical faults in power systems using dynamic phasors". *IEEE Transactions on Power Systems,* 15(3), 1062 - 1068.

Tamura, J., Takeda, I., Kimura, M. and Yonaga, S., 1997. "A synchronous machine model for unbalanced analyses". *Electr Eng Jpn,* 119(2), 46–59.

Zimmerman, R. D., Murillo-Sanchez, C. E. and Thomas, R. J., 2011. "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education". *IEEE Trans on Power Systems,* 26(1), 12 - 19.

AUTHOR BIOGRAPHIES

MICHAEL KYESSWA is a Ph.D student at the Institute for Automation and Applied Informatics at Karlsruhe Institute of Technology. His main areas of research are power system dynamic modeling, transient stability and analysis of the dynamic behavior of large power systems. His e-mail address is: michael.kyesswa@kit.edu.

HÜSEYIN KEMÂL ÇAKMAK is scientific staff at the Institute for Automation and Applied Informatics at Karlsruhe Institute of Technology. His research interests are modeling, simulation, visualization, 3d/VR/AR, big data analysis, web systems, parallel computing and software development for energy system analysis. His e-mail address is: hueseyin.cakmak@kit.edu.

UWE G. KÜHNAPFEL is head of the working group energy systems simulation and visualization at the Institute for Automation and Applied Informatics at Karlsruhe Institute of Technology. His main research fields are energy systems (modeling, simulation, monitor ing, analysis), mechatronics, virtual and augmented reality. His e-mail address is: uwe.kuehnapfel@kit.edu.

VEIT HAGENMEYER is director of the Institute for Automation and Applied Informatics at Karlsruhe Institute of Technology. His main field of research is automation technology, control engineering and energy informatics. His e-mail address is: veit.hagenmeyer@kit .edu.

Simulation and Optimization

PROCESS OPTIMIZATION IN "SMART" COMPANIES THROUGH CONDITION MONITORING

Frank Morelli
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
T +49 7231 28-6697,
frank.morelli
@hs-pforzheim.de

Jan-Felix Mehret
ProSeS BDE GmbH
Richard-Wagner-Allee 10c,
75179 Pforzheim
T +49 7231 147 37-64,
j.mehret
@proses.de

Thorsten Weidt
BridgingIT GmbH
Marienstr. 17,
70178 Stuttgart
T +49 151 52 66 93 98,
thorsten.weidt
@bridging-it.de

Moustafa Elazhary
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
T +49 7231 28-6526,
moustafa.elazhary
@hs-pforzheim.de

KEYWORDS
Condition monitoring, process optimization, smart companies, process innovation

ABSTRACT

The objective of "Condition Monitoring" (CM) is to ensure smooth production operations. Traditional reactive and proactive CM processes are not enough. Horizontal and vertical IT integration form the basis for process optimization through predictive and prescriptive management. Digital concepts of industry 4.0 enable innovative business models and flexible business processes.

APPLICATION POTENTIAL OF CONDITION MONITORING

The desire for more resource-efficient processes arises in particular in connection with digital, disruptive technologies. In this context, both established and innovative methods are taken into account in corporate practice. A future-oriented approach is condition monitoring. CM is based on a regular collection of plant and machine data, by recording and analyzing various parameters such as vibration, temperature, quantity as well as wear and tear.

The simplest form of CM follows the reactive principle. The machine or plant can be used productively until a component becomes defective or damaged. In contrast to this, a proactive approach is based on the precautionary principle when wear parts are replaced as a preventive measure according to the manufacturer's instructions or the experience of a maintenance operator. This is to avoid unscheduled downtime, line stoppage and costs incurred from this.

For "smart" companies that rely on digital and networked manufacturing within the framework of Industry 4.0, there is ample room for improvement beyond reactive and proactive CM. The implementation can take place through real-time-based, predictive and possibly prescriptive IT solutions. An innovative process that combines an integrated and a flexible approach is able to eliminate limitations and problems of reactive and proactive CM, optimize resource efficiency and at the same time guarantee sufficient reliability. Based on this, additional design optimization can also be created in the associated processes and, if valuable, in the business models.

By using an advanced version of CM, certain operational activities in the environment of error detection, identification, diagnosis, process recovery/intervention, maintenance and error avoidance can be supported in the decision making process (Laakso et al. 2002) or can be taken over partially or completely autonomously. Algorithms implement the repetitive (routine) acquisition and evaluation in real-time. Furthermore, we are able to predict future IT conditions and different scenarios can be proposed. Management innovation, defined as the implementation of new state of the art management practice, process or technique, has an important role in the overall approach (Birkinshaw 2008). Such management envisions innovative business models where it can select optimal scenarios and more refined adjustment of the parameters in real-time to control the target processes with the resultant increased efficiency as well as improvement in the technical condition of the objects. On the level of strategic decisions, a higher diagnostic effectiveness can be achieved by means of "Smart Data" (Laakso et al. 2002). This is about planning, monitoring and controlling the life cycles in a factory that are associated with condition monitoring of production equipment and maintenance (Hoppe 2014).

In practice, the definition of optimum maintenance intervals proves to be extremely complex. However, this definition forms the basis for intensive coordination between planning, production and maintenance areas in the company. Suboptimal solutions inevitably lead to conflicts between the planning of production orders and the associated output on the one hand and the fixed maintenance intervals on the other. A synergy of the knowledge of workers, maintenance operators and the manufacturer's specifications is the central basis for improvement.

PROCESS-BASED CONDITION MONITORING

Traditional Condition Monitoring

Maintenance is often referred to as having a very short time frame to address actual issues or further optimizations, because of its reactive character. The

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

technical analysis of the system´s condition is focused on interweaving empirical values in connection with selective point checks. Its decision is based on previous experiences in regards to the scheduled maintenance interval or the need for maintenance. If the optimum degree of wear is not precisely met, this usually results in avoidable maintenance or a short running time of the machine or plant. These are associated with more frequent downtimes, which in turn lead to a lower production output, production delays and related cost effects. Other environmental conditions such as heat or vibration can also lead to deviations from the optimally planned maintenance intervals.

The best possible time for maintenance is not precisely predictable and is based on the experience and assessment of a central maintenance department guided by known wear and tear schedule. In the positive case, this leads to a better learning process with a reactive diagnostic characteristic ("What just happened?"). This enables the worker to learn and expand his or her know-how. Experience acquired by the workers and maintenance staff over time cannot be easily transferred. Furthermore, in case of non-availability or withdrawal from the company, this know-how is lost, which makes condition monitoring problematic. If necessary, companies can use other sub-areas/plants or unconditionally follow the manufacturer's specifications for the corresponding maintenance interval.

The objective of systematic condition monitoring in a company is to relieve workers, maintenance staff and other departments involved (e.g. quality assurance, sales, purchasing and human resources management) and to exclude them as a possible source of error. Possibilities for optimization arise from an orientation to the ITIL approach within the area of service operations or to ISO/IEC 20000-1 within the "Resolution Processes". Here, a distinction is made between incident and problem management:

An "incident" describes an unplanned interruption or quality reduction or an event that could cause impairment in the future. Incident management is responsible for rectifying faults as quickly as possible and in the specified time. It manages all incidents throughout their entire lifecycle. The term "problem" represents the cause of one or more incidents. An associated process systematically takes preventive measures to avoid incidents or - if this is not possible - to keep incidents to a minimum.

In traditional condition monitoring, the company does not have any further indicators for the degree of wear and tear or expected known failures which may lead to complications in production. Furthermore, the unplanned downtime of a machine or plant can cause further damage to the objects. It is a complex task for companies to implement a methodology within this framework that brings them closer to the optimal degree of wear and tear. The integration of a vertical and horizontal IT landscape is helpful, in which information technology supports the affected parties.

Optimization through Increasing IT Support

In comparison to traditional condition monitoring with low networked IT systems and centered more on people, the manufacturers of machines and plants have equipped their products with additional sensors in recent years. A dynamic new development in regards to size, functionality and possible connection/networking can be detected. Standardization initiatives in the area of interfaces have significantly improved the ability to connect to Manufacturing Execution Systems (MES).

MES represent IT-based production control systems that manage production orders and production resources. They integrate additional modules such as quality and human resources management and can be connected to ERP or similar systems. Within the recent years, MES development has experienced a strong functional improvement by extending the integration of production-related areas. The company- and plant-specific configuration of MES components has also been extended. This development accompanied by digital technological innovations made it possible to have better use of CM functionalities such as remote access to machine controls and control surfaces (e. g. mirroring). The increased networking of production-related IT systems enables the company's own or external employees to proactively operate CM in real-time. In the context of plant order planning, it is possible to monitor the status of several machines with their respective production orders via a graphical user interface.

There is a wide range of optimization options available within the scope of production order planning. In the case of traditional CM, orders are often still calculated manually and visualized on a blackboard. Coordination between the planning, production and maintenance departments is typically limited. In the event of a constantly changing order situation, taking into account unforeseen maintenance tasks, this approach does not meet the requirements. As a consequence, the processing of production orders is delayed, production may stagnate and the overall effectiveness of the process may be affected by “informational insufficiencies”.

CM is able to support the integrated planning process and to facilitate optimization (Kletti et al. 2015). After production orders have been either generated manually or transferred by the leading ERP system, detailed planning of the production orders can be carried out. This usually takes place on a graphical user interfaces, which display all machines with the respective production orders over a selected period of time. The production order duration is automatically calculated in real-time on the basis of sensor data recorded within the respective machines. In addition, the planning department receives data on existing resource bottlenecks, tool locks, and so on. In

combination with the managed master data, optimized production operations management is possible to reduce throughput times, stock levels and setup times supported by the system. This enables determination of the completion date of the respective production order more precisely with increased capacity utilization while taking into account the necessary maintenance tasks.

System integration leads to stronger networking and more intense cooperation: For example, when the planning department reschedules production orders, they automatically receive a message in case of availability conflicts. This occurs for example, if the tool is blocked by maintenance within the intended period of time. If the related order has a higher priority, the planning department can immediately contact the maintenance department to postpone the maintenance at short notice. In the event of unforeseen repairs, the planning department is informed in real-time so that any necessary changes can be made immediately.

CM makes it possible to simultaneously plan, monitor and control the condition of numerous machines based on real-time data. Furthermore, the associated management activities are still strongly centered on people and roles. In addition, numerous parameters are monitored with regard to limit value over- or undershooting, but no (partially) automated optimization of the operating status is carried out. This is an important potential for the expansion of condition monitoring to a future-oriented CM.

Future-Oriented Design Potential

If conventional condition monitoring is based on descriptive analysis ("What happened?") and real-time diagnosis ("Why did it happen?"), an expanded understanding then allows the inclusion of predictive and prescriptive questions. Furthermore, the design of business rules to support real-time decisions is an important option for the optimized interaction between man and machine. This requires a high level of expertise in strategic planning and operational management of processes as well as a consistent and congruent allocation of responsibility. Depending on the complexity and scope of the project, certain areas can also be automated in this context.

Predictive analysis ("What will happen or what could happen?") is based on a suitable mathematical model in order to be able to make forward-looking predictions. Patterns and trends are generated from quantitative data sets using inductive statistical methods. The focus is on the analysis of the relationships between already known and predicted variables using past events. The result is then used for the forecast. Accordingly, the quality of the data analysis and the selection of the assumptions made have a direct impact on the accuracy and usefulness of a forecast.

Barriers to the use of predictive analyses are posed by the following circumstances, among others (Eckerson 2007):

- Content-related complexity: The development of realistic models typically proves to be a slow, iterative and labor-intensive process.
- Data quality: Incorrect or missing data is a key problem for real-time decisions.
- IT performance: Complex analytical queries and evaluations can adversely affect network and database performance. Interfaces between different IT systems may also play an important role.
- Total (transactional) costs: Direct and indirect costs must be taken into account, both in the personnel area and within IT.

Forecasts provide well-founded statements about events, conditions or developments in the future. This can take the form of a simulation with “What-If” or “How-to-achieve” questions, for example. Projections can be characterized by the fact that they are not based exclusively on historical data, but also on the use of subjective assessments.

The term “data mining” is closely linked to predictive analysis. This means the systematic application of descriptive methods or inductive test methods to a database in order to identify new patterns. Pattern recognition is primarily concerned with recognizing regularities, repetitions or similarities in a set of data.

On the other hand, prescriptive models or analyses ("What should happen?") aim to provide guidance for practical implementation by generating recommendations that change, suppress or steer predicted events in the desired direction. Such a model typically describes the following factors:

- A defined set of options for action and decision making
- A target or key performance indicators for measuring success
- A set of rules that defines which selection is allowed.

Corresponding models are suitable for real-time decisions, which characterize the analytical process (Panian 2009): The aim here is to analyze events that have occurred and to propose actions that will achieve the previously defined goal with the highest possible probability. Related interim and final results are systematically recorded and analyzed in order to improve future recommendations. On the IT side, "decision management" systems can be used for this purpose. One field of application in this area is business rules, which comprise a collection of if-then conditions. These will check the attributes for the respective values and execute corresponding actions if the conditions are fulfilled.

Machine learning can also be used in this context: The aim is to apply processes that enable IT systems to absorb and expand knowledge independently. The algorithms used learn from training data and test cases of the past in order to find correct hypotheses and thus transfer them to new data constellations. There is also the possibility of active learning, where the algorithm generates questions about correct output based on predefined input values. Machine learning is typically suitable for unknown, difficult to describe or difficult to calculate problems. An example of future-oriented condition monitoring based on digital networking is provided by the maintenance department, for example by remote monitoring and maintenance. In the context of Industry 4.0, "Internet of Things (IoT)" and the use of "Cyber-physical Systems (CPS)", maintenance and servicing are becoming increasingly important for the optimization of overall costs. Industry 4.0 addresses the individualization or hybridization of products as well as the integration of customers and business partners into business processes.

The future-oriented concept is moving away from failure and time-dependent maintenance to condition-based maintenance (CBM). In the event of malfunctions, this means that real-time rescheduling is necessary. The aim is to optimize the expenditure, material usage, service life and reliability of the overall system.

For prognostic maintenance, data on faults that have occurred in the past have to be systematically collected, analyzed and correlated with historical sensor readings and observed effects. On the IT side, several components are required for this.

- Integrated data in sufficient quality: Measuring equipment, monitoring and controlling systems must be consolidated to form an overall picture of the machine and plant performance as well as the logistic processes. This includes, among other things, order data, maintenance plans, standards and legal regulations as well as the availability of personnel and material resources.

- Automatic monitoring and event-based signals: The permanent real-time monitoring in combination with the evaluation of historical data indicates possible problem states or maintenance cases and triggers an alarm when thresholds are exceeded. The detection of defect patterns can be linked to the automatic creation and scheduling of a maintenance order or maintenance task.

- Web- and KPI-based dashboards, analysis tools and augmented reality: These serve as the basis for mobile and flexible human-machine interaction. Depending on the stakeholders, different cockpits or solutions have to be designed.

- Dynamic model development: The automatic recognition of error patterns and the derivation of associated tolerance intervals have to be checked over time and optimized when indicated.

The establishment of a future-oriented CM requires corresponding measures to optimize data, to connect and network technical components as well as to improve supporting analysis tools. This enables vertical IT integration, from the field level with CPS to management information systems. Furthermore, the organizational structure has to be redesigned in order to support the optimized business processes.

"SMART" CONDITION MONITORING MANAGEMENT

Companies that have successfully implemented approaches such as Lean Management and on their way to Industry 4.0 can extend their business processes with an IT-based CM and implement additional services from this. A central challenge is to actively monitor the "market of technical possibilities" and to evaluate both the technical and economic maturity of the respective services. SAP for example offers a rapid-deployment solution for condition-based maintenance (CBM) that enables companies to perform maintenance only when necessary with sophisticated intelligence (Lange 2014).

For "smart" factories, there is the opportunity to expand existing systems to use CM of technical components and processes with the help of IoT elements: There is an ongoing development in the field of sensor technology, in monitoring and controlling of actuators, in the automation of process sequences and in the processing and storage of processed data. This is accompanied by the development of decision support tools that are more user-friendly, and have far more powerful adaptable analytic capabilities which create numerous different and more efficient possibilities for process optimization. The technological feasibility is typically accompanied by an economically justifiable effort and the optimization potential makes a rapid payback period seem realistic (Kletti et al. 2015).

Digitalization offers application companies the opportunity to provide their own customers value-added services that in turn improve their work processes. The use of digital data eyeglasses optimizes the possibilities for remote support by maintenance experts in the event of a malfunction and, if necessary, reduces the costs of worldwide product support. Production-relevant containers, their condition and level can be monitored via real-time location systems and container sensors, which reduce the risk of production delays. For example, the monitoring of the historical, hourly and expected future condition of aircraft engines allows a "ground time"-optimal execution of engine maintenance considering the availability of parts, expertise and available time slots. Real-time monitoring of building elevators forms the basis for safe and economical operation. Automated remote diagnosis of household and office equipment opens up potential for the early detection of future malfunctions as well as automated procurement of spare

parts and consumables.

It is an open question whether this development will lead to a reduction or polarization of the associated jobs. The discussion of the consequences from digital transformation is ambivalent in various disciplines and areas. In principle, a higher degree of automation means, for example, that the possibility exists of further shifting the tasks from the workers to essential, quality assurance or monitoring work and, if necessary, of employing workers who have been trained on an ad hoc basis in a flexible manner. However, the new work profiles will tend to place higher demands on employees with an impact on the associated integrated business processes. An important question in this context is the future interaction between man and machine. In the sense of an optimized distribution and coordination, it is a question of who will assume which tasks and roles within human-machine interaction.

In the context of CM management, the authors believe that it is primarily about evolutionary changes. Regarding business practice, CM finds its limits if the technical possibilities and the services offered do not meet the expectations of effective and efficient process monitoring or decision support. In strategic CM, creative ideas for innovative business models are in demand, which can only be achieved by highly qualified and motivated participants. To this end, creative management practices and innovative business processes have to be developed. Targeted investments in advanced CM-enabling technology and functionalities can form the basis for placing value-added services, which support the relationships and interactivity of the CM solution.

REFERENCES

Birkinshaw, J.; G. Hamel; and M.J. Mol. 2008. Management innovation. Academy of management Review, 33(4), 825-845.

Eckerson, W.W. 2007. Predictive analytics. "Extending the Value of Your Data Warehousing Investment". TDWI Best Practices Report, 1 (1-36).

Hoppe, G. 2014. High-Performance Automation verbindet IT und Produktion. *Industrie 4.0 in Produktion, Automatisierung und Logistik*, 249-275.

Kletti, J.D.; M. Diesner; W. Kletti; J.P. Lübbert; J. Schumacher; and T. Strebel. 2015. MES als Werkzeug für die perfekte Produktion. *MES-Manufacturing Execution System*, 19-29.

Laakso, K.; T. Rosqvist; and J.L. Paulsen. 2002. The use of condition monitoring information for maintenance planning and decision-making.

Lange, J. 2014. *SAP Condition-Based Maintenance Rapid Deployment Solution*. Retrieved December 2017, from SAP: https://www.sap.com/documents/2014/05/1c41b250-737c-0010-82c7-eda71af511fa.html

Panian, Z. 2009. Just-in-time business intelligence and real-time decisioning. *Proceedings of AIC. 9*, pp. 106-111. Recent Advances in Applied Informatics and Communications.

AUTHOR BIOGRAPHIES

FRANK MORELLI is a Professor at Pforzheim University of Applied Sciences. He is director of the Master Information Systems program. He carries out research and practice-based projects in business process management, business intelligence, SAP S/4 HANA, project management, and IT organization.

JAN-FELIX MEHRET studies Information Systems in the master program at the Pforzheim University of Applied Sciences. He also acts as a consultant for ProSeS BDE GmbH, where he carries out projects in business process management and manufacturing execution systems.

THORSTEN WEIDT is an adjunct professor for Business Process Management at Pforzheim University of Applied Sciences. He is one of the co-founders of BridgingIT GmbH. In his role as a management consultant he implements IoT/Industry 4.0 technologies and other innovations for the company's customers.

MOUSTAFA ELAZHARY was the MIS program leader for undergraduate studies at MSA University, Cairo, Egypt. He also acted as a business information systems consultant and auditor. Currently, at Pforzheim University of Applied Sciences, he carries out research in IT governance of cloud computing and Internet of Things (IoT).

A DOMAIN-SPECIFIC LANGUAGE FOR ROUTING PROBLEMS

Benjamin Hoffmann, Michael Guckert,
Thomas Farrenkopf
KITE - Kompetenzzentrum für
Informationstechnologie
Technische Hochschule Mittelhessen, Germany
{benjamin.hoffmann,michael.guckert,thomas.farrenkopf}
@mnd.thm.de

Kevin Chalmers,
Neil Urquhart
School of Computing
Edinburgh Napier University, Scotland
{k.chalmers,n.urquhart}
@napier.ac.uk

KEYWORDS

Domain-specific language, Agent-based modelling, Ant algorithms, Dynamic TSP.

ABSTRACT

Vehicle Routing Problems (VRPs) are commonly used as benchmark optimisation problems and they also have many applications in industry. Using agent-based approaches to solve VRPs allows the analysis of dynamic VRP instances that incorporate congestion effects. By using a domain-specific language as part of a model-driven approach, routing problems can be modelled in an abstract form that does not contain implementation and other technical details. With such a tool domain experts can concentrate on the actual modelling task without being distracted by low-level intricacies. We present the DSL Athos in which computational and platform independent routing problems can be defined. The DSL offers an efficient way to model problems with seamless integration of established optimisation methods. Generators create executable code for several agent based platforms. Proof of concept is given by applying the tools to the Oliver 30 TSP and an instance of a dynamic TSP.

INTRODUCTION

The planning and optimisation of logistics networks and other VRP problems have many industrial applications. Planning and optimisation of resources is necessary in a world in which sustainability and efficiency are important for organisations that wish to remain competitive. A typical example is where algorithms and models have to find optimal, or near-optimal, solutions for vehicles in delivery processes. Creating a schedule in which all customers are visited in an efficient order means solving a Travelling Salesman Problem (TSP) (see Schwab, Guckert, and Willems 2017).

TSP is NP-hard, and thus heuristic approaches are commonly used to find solutions. Nature inspired techniques are such an approach (Afaq and Saini 2011). Schwab et. al also describe how a solution for TSPs with additional time constraints (see Savelsberg 1985) based on the ant algorithm was integrated into an off-the-shelf transportation management system (Schwab, Guckert, and Willems 2017).

In real-world applications, solutions should consider the current traffic situation and respect congestion effects. Adding a dynamic element increases the complexity of the problem. Such problems belong to a special class of TSPs known as Dynamic TSP (DTSP) (Cheong and White 2012). By modelling the problem by means of an agent based approach in which the behaviour of the agents can be dynamically adjusted to reflect the current level of traffic congestion, it is possible to analyse and optimise instances of the DTSP.

In this paper, we present the DSL Athos that allows domain experts to specify traffic and transport related optimisation problems in a declarative and concise way. In a program written in Athos, agents move through a network of roads. The agents in the network mutually influence the speed with which they can travel the routes of the given network. Agent behaviour can be defined in various ways. While it is possible to let agents travel a pre-defined list of nodes, they can also be assigned the task of travelling an optimised route. For finding an optimised tour (measured by distance) that visits each node of a given set of nodes the agents must solve a static TSP instance. Agents could also be instructed to optimise for the *fastest* route that visits each node in their list. Since agents increase the time it takes to pass a given road in the network by using it themselves, finding the fastest tour for a given set of nodes requires to solve a *dynamic* version of the TSP. Pillac et al. refer to this as the *evolution of information* (Victor Pillac et al. 2013): agents can only plan their tour based on the information they have at the moment the plan is created. The best they can hope for is that this information represents the current traffic situation in the network. However, a single time step later, the traffic situation has changed, and the more time passes, the more likely it is that the created tour is no longer the optimal one. Agents therefore have to recalculate and adjust their plan regularly.

We will describe Athos and then, as a proof of concept, apply it to the published Oliver30 TSP based on the problem published by Hopfield and Tank (Hopfield and Tank 1987) and on an instance of a dynamic TSP which will be solved by using *Ant Colony System (ACS)* (Dorigo and Gambardella 1997).

RELATED WORK

The TSP is an NP-hard benchmark combinatorial optimisation problem. The problem requires a route to be determined for a

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

salesman who must find the shortest tour to visit a number of cities. Each city must be visited once and once only within a tour so that the solution is a Hamiltonian circuit that starts form and ends in a designated starting point. The number of possible tours is calculated as $(n-1)!$. However if the distances between the cities are bi-directional then the number of possible tours reduces to $(n-1)!/2$.

Problems related to the TSP were discussed by William Hamilton in the 1800s, but the first published work proposing a method for solving the TSP appeared in 1954 (Dantzig, Fulkerson, and Johnson 1954). Subsequently, Chvátal (Chvátal et al. 2010) proposed the cutting plane method based on linear equations and solved a 49 city TSP instance. Subsequent notable methods for solving TSP instances include the 2-opt (Croes 1958), 3-opt (Shen Lin 1965) and Lin-Kernighan (S. Lin and Kernighan 1973) heuristics. Stochastic and nature inspired methods applied to the TSP include genetic algorithms Grefenstette et al. 1985 and ant colony optimisation (Dorigo and Gambardella 1997).

There exists a number of variants of the basic TSP. These include TSP with time windows (TSPW), multiple TSP (MTSP) and dynamic TSP (DTSP). TSPW (Baker 1983) allocates a time window to each city. Cities may only be visited if the salesman arrives within the respective time window. In the MTSP (Laporte and Nobert 1980), with multiple start/end points, the solution has to contain multiple Hamiltonian circuits. The DTSP adds and removes cities at run time (Gharehchopogh, Maleki, and Khaze 2013). The challenge is not to produce one solution, but to produce a series of updated solutions in response to cities being added or deleted from the problem. Beyond that, our definition of a DTSP allows dynamic changes of the weights (i.e. length) of the edges in the network (compare Tinos 2015).

A DSL named Turn was developed by Steil et al. (Steil et al. 2011) in order to specify how vehicles should be routed with a specific real-world VRP instance. A reference to a DSL is made by Pigden et. al. in (Pigden et al. 2012), but no details are supplied. It becomes apparent that whilst some work has been undertaken in relation to the application of DSLs to VRP type problems, there remains a significant requirement of a DSL that can be adapted to a range of VRP instances.

A DSL FOR ROUTING PROBLEMS

Athos is a DSL that allows researchers to define models for traffic-related optimisation problems at a computational and platform independent level. Domain experts can take a declarative approach, instead of imperatively coding agent behaviours. A convention over configuration approach is taken that allows, but does not require, users to control certain aspects of the modelled simulation. For many aspects of the simulation, Athos assumes reasonable default values and leaves it open to the language users to override these defaults with their desired values.

The problems that we wish to simulate comprise agents that must try to optimise routes within a network of roads. Each road (edge) in the network has a capacity attribute that determines the extent to which the road is affected by congestion. These congestion effects are also dependent on the number and types of agents on the road. Some agents have a greater congestion effect on roads than others. For example, an agent that represents a slow-moving large tractor is far more likely to congest a low-capacity road than an agent that represents a small motorcycle. Agents enter the network from arbitrary nodes, meaning that any type of vehicle may potentially enter from any node. It is possible to define distribution functions that represent the probability with which a given place in the network is the origin of a given type of agent. These distributions can be calibrated with data taken from empirical observations. Agents may be specified to differ in their travelling behaviour, e.g. some agents start at a given node and seek to reach a given destination node, visiting a given list of nodes en route. Other agents exhibit a circling behaviour. They repeatedly travel along a given route within the network. Similar to circling agents are shuttle agents which shuttle along a given route of points. The agents with circling and shuttling behaviour are used to create noise and traffic in the network thus simulating high rates of traffic.

Architecture of Athos' generator

The architecture of the generator used to transform Platform-Independent Models (PIMs) into Platform-Specific Models (PSMs) is depicted in Figure 1. The Athos-generator comes in the form of an Eclipse plug-in. Once the generator is given a program written in Athos, it uses a set of templates to generate a traffic model for the NetLogo platform. At this point, it is important to note that Athos is not a mere interface to facilitate the creation of NetLogo models. It is also possible to generate models for other ABM platforms like Repast (a version with different and reduced functionality for Repast Simphony has already been implemented) or Jadex. Together with the generator and its templates, we have developed a NetLogo extension that can be accessed from the generated NetLogo models via the NetLogo-Extension-API. This extension contains several optimisation algorithms for graph structures (e.g. Dijkstra's algorithm) by using available frameworks (e.g. JUNG http://jung.sourceforge.net/).

The extension also features a meta-heuristic known as *Ant Colony System (ACS)* (Dorigo and Gambardella 1997). We have implemented ACS as a NetLogo extension so that agents in the network can find solutions of sufficient quality for the given TSP. ACS is one out of many other possible approaches to solve TSPs heuristically. It was chosen to demonstrate how the philosophy of our architecture can integrate optimisation algorithms. As Athos matures we will possibly implement other algorithms if that widens the applicability. According to Athos' philosophy of providing sensible defaults, if not specified differently, ACS parameter values are those obtained from Dorigo and Gambardella 1997 (exception: $t_0 = 10$).

The ACS implementation

From a theoretical point of view, solving a TSP means finding a Hamiltonian cycle of shortest length (see e.g. Laporte and Osman 1995). As is noted by Laporte and Nobert (Laporte and Nobert 1987), this definition does not cause problems in

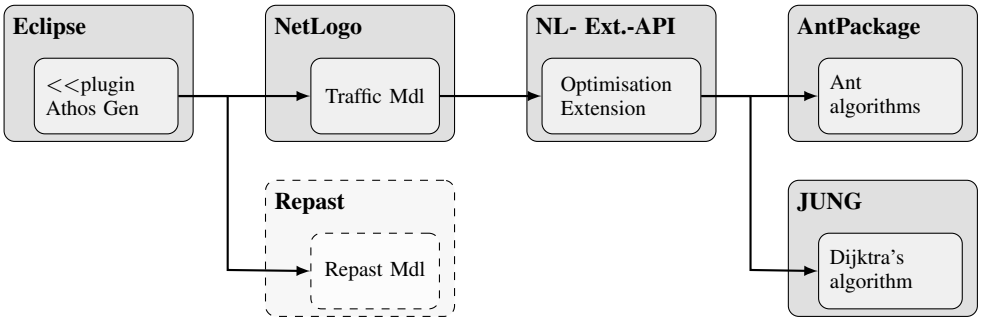


Fig. 1. Athos' modelling approach

complete graphs in which the triangle inequality holds (i.e. in which there is no shorter path between any two nodes in the network than their direct link). Athos does not require the underlying graph to be complete, because real world traffic networks are often incomplete. Therefore, for any two nodes in the network their distance is defined as the length of the shortest path from one to the other. Furthermore, the constraint that each node is to be visited exactly once is relaxed by distinguishing between *service* visits and *crossing* visits. Each node then has to be serviced exactly once but may be crossed an arbitrary number of times.

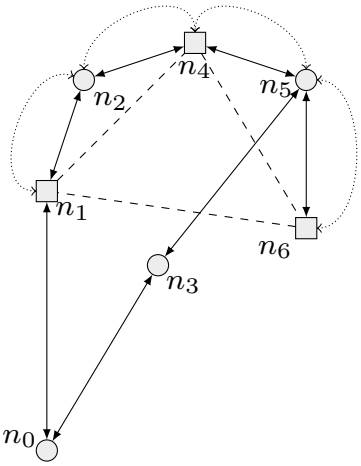


Fig. 2. Behaviour of the salesman-agent in an Athos-generated simulation. Solid lines represent edges, dashed lines represent calculated distances, dotted arcs represent the salesman's moving pattern.

In Figure 2, the graph consists of nodes $N = \{n_0, \ldots, n_6\}$. However, the actual tour of the salesman consists only of $T = \{n_1, n_4, n_6\}$ with $T \subset N$. Following Laporte and Norbert, in our ACS implementation, we first calculate the minimal distance between any two nodes of the tour by means of Dijkstra's algorithm. Take $\overline{n_1 n_6}$ as an example. In our implementation, we would use Dijkstra's to introduce an artificial edge and define $\overline{n_1 n_6} \coloneqq \overline{n_1 n_2 n_4 n_5 n_6}$.
The arcs in Figure 2 then show the solution for the incomplete graph, given that the salesman starts in n_1. The salesman will first service n_4 by going from n_1 to n_2 and then to n_4. Next, he will go to n_5 and then service n_6. From there, he will cross the nodes n_5, n_4, n_2 to return to node n_1. Note that another feasible solution would have the salesman to first service node n_6 and then service node n_4 on the way back to node n_1.
Our ACS implementation works on two-dimensional arrays that contain the mutual distances between any two tour nodes and the pheromone values assigned to the edges. The implemented ACS closely follows the description described in (Dorigo and Gambardella 1997).
The ants work with two lists: one in which the indices of the tour are stored and one in which the indices of the cities yet to be visited are kept. By default, ten ants are used. The algorithm performs a given number of iterations in which the ants construct their tour incrementally beginning with an empty list for the tour. The ants are placed randomly at one of the nodes of the tour. In sequence, the ants are then asked to perform a *state transition*, i.e. add a node to their tour. In order to determine which node to service next, ants execute a *pseudo-random-proportional rule* which is a centrepiece of the ACS algorithm. Depending on the outcome the ant either chooses to *exploit* current length and pheromone information or to *explore* new links by application of a *probabilistic state transition* rule. The pheromone-value for the edge that connects an ant's current and next city is then updated *locally*.
When the list of the yet to be visited cities is empty, the other list in which the tour is stored must be complete. It is important to note that to this point the tour list of an ant only stores each node index exactly once. This means that the node in which the ant is supposed to end the tour must be added in a final step to have a complete tour. In a Hamiltonian circle, the ant must finish its tour at its starting node. However, in order to allow re-optimisation during tour execution, it must be possible to explicitly define a node where the tour is supposed to end. As an example, consider a salesman that found the tour of nodes a, b, c, d to be optimal. When the salesman enters node b, the remaining nodes consist of nodes b, c, d. If the salesman wants to re-optimise this tour, the tour still must be finished in a – and not in b which is the current location of the salesman. For this reason, we modified the algorithm in a way that allows to explicitly define a node where the tour is supposed to end. It is still possible to define an end node that is also the starting node, in which case the outcome will be a Hamiltonian cycle.

The nodes are permuted so that they start at the current location of the salesman (and not at the current location of the ant). Then, the final city is added to the list.

In our implementation, ants possess a `tourCost` attribute. A `reset` method is executed at the beginning of each iteration, and the `tourCost` of each ant is set to a negative value. In this way an ant has to calculate the cost for its tour only when the value of the `tourCost` attribute is negative, otherwise it can simply return the value of the `tourCost` attribute. In a final iteration step, the ants are sorted according to the length of their tour. Over all iterations a `globalBestTourIndices` list is stored, and at the end of each iteration a *global pheromone update* on all edges that connect nodes of these list is performed.

EXAMPLES

Oliver 30 TSP

In this section we show how Athos can be applied to the Oliver 30 TSP instance (http://stevedower.id.au/blog/research/oliver-30/).

```
model oliver30 world xmax 100 xmin 0 ymax 100 ymin 0
functions
durationFunction normal length default
complete network
nodes
node n0 (54.0, 67.0) node n1 (54.0, 62.0) node n2 (37.0, 84.0)
node n3 (41.0, 94.0) node n4 (2.0, 99.0) node n5 (7.0, 64.0)
node n6 (25.0, 62.0) node n7 (22.0, 60.0) node n8 (18.0, 54.0)
node n9 (4.0, 50.0) node n10 (13.0, 40.0) node n11 (18.0, 40.0)
node n12 (24.0, 42.0) node n13 (25.0, 38.0) node n14 (44.0, 35.0)
node n15 (41.0, 26.0) node n16 (45.0, 21.0) node n17 (58.0, 35.0)
node n18 (62.0, 32.0) node n19 (82.0, 7.0) node n20 (91.0, 38.0)
node n21 (83.0, 46.0) node n22 (71.0, 44.0) node n23 (64.0, 60.0)
node n24 (68.0, 58.0) node n25 (83.0, 69.0) node n26 (87.0, 76.0)
node n27 (74.0, 78.0) node n28 (71.0, 71.0) node n29 (58.0, 69.0)
edges
sources
    n0 sprouts ( congestionFactor 2.0 route (n0,n1,n2,n3,n4,n5,n6,n7,n8,n9,n10,n11,
          n12,n13,n14,n15,n16,n17,
n18,n19,n20,n21,n22,n23,n24,n25,n26,n27,n28,n29) optimise )frequency 1.0 every 1
      until 1
```

Athos is purely declarative and neither contains information about the execution platform nor the method with which the problem is going to be solved. Therefore it is computationally independent and platform independent. The Athos generator generates an executable NetLogo program, which will use Ant Colony System (ACS) to solve the problem. By modifying the generator the optimisation heuristic can be changed.

ACS is a stochastic algorithm which can produce varying results with each execution. The generated solutions tend to have a length in the range of 430 to 480. This is achieved by a default of 30 iterations, for which our implementation requires less than one second on a machine equipped with an Intel i7-6820HQ CPU running at 2.70GHz. One example route generated by our implementation has a length of 448.834. The optimal solution published on the Oliver30 website is 423.741. It is not our aim to find optimal solutions or to outperform current published solutions. Our goal is the development of a tool that allows efficient definition of problems and integrates appropriate solving methods.

Dynamic vs. static agents

Our second example shows how Athos can be used to study the implications of dynamic decision making based on complete information within in a simple road network. For this purpose, we define a network that consists of nine nodes and eleven bidirectional edges between selected nodes. Within this network, there are five agents, two of which travel pre-defined paths and three who aim to find an optimal tour for a given set of nodes. In this example, the optimisation function is not defined in terms of the travelled distance but in the amount of time required to complete a tour. More precisely, the value of interest is the time it takes for the three agents to complete a total of 100 tours.

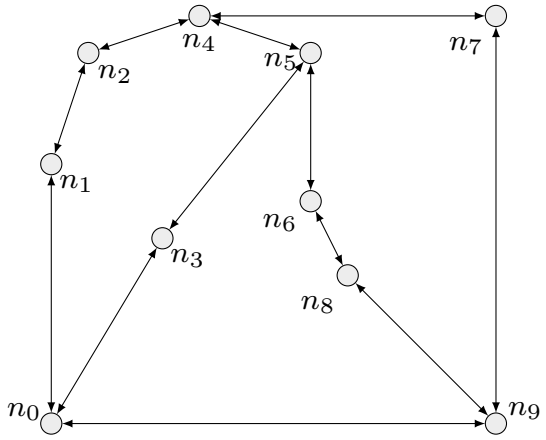


Fig. 3. Network for dynamic problem.

Figure 3 illustrates the network. Agent 1 is assigned a tour that consists of the nodes n_1, n_4, n_6. Agent 2 is to service nodes n_2, n_6, n_7, n_9 and Agent 3 services nodes n_8, n_3, n_1. The three agents aim at tour completion in a minimum amount of time. In order to intensify congestion effects, Agent 4 and Agent 5 travel the predefined route $n_4, n_5, n_3, n_0, n_1, n_2$ and $n_3, n_0, n_1, n_2, n_4, n_5$, respectively. Depending on the simulation configuration, the tour-optimising agents either exhibit a static or dynamic optimising behaviour.

In this context, static behaviour allows the agent to calculate the optimal route at the beginning of the simulation, based on the traffic situation at the moment the calculation is performed. Once a route is calculated, the agent sticks to this route throughout the entire simulation. Dynamic optimisation implies that each time an agent reaches a node within its tour it calculates a new optimised tour that comprises the unserviced nodes. Agents in this dynamic scenario only recalculate their tour when they have serviced a node. They do not recalculate when travelling through a node. If an agent has just recalculated its tour, and the next node to be serviced is not directly connected to the current node, the path to the next node to be serviced is calculated based on the data used when recalculating the tour. The agent will then follow the path and not perform any recalculation until it arrives at the node to be serviced.

TABLE I. RESULTS OF DYNAMIC TSP EXPERIMENT IN THE NETWORK. ALL EDGES IN THE NETWORK WHERE ASSIGNED THE TRAVEL DURATION FUNCTION $t = l + 2AC$ (l : LENGTH OF EDGE, AC : ACCUMULATED CONGESTION FACTOR). OPTIMAL TOURS WERE CALCULATED BY MEANS OF THE ACS ALGORITHM WITH THE FOLLOWING PARAMETERS $\alpha = 0.1, \rho = 0.1, q_0 = 0.9, \beta = 2.0$, 10 ANTS WERE USED IN 30 ITERATIONS.

Agent	Agent 1		Agent 2		Agent 3		Agent 4		Agent 5		
Mode	Optimising		Optimising		Optimising		Static		Static		
Tour	n_1, n_4, n_6		n_2, n_6, n_7, n_9		n_8, n_3, n_1		$n_4, n_5, n_3, n_0, n_1, n_2$		$n_3, n_0, n_1, n_2, n_4, n_5$		
Config	dyn?	CF	dyn?	CF	dyn?	CF	dyn?	CF	dyn?	CF	Avg. time 100 tours
C_1	no	50	no	50	no	50	no	50	no	50	117,783.4
C_2	yes	50	yes	50	yes	50	no	50	no	50	109,593.0
C_3	no	250	no	250	no	250	no	50	no	50	589,651.4
C_4	yes	250	yes	250	yes	250	no	50	no	50	555,630.9
C_5	no	1250	no	1250	no	1250	no	50	no	50	2,706,275.7
C_6	yes	1250	yes	1250	yes	1250	no	50	no	50	2,691,348.8

The listing below demonstrates how the intended scenario is modelled in Athos. For Agents 1 – 3 the model only uses the keyword `optimise` – it does not go into any computational detail as to *how* this optimisation is to be achieved. Also note how each `edge` is associated with a `durationFunction` that defines the time it takes to travel an edge in terms of its `length`, congestion factor (`cfactor`) and its current traffic (accumulated congestion factor of all agents on the respective road `accCongestionFactor`). If the `optimise` keyword is directly followed by the `dynamic` keyword, the respective agent behaves in the dynamic manner described above.

```
model incomplete world xmax 30 xmin 0 ymax 30 ymin 0
functions
durationFunction normal  length + cfactor * accCongestionFactor default
network
nodes
node n0 (1.0, 1.0) node n1 (1.0, 8.0) node n2 (2.0, 11.0)
node n3 (4.0, 6.0) node n4 (5.0, 12.0) node n5 (8.0, 11.0)
node n6 (8.0, 7.0) node n7 (13.0, 12.0) node n8 (9.0, 5.0)
node n9 (13.0, 1.0)
edges
edge undirected e0 from n0 to n1 length 0.0 cfactor 2.0 function normal
edge undirected e1 from n1 to n2 length 0.0 cfactor 2.0 function normal
edge undirected e2 from n2 to n4 length 0.0 cfactor 2.0  function normal
edge undirected e3 from n4 to n5 length 0.0 cfactor 2.0  function normal
edge undirected e4 from n5 to n3 length 0.0 cfactor 4.0 function normal
edge undirected e5 from n6 to n5 length 0.0 cfactor 2.0 function normal
edge undirected e6 from n3 to n0 length 0.0 cfactor 4.0 function normal
edge undirected e7 from n7 to n4 length 0.0 cfactor 2.0 function normal
edge undirected e8 from n7 to n9 length 0.0 cfactor 2.0 function normal
edge undirected e9 from n9 to n0 length 0.0 cfactor 2.0 function normal
edge undirected e10 from n9 to n8 length 0.0 cfactor 2.0 function normal
edge undirected e11 from n8 to n6 length 0.0 cfactor 2.0 function normal
sources
   n1 sprouts ( congestionFactor 250.0 route (n1, n4, n6) optimise dynamic)
         frequency 1.0 every 1 until 1 // Agent 1
   n2 sprouts ( congestionFactor 250.0 route (n2, n6, n7, n9) optimise dynamic)
         frequency 1.0 every 1 until 1 // Agent 2
   n8 sprouts ( congestionFactor 250.0 route (n8, n3, n1) optimise dynamic)
         frequency 1.0 every 1 until 1  // Agent 3
   n4 sprouts ( congestionFactor 250.0 route (n4, n5, n3, n0, n1, n2) mode 1 )
         frequency 1.0 every 1 until 1    //Agent 4
   n3 sprouts ( congestionFactor 250.0 route (n3, n0, n1, n2, n4, n5) mode 1)
         frequency 1.0 every 1 until 1    //Agent 5
```

Table I shows the results of the simulation experiments. The columns of the table show the agents that were used in the simulations and their general behaviour as well as their respective route. For the experiment, six different configurations, with differing agent configurations, were executed. Ten simulations were executed for each configuration and the amount of time in simulation ticks was recorded. The tour-optimising agents had to complete a total of 100 tours. Three different congestion factor (CF) values were used for all agents: 50, 250 and 1250. For each of those values, 10 simulations were executed, in which all tour-optimisation agents behaved in a static way and ten simulations were executed in which they utilised dynamic behaviours. The right column of the table presents the mean time (measured in simulation ticks) the tour-optimising agents required to perform 100 tours.

Table I shows that in the underlying network, the agents performed better when they acted in a dynamic way. With a congestion factor of 50, the dynamic agents required approx. 6.9% less time than the static agents. Surprisingly, this value decreased to only approx. 5.7%, when we intended to increase congestion effects by altering the agents' congestion factor to 250. When the congestion factor was increased to 1,250, this saw the advantage of the dynamic agents decline to 0.55%.

The reason for the diminishing performance advantage of the dynamic agents with increased congestion factors requires further investigation. However, we decided to modify the underlying network in a way that it contained one road of increased length so that dynamic agents could avoid this road when they identified congestion on it. To this end, we created the network depicted in Figure 4. The network features an additional edge between nodes n_5 and n_7 and the undirected edges $n_0, n_3, n_3, n_5, n5, n7$ form a coherent path. The congestion factor of an agent that is on an edge of such a coherent path, is also considered in the calculations of all other edges that belong to the same coherent path. Athos allows the specification of such a coherent path of edges simply by adding the `path` keyword followed by the name of the path to any edge in a network. For this scenario, we also wanted some more influence on the underlying ant algorithm and thus chose to leave the computationally independent level. Due to this step, we were able to explicitly define the parameters of the ACS algorithm. Note how the Athos model below increases the iterations to 60.

```
model incomplete world xmax 30 xmin 0 ymax 30 ymin 0
functions
durationFunction normal  length + cfactor * accCongestionFactor default
network nodes
node n0 (1.0, 1.0) node n1 (1.0, 8.0) node n2 (2.0, 11.0)
node n3 (4.0, 6.0) node n4 (5.0, 12.0) node n5 (8.0, 11.0)
node n6 (8.0, 7.0) node n7 (13.0, 12.0) node n8 (9.0, 5.0)
node n9 (13.0, 1.0)
edges
edge undirected e0 from n0 to n1 length 0.0 cfactor 2.0 function normal
edge undirected e1 from n1 to n2 length 0.0 cfactor 2.0 function normal
edge undirected e2 from n2 to n4 length 0.0 cfactor 2.0  function normal
edge undirected e3 from n4 to n5 length 0.0 cfactor 2.0  function normal
edge undirected e4 from n5 to n3 length 0.0 cfactor 4.0 path "ab" function normal
edge undirected e5 from n6 to n5 length 0.0 cfactor 2.0 function normal
edge undirected e6 from n3 to n0 length 0.0 cfactor 4.0 path "ab" function normal
edge undirected e7 from n7 to n4 length 0.0 cfactor 2.0 function normal
edge undirected e8 from n7 to n9 length 0.0 cfactor 2.0 function normal
edge undirected e9 from n9 to n0 length 0.0 cfactor 2.0 function normal
edge undirected e10 from n9 to n8 length 0.0 cfactor 2.0 function normal
edge undirected e11 from n8 to n6 length 0.0 cfactor 2.0 function normal
edge undirected e12 from n5 to n7 length 0.0 cfactor 2.0 path "ab" function normal
sources
   n0 sprouts (congestionFactor 50.0 route (n0, n8, n4) ant dynamic ants 10 alpha
         0.1 rho 0.1 iterations 120 q0 0.9 beta 2.0) frequency 1.0 every 1 until 1
   n9 sprouts (congestionFactor 50.0 route (n9, n4, n0) ant dynamic ants 10 alpha
         0.1 rho 0.1 iterations 120 q0 0.9 beta 2.0) frequency 1.0 every 1 until 1
   n7 sprouts (congestionFactor 50.0 route (n7, n0) ant dynamic ants 10 alpha 0.1
         rho 0.1 iterations 120 q0 0.9 beta 2.0) frequency 1.0 every 1 until 1
   n4 sprouts ( congestionFactor 50.0 route (n4, n5, n3, n0, n1, n2) mode 1 )
         frequency 1.0 every 1 until 1
   n3 sprouts ( congestionFactor 50.0 route (n3, n0, n1, n2, n4, n5) mode 1)
         frequency 1.0 every 1 until 1
```

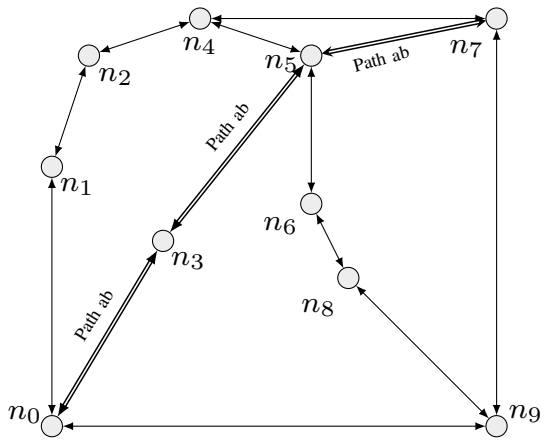


Fig. 4. Network for dynamic problem.

The right column of Table II shows the average time required by the agents for the new network. When assigned a congestion factor of 50, the static agents exhibited better performance. While the dynamic agents required an average of 99,027 ticks (over ten simulation runs), their static counterparts only required 92,928.3 ticks. However, when the congestion factor was increased to 250, the dynamic agents outperformed the static agents by 4.9%. With a congestion factor of 1.250, the performance advantage dropped to 2.6%.

The numbers indicate that dynamism alone is no guarantee for better performance. Other factors like the topology of the underlying network have to be considered. As the presented problems are highly dynamic, even an updated tour is outdated very quickly. This example showed how Athos allows for the convenient definition of dynamic traffic simulations. From these simulations further data can be derived that allow to analyse traffic and routing related problems.

CONCLUSIONS AND FUTURE WORK

We have presented the DSL Athos with which computational and platform independent descriptions of traffic simulations and routing problems can be created. These models are free of implementation details and do not explicitly specify the method to be applied to solve the problem. All necessary details are generated into the code to be executed in the target environment. We have shown how Athos can be applied to different traffic-related optimisation problems. In a first example, we have demonstrated how Athos describes and solves a popular TSP benchmark problem. In a second example, we applied Athos to analyse a traffic scenario where multiple dynamic TSP problems occurred.

Our goal is to develop a language in which models can be described more efficiently than with conventional methods and not to improve algorithms and solutions. In the current version, we generate code that implements the ACS heuristic but this can and will be extended to other approaches. There is the potential to create a more intelligent solution generator which can analyse given models and chose appropriate methods and heuristics. Doing this in the generator, rather than in a more interpretative way at runtime, results in lean best-practice implementations which can scale in size and usability by choosing the best platform and algorithmic approach for a problem. The next steps in the development of Athos are to extend the expressiveness of the language (to encompass more problem types) and intelligent features of the generator.

REFERENCES

Afaq, Hifza and Janjay Saini (2011). “On the solutions to the Travelling Salesman Problem using Nature Inspired Computing Techniques”. In: *IJCSI International Journal of Computer Science Issues* 8.2, pp. 326–334.

Baker, E K (1983). “An exact algorithm for the time-constrained travelling salesman problem”. In: *Operations Research* 31.5, pp. 938–945.

Cheong, T and C.C. White (2012). “Dynamic Traveling Salesman Problem: Value of Real-Time Traffic Information”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.2, pp. 619–630.

Chvátal, Vašek et al. (2010). “Solution of a Large-Scale Traveling-Salesman Problem”. In: *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Ed. by Michael Jünger et al. Springer Berlin Heidelberg, pp. 7–28.

Croes, G. A. (1958). “A method for solving traveling salesman problems”. In: *Operations Research* 6, pp. 791–812.

Dantzig, George B., Delbert R. Fulkerson, and Selmer M. Johnson (1954). “Solution of a large-scale travelling-salesman problem”. In: *Technical Report P-510, RAND Corporation, Santa Monica, California, USA*.

Dorigo, M. and L. M. Gambardella (1997). “Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem”. In: *Trans. Evol. Comp* 1.1, pp. 53–66.

Gharehchopogh, F. S., I. Maleki, and S.R Khaze (2013). “A New Optimization Method for Dynamic Travelling Salesman Problem with Hybrid Ant Colony Optimization Algorithm and Particle Swarm Optimization”. In: *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)* 2.2, pp. 352–358.

TABLE II. RESULTS OF DYNAMIC TSP EXPERIMENT IN THE NETWORK. ALL EDGES IN THE NETWORK WHERE ASSIGNED THE TRAVEL DURATION FUNCTION $t = l + 2AC$ (l : LENGTH OF EDGE, AC : ACCUMULATED CONGESTION FACTOR). OPTIMAL TOURS WERE CALCULATED BY MEANS OF THE ACS ALGORITHM WITH THE FOLLOWING PARAMETERS $\alpha = 0.1, \rho = 0.1, q_0 = 0.9, \beta = 2.0$, 10 ANTS WERE USED IN 60 ITERATIONS.

Agent	Agent 1		Agent 2		Agent 3		Agent 4		Agent 5		
Mode	Optimising		Optimising		Optimising		Static		Static		
Tour	n_0, n_8, n_4		n_9, n_4, n_0		n_7, n_0		$n_4, n_5, n_3, n_0, n_1, n_2$		$n_3, n_0, n_1, n_2, n_4, n_5$		
Config	dyn?	CF	dyn?	CF	dyn?	CF	dyn?	CF	dyn?	CF	Avg. time 100 tours
C_1	no	50	no	50	no	50	no	50	no	50	92,928.3
C_2	yes	50	yes	50	yes	50	no	50	no	50	99,027.0
C_3	no	250	no	250	no	250	no	50	no	50	491,478.8
C_4	yes	250	yes	250	yes	250	no	50	no	50	467,292.3
C_5	no	1250	no	1250	no	1250	no	50	no	50	2,356,096.0
C_6	yes	1250	yes	1250	yes	1250	no	50	no	50	2,294,698.8

Grefenstette, John J. et al. (1985). "Genetic Algorithms for the Traveling Salesman Problem". In: *Proceedings of the 1st International Conference on Genetic Algorithms*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., pp. 160–168.

Hopfield, J.J. and D.W. Tank (1987). "A study of permutation crossover operators on the travelling salesman problem". In: *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*, pp. 224–230.

Laporte, Gilbert and Yves Nobert (1980). "A Cutting Planes Algorithm for the m-Salesmen Problem". In: *Journal of the Operational Research Society* 31.11, pp. 1017–1023.

Laporte, Gilbert and Yves Nobert (1987). "Exact algorithms for the vehicle routing problem". In: *North-Holland Mathematics Studies* 132, pp. 147–184.

Laporte, Gilbert and Ibrahim H. Osman (1995). "Routing problems: A bibliography". In: *Annals of Operations Research* 61.1, pp. 227–262.

Lin, S. and B. W. Kernighan (1973). "An Effective Heuristic Algorithm for the Traveling-Salesman Problem". In: *Operations Research* 21.2, pp. 498–516.

Lin, Shen (1965). "Computer Solutions of the Traveling Salesman Problem". In: *Bell System Technical Journal* 44.10, pp. 2245–2269.

Pigden, Tim et al. (2012). "VeRoLog (Vehicle Routing and Logistics Conference)". In: *EURO Working Group on Vehicle Routing and Logistics Optimisation.*

Savelsberg, M.W.P. (1985). "Local search in routing problems with time windows". In: *Annals of Operations-Research* 4.1, pp. 285–305.

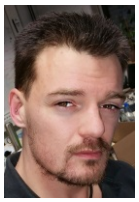
Schwab, J., M. Guckert, and M. Willems (2017). "Tourenoptimierung". In: *Prozesse, Technologie, Anwendungen, Systeme und Management 2017: Angewandte Forschung in der Wirtschaftsinformatik*. Ed. by Thomas Barton. Heide, Deutschland: Mana Buch, pp. 1–33.

Steil, D. A. et al. (2011). "Patrol Routing Expression, Execution, Evaluation, and Engagement". In: *IEEE Transactions on Intelligent Transportation Systems* 12.1, pp. 58–72.

Tinos, R. (2015). "Analysis of the dynamic traveling salesman problem with weight changes". In: *Proceedings of the 2015 Latin America Congress on Computational Intelligence (LA-CCI)*.

Victor Pillac et al. (2013). "A review of dynamic vehicle routing problems". In: *European Journal of Operational Research* 225.1, pp. 1–11.

AUTHOR BIOGRAPHIES



BENJAMIN HOFFMANN is a research assistant at Technische Hochschule Mittelhessen in Friedberg from which he also received his master's degree. He is also a PhD student at Edinburgh Napier University. His research activities are in domain-specific languages, model-driven software development and optimisation problems.

MICHAEL GUCKERT is a Professor of Applied Informatics at Technische Hochschule Mittelhessen and head of KITE - AACC (Kompetenzzentrum für Informationstechnologie - Advanced Analytics Cognitive Computing) . He received a degree in Mathematics from Justus Liebig University Giessen and a PhD in Computer Science from Philipps University Marburg. His research areas are multi agent systems, model driven software development and applications of artificial intelligence.

THOMAS FARRENKOPF is a lecturer at Technische Hochschule Mittelhessen, Friederg. He completed a PhD in the use of software agents and ontologies for business simulation at School of Computing, Edinburgh Napier University. Alongside his research activities, he currently teaches modules on software engineering and algorithms.

KEVIN CHALMERS is a senior lecturer at Edinburgh Napier University where he leads the Computer Science and Software Engineering subject area. He gained his PhD from Edinburgh Napier University in 2009, examining the application of mobile concurrency models to ubiquitous computing. His research is focused primarily on concurrency and parallelism and how different technologies can support this.

NEIL URQUHART is a lecturer in Computing Science at Edinburgh Napier University where he is Programme Leader for the Computing Science. He gained is PhD from Edinburgh Napier University in 2002, writing a thesis examining the use of Software Agents and Evolutionary Algorithms to solve a real-world routing optimisation problem. His research interests include Evolutionary Computation and Agent-based Systems and their application to real-world optimisation problems

POSITIVITY AND STABILITY OF DESCRIPTOR CONTINUOUS-TIME LINEAR SYSTEMS WITH INTERVAL STATE MATRICES

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

ABSTRACT

Necessary and sufficient conditions for the positivity and stability of descriptor continuous-time linear systems with interval state matrices are established. The convex linear combination of the Hurwitz polynomials of positive descriptor linear systems is analyzed. It is shown that the convex linear combination of the Hurwitz polynomials of positive linear systems is also the Hurwitz polynomial. The Kharitonov theorem is extended to the positive descriptor linear systems with interval state matrices.

KEY WORDS

interval, positive, descriptor, linear, continuous-time, system, stability, Kharitonov theorem.

INTRODUCTION

A dynamical system is called positive if its state variables take nonnegative values for all nonnegative inputs and nonnegative initial conditions. The positive linear systems have been investigated in (Berman and Plemmons 1994, Farina and Rinaldi 2000, Kaczorek 2002 and 2008) and positive nonlinear systems in (Kaczorek 2014, 2015a, 2015b, 2015c and 2016).

Examples of positive systems are industrial processes involving chemical reactors, heat exchangers and distillation columns, storage systems, compartmental systems, water and atmospheric pollution models. A variety of models having positive linear behavior can be found in engineering, management science, economics, social sciences, biology and medicine, etc.

Positive linear systems with different fractional orders have been addressed in (Busłowicz 2012, Kaczorek 2010 and 2011). Descriptor (singular) linear systems have been analyzed in (Kaczorek 1993, 1997, 2012, 2014, 2018a) and the stability of a class of nonlinear fractional-order systems in (Kaczorek 2015c and 2016, Xiang-Jun et al. 2008). Application of Drazin inverse to analysis of descriptor fractional discrete-time linear systems has been presented in (Kaczorek 2013) and stability of discrete-time switched systems with unstable subsystems in (Zhang et al. 2014a). The robust stabilization of discrete-time positive switched systems with uncertainties has been addressed in (Zhang et al. 2014b). Comparison of three method of analysis of the descriptor fractional systems has been presented in (Sajewski 2016a). Stability of linear fractional order systems with delays has been analyzed in (Busłowicz 2008) and simple conditions for practical stability of positive fractional systems have been proposed in [4]. The stability of interval positive continuous-time linear systems has been addressed in (Kaczorek 2018b).

In this paper the positivity and the asymptotic stability of descriptor continuous-time linear systems with interval state matrices will be investigated.

The paper is organized as follows. In section 2 some basic definitions and theorems concerning descriptor linear systems and elementary row and column operations are recalled. Necessary and sufficient conditions for the positivity of the descriptor linear systems are established in section 3. The convex linear combination of Hurwitz polynomials of positive linear systems and an extension of the Kharitonov theorem are given in section 4. The stability of positive descriptor linear systems with interval state matrices is addressed in section 5. Concluding remarks are given in section 6.

The following notations will be used: $\Re$ - the set of real numbers, $\Re^{n\times m}$ - the set of $n\times m$ real matrices, $\Re_+^{n\times m}$ - the set of $n\times m$ real matrices with nonnegative entries and $\Re_+^n = \Re_+^{n\times 1}$, 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.

PRELIMINARIES

Consider the autonomous descriptor continuous-time linear system

Consider the autonomous descriptor continuous-time linear system

$$E\dot{x} = Ax, \tag{1}$$

where $x = x(t) \in \Re^n$ is the state vector and $E, A \in \Re^{n\times n}$.

It is assumed that

$$\det[Es - A] \neq 0 \text{ for some } s \in \mathbf{C}, \tag{2}$$

where $\mathbf{C}$ is the field of complex numbers and the system (1) has unique solution for admissible initial conditions $x_0 = x(0) \in \Re^n$.

It is well-known (Kaczorek 1993) that if (2) holds then there exists a pair of nonsingular matrices $P, Q \in \Re^{n\times n}$ such that

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

$$P[Es-A]Q=\begin{bmatrix} I_{n_1}s-A_1 & 0 \\ 0 & Ns-I_{n_2} \end{bmatrix},$$

$$A_1 \in \Re^{n_1 \times n_1},\ N \in \Re^{n_2 \times n_2}, \tag{3}$$

where $n_1 = \deg\{\det[Es-A]\}$ and N is the nilpotent matrix, i.e. $N^{\mu}=0$, $N^{\mu-1}\neq 0$ (μ is the nilpotency index).

To simplify the considerations it is assumed that the matrix N has only one block. The nonsingular matrices P and Q can be found for example by the use of elementary row and column operations [20]:

1) Multiplication of any i-th row (column) by the number $c\neq 0$. This operation will be denoted by $L[i\times c]$ ($R[i\times c]$).
2) Addition to any i-th row (column) of the j-th row (column) multiplied by any number $c\neq 0$. This operation will be denoted by $L[i+j\times c]$ ($R[i+j\times c]$).
3) Interchange of any two rows (columns). This operation will be denoted by $L[i,j]$ ($R[i,j]$).

POSITIVITY OF DESCRIPTOR LINEAR SYSTEMS

Definition 1. The descriptor system (1) is called (internally) positive if $x(t)\in\Re_+^n$, $t\geq 0$ for all admissible nonnegative initial conditions $x(0)\in\Re_+^n$.

Definition 2. A real matrix $A=[a_{ij}]\in\Re^{n\times n}$ 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. The descriptor system (1) is positive if and only if the matrix E has only linearly independent columns and the matrix $A_1\in M_{n_1}$.

Proof. Knowing $n_1=\deg\{\det[Es-A]\}$ and $\operatorname{rank}E$ we may find the nilpotency index $\mu=\operatorname{rank}E-n_1+1$ of the matrix N. Using column permutation of E we choose its n_1 linearly independent columns as its first columns. Next using elementary row operations we transform the matrix E to the form $\begin{bmatrix} I_{n_1} & 0 \\ 0 & N \end{bmatrix}$ and the matrix A to the form $\begin{bmatrix} A_1 & 0 \\ 0 & I_{n_2} \end{bmatrix}$.

From (3) it follows that the system (1) has been decomposed into two independent subsystems

$$\dot{x}_1 = A_1x_1,\ x_1\in\Re^{n_1} \tag{4}$$

and

$$N\dot{x}_2 = x_2,\ x_2\in\Re^{n_2}, \tag{5}$$

where

$$Q^{-1}x=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{6}$$

and Q and Q^{-1} are permutation matrices.

It is well-known (Farina and Rinaldi 2000, Kaczorek 2002) that the solution $x_1=e^{A_1t}x_1(0)$ of (4) is not negative if and only if $A_1\in M_{n_1}$ and the solution x_2 of (5) is zero for $t>0$. □

Definition 3. (Farina and Rinaldi 2000, Kaczorek 2002) The positive system (4) is called asymptotically stable if

$$\lim_{t\to\infty} x_1(t)=0 \text{ for all admissible } x_1(0)\in\Re_+^{n_1}. \tag{7}$$

Theorem 2. (Farina and Rinaldi 2000, Kaczorek 2002 and 2018a) The positive system (4) is asymptotically stable if and only if one of the equivalent conditions is satisfied:

1) All coefficients of the polynomial

$$\det[I_{n_1}s-A_1]=s^{n_1}+a_{n_1-1}s^{n_1-1}+...+a_1s+a_0 \tag{8}$$

are positive, i.e. $a_k>0$ for $k=0,1,...,n_1-1$.

2) All principal minors $\overline{M}_i$, $i=1,...,n_1$ of the matrix $-A_1$ are positive, i.e.

$$\overline{M}_1=\left|-a_{11}\right|>0,\ \overline{M}_2=\begin{vmatrix} -a_{11} & -a_{12} \\ -a_{21} & -a_{22} \end{vmatrix}>0,\ ...,\quad \overline{M}_{n_1}=\det[-A_1]>0 \tag{9}$$

3) There exists a strictly positive vector $\lambda=[\lambda_1\ \cdots\ \lambda_{n_1}]^T$, $\lambda_k>0$, $k=1,...,n_1$ such that

$$A_1\lambda<0 \text{ or } A_1^T\lambda<0. \tag{10}$$

If $\det A\neq 0$ then we may choose $\lambda=-A_1^{-1}c$, where $c\in\Re^{n_1}$ is any strictly positive vector.

Example 1. Consider the descriptor system (1) with the matrices

$$E=\begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & -2 \\ 1 & -2 & 0 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix},\ A=\begin{bmatrix} 0 & 1 & 0 & -4 \\ 1 & -4 & 0 & 4 \\ 0 & 6 & 1 & 0 \\ 1 & -1 & 0 & 4 \end{bmatrix}. \tag{11}$$

The condition (2) for (11) is satisfied since

$$\det[Es-A]=\begin{vmatrix} 0 & -1 & 0 & 2s+4 \\ -1 & s+4 & 0 & -2s-4 \\ s & -2s-6 & -1 & 0 \\ -1 & 1 & 0 & -2s-4 \end{vmatrix} \tag{12}$$

$$=-2s^2-10s-12$$

and $n_1=2$. In this case $\operatorname{rank}E=3$ and $\mu=\operatorname{rank}E-n_1+1=2$.

To transform the matrix $Es-A$ with (11) to the desired form

$$\begin{bmatrix} I_2 s - A_1 & 0 \\ 0 & Ns - I_2 \end{bmatrix}$$

$$\text{with } A_1 = \begin{bmatrix} -2 & 1 \\ 0 & -3 \end{bmatrix},\ N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \qquad (13)$$

The following elementary column operations $R[4\times\frac{1}{2}]$, $R[4,1]$ and elementary row operations $L[2+4\times(-1)]$, $L[4+1\times1]$, $L[3+2\times2]$ have been performed.

In this case the matrices Q and P have the form

$$Q = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{bmatrix},\ P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 2 & 1 & -2 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \qquad (14)$$

Note that the matrix A_1 defined by (13) is the stable Metzler matrix and the descriptor system with (11) is positive and asymptotically stable.

CONVEX LINEAR COMBINATION OF HURWITZ POLYNOMIALS AND EXTENSION OF KHARITONOV THEOREM

Consider the set (family) of the n-degree polynomials

$$p_n(s) := a_n s^n + a_{n-1} s^{n-1} + ... + a_1 s + a_0 \qquad (15a)$$

with the interval coefficients

$$\underline{a_i} \le a_i \le \overline{a_i},\ i = 0,1,...,n. \qquad (15b)$$

Using (15a) we define the following four polynomials

$$\begin{aligned} p_{1n}(s) &:= \underline{a_0} + \underline{a_1} s + \overline{a_2} s^2 + \overline{a_3} s^3 + \underline{a_4} s^4 + \underline{a_5} s^5 + ..., \\ p_{2n}(s) &:= \underline{a_0} + \overline{a_1} s + \overline{a_2} s^2 + \underline{a_3} s^3 + \underline{a_4} s^4 + \overline{a_5} s^5 + ..., \\ p_{3n}(s) &:= \overline{a_0} + \underline{a_1} s + \underline{a_2} s^2 + \overline{a_3} s^3 + \overline{a_4} s^4 + \underline{a_5} s^5 + ..., \\ p_{4n}(s) &:= \overline{a_0} + \overline{a_1} s + \underline{a_2} s^2 + \underline{a_3} s^3 + \overline{a_4} s^4 + \overline{a_5} s^5 + ... \end{aligned} \qquad (16)$$

Theorem 3. (Kharitonov Theorem) The set of polynomials (15) is asymptotically stable if and only if the four polynomials (16) are asymptotically stable.

Proof. The proof is given in (Kharitonov 1987, Kaczorek 1993).

The polynomial

$$p(s) := s^n + \overline{a_{n-1}} s^{n-1} + ... + \overline{a_1} s + \overline{a_0} \qquad (17)$$

is called Hurwitz if its roots s_i, $i = 1,...,n$ satisfy the condition $\mathrm{Re}\, s < 0$ for $i = 1,...,n$.

Definition 4. The polynomial

$$p(s) := (1-k) p_1(s) + k p_2(s) \text{ for } k \in [0,1] \qquad (18)$$

is called convex linear combination of the polynomials

$$\begin{aligned} p_1(s) &= s^n + a_{n-1} s^{n-1} + ... + a_1 s + a_0, \\ p_2(s) &= s^n + b_{n-1} s^{n-1} + ... + b_1 s + b_0. \end{aligned} \qquad (19)$$

Theorem 4. The convex linear combination (18) of the Hurwitz polynomials (19) of the positive linear system is also a Hurwitz polynomial.

Proof. By Theorem 2 the polynomials (19) are Hurwitz if and only if

$$a_i > 0 \text{ and } b_i > 0 \text{ for } i = 0,1,...,n-1. \qquad (20)$$

The convex linear combination (18) of the Hurwitz polynomials (19) is a Hurwitz polynomial if and only if

$$(1-k)a_i + k b_i > 0 \text{ for } k \in [0,1] \text{ and } i = 0,1,...,n-1. \qquad (21)$$

Note that the conditions (20) are always satisfied if (21) holds.

Therefore, the convex linear combination (18) of the Hurwitz polynomials (19) of the positive linear system is always the Hurwitz polynomial. □

Example 1. Consider the convex linear combination (18) of the Hurwitz polynomials

$$\begin{aligned} p_1(s) &= s^2 + 5s + 2, \\ p_2(s) &= s^2 + 3s + 4. \end{aligned} \qquad (22)$$

The convex linear combination (18) of the polynomials (22) is a Hurwitz polynomial since

$$\begin{aligned} &(1-k)5 + 3k = 5 - 2k > 0 \text{ and} \\ &(1-k)2 + k4 = 2 + 2k > 0 \text{ for } k \in [0,1]. \end{aligned} \qquad (22)$$

The above considerations for two polynomials (19) of the same order n can be extended to two polynomials of different orders (Kaczorek 2018b).

Consider the set of positive interval linear continuous-time systems with the characteristic polynomials

$$p(s) = p_n s^n + p_{n-1} s^{n-1} + ... + p_1 s + p_0, \qquad (23a)$$

where

$$0 < \underline{p_i} \le p_i \le \overline{p_i},\ i = 0,1,...,n. \qquad (23b)$$

Theorem 5. The positive interval linear system with the characteristic polynomial (23a) is asymptotically stable if and only if $\underline{p_i} > 0$ for $i = 0,1,...,n.$.

Proof. By Kharitonov Theorem the set of polynomials (23) is asymptotically stable if and only if the polynomials (16) are asymptotically stable. Note that the coefficients of polynomials (16) are positive if $\underline{p_i} > 0$ for $i = 0,1,...,n$. Therefore, by Theorem 2 the positive interval linear system with the characteristic polynomials (23a) is asymptotically stable if and only if $\underline{p_i} > 0$ for $i = 0,1,...,n.$ □

Example 2. Consider the positive linear system with the characteristic polynomial

$$p(s) = a_3 s^3 + a_2 s^2 a_1 s + a_0 \tag{24a}$$

with the interval coefficients

$$\begin{aligned} &0.5 \le a_3 \le 2,\ 1 \le a_2 \le 3, \\ &0.4 \le a_1 \le 1.5,\ 0.3 \le a_0 \le 4. \end{aligned} \tag{24b}$$

By Theorem 5 the interval positive linear system with (24) is asymptotically stable since the coefficients $a_k, k = 0,1,2,3$ of the polynomial (24a) are positive, i.e. the lower and upper bounds are positive.

STABILITY OF DESCRIPTOR POSITIVE LINEAR SYSTEMS WITH INTERVAL STATE MATRICES

Consider the autonomous descriptor positive linear system

$$E\dot{x} = Ax, \tag{25}$$

where $x = x(t) \in \Re^n$ is the state vector, $E \in \Re^{n \times n}$ is constant (exactly known) and $A \in \Re^{n \times n}$ is an interval matrix defined by

$$\underline{A} \le A \le \overline{A} \text{ or equivalently } A \in [\underline{A}, \overline{A}]. \tag{26}$$

It is assumed that

$$\det[Es - \underline{A}] \ne 0 \text{ and } \det[Es - \overline{A}] \ne 0 \tag{27}$$

and the matrix E has only linearly independent columns. If these assumptions are satisfied then there exist two pairs of nonsingular matrices (P_1, Q_1), (P_2, Q_2) such that

$$P_1[Es - \underline{A}]Q_1 = \begin{bmatrix} I_{\underline{n}_1} s - \underline{A}_1 & 0 \\ 0 & \underline{N}s - I_{\underline{n}_2} \end{bmatrix},$$

$$\underline{A}_1 \in \Re^{\underline{n}_1 \times \underline{n}_1},\ \underline{N} \in \Re^{\underline{n}_2 \times \underline{n}_2},\ \underline{n}_1 + \underline{n}_2 = n, \tag{28a}$$

and

$$P_2[Es - \overline{A}]Q_2 = \begin{bmatrix} I_{\overline{n}_1} s - \overline{A}_1 & 0 \\ 0 & \overline{N}s - I_{\overline{n}_2} \end{bmatrix},$$

$$\overline{A}_1 \in \Re^{\overline{n}_1 \times \overline{n}_1},\ \overline{N} \in \Re^{\overline{n}_2 \times \overline{n}_2},\ \overline{n}_1 + \overline{n}_2 = n, \tag{28b}$$

where $\underline{n}_1 = \deg\{\det[Es - \underline{A}]\}$ and $\overline{n}_1 = \deg\{\det[Es - \overline{A}]\}$.

Theorem 6. If the assumptions are satisfied then the interval descriptor system (25) is positive if and only if

$$\underline{A}_1 \in M_{\underline{n}_1} \text{ and } \overline{A}_1 \in M_{\overline{n}_1}. \tag{29}$$

Proof. The proof is similar to the proof of Theorem 1.

Definition 5. The descriptor interval positive system (25) is called asymptotically stable (Hurwitz) if the system is asymptotically stable for all matrices E, A, $A \in [\underline{A}, \overline{A}]$.

Theorem 7. If the matrices $\underline{A}$ and $\overline{A}$ of the positive system (25) are asymptotically stable then their convex linear combination

$$A = (1-k)\underline{A} + k\overline{A} \text{ for } 0 \le k \le 1 \tag{30}$$

is also asymptotically stable.

Proof. By condition (10) of Theorem 2 if the positive systems are asymptotically stable then there exists strictly positive vector $\lambda \in \Re_+^n$ such that

$$\underline{A}\lambda < 0 \text{ and } \overline{A}\lambda < 0. \tag{31}$$

Using (30) and (31) we obtain

$$A\lambda = [(1-k)\underline{A} + k\overline{A}]\lambda = (1-k)\underline{A}\lambda + k\overline{A}\lambda < 0 \text{ for } 0 \le k \le 1. \tag{32}$$

Therefore, if the matrices $\underline{A}$ and $\overline{A}$ are asymptotically stable and (31) hold then the convex linear combination is also asymptotically stable. □

Theorem 7. The interval descriptor positive system (25) with (26) and matrix E with only linearly independent columns is asymptotically stable if and only if there exists a strictly positive vector $\lambda \in \Re_+^n$ such that

$$P_n\underline{A}\lambda < 0 \text{ and } P_n\overline{A}\lambda < 0, \tag{33}$$

where P_n is the submatrix of P consisting of its first n rows.

Proof. If by assumption the matrix E has only linearly independent columns then $\lambda = Q\lambda_q \in \Re_+^n$ with all positive components for any $\lambda_q \in \Re_+^n$ with all positive components. By condition (10) of Theorem 2 and Theorem 6 the interval descriptor positive system (25) with (26) is asymptotically stable if and only if the conditions (32) are satisfied. □

Example 3. (Continuation of Example 1) Consider the descriptor positive system (25) with the matrix E of the form (11) and the interval matrix A with

$$\underline{A} = \begin{bmatrix} 0 & -1 & 0 & -2 \\ 1 & -3 & 0 & 2 \\ 0 & 4 & 1 & 0 \\ 1 & -1 & 0 & 2 \end{bmatrix},\ \overline{A} = \begin{bmatrix} 0 & -1 & 0 & -6 \\ 1 & -5 & 0 & 6 \\ 0 & 8 & 1 & 0 \\ 1 & -1 & 0 & 6 \end{bmatrix}. \tag{34}$$

The matrices $(E, \underline{A})$ and $(E, \overline{A})$ satisfy the assumptions (27) and the matrix E given by (11) has only linearly independent columns.

In this case

$$n=\deg\{\det[Es-\underline{A}]\}=\begin{vmatrix}0 & -1 & 0 & 2s+2\\ -1 & s+3 & 0 & -2s-2\\ s & -2s-4 & -1 & 0\\ -1 & 1 & 0 & -2s-2\end{vmatrix} \quad (35a)$$

$$=\deg(-2s^2-6s-4)=2,$$

$$n=\deg\{\det[Es-\overline{A}]\}=\begin{vmatrix}0 & -1 & 0 & 2s+6\\ -1 & s+5 & 0 & -2s-6\\ s & -2s-8 & -1 & 0\\ -1 & 1 & 0 & -2s-6\end{vmatrix} \quad (35b)$$

$$=\deg(-2s^2-14s-24)=2$$

and from (14) we have

$$P_2=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}. \quad (35)$$

Using (29) and (35) for $\lambda=[1 \quad 1 \quad 1 \quad 1]^T$ we obtain

$$P_2\underline{A}\lambda=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\begin{bmatrix}0 & -1 & 0 & -2\\ 1 & -3 & 0 & 2\\ 0 & 4 & 1 & 0\\ 1 & -1 & 0 & 2\end{bmatrix}\begin{bmatrix}1\\ 1\\ 1\\ 1\end{bmatrix} \quad (36a)$$

$$=\begin{bmatrix}-3\\ -2\end{bmatrix}<\begin{bmatrix}0\\ 0\end{bmatrix}$$

and

$$P_2\overline{A}\lambda=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\begin{bmatrix}0 & -1 & 0 & -6\\ 1 & -5 & 0 & 6\\ 0 & 8 & 1 & 0\\ 1 & -1 & 0 & 6\end{bmatrix}\begin{bmatrix}1\\ 1\\ 1\\ 1\end{bmatrix} \quad (36b)$$

$$=\begin{bmatrix}-7\\ -4\end{bmatrix}<\begin{bmatrix}0\\ 0\end{bmatrix}.$$

Therefore, by Theorem 7 the interval positive descriptor system is asymptotically stable.

CONCLUDING REMARKS

The positivity and asymptotic stability of descriptor linear continuous-time systems with interval state matrices have been investigated. It has been shown that the descriptor system is positive if and only if the matrix E has only linearly independent columns and the matrix A_1 is a Metzler matrix (Theorem 1) and the convex linear combination of the Hurwitz polynomials of positive linear systems is also the Hurwitz polynomial (Theorem 3). The Kharitonov theorem has been extended to positive descriptor linear systems with interval state matrices (Theorem 5). Necessary and sufficient conditions for the asymptotic stability of descriptor positive linear systems has been also established (Theorem 7). The considerations have been illustrated by numerical examples.

The above considerations can be extended to positive linear discrete-time systems and to fractional linear systems. An open problem is an extension of these considerations to standard (non-positive) descriptor linear systems.

ACKNOWLEDGMENT

The studies have been carried out in the framework of work No. S/WE/1/2016 and financed from the funds for science by the Polish Ministry of Science and Higher Education.

REFERENCES

Berman, A. and R.J. Plemmons. 1994.: *Nonnegative Matrices in the Mathematical Sciences*, SIAM.

Busłowicz, M. 2008. "Stability of linear continuous-time fractional order systems with delays of the retarded type". *Bull. Pol. Acad. Sci. Tech.*, vol. 56, no. 4, 319-324.

Busłowicz, M. 2012. "Stability analysis of continuous-time linear systems consisting of n subsystems with different fractional orders". *Bull. Pol. Acad. Sci. Tech.*, vol. 60, no. 2, 279-284.

Busłowicz, M. and T. Kaczorek. 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 S. Rinaldi. 2000. *Positive Linear Systems; Theory and Applications*, J. Wiley, New York.

Kaczorek, T. 1993. *Theory of Control and Systems*, PWN, Warszawa, (in Polish).

Kaczorek, T. 1997. "Positive singular discrete-time linear systems". *Bull. Pol. Acad. Sci. Tech.*, vol. 45, no. 4, 619-631.

Kaczorek, T. 2002. *Positive 1D and 2D Systems*, Springer-Verlag, London.

Kaczorek, T. 2008, "Fractional positive continuous-time linear systems and their reachability". *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 2, 223-228.

Kaczorek, T. 2010. "Positive linear systems with different fractional orders", *Bull. Pol. Acad. Sci. Techn.*, vol. 58, no. 3, 453-458.

Kaczorek, T. 2011, "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. 2013. "Application of Drazin inverse to analysis of descriptor fractional discrete-time linear systems with regular pencils". *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 1, 29-34.

Kaczorek, T. 2014. "Descriptor positive discrete-time and continuous-time nonlinear systems". *Proc. of SPIE*, vol. 9290.

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.

Kaczorek, T. 2015b. "Positivity and stability of discrete-time nonlinear systems". *IEEE 2nd International Conference on Cybernetics*, 156-159.

Kaczorek, T. 2015c. "Stability of fractional positive nonlinear systems". *Archives of Control Sciences*, vol. 25, no. 4, 491-496.

Kaczorek, T. 2016. "Analysis of positivity and stability of fractional discrete-time nonlinear systems". *Bull. Pol. Acad. Sci. Tech.*, vol. 64, no. 3, 491-494.

Kaczorek, T. 2018a. "Positivity and stability of standard and fractional descriptor continuous-time linear and nonlinear systems". *Int. J. of Nonlinear Sciences and Num. Simul.*, 2018 (in press)

Kaczorek, T. 2018b. "Stability of interval positive continuous-time linear systems". *Bull. Pol. Acad. Sci. Techn.*, vol. 66, no. 1.

Kharitonov, V.L. 1978. "Asymptotic stability of an equilibrium position of a family of systems of differential equations". *Differentsialnye uravneniya*, vol. 14, 2086-2088.

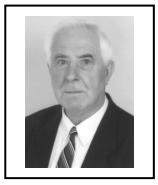
Sajewski. Ł. 2016a. "Descriptor fractional discrete-time linear system and its solution – comparison of three different methods, *Challenges in Automation, Robotics and Measurement Techniques*, Advances in Intelligent Systems and Computing, vol. 440, 37-50.

Sajewski. Ł. 2016b. "Descriptor fractional discrete-time linear system with two different fractional orders and its solution". *Bull. Pol. Acad. Sci. Tech.*, vol. 64, no. 1, 15-20.

Zhang, H., D. Xie, H. Zhang and G. Wang 2014a. "Stability analysis for discrete-time switched systems with unstable subsystems by a mode-dependent average dwell time approach". *ISA Transactions*, vol. 53, 1081-1086.

Zhang, J., Z. Han, H. Wu and J. Hung. 2014b. "Robust stabilization of discrete-time positive switched systems with uncertainties and average dwell time switching". *Circuits Syst. Signal Process.*, vol. 33, 71-95.

Xiang-Jun, W., W. Zheng-Mao and L. Jun-Guo. 2008. "Stability analysis of a class of nonlinear fractional-order systems". *IEEE Trans. Circuits and Systems-II*, Express Briefs, vol. 55, no. 11, 1178-1182.



TADEUSZ KACZOREK, born 27.04.1932 in Elzbiecin (Poland), received the MSc., PhD and DSc degrees from Electrical Engineering of Warsaw University of Technology in 1956, 1962 and 1964, respectively. In the period 1968 - 69 he was the dean of Electrical Engineering Faculty and in the period 1970 - 73 he was the prorector of Warsaw University of Technology. Since 1971 he has been professor and since 1974 full professor at Warsaw University of Technology. In 1986 he was elected a corresp. member and in 1996 full member of Polish Academy of Sciences. In the period 1988 - 1991 he was the director of the Research Centre of Polish Academy of Sciences in Rome. In June 1999 he was elected the full member of the Academy of Engineering in Poland. In May 2004 he was elected the honorary member of the Hungarian Academy of Sciences. He was awarded by the title doctor honoris causa by 13 Universities.

His research interests cover the theory of systems and the automatic control systems theory, specially, singular multidimensional systems, positive multidimensional systems and singular positive and positive 1D and 2D systems. He has initiated the research in the field of singular 2D and positive 2D linear systems and in the fractional positive linear and nonlinear systems. He has published 28 books (7 in English) and over 1100 scientific. He has given invited lectures in more than 50 universities in USA, Canada, UK, Germany, Italy, France, Japan, Greece etc. He has been a member of many international committees and programme committees.

He supervised 69 Ph.D. theses. More than 20 of this PhD students became professors in USA, UK and Japan. He is editor-in-chief of Bulletin of the Polish Academy of Sciences, Techn. Sciences and editorial member of about ten international journals.

Web-based Simulation of Production Schedules with High-level Petri Nets

Carlo Simon
Fachbereich Informatik
Hochschule Worms
Erenburgerstrasse 19, 67549 Worms, Germany
E-mail: simon@hs-worms.de

KEYWORDS

Petri Nets, Simulation Tool, Production Processes, Business Processes, Design Science Research

ABSTRACT

The paper addresses a problem, practitioners in production and logistics are faced with every day: Although several mathematical solutions including Petri net approaches exist for the simulation of complex processes, appropriate and simple to use tools do not exist. In the following, the results of a two year running research project are presented which addresses this problem. Guided by the Design Science Research approach, in this project a programming language and a programming and simulation environment for higher Petri nets has been developed, tested and evaluated. Method and tool are applied to complex simulation problems in production planning and control now. The following results had to be achieved:

1) Schedule-related data like customer orders and their prioritization can be considered now. 2) Various production strategies such as push or pull can be applied. 3) Complex data types including date and time information are supported. 4) Working time and downtime of machines can be imported from external data sources like an MES in order to calculate productivity measures like an OEE index.

The tool has been developed as a web-based system in order to support mobile use even on the shop floor. The application of language, tool and simulation results are demonstrated here introductorily at the example of a (simple) production control. The problems handled in current industry projects are more complex of course.

INTRODUCTION

The management of processes is one of the central topics in business informatics (see the remarks in (Hansen et al., 2015)): Process modeling languages are widely used to illustrate processes for analyics and optimization, as work instructions for employees, or for the specification of process aspects in software development. Popular languages for these tasks are event-driven process chains (Staud, 2006), BPMN (Allweyer, 2005), or equivalent UML languages (Randen and Fieml, 2016).

Petri nets differ significantly from these languages by their build in ability to simulate and analyze process models. For working with complex problems and models appropriate tools are needed. However, this is the core problem of Petri nets today: These tools do not exist (anymore), since the Petri net researchers of the early days developed tools mainly to answer their specific research problems but not for a broader use in engineering or software development.

The current situation is well documented looking at `https://www.informatik.uni-hamburg.de/TGI/PetriNets/tools/quick.html`, the central Petri net community website in Germany. For the 91 Petri net tools listed, the following balance can be made up:

- For 47 tools, the referenced website is no longer available or does not contain any hint on the tool anymore.
- For the next 23 tools, no development progress has been reported since 2013 or (much) earlier. Maintenance is not offered anymore and some of these tools are still MS-DOS projects.
- Another seven tools address singular research problems and can only be used by shell-scripting or as an Eclipse plug-in. They can hardly be used by others than their developers.
- In the group of the remaining 14 tools, only TimeNet supports date and time as data types (`http://www.tu-ilmenau.de/sse/timenet/`). Although the tool was still maintained until 2015, it's core bases on a project finished in 2007. The user interface is outdated and the authors do not give any hint on a mobile version or other further development.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

The simulation of production processes, however, needs tools which fulfill the following requirements:

- The ability to simulate timed high-level Petri nets where the tokens carry real-time information in addition to other freely definable attributes such as batch size or order priority.
- Theability to prioritize schedules, for example with regard to order priorities or the duration orders are already in progress.
- An API for accessing external data like order lists or the disposability of machines as a base for running simulations. Also there is a need to export the simulation results.
- The ability to access the application via mobile devices in order to answer optimization problems on site.
- An environment for defining processes specifications which at once generates visualizations of these processes.

This paper reports the current state of development of a novel Petri net specification language and of a web-based programming and simulation environment using this language. The illustrations shown in the following are taken from the tool. The technology base of this tool is JavaScript and SVG.

The research follows the Design Science Research approach of Hevner et al. (2004). The research project is beyond the prototype phase now and the first customers use the results from the tool for the optimization their business processes and the underlying business rules. These first results of these industry projects are reported here, too.

RESEARCH METHOD

According to (Hevner et al., 2004), there are seven guidelines regarding Design Science Research. In the following, these are briefly explained and it is shown how they are implemented within the framework of the project:

- The Design Science Research process must generate a viable artifact in form of a construct, model, method or description (*Design as an Artifact*): The artifact in this project is a grammar for the specification language and the implementation of a programming and simulation tool for this language.
- The goal of Design Science Research must be the development of a technology and / or methodology-based solution to a significant and relevant business problem (*Problem relevance*): Although several mathematical approaches for simulation of processes exist, practitioners in logistics and production often improvise using Excel or other calculation tools which are not developed for this specific problem. The approach presented here is generic and process-oriented at the same time. It facilitates rapid development of process models that can be simulated at once. Even more, real business and production data can be imported into the simulation very easily.
- The usability, quality and effectiveness of a design artifact must be evidenced by evaluation methods (*Design evaluation*): After the requirements of a simulation environment have been discussed with practitioners and a number of prototypes have been developed in the past, the simulation environment is now used in teaching and in its first customer projects. The achievements made are evaluated continuously and further requirements for future versions are derived from this.
- Design Science Research must provide a verifiable and well-structured contribution in the areas of Design Artifact, Design Foundations and Design Methodologies (*Research Contributions*): This requirement is met by the theoretical foundation based on Petri nets. An introduction to Petri nets can be found in (Baumgarten, 1996). Higher Petri net concepts used in the specification language and in the tool go back to the work of Genrich and Lautenbach (1981) and Lautenbach and Simon (1999).
- The result of Design Science Research relies heavily on the application of accurate and precise methods for creating and evaluating a design artifact (*Research Rigor*): Specification language and simulation environment are used and compared to alternative simulation environments, as shown in the introduction of this paper.
- The search for an effective artifact requires the use of available resources to achieve the desired goal while adhering to established guidelines within the considered problem environment (*design as a search process*): Three prototypes have already been developed in the past 2 years (see (Behnert, 2016), (Simon and Behnert, 2016), (Simon, 2017)). The current status can be regarded as the first productive system and is currently in use with industry partners from the fields of logistics, production, and automation. They use the simulation results to improve their business processes and business rules.
- Design Science Research and the resulting results must be disclosed and presented to both technology-oriented employees and the management (*Communication of*

Research): This requirement of the Design Science Research approach is fulfilled, since the simulation environment and the results that can be achieved with it are presented to the scientific community, students and industry partners.

FUNDAMENTAL LANGUAGE CONCEPTS AND EXAMPLES

Petri nets are bipartite graphs with places and transitions as nodes connected by directed arcs. Places are carriers of information and can be marked.

The first specification defines the behavior of a simple machine which is *preparing* to produce some product after an *init* event. After a *run* event the machine is actualle *producing* until a *stop* event occurs. The machine then turns into a *down* phase until the entire process *ends*. In case of an *error* event, the production is prohibited.

```
N Machine {
  T init;
  P preparing;
  T run;
  P producing;
  T stop;
  P down;
  T end;

  A (init, preparing);
  A (preparing, run);
  A (run, producing);
  A (producing, stop);
  A (stop, down);
  A (down, end);

  T error;
  A (preparing, error);
  A (error, down);

  P off;
  A (off, init);
  A (end, off);

  M (off=1);
}
```

This specification results in the Petri net of Figure 1. Its layout is generated automatically. Because of the marked place *off*, transition *init* is enabled indicated by its green color. If it occurs, the token is taken from place *off* and a token is put on *preparing*. This indicates the new state of the machine and the next events *run* or *error* may occur.

The attributes of the Petri net elements can easily be modified. This is demonstrated by the second example where the automatic layout is switched of and the positions of the nodes are specified with the aid of a grid. Furthermore, transition *error* is emphasized by using the color red when the transition is enabled.

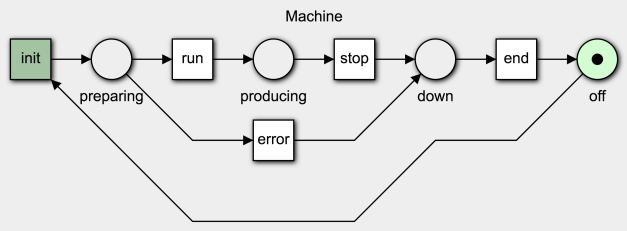


Figure 1: Petri net model of the first specification

```
N Machine(layout=false) {
  T init(col=1);
  P preparing(col=2);
  T run(col=3);
  P producing(col=4);
  T stop(col=5);
  P down(col=6);
  T end(col=7);

  A (init, preparing);
  A (preparing, run);
  A (run, producing);
  A (producing, stop);
  A (stop, down);
  A (down, end);

  T error(col=4, row=1, fillEnabled='red');
  A (preparing, error);
  A (error, down);

  P off(col=4, row=2);
  A (off, init, detour='1,2');
  A (end, off, detour='7,2');

  M (preparing=1);
}
```

The yellow color of place *preparing* indicates that this place is in conflict, i.e. either transition *run* or *error* can fire but not both of them.

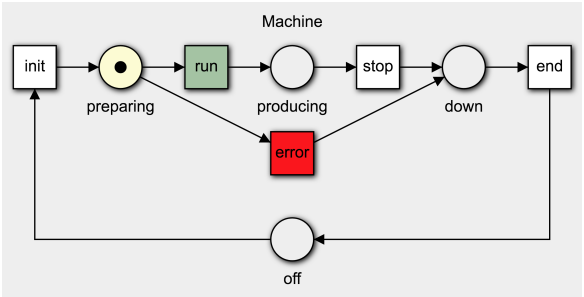


Figure 2: Petri net of the second specification

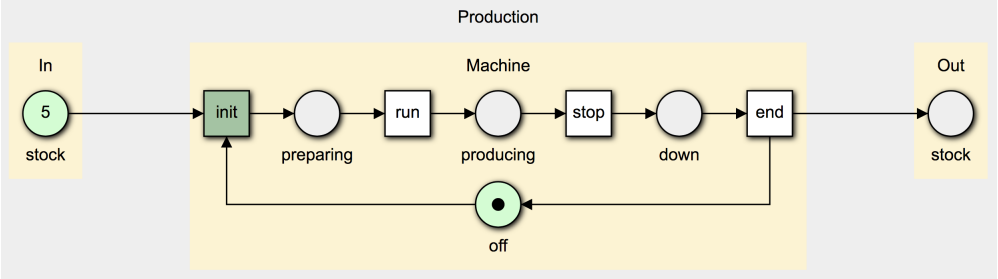


Figure 3: Petri net of two stocks and production

HIERARCHICAL MODELS

A key feature for the development of complex simulation models is the ability to group subsystems and to recycle them in another context as already stated by Stachowiak (1969). The Petri net language presented here and its specification and simulation environment allows to define hierarchical structures and to combine them as illustrated by the following example.

```
N Production {
  N In(col=1, icon='Stock') {
    P stock;
  }

  N Machine(layout=false,col=2,icon='Gear')
      {
    T init (col=1);
    P preparing (col=2);
    T run (col=3);
    P producing (col=4);
    T stop (col=5);
    P down (col=6);
    T end (col=7);
    A (init, preparing);
    A (preparing, run);
    A (run, producing);
    A (producing, stop);
    A (stop, down);
    A (down, end);
    P off (col=4, row=1);
    A (off, init, detour='1,1');
    A (end, off, detour='7,1');

    M (off=1);
  }

  N Out = In(col=11);

  A (In.stock, Machine.init);
  A (Machine.end, Out.stock);

  M (In.stock=5);
}
```

The main net *Production* consists of three subnets *In*, *Machine* and *Out* where *Out* is a copy of *In* moved into another column according to the underlying grid.

The out net is used to connect the inner nets by arcs. The name of the respective subnet is used as qualifier.

Also the initial marking of the *In* is defined in the outer net. Five tokens are put on its place which are represented as a number instead of five bullets. If the initial marking would have been defined in the definition of *In*, also this marking would have been copied to place *stock* of net *Out*.

Finally, for each of the subnets an alternative symbol is defined. The tool allows to switch between a Petri net view and a symbol view. The symbol view for this example is shown in figure 4.

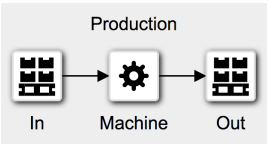


Figure 4: Alternative view on the Petri net of figure 3

MODELING OF PRODUCTION DATA

In the past, most Petri net tools have been restricted to anonymous tokens - like in the first examples of this paper - where the tokens cannot be distinguished. However then in a simulation of business or production processes important data like different orders or their priority cannot be used. Simulation is then limited to the fundamental process structure.

The Petri net language introduced here supports the definition of data types called records. A record associated with a place means that the marking of this place is restricted to tokens that correspond to this data type. This interpretation is comparable to the definition of tables in relational database systems. An example explains this concept.

The following specification begins with the definition of a record *RecordOrder* which contains an *OId* and a priority

(*Prio*), both of type *int*, and the date the order was receipt (*Receipt*) of type *date*. Both places, *oIn* and *oOut* are typed with the aid of this record definition. Moreover, the last line of the specification defines the initial marking of place *oIn*. Three tokens are put on this place.

```
N Orderreleases (layout=false) {
  R ROrder (
    OId     : int,
    Receipt : date,
    Prio    : int
  );

  P oIn(type=ROrder,col=1);
  T take(select='min(O.OId)',rcy=32,col=3);
  P oOut(type=ROrder,col=5);

  A (oIn, take, label='O');
  A (take, oOut, label='O');

  M (oIn='(1,"2018.01.24",4);
          (2,"2018.01.24",1);
          (3,"2018.01.26",2)');
}
```

Figure 5 shows the Petri net resulting from this specification. Since even a small number of individual tokens does not fit into the space given for a place, a marking is indicated by the symbol *(..)* and by colorizing the place.

Tokens in high-level Petri nets are referenced by a transition through variables annotated at the incident labels. In the example, variable *O* is used to pick up tokens from place *oIn* and to drop it on place *oOut*.

In opposite to a database, the Petri net tool presented here does not occur for sets of tokens but operates tokens in sequence. For this, a transition can be enriched by a *select*-attribute to define the order in which tokens are taken from a place. In the given example, the tokens are operated in the sequence of their *oId*.

By default, the selection condition in drawn in the center of the transition. If it does not fit into the given space, it can be moved with the aid of attribute *rcy*. Further attributes exist to modify any position of each label.

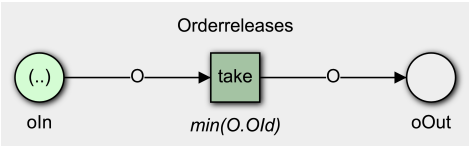


Figure 5: Typed Petri net for releasing orders

The next example demonstrates the possibility to model different selection strategies. For this record *ROrder* is reduced to two parameters: the number of an order *OId* and an identifier *CId* for the customer of this order. Place *oIn* is initially marked with three orders.

Record *RCustomer* contains two attributes: *CId* is a unique identifier for each customer and *ABC* is an indicator for her/his importance.

```
N Orderreleases (layout=false) {
  R ROrder (
    OId     : int,
    CId     : int
  );

  R RCustomer (
    CId     : int,
    ABC     : char
  );

  P oIn(type=ROrder, col=1, row=1);
  P customer(type=RCustomer,col=1,row=3);
  T take(select='min(abc)min(o)',
      col=3,row=2,rcx=16,rcy=32);
  P oOut(type=ROrder,col=5,row=2);

  A (oIn, take, label='(o,c)');
  A (customer, take, label='(c,abc)');
  A (take, customer, label='(c,abc)');
  A (take, oOut, label='(o,c)');

  M (oIn='(1,9);(2,8);(3,7)');
  M (customer='(9,"a");(8,"b");(7,"a")');
}
```

In the specification and its net shown in figure 6 a second possible way to access the tokens of a place is presented. Instead of a tuple variable which accesses an entire record at once, tuples are annotated to the arcs. This notation is used to access the single attributes of the records immediately. Label *(o,c)* of arc *oIn* $\rightarrow$ *take* references to the order number *o* and *c* to the identifier of the customer. Label *(c,abc)* of arc *customer* $\rightarrow$ *take* references to the identifier *c* of the customer and her/his priority *abc*.

Now, transition *take* is enabled, if tuples exist on both places with the same variable value for *c*. Hence a natural join of the records of places *oIn* and *customer* is built.

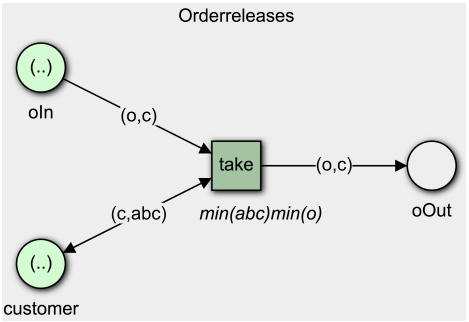


Figure 6: Releasing orders dependent on the customers' priority

Moreover this example demonstrates how to extend business rules. If the customer's priority is the main criterium for releasing an order this is expressed by putting *min(abc)* at the beginning of the selection. Since several customers might have the same priority, a second selection criterium is needed: in the example, the order number *o* is chosen.

Since the customer information must remain in the modeled system, place *customer* and transition *take* are connected by a double arc. Hence, the tuples of place *customer* are taken for the decision but are put back afterwards.

Obviously, the example is still incomplete. In a real world simulation other information like the customers' or products' names are relevant, too. The selection of these specific attributes was made to keep the example for this presentation as simple as possible.

CONDITIONAL BEHAVIOUR

The final example explains how to use transition conditions to choose the records for which a transition can occur.

```
N Orderreleases (layout=false) {
  R ROrder (
    OId     : int,
    CId     : int
  );

  R RCustomer (
    CId     : int,
    ABC     : char
  );

  P oIn(type=ROrder, col=1, row=1);
  P customer(type=RCustomer,col=1,row=3);
  T take(select='min(abc)min(o)',
      condition='abc!="x"',
      col=3,row=1,rcx=40,rcy=32);
  T leave(select='min(o)',
      col=3,row=3,rcy=32);
  P oOut(type=ROrder,col=5,row=1);

  A (oIn, take, label='(o,c)');
  A (customer, take, label='(c,abc)',
      rlx=16,rly=-16);
  A (take, customer, label='(c,abc)',
      rlx=16,rly=-16);
  A (take, oOut, label='(o,c)');

  A (oIn, leave, label='(o,c)',
      rlx=32,rly=32);
  A (customer, leave, label='(c,"x")');
  A (leave, customer, label='(c,"x")');

  M (oIn='(1,9);(2,8);(3,7)');
  M (customer='(9,"b");(8,"x");(7,"a")');
}
```

For this purpose the customer data was modified slightly. In addition to an ABC-classification of the customers a fourth classification is added which indicates suspended customers. These customers are marked with an attribute value *"x"* instead of *"a"*, *"b"*, or *"c"*.

Now, an order of a customer is taken for its release or it is left away based on the classification stored in attribute *ABC*. For this, transition *take* is extended by a condition saying that the value of variable *abc* (which corresponds to attribute *ABC*) must be different from *"x"*. For transition *leave* the selection of suspended customers is realized by using a constant in the label of *customer* $\rightarrow$ *leave*. This transition is only enabled for customers' tuples where the second attribute has value *"x"*. If it occurs, the joined order of such a customer is removed.

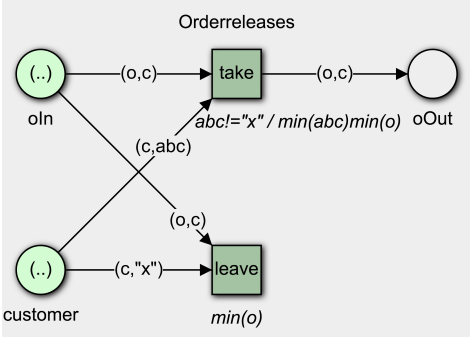


Figure 7: Conditional selection of tuples

CONCLUSIONS

What began as a small project to illustrate Petri net concepts in a Manufacturing Execution Systems course became a programming and simulation environment for Petri nets. And after a certain level of maturity was reached, the students voluntarily demonstrated the tool in their companies and reported a customers' need for this kind of a process simulation environment including a simple to use specification environment. Nowadays, the first customer projects are initiated.

Moreover, as a result of student projects and an extensive use of the tool in teaching a number of requirements have been found. The following concepts are already implemented however could not not demonstrated here in detail:

- The tool implements a layout algorithm. Its results are meaningful enough to support a rapid specification process in which users can concentrate on the internal process structure and the actual problem.
- It also implements a CSV import and export interface in order to import real business and production data

as the initial starting point for simulation and for the export of the simulation results.

- Symbols can also be used in the Petri net view and since the symbols can be animated, a meaningful process representation can be produced even for employees who are unfamiliar with process thinking.

In customers' projects the simulation environment is used to optimize business rules in logistics and production. Although some researchers assume that in such projects BPMN would be of importance, this is actually not the case. Instead of this, the mathematical foundation of Petri nets plays a more important role.

One central topic for the future development of the tool is to expand the simulation component by a controller component. In a first step, an interface for the Raspberry Pi GPIO is implemented to control machines. The further development will extend workflow management capabilities. Hence the tool will be an integrated environment for the specification, simulation and control of processes from the business level to the shop floor and back.

References

Allweyer, T. (2005): *Geschäftsprozessmanagement: Strategie, Entwurf, Implementierung, Controlling*. W3L-Verlag, Herdecke.

Baumgarten, B. (1996): *Petri-Netze: Grundlagen und Anwendungen*. Spektrum Akademischer Verlag, Heidelberg, 2nd edn.

Behnert, T. (2016): *Simulator für h¨here Petri-Netze*. Master's thesis, Provadis School of Management and Technology.

Genrich, H. J. and K. Lautenbach (1981): System Modelling with High-Level Petri Nets. *Theoretical Computer Science*, 13.

Hansen, H. R.; J. Mendling; and G. Neumann (2015): *Wirtschaftsinformatik*. De Gruyter - Oldenbourg, Berlin, 11th edn.

Hevner, A. R.; S. T. March; J. Park; and S. Ram (mar 2004): Design Science in Information Systems Research. *MIS Q.*, 28(1):75–105.

Lautenbach, K. and C. Simon (1999): *Erweiterte Zeitstempelnetze*. Fachberichte Informatik 03–99, Universität Koblenz-Landau, Institut für Informatik, Rheinau 1, D-56075 Koblenz.

Randen, H. J.and C. Bercker, van and J. Fieml (2016): *Einführung in UML: Analyse und Entwurf von Software*. Springer-Vieweg, Wiesbaden.

Simon, C. (2017): Eine Petri-Netz-Programmiersprache und Anwendungen in der Produktion. In: *Tagungsband AKWI*, eds. T. Barton; F. Herrmann; V. G. Meister; C. Müller; and C. Seel. pp. 61–70.

Simon, C. and T. Behnert (2016): Petri-Netz-Simulationen zur Theory of Constraints. In: *Tagungsband AKWI*, eds. T. Barton; F. Herrmann; V. G. Meister; C. Müller; and C. Seel. pp. 78–87.

Stachowiak, H. (1969): *Denken und Erkennen im kybernetischen Modell*. Springer, Wien, 2nd edn.

Staud, J. L. (2006): *Geschäftsprozessanalyse: Ereignisgesteurte Prozessketten und objektorientierte Geschäftsprozessmodellierung für Betriebswirtschaftliche Standardsoftware*. Springer, Berlin, 3rd edn.

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 Professor for Information Systems, first at the Provadis School of Technolgy and Management and since 2015 at the Hochschule Worms. His e-mail address is: simon@hs-worms. The tool described here can be tested on `http://www.prozesseigner.de`

Minimisation of Network Covering Services with Predefined Centres

Milos Šeda[1], Pavel Šeda[2]
[1]Institute of Automation and Computer Science,
Brno University of Technology, Technická 2, 616 69 Brno, Czech Republic
Email: seda@fme.vutbr.cz
[2]Department of Telecommunications, Brno University of Technology, Brno, Czech Republic

KEYWORDS

Set covering, unicost problem, threshold, reachability matrix, genetic algorithm, repair operator

ABSTRACT

In this paper, we deal with a special version of the set covering problem, which consists in finding the minimum number of service centres providing specialized services for all customers (or larger units, such as urban areas) within a reasonable distance given by a threshold. If a suitable threshold is found that makes it possible to determine a feasible solution of the task, the task is transformed into a general set covering problem. In order to reflect the importance of the centers, we assign weights to them and, if some centers must be contained in the result, we can either add columns in the reachability matrix with link to these centres or add special constraints in the mathematical model. However, this is of a combinatorial nature and, because it belongs to the class of NP-hard problems, for a large instance of the problem, it cannot be used to find the optimal solution in a reasonable amount of time. In the paper, we present a solution that uses two heuristic methods: genetic algorithm and tabu search.

INTRODUCTION

There are numerous discussions on how to optimise a network of public facilities (e.g. hospitals and schools) that provide essential services (health, education) for the population so that the cost of their operation is as low as possible and each inhabitant or an urban district has at least one of the service centres in an affordable distance. It is clear that the question of what is an affordable distance is debatable and could be determined by agreement of the ruling political parties. In this text, however, we ignore the political aspects and address a formal mathematical approach to solve such tasks.

In the literature, the general set covering problem is studied that does not address any threshold of availability, but it is directly given by the matrix of binary values and a covering of all columns by suitable choice of rows is looked for. This task is an NP-hard problem [3] and, for a larger problem instance, can be solved in a reasonable time only by heuristic methods.

The problem that we investigate can be converted to a set covering problem because, by using a threshold, the distance matrix is changed to binary reachability matrix. However, if the threshold is chosen inadequately, the original task may have a number of degenerative cases, described in the following section, and we will show how setting an appropriate threshold makes it possible to find a solution using genetic algorithms and tabu search.

We also present modified data structures and model to guarantee that the results could not omit important service centres. Instead of the traditional weights, which only increase the probability that the important service centres will be included in the results, we propose additional colums in the reachability matrix, or additional constraints in the model.

PROBLEM FORMULATION

Assume that the transport network contains m vertices, that can be used as operating service centres, and n vertices to be served, and for each pair of vertices v_i (considered as service centres) and v_j (serviced vertex) their distance d_{ij} is given and D_{max} is the maximum distance which will be accepted for operation between the service centres and serviced vertices (Seda and Seda, 2015).

The aim is to determine which vertices must be used as service centres for each vertex to be covered by at least one of the centres and for the total number of operating centres to be minimal.

Remark 1.
1. A condition necessary to solve the task is that all of the serviced vertices are reachable from at least one place where an operating service centre is considered.
2. Serviced vertex v_j is reachable from vertex v_i, which is regarded as an operating service centre if $d_{ij} \leq D_{max}$. If this inequality is not satisfied, vertex v_j is unreachable from v_i.

Here, $a_{ij} = 1$ means that vertex v_j is reachable from v_i and $a_{ij} = 0$ means that it is not if v_i is operating service centre i. Similarly, $x_i = 1$ means that service centre i is selected while $x_i = 0$ means that it is not selected.

Then, the set covering problem can be described by the following mathematical model:

Minimise

$$z = \sum_{i=1}^{m} x_i \tag{1}$$

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

subject to

$$\sum_{i=1}^{m} a_{ij} \cdot x_i \geq 1, j = 1, \ldots, n \tag{2}$$

$$x_i \ \in \{0, 1\}, i = 1, \ldots, m \tag{3}$$

The objective function 1 represents the number of operating centres, constraint 2 means that each serviced vertex is assigned at least to one operating service centre. The parameter D_{max} represents a threshold of service reachability.

Example 1. Consider the following *distance matrix* which expresses service centres and serviced vertices (=customer locations) and D_{max}=40. Rows are service centres and columns are serviced vertices (customer locations).

	serviced vertices (customer locations)							
	1	2	3	4	5	6	7	8
centres 1	5	41	50	26	38	60	44	59
2	49	82	13	67	68	20	32	31
3	45	17	61	45	67	48	53	127
4	37	170	195	32	77	88	90	30
5	58	42	25	101	133	32	21	78

From $D_{max} = 40$ we get the reachability matrix of serviced vertices from sevice centres.

	1	2	3	4	5	6	7	8
1	1	0	0	1	1	0	0	0
2	0	0	1	0	0	1	1	1
3	0	1	0	0	0	0	0	0
4	1	0	0	1	0	0	0	1
5	0	0	1	0	0	1	1	0

△

Special Cases

In this section, we will summarise cases for which the problem has no solution, or specified data need a modification. We will show this directly by using the below reachability matrices.

	1	2	3	4	5	6	7	8
1	0	0	1	1	0	0	0	0
2	0	0	0	0	0	1	0	1
3	1	0	0	1	1	0	0	0
4	1	0	0	0	0	0	1	1
5	0	0	1	0	0	1	0	0

In the 2nd column of the previous matrix, we can see that the threshold distance is too low and the 2nd customer has no chance to visit a centre in a reachable distance. The threshold must be increased to get at least one 1 in each column. In the 3rd row of the previous matrix, is shown that service centre 3 can be omitted because it exceeds the threshold distance to all customers and nobody would visit it.

	1	2	3	4	5	6	7	8
1	1	0	1	1	1	0	0	0
2	1	0	1	0	0	1	1	0
3	0	1	0	1	0	0	0	1
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	1	0

In the 4th row of the previous matrix, we can see that service centre 4 can be omitted because it exceeds the threshold distance to all customers and nobody would visit it.

	1	2	3	4	5	6	7	8
1	1	0	0	1	0	0	0	0
2	0	0	1	0	0	1	1	1
3	0	0	0	0	0	0	0	0
4	1	0	0	1	0	0	0	1
5	0	0	1	0	0	1	1	0

If a service centre must not be omitted, it represents only one centre for a customer, i.e., in the customer column, there is only one 1. Of course, we can have more necessary centres which cannot be omitted. However, if necessary centres cover all customers, then no centre needs to be added to the necessary ones and we immediately have a solution.

Computational Results

Since the mathematical model is simple, it seems that the problem could be solved by one of the optimisation toolboxes such as in GAMS (General Algebraic Modelling System) with the main part of code as follows:

```
LOOP(I,
      LOOP(J,
            IF (D(I,J) <= Dmax,
                  A(I,J)=1;
                ELSE
                  A(I,J)=0;
              );
         );
   );

VARIABLES
  X(I) decision variables
  XSum objective function;
BINARY VARIABLE X;

EQUATIONS
  EQ2(J) cover conditions
  EQ1   objective function (number of selected
       centres);
  EQ2(J) .. SUM(I,A(I,J)*X(I)) =G= 1;
  EQ1   .. XSum =E= SUM(I,C(I)*X(I));

MODEL COVER /ALL/;
SOLVE COVER USING MIP MINIMIZING XSum;
DISPLAY XSum.L, X.L;
```

This simple code in GAMS was tested here (and also in Excel Solver) for several cases such as pharmacies, employment offices and language schools in (Trchalíková, 2015).

However, if we apply the above procedure to minimising a network of service centres, we could get a solution where service centres in cities would be omitted.

Fig. 1: Network of employment offices.

Fig. 2: Minimal cover in a reachable distance..

The first two figures taken from (Trchalíková, 2015) show the locations of employment offices in the city of Olomouc and its surroundings and the minimal cover of this area in a threshold distance. We can see that the number of these offices may be significantly reduced, but the office in the regional centre at Olomouc has also been cancelled. Of course, this situation is undesirable and, therefore, the model needs a modification.

In this case, it is appropriate to consider the importance of locations given by their size or necessity. As the objective function is minimised, it is necessary to determine the weights so that the lower the weight, the higher the priority.

It could even be suitable to classify facilities with high importance as necessary as if they represented the only choice for at least one of the customers.

If weights of service centres are expressed by coefficients c_j, the corresponding mathematical model would change as follows:

Minimise

$$z = \sum_{i=1}^{m} c_i \cdot x_i \tag{4}$$

subject to

$$\sum_{i=1}^{m} a_{ij} \cdot x_i \geq 1, j = 1, \ldots, n \tag{5}$$

$$x_i \in \{0, 1\}, i = 1, \ldots, m \tag{6}$$

From the point of view of the problem representation and parameter settings, there is no change with the exception of the objective function, for which equation 4 is used rather than equation 1.

However, if we want to ensure that the important centres in solving the problem will never be omitted, then the safer way is to extend the reachability matrix by columns containing only a single 1 in rows corresponding to these centres. Let us assume that the regional centres 2 and 5 must be contained in the result, then we will extend the reachability matrix by two additional columns (they represent dummy serviced vertices) as follows:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left(\begin{array}{cccccccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \end{array}$$

The advantage of this approach is that the model remains the same, only with the reachability matrix adapted. However, we can achieve the same result more simply without increasing the data structures by adding constraints to the model, assigning values 1 to the corresponding decision variables, here $x_2 = 1$ and $x_5 = 1$.

Another problem is that the GAMS software tool is useable only for "small" instances as in Figure 1 and in Figure 2. All computations leading to an optimum were performed in a few seconds, but for larger instances, they ended with a run time error with GAMS indicating "insufficient space to update U-factor ...". It is caused by the fact that time complexity of the problem with m rows is $O(2^m)$ and, say, for an instance with 200 rows and 2000 columns tested in the following sections, its

searching space has 2^{200} possible selections and $2^{200} = (1024)^{20} \approx 10^{60}$.

Therefore, for these cases, heuristics must be used. Two of them, genetic algorithm and tabu search, have been implemented and recommendations for their parameter settings are presented, based on many tests with various sets of possible operators (selection, crossover, mutation, etc).

Tabu-search

Definition 1. Tabu-search is a stochastic algorithm containing following parameters (Glover, 1989, 1990):

$$TS = (M, x_0, \Theta, f, t_{max}, TL, k), \tag{7}$$

where:
- M is a solution space,
- x_0 is the initial solution. If x is determined randomly, then the local search method is the stochastic algorithm,
- Θ denotes the set of permissible transformations generating a plurality of adjacent solutions,
- f is the objective function
- TL stands for a tabu list of forbidden transformations, and
- k is the size of TL, i.e., the capacity of the short-term memory.

■

The basic version of tabu search is represented by the following pseudopascal code 1. The empty list is denoted by $\varnothing$. The symbol $\oplus$ in binary operation between two lists represents the operation of connecting the second list to the end of the first one and, vice versa, $\ominus$ removes the first symbol of TL. The ϑ_1 symbol indicates the first element of the list (Glover and Taillard, 1993).

Obviously, the size of the forbidden transformation list affects the quality of the resulting solutions. With a small k, it may occur as a climbing algorithm, but not in the adjacent two steps. With a large k on the other hand, there is a risk of skipping promising local minima, among which there might be a local minimum. One of the possible modifications to the algorithm is *adapting the length of the tabu list*. Another modification of the tabu search algorithm is using a long-term memory. In proportion to the value, transformations are penalized. There are many other tabu search modifications. One of them is the reactive tabu-search described in (Battitti and Tecchollli, 1994; Qingfu, 1993).

As to the neighbourhood operation in tabu search, we use the principle of the genetic algorithm shift mutation operator from the following paragraph with the length of the tabu list being 5.

Algorithm 1 Tabu-search

```
1:  procedure TabuSearch
2:      t ← 1;
3:      randomly select initial solution x_0t;
4:      x* ← x_0t; f_min ← f(x_0t); TL ← ∅;
5:      while t ≤ t_max do
6:          x_loc ← x_0t; f_loc ← f(x_0t);
7:          for all ϑ ∈ Θ do
8:              y ← ϑ(x_0t);
9:              if (f(y) < f_loc) and ((ϑ ∉ TL) or
    (f(y) < f_min)) then
10:                 x_loc ← y;
11:                 f_loc ← f(y);
12:                 ϑ_loc ← ϑ;
13:             end if
14:         end for
15:         if f_loc < f_min then
16:             f_min ← f_loc;
17:             f_min ← f_loc;
18:         end if
19:         if |TL| < k then
20:             TL ← TL ⊕ (ϑ_loc^-1);
21:         else
22:             TL ← TL ⊖ (ϑ_1) ⊕ (ϑ_loc^-1);
23:         end if
24:         t ← t + 1;
25:         x_0t ← x_loc;
26:     end while▷ {x_min aproximation of minimal
    cover}
27: end procedure
```

Genetic Algorithms

Since the principles of GA's are well-known, we will only deal with GA parameter settings for the problems to be studied. Now we describe the general settings and the problem-oriented setting used in our application.

Individuals in the population (*chromosomes*) are represented as binary strings of length n, where a value of 0 or 1 at bit i (*gene*) implies that $x_i = 0$ or 1 in the solution respectively.

The *population size* is usually set in the range [50, 200], in our programme, implemented in Java, 200 individuals in the population were used, because 50 individuals led to a reduction chromosome diversity and premature convergence.

Initial population is obtained by generating random strings of 0s and 1s in the following way: First, all bits in all strings are set to 0, and then, for each of the strings, randomly selected bits are set to 1 until the solutions (represented by strings) are feasible.

The *fitness function* corresponds to the objective function to be maximised or minimised; here, it is minimised.

Three of the most commonly used methods of *selection* of two parents for *reproduction*, roulette selection, ranking selection, and tournament selection, were tested.

As to *crossover*, uniform crossover, one-point and two-point crossover operators were implemented.

Mutation was set to 5, 10 and 15 %, exchange mutation, shift mutation, and mutation inspired by well-known *Lin-2-Opt change operator* usually used for solving the travelling salesman problem (Gutin and Punnen, 2007) were implemented.

In *replacement* operation two randomly selected individuals with below-average fitness were replaced by

the children generated.

Termination of a GA was controlled by specifying a maximum number of generations t_{max}, e.g. $t_{max} \leq 10000$.

Repair Operator

The chromosome is represented by an m-bit binary string S where m is the number of columns in the SCP. A value of 1 for bit i implies that service centre i is in the solution and 0 that it is not. Since the SCP is a minimisation problem, the lower the fitness value, the more fit the solution is. The fitness of a chromosome for the unicost SCP is calculated by 8.

$$f(S) = \sum_{i=1}^{m} S_i \tag{8}$$

The binary representation causes problems with generating infeasible chromosomes, e.g., in the initial population, in crossover, and/or mutation operations. To avoid infeasible solutions, a *repair operator* (Seda and Seda, 2015) is applied.

Algorithm 2 Repair Operator for Set Covering Problem

Input: $I = \{1, \ldots, m\}$ = the set of all rows; $J = \{1, \ldots, n\}$ = the set of all columns; S = the set of rows in a solution; U = the set of uncovered columns; w_i = the number of rows that cover column $j, j \in J$ in S; $\alpha_j\{i \in I \ |a_{ij} = 1\}$ = the set of rows that cover column $j, j \in J$; $\beta_i = \{j \in J | a_{ij} = 1\}$ = the set of columns that are covered by row i, $i \in I$;

1: **procedure** REPAIROPERATOR
2: initialise $w_j = |S \cap \alpha_j|, \forall j \in J$;
3: initialise $U = \{j | w_j = 0, \forall j \in J\}$;
4: **for all** column j in U (increasing order of j) **do**
5: find the first row i (in increasing order of i) in α_j that minimises $1/—U \cap \beta_i—$;
6: $S \leftarrow S + i$;
7: $w_j \leftarrow w_j + 1, \forall j \in \beta_i$;
8: $U \leftarrow U - \beta_i$;
9: **end for**
10: **for all** row i in S (in decreasing order of i **do**
11: **if** $w_j \geq 2, \forall_j \in \beta_i$ **then**
12: $S \leftarrow S - i$;
13: $w_j \leftarrow w_j - 1, \forall_j \in \beta_i$;
14: **end if**
15: **end for**
16: **end procedure** ▷ {S is now a feasible solution to the SCP and contains no redundant rows }

The initialising steps identify the uncovered columns. Since the statements are "greedy" heuristics in the sense that in the 1st **for**, rows with low cost-ratios are being considered first and in the 2nd **for**, rows with high costs are dropped first whenever possible.

RESULTS

It is obvious that the tabu search converges to a very close approximation of the optimal solution, which is consistent with the well-known "No free lunch theorem". GA started with the objective function value 1225 and finished with the best value 166, and TS found the best value 172. Because the tabu search is a one-point method, only the dependence of the objective function for gradually updated points (centres for generating neighbours) in the search space is plotted. As in the genetic algorithm, it is necessary to apply the repair operator for the selected solution in the neighbourhood, as described in the previous section. The computational time of GA for the tested instance with 200 rows was only 19 seconds on a computer with a processor frequency of 2.4 GHz and operating memory of 4 GB while SA takes more than 1 minute. The reason is that, in each GA iteration, we generate only two children while TS creates the neighbourhood with 200 neighbours in each iteration.

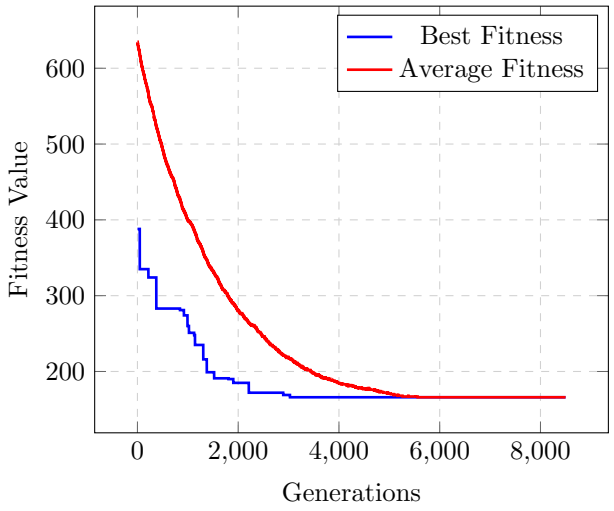


Fig. 3: Results from GA. 8500 generations, 5 % mutation, Roulette wheel selection

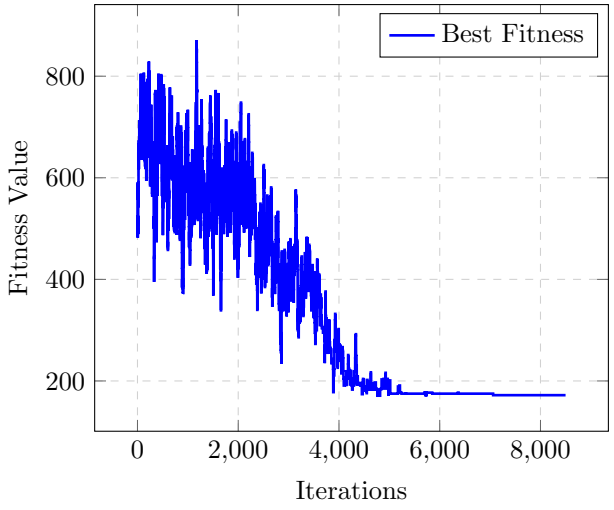


Fig. 4: Results from TS. TS – 8500 iterations

CONCLUSIONS

In this paper, we studied the set covering problem in a special case, in which a threshold is defined. This task may be used for optimising networks providing public services with operation costs being minimal.

We have shown how to increase the probability of selecting important centres with the addition of their weights, or how to directly ensure that some centres will not be missing in the result by adding columns to the reachability matrix, or by adding constraints to the model.

Due to the exponential time complexity, classical optimisation programs, often based on a branch and bound method, cannot be used to solve larger instances of (mixed-)integer programming problems. Therefore, a heuristic approach was proposed. The programme for solving this problem was implemented and parameter settings recommended based on testing many combinations of possible selections of their operators. It was shown that these methods yield very similar results when executed tens of times. In the future, we will try to implement other modern heuristic methods.

REFERENCES

R. Battitti and G. Teccholli. Simulated annealing and tabu search in the long run: A comparison on qap tasks. *Computers and Mathematics with Applications*, 28(6): 1–8, 1994. ISSN 08981221. doi: 10.1016/0898-1221(94) 00147-2. URL `http://linkinghub.elsevier.com/retrieve/pii/0898122194001472`.

F. Glover. Tabu search—part i. *ORSA Journal on Computing*, 1(3):190–206, 1989. ISSN 0899-1499. doi: 10.1287/ijoc.1.3.190. URL `http://pubsonline.informs.org/doi/abs/10.1287/ijoc.1.3.190`.

F. Glover. Tabu search—part ii. *ORSA Journal on Computing*, 2(1):4–32, 1990. ISSN 0899-1499. doi: 10.1287/ijoc.2.1.4. URL `http://pubsonline.informs.org/doi/abs/10.1287/ijoc.2.1.4`.

F. Glover and E. Taillard. A user's guide to tabu search. *ORSA Journal on Computing*, 41(1):1–28, 1993. ISSN 0254-5330. doi: 10.1007/BF02078647. URL `http://link.springer.com/10.1007/BF02078647`.

G. Gutin and A. P. Punnen. *The Traveling Salesman Problem and Its Variations*. Discrete Mathematics in Computer Science. Springer US, 2007. ISBN 978-0-387-44459-8. doi: 10.1007/b101971.

Z. Qingfu. A user's guide to tabu search. *IEEE Computational Intelligence Magazine*, 5(1):59–60, 1993. ISSN 1556-603x. doi: 10.1109/MCI.2009.935312. URL `http://ieeexplore.ieee.org/document/5386108/`.

M. Seda and P. Seda. *A Minimisation of Network Covering Services in a Threshold Distance*. In Recent Advances in Soft Computing. Advances in Intelligent Systems and Computing. Heidelberg, Germany: Springer, 2015. ISBN 978-3-319-19823-1.

P. Trchalíková. *Optimisation of Coverage of Public Services (in Czech)*. Diploma project. Přerov: VŠLG, 2015.

FINITE ELEMENT MODELLING OF PACEMAKER ELECTRODE FOR TIME VARYING EXCITATION

Shifali Kalra, M. Nabi
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India
Email: shifalikalra@gmail.com

KEYWORDS

3D modelling; Pacemaker electrode; Finite element method; Time varying excitation

ABSTRACT

In an adult person, heart normally beats 60 to 100 times in resting state. But there are cases when the heart does not beat rhythmically. This may happen either due to blockage or slow electrical conduction system of the heart. Such conditions can be fatal and thus gives rise to the need of deploying an artificial pacemaker in the patient to maintain normal rhythmic activity of the heart. An artificial pacemaker delivers electrical impulses to the heart muscles via electrode which is also called pacemaker electrode. A three-dimensional computer simulation of a pacemaker electrode for time varying excitation, has been presented in this paper for modelling the voltage and current distributions in human heart using Finite Element Method.

INTRODUCTION

An artificial pacemaker is a medical device which is used to maintain normal rhythmic activity of the heart. It consists of a pulse generator and one or more leads. The pulse generator is an electronic device which is capable of generating electric pulses for stimulating the heart muscles as given in [1][2].

There are different techniques that can be used for the analysis of the pacemaker electrode. These are the analytical method, experimental method and the numerical method. The analytical method is used for simple geometries as shown in [3] whereas, the cost of setting up the laboratory equipment is high in case of experimental method, and also it is difficult to create complex physical phantoms [4]. Keeping in view these constrains, finite element method, which is one of the numerical methods, has been preferred increasingly [5]-[9]. Analysis of the optimal electrode placement, size and electrolyte resistivity using finite element method is given in [10]. A 2D finite element model to access the current density distribution in atrial pacing is shown in [11]. 3D modelling of a pacemaker electrode with different tip size showing the effect of tip radius on the voltage distribution on the surface of the electrode and current distribution inside the heart is shown in [12].

Finite element method (FEM) is known for its flexibility and versatility. FEM uses three steps for solving a problem: pre-processing, solution and post-processing. Pre-processing involves defining geometry, application of boundary conditions and mesh generation. In mesh generation, the complex geometry is broken into a number of small elements called finite elements that are governed by the finite element equations. The finite element equations are solved to obtain the solution. Post-processing enables visualisation of results and obtaining secondary quantities. In this paper 3D modelling of pacemaker electrode is done which is better than 2D modelling since it provides more realistic results.

This paper is organised in 5 sections. The following section gives the 3D geometrical design of the pacemaker electrode along with the description of the boundary conditions. The section after that gives the modelling and meshing of the pacemaker electrode. Proceeding that there is a section that gives the explanation of the results obtained and discussion, before the conclusion section.

3D GEOMETRICAL DESCRIPTION OF PACEMAKER ELECTRODE

The Fig. 1 shows an artificial pacemaker with a pulse generator and a pacemaker electrode. The leads connect the pulse generator to the heart muscles via electrodes. These electrodes deliver electrical pulses, generated by the pulse generator, to the heart muscles in order to regulate the beating of the heart.

The geometry of a pacemaker electrode consists of a supporting cylinder at the tip of which an active electrode is placed in form of a small hemisphere. Figure 2 shows a pacemaker electrode with its surrounding modelling domain. The counter electrode is placed at the waist of the supporting cylinder. Hence, a pacemaker electrode basically consists of two electrodes, tip acts as anode and the waist acts as cathode. The geometry of pacemaker electrode also has four thin cylindrical structures extending outwards. These thin cylinderical structures are called tines. Tines get entangled to the netlike lining of the heart called trabeculae and provides passive fixation [2]. The domain around the pacemaker electrode consists of blood and tissue. The do-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

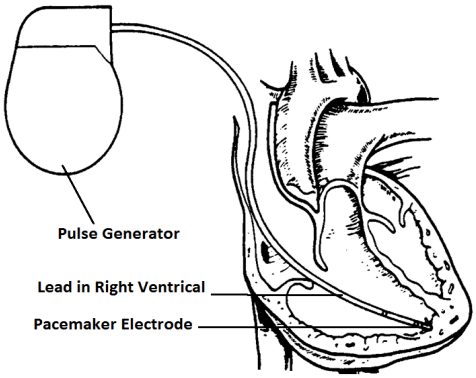


Fig. 1. Artificial pacemaker with pulse generator, lead and electrode.

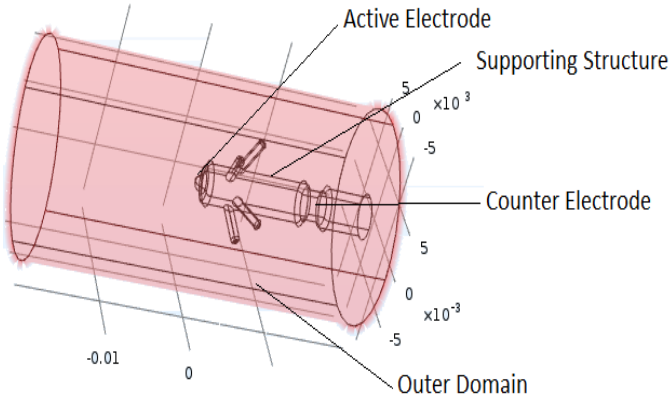


Fig. 2. Pacemaker electrode with surrounding domain.

main of the pacemaker electrode carries current which follows the Maxwell's equations

$$- \bigtriangledown .(\sigma \bigtriangledown V) = 0 \tag{1}$$

where, V is the voltage and σ is the conductivity of the human heart. The current density and the electric field are respectively given by the following two equations

$$J = (\sigma + \epsilon_o \epsilon_r \frac{\partial}{\partial t})E \tag{2}$$

$$E = - \bigtriangledown V \tag{3}$$

σ and ϵ_r are taken from the material of the heart. Conductivity, σ is taken to be $5000 S/m$ and relative permittivity, ϵ_r is taken to be 1.

A time dependent study is performed on the pacemaker electrode. The excitation is applied to the active electrode and the counter electrode is grounded. The remaining part of the geometry is electrically insulated.

MODELLING AND MESHING

Comsol Multiphysics software is used to perform 3D modelling of the pacemaker electrode [22]. Figure 3 shows a 3D model of the pacemaker electrode. The tip of the electrode i.e. the active electrode is applied with a potential of 5V for duration of 0.6ms. This value

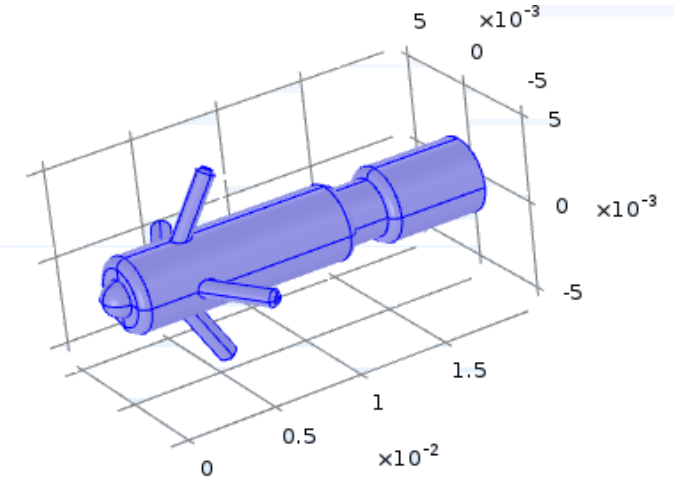


Fig. 3. 3D model of the pacemaker electrode.

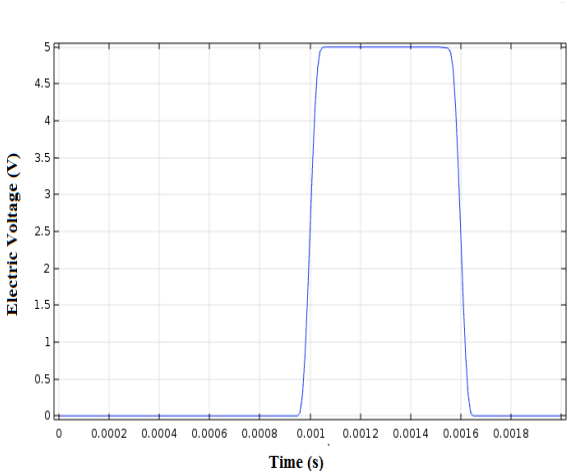


Fig. 4. Voltage pulse of duration 0.6ms applied to the electrode tip.

is chosen because modern pulse generators produce an output pulse of amplitude from 0.8 to 5V. The duration of this pulse is chosen in the interval 0.1 to 1.5ms as given in [13][14]. This pulse acts as input to the active electrode of the pacemaker. Figure 4 shows the voltage pulse applied to the active electrode. To the counter electrode which is placed at the thinner waist of the structure, ground potential is applied as boundary condition. The rest of the boundaries are electrically insulated which is given by the following equation

$$n \cdot J = 0 \tag{4}$$

where n is the outward normal. The boundaries of the outer domain as shown in Fig. 2, are placed far enough from the electrode such that it gives a very small impact on the current and potential distribution.

Once the 3D model of the pacemaker is made and appropriate boundary conditions are applied, meshing is done with tetrahedral elements. Finite element modelling was carried out using different mesh sizes. Figure 5 shows the mesh plot of pacemaker electrode with 27865 nodes.

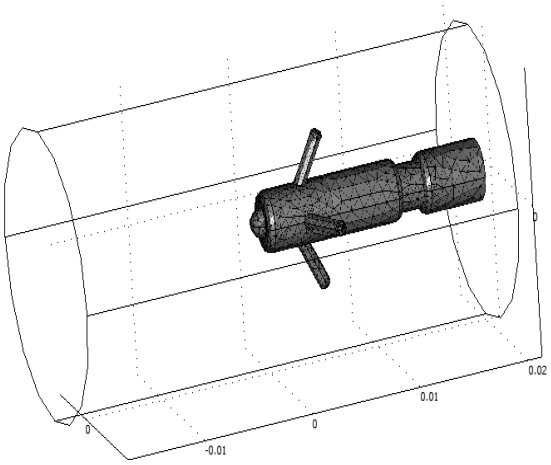


Fig. 5. Mesh plot of the pacemaker electrode with 27865 nodes.

RESULTS

3D simulation of pacemaker electrode gives voltage distribution at the surface of the electrode and the current distribution inside the heart. Figure 6 shows the potential distribution on the surface of the electrode and Fig. 7 shows the total current density by the streamlines around the pacemaker electrode. These plots have been shown at time 0.0012s, which is a representative point in time during the interval of the input pulse, as shown in Fig. 4. From Fig. 7, it can be seen that the current density is highest at the active electrode. This causes the excitation of the heart. Also, the current density is uniform at the active electrode. The counter electrode is large and therefore there is larger variation in the current density on its surface.

The middle layer of the heart muscle is called myocardium and it is responsible for conduction of electric impulses. Therefore, for optimized performance of the pacemaker electrode, the radius of the active electrode should be chosen such that there is maximum current density at the myocardium of the heart [12]. Too much increasing of the tip size would decrease the electric influence on the heart cells whereas, decreasing it too much would cause localised tissue traumas. The optimal radius of the electrode tip is from 0.8mm to 1.4mm as given in [15]. The tip radius of the pacemaker electrode is chosen to be 1mm in this paper.

Figure 8 shows few representative points numbering, 1, 2, 3, 4, and 5, at different locations on the surface of the pacemaker electrode geometry. Figure 9 shows the time history plots of these selected points. This plot resembles the voltage pulse plot that is fed to the tip of the electrode. This plot also gives a view of how voltage magnitude changes over the surface of the pacemaker electrode.

Discussion

Three-dimensional cardiac and implantable pacemaker modelling is getting popular and many researchers are adopting this technique to do various analysis such as understanding electrophysiology of the

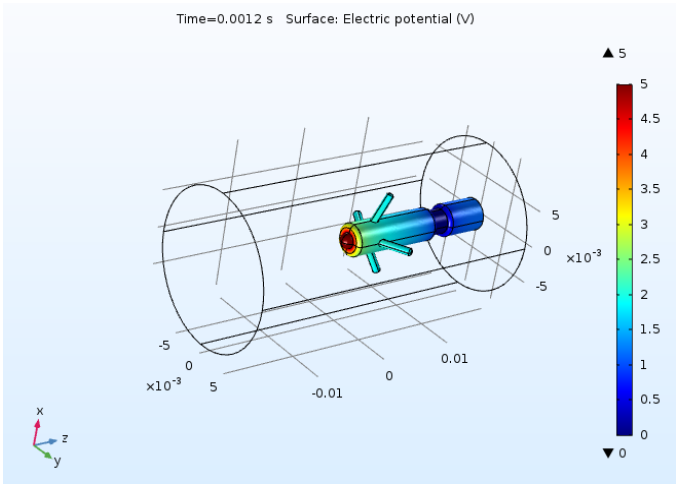


Fig. 6. Surface plot showing electric potential (V) at Time=0.0012s.

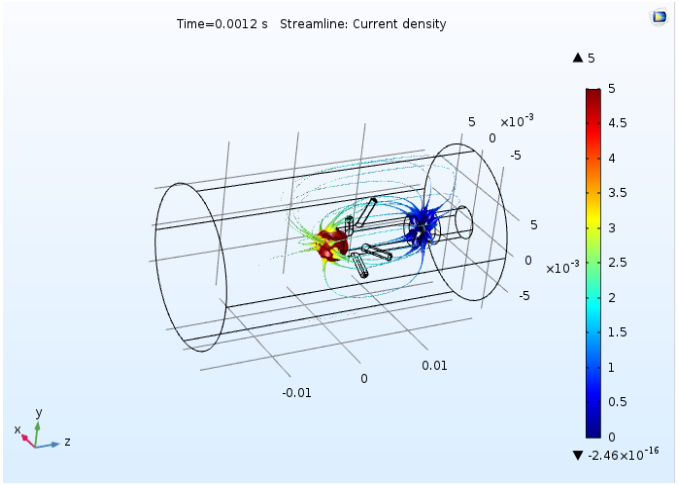


Fig. 7. Streamline plot showing current density at Time=0.0012s.

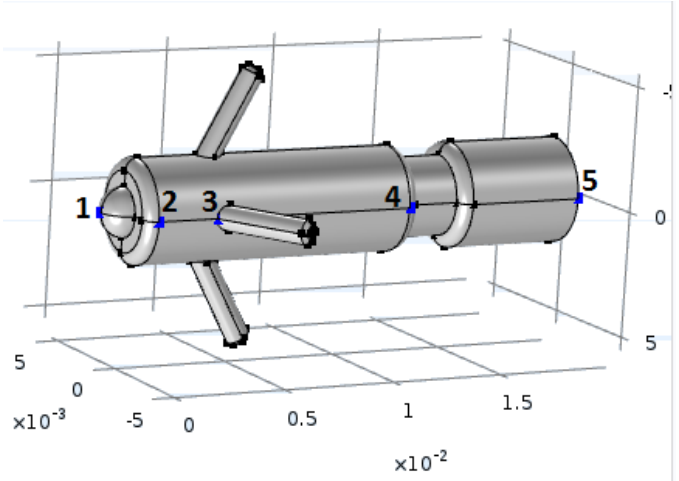


Fig. 8. Selected points on the surface of pacemaker electrode.

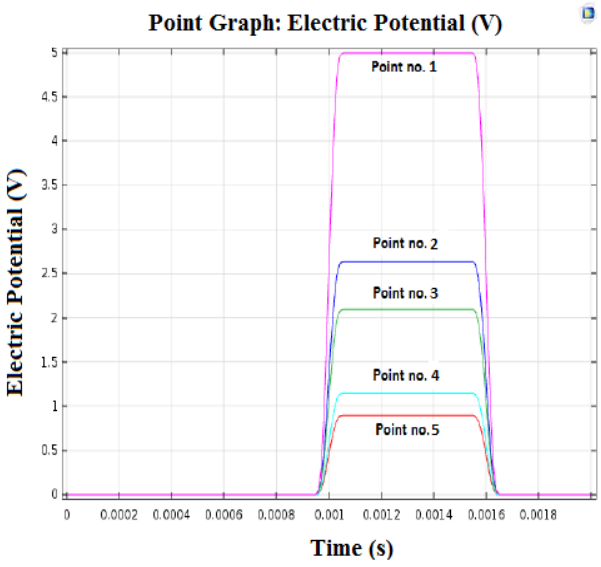


Fig. 9. Time history plot of few selected points on the geometry.

heart, complex cardiac diseases, patient specific excitations for the heart etc. [16]. The importance of developing virtual 3D heart models for understanding the structural complexity of paediatric hearts is given in [17]. It shows how the models of the human thorax can be used to predict the defibrillator's electric field intensity and the amount of current to reach the human heart. Thorough analysis and verification of a dual chamber pacemaker has been performed in [18]. It verified a dual chamber pacemaker model against a set of safety properties. In [19] a novel hybrid heart model was developed in Simulink that worked directly with the action potential signal from specific cardiac cells and was suitable for quantitive verification of implantable cardiac pacemakers. A model based framework was developed that supported quantitative verification of implantable cardiac pacemakers over hybrid heart models [20]. Use of genetic algorithms for improving pacemaker efficiency of defective heart was shown in [21].

Though, lot of work has been done on the simulation and analysis of various heart models, it is hard to find the modelling and analysis of artificial pacemakers, especially the pacemaker electrode which is actually responsible for transmitting the electrical impulses to the heart tissue. The simulation results of pacemaker electrode in this paper demonstrates the voltage and current distributions in heart. The electrode tip radius is chosen such that it produces optimum electric field distribution in the heart tissue. The consequences of choosing any pacemaker electrode tip radius have been discussed in [12].

CONCLUSIONS

Artificial pacemakers are used to regulate the heart rate by stimulating the heart cells and causing the contraction of the heart muscles. For time varying excitation, this paper gives a 3D finite element simulation of the pacemaker electrode using COMSOL Multiphysics software. In particular, it gives the potential distribution on the surface of the electrode and the current distribution inside the human heart. The voltage and current distributions are shown in the plots.

References

[1] Christoffels VM, Moorman AFM. 2009. Development of the cardiac conduction system: why are some regions of the heart more arrhythmogenic than others? Circulation: Arrhythmia and electrophysiology. 2:195–207.

[2] Stokes K. 1990. Implantable pacing lead technology. Engineering in Medicine and Biology Magazine, IEEE. 9:43–49.

[3] Gaffour L. 1998. Analytical method for solving the one dimensional wave equation with moving boundary. Journal of Electromagnetic Waves and Applications. 20(11):1429–1430.

[4] Ibrahim WMA, Algobroum HM, Almaqtari MT. 2008. Short review on the used recipies to simulate the bio-tissue at microwave frequencies. Proceedings of 4^{th} Kuala Lumpur International Conference on Biomedical Engineering; June 25–28; Kuala Lumpur, Malaysia. 21:234–237.

[5] Rao SS. 2005. The finite element method in engineering. Elsevier Butterworth-Heinemann, 4^{th} edition.

[6] Ciarlet PG, Lions JL. (1995). Handbook of numerical analysis: finite element methods (Part 2), numerical methods for solids (Part 2). North Holland. 4.

[7] Reddy J. 2005. An introduction to finite element method, 3^{rd} edition. McGraw-Hill Education.

[8] Chandrupatla TR. 2004. Finite element analysis. Engineering and Technology.

[9] Kolston PJ. 2000. Finite elememt modelling: a new tool for the biologist. Philosophical Transaction of the Royal Society A. 358:611–631.

[10] Panescu D, Webster JG, Tompkins WJ, Stratucker RA. 1995. Optimising of transcutaneous cardiac pacing by three dimensional finite element modelling of the human thorax. Medical and Biological Engineering and Computing. 33:769–775.

[11] Krasteva V. 2003. Finite element modelling approach for optimal electrode configuration in atrial pacing. Computers in Cardiology. 30:441–444.

[12] Kalra S, Marwaha S. 2013. Design and optimisation of pacemaker electrode using finite element method. International Journal of Advanced Electrical and Electronics Engineering, 2(6):125–130.

[13] Forde M, Ridgely P. 2000. Implantable cardiac pacemakers. The Biomedical Engineering Handbook: 2^{nd} Edition.

[14] Sanders RS. 2008. The pulse generator. Cardiac Pacing for the Clinician. 47–71.

[15] Wagner BK. 1995. Design of Cardiac Pacemaker. IEEE. 12:132–160.

[16] Lopez-Perez A, Sebastian R, Ferrero JM. 2015. Three-dimensional cardiac computational modelling: methods, features and applications. BioMedical Engineering Online, 14(35):1–31.

[17] Trayanova NA. 2014. Virtual 3D heart models to aid pacemaker implantation in children. Future Cardiol, 10(1):5–8.

[18] Jiang Z, Pajic M, Moarref S, Alur R, Mangharam R. 2012. Modeling and verification of a dual chamber implantable pacemaker. International Conference on Tools and Algorithms for the Construction and Analysis of Systems, March 24–april 1; Tallinn, Estonia. 4(16):188-203.

[19] Chen T, Diciolla M, Kwiatkowska M, Mereacre A. 2013. A simulink hybrid heart model for quantitative verification of cardiac pacemakers. HSCC'13, April 8–11, Philadelphia, Pennsylvania, USA. 131–136.

[20] Chen T, Diciolla M, Kwiatkowska M, Mereacre A. 2014. Quantitative verification of implantable cardiac pacemakers over hybrid heart models. Information and Computation, 236:87–101.

[21] Dumas L, Alaoui L. 2007. How genetic algorithms can improve a pacemaker efficiency. GECCO'07, July 7–11; London, England, United Kingdom, 2681–2685.

[22] Comsol Multiphysics User Guide. Available from: https//:www.comsol.co.in

Shifali Kalra is a research scholar in electrical engineering department at Indian Institute of Technology Delhi. Her research interests includes finite element modelling and advance simulation strategies based on both linear and non-linear model order reduction applied to biomedical devices and electrical circuits.

M. Nabi is Associate Professor in electrical engineering department at Indian Institute of Technology Delhi. His research interests include model order reduction of nonlinear and parametric systems, computational algorithms for modelling and simulation of distributed systems and finite element modelling of electromagnetic and coupled systems.

IMPROVED TPWL BASED NONLINEAR MOR FOR FAST SIMULATION OF LARGE CIRCUITS

Ammu Chathukulam
Debashree Sarkar
Shifali Kalra
M.Nabi
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India
Email: achathukulam@gmail.com

KEYWORDS

Model Order Reduction(MOR); Trajectory Piecewise Linearisation(TPWL); Principal Angles; Linearisation Points(LPs)

ABSTRACT

Trajectory Piecewise Linearisation(TPWL) is one of the nonlinear Model Order Reduction(MOR) techniques that involves multiple linearisations around suitably selected linearisation points(LPs). Selection of linearisation points play a major role in the accuracy of the reduced order model in TPWL technique. Standard linearisation point selection methods based on uniform distance criteria are found to be less efficient. This paper proposes a new adaptive scheme for the selection of LPs by taking into account of the 'sensitivity' of the nonlinear system towards the state. The proposed method has been implemented on two nonlinear benchmark problems, nonlinear transmission line circuit and nonlinear chain of inverter circuit and the simulation results are compared with the conventional schemes.

INTRODUCTION

Mathematical modelling of physical systems results in systems of very large order. The storage and simulation of these higher order systems is a challenge due to the huge requirement of memory and time[1]. This points towards the necessity of model order reduction(MOR) where the reduced order models with similiar input-output mappings are generated.

Common MOR methods for nonlinear systems are Proper Orthogonal Decomposition(POD)[2], DEIM(Discrete Empirical Interpolation)[3], Linear or Polynomial expansion of the system nonlinearity[4], bi-linearisation [5], or Volterra series expansion [6] and Trajectory Piecewise Linearisation(TPWL).

The idea of TPWL was first introduced by Michal Rewienski and Jacob White[7] in which it was implemented on a nonlinear transmission line circuit model and micromachined switch[8][9]. Later this method was applied to diverse areas including fluid dynamics[15],nonlinear electronic circuit simulation[17][13][14],electromagnetics[16], etc.
Since this method involves considering multiple linearisations about suitably selected states of the system, judicious location of LPs play an important role in the accuracy of the generated reduced models. The LP selection methods proposed by Rewienski[7] takes distances from the previously selected LPs as the crieteria for location of LPs on either the exact or the approximate trajectory. This method was improved by S.A. Nahvi[18] giving the idea of FAS which includes a 'coarse run' over the whole trajectory, followed by selective refinement in areas where more LPs are required.

A new method has been proposed in this paper that adaptively decide the placement of the LPs along the trajectory. This is inspired by[11], where the authors have used adaptive sampling technique in parametric linear MOR with the example of a linear beam. In this paper, this adaptive sampling idea is incoorporated into TPWL method to locate the LPs. The regions where the nonlinear system is more sensitive to the state are identified based on the concept of distance between the subspaces or principal angles.The proposed method has been implemented on two nonlinear benchmark problems,nonlinear transmission line circuit and nonlinear chain of inverter circuit and the simulation results are compared with the conventional schemes.

The next two sections describe the basic TPWL process and the conventional methods for LP selection and their drawbacks respectively. The last section introduces the proposed adaptive scheme followed by its implementation on benchmark problems and error comparison.

TRAJECTORY PIECEWISE LINEARISATION

The TPWL process includes the following steps:

- The higher order system is simulated for a training input to obtain the nonlinear system trajectory.
- Location of suitable Linearisation Points along the exact or approximate trajectory[7] followed by the sub-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

sequent linearisations around these LPs.
• Finding out the dominant subspace and projecting the local-submodels into this dominant subspace.
• Combine the local reduced models using proper weight assignment stategies to form the final reduced model.

Mathematical Formulation

For the nonlinear dynamical system in the following state space form:

$$\frac{dx}{dt} = f(x) + Bu \quad (1)$$
$$y = Cx \quad (2)$$

Where $x \in R^n$ is a vector of system states, $f : R^n \rightarrow R^n$ is the nonlinear vector field, $B \in R^{n \times p}$ is the input matrix, $C \in R^{q \times n}$ is the output matrix, $u \in R^k$ is the input and $y \in R^q$ is the output. The Taylor series first-order approximation of $f(x)$ about an initial state x_0 is given by

$$\overline{f}(x) = f(x_0) + A_0(x - x_0) \quad (3)$$

Where A_0 is the Jacobian of $f(x)$ evaluated at x_0. Clearly, the error in this approximation is:

$$e(x) = f(x) - \overline{f}(x) \quad (4)$$

The linearised system at x_0 is hence

$$\frac{dx}{dt} = f(x_0) + A_0(x - x_0) + Bu \quad (5)$$
$$y = Cx \quad (6)$$

The dynamics are restricted to the dominant subspace by the state transformation $x = Vz$. $V \in R^{n \times r}$ is the projection matrix spanning the dominant subspace, $z \in R^r$, $V^T V = I$ and $r \ll n$. The reduced model at x_0 is:

$$\frac{dz}{dt} = A_{0r}z + V^T(f(x_0) - A_0x_0)) + B_ru \quad (7)$$
$$y = C_rz \quad (8)$$

where $A_{0r} = V^T A_0 V, B_r = V^T B$, and $C_r = CV$. Assuming m similar linearised models generated along the training trajectory about the states $x_0, ..x_i, ..x_{m-1}$, the final TPWL model is expressed as the weighted sum:

$$\frac{dz}{dt} = \sum_{i=0}^{m-1} w_i(z)(A_{ir}z + V^T(f(x_i) - A_ix_i)) + B_ru \quad (9)$$
$$y = C_rz \quad (10)$$

where A_i is the Jacobian of $f(x)$ evaluated at x_i and $A_{ir} = V^T A_i V$.

CONVENTIONAL METHODS FOR LP SELECTION

Uniform Division of Trajectory

[10]The first proposed and most widely used methods for LP selection in TPWL based MOR are the ones which divide the system trajectory uniformly to create linearised systems. Two such methods, very similar to each other, but differing in whether the nonlinear system trajectory is exact or approximate, are given in [7]. The first method can be found in [7, p.45]. In it, given a training input $u(t)$ and an initial state x_0, the nonlinear system is simulated and a finite number of linear systems are generated on its trajectory. The linear systems are created by a uniform division of the trajectory at a constant, pre-selected euclidean distance δ. The second method given in [7, p.53], is similar to first, but it circumvents simulating the full-order nonlinear system. It is called Fast Approximate Simulation (FAS), and in it successively selected sub-models of the nonlinear system are simulated to get an approximate trajectory.

The uniform division of the trajectory suffers from two major drawbacks as pointed out in[12]. The first one is the heuristic nature of the δ being chosen. Selection of δ is mainly a game of 'hit and trial' and is unrelated with the nonlinearity involved. Moreover, these algorithms rely on the fact that the linearisation of a nonlinear function f at the state x_i is an accurate enough approximation at some other state x, provided that x is close enough to x_i, i.e., $||x - x_i|| < \delta$, or x lies within a ball of radius δ centered at x_i. The procedure contains an implicit assumption that dividing the trajectory uniformly would lead to acceptable sampling. In other words, it is assumed that δ is a constant throughout the trajectory, which is not necessarily true. There can be situations in which the trust regions don't overlap resulting in inadequate sampling and the trust region around one point covering the other trust region resulting in over-sampling as shown in Fig[1],[2].

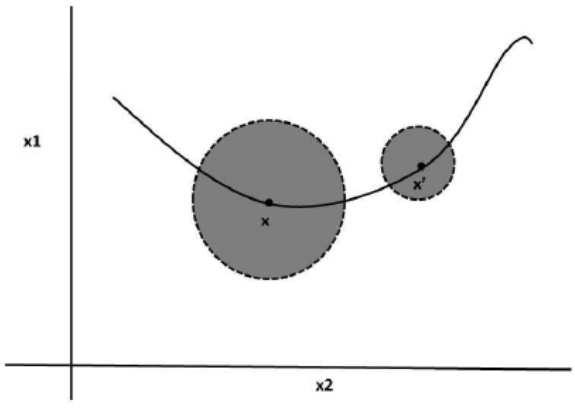


Fig. 1. The trajectory is inadequately sampled between the two trust regions

PROPOSED METHOD FOR ADAPTIVE SELECTION OF LPS

Uniform Division of the trajectory has proven to be a less efficient method for LP selection because of the above mentioned reasons. This section presents an algorithm to adaptively decide the number and placement of the linearisation points along the trajectory by

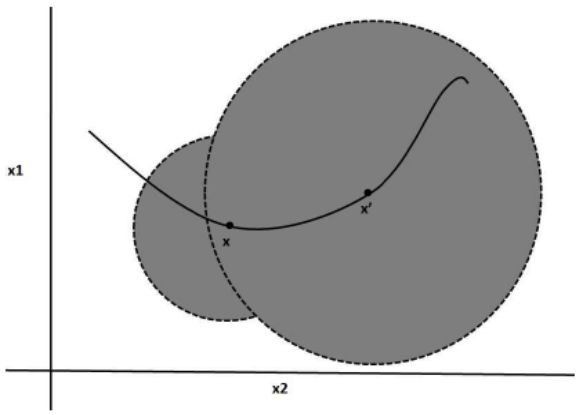


Fig. 2. The trust region around x covers x^1

placing more LPs in highly sensitive zones and lightly gridding the less sensitive zones. The high and low 'sensitive' regions are identified based on the concept of distance between the subspaces or principal angles.

Principal Angle Concept

Let V_1 and V_2 be the orthonormal projection matrices at the linearization points x_1 and x_2. The largest subspace angle across these two points is computed as

$$\theta_{12} = arcsin(\sqrt{(1-\sigma_r^2)}) = arccos(\sigma_r)$$

σ_r is the smallest singular value of $V_1^T V_2$.[11]

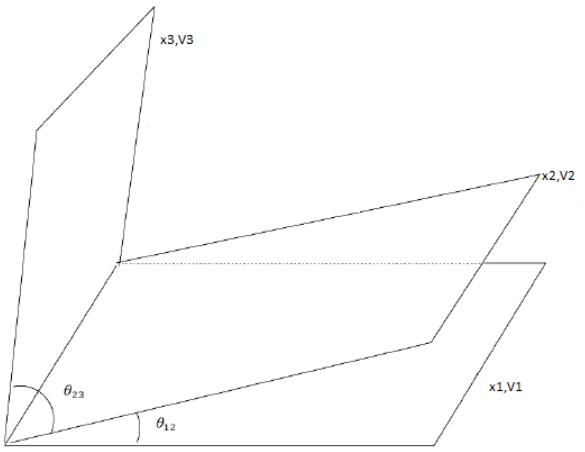


Fig. 3. Principal angles between V_1, V_2, V_3

Higher the angle between the subspaces, higher the sensitivity of the ROM and hence more LPs should be considered. Hence we have exploited the concept of subspace angles as a crieteria for deciding the selection the linearization points.This method hence assures more judicious placement of the LPs when compared with the uniform trajectory division.The new scheme for adaptive selection of LPs first perform a rough uniform sampling by selectiong a very high value of δ followed by the refinement in those regions where the system is more sensitive to state. In other words, the method adds more LPs to those regions where the subspace angle is more than θ_{max}, the maximum tolerable subspace angles. The proposed adaptive sampling algorithm is given in the following section.

Automatic Adaptive Sampling Algorithm

- Input θ_{max}, maximum error between the subspace angles that can be tolerated.
- Select linearization points $x_1, x_2, ...x_k$ by initial rough sampling by selecting a very high value for δ.
- While all $l_{i,i+1} > 1$

a) Calculate the projection matrices $V_1, V_2 ... V_k$ corresponding to each of these values $x_1, x_2, ...x_k$.

b) Compute subspace angles $V_{12}, V_{23}, ...V_{k-1,k}$ between these V_is, each taken pairwise.

c) Calculate $l_{12} = \frac{\theta_{12}}{\theta_{max}}, l_{23} = \frac{\theta_{23}}{\theta_{max}}, ...$

d) Divide the interval between x_1 and x_2 into l_{12} further intervals. Likewise do the same for all the other intervals.

IMPLEMENTATION ON NONLINEAR BENCHMARK PROBLEMS

The adaptive method for LP selection is implemented on the two nonlinear bench mark problems ie, Nonlinear Transmission Line and Inverter Chain Circuit. The results are compared with the LP selection method using uniform division of trajectory.

Example-1 : Nonlinear Transmission Line

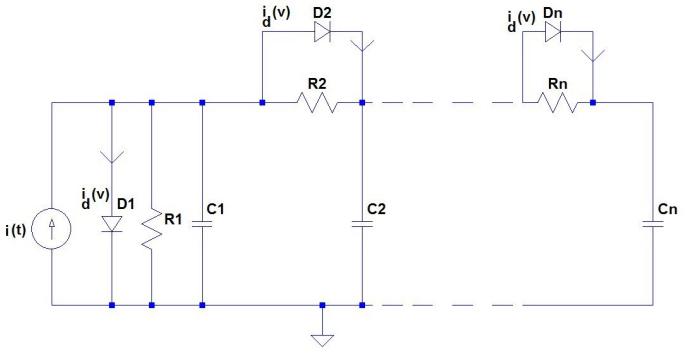


Fig. 4. Nonlinear transmission line circuit

The first example considered is a nonlinear transmission line circuit model shown in Figure 4. The circuit consists of resistors, capacitors, and diodes with a constitutive equation $i_d(v) = exp(40v) - 1$, where v is the voltage between diodes terminals. For simplicity we assume that all the resistors and capacitors have unit resistance and capacitance, respectively $(R_1 = R_2 = ... = R_n = 1, C_1 = C_2 = ... = C_n = 1)$. The input is the current source entering at node 1: $u(t) = i(t)$, and the output is chosen to be the voltage at node 1: $y(t) = v_1(t)$. Using constitutive relations for capacitors, resistors, and diodes, as well as kirchhoffs current law, we obtain an input affine nonlinear dynamical system given by equations 1 and 2. This

higher order nonlinear circuit is reduced using TPWL method with adaptive selection of LPs and the results are compared with the uniform trajectory division case. In both the cases the reduced order and the number of LPs are kept same. In this case, the value of δ is 1 for uniform trajectory division and while using adaptive case the trajectory is first roughly sampled with δ=3 and then used the adaptive algorithm keeping maximum error tolerance angle, θ_{max} as 5°. As a result, the number of LPs came out to be 5 in both the cases. The results show that the proposed method gives only 4.2% error at the output voltage compared to 6.74% error in the conventional case.

Error Analysis

The error in the output voltage between Full Order Model(FOM) and Reduced Order Model(ROM)

$$e(t) = (y_r(t) - y(t)) \tag{11}$$

where y_r and y are the output voltages of ROM and FOM respectively. The absolute error for the entire time span can be calculated as

$$E = \int_0^T |e(t)|dt \tag{12}$$

Using the above equations, the absolute error for uniform trajectory division, $E_u = 0.6233$ V while in the adaptive case, the error is $E_a = 0.3129$ V, clearly showing $E_a < E_u$.
The output voltages for both the cases are plotted in Fig 5 and 6 along with the FOM response. It is clearly visible from the graphs that the error peak during time 3.5 sec to 4.5 sec is reduced drastically in adaptive sampling case.

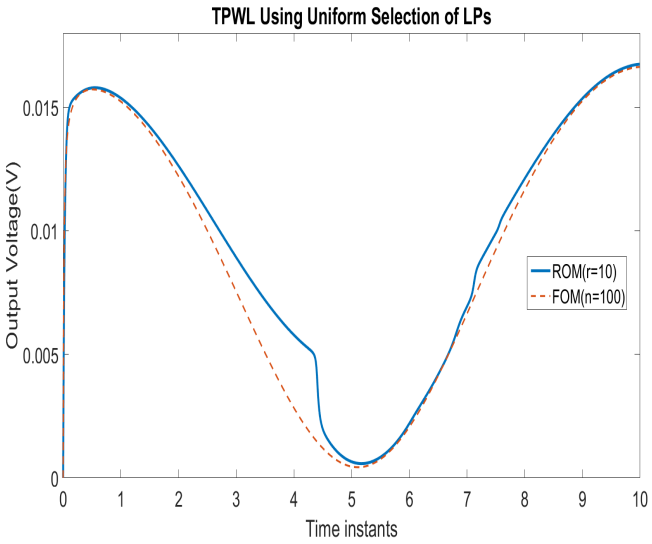


Fig. 5. TPWL Using Uniform Division of Trajectory for Nonlinear Transmission Line Model

Example-2 : Inverter Chain Circuit

The inverter circuit shown in Fig.7 is a nonlinear circuit consisting of an input voltage source, capacitors,

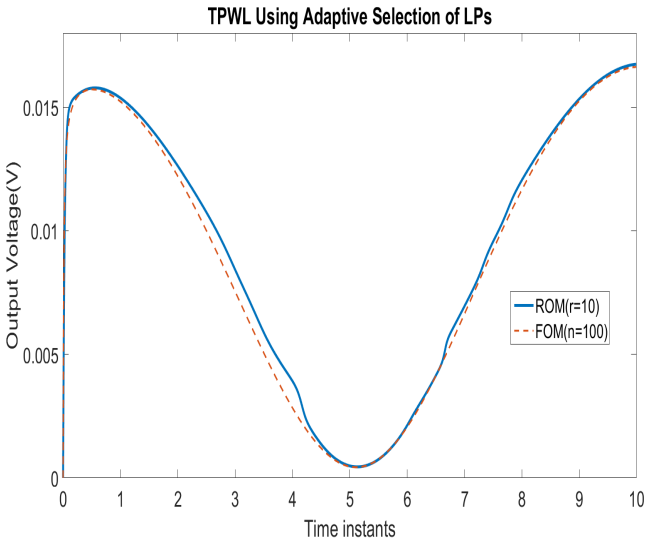


Fig. 6. TPWL Using Adaptive Selection of LP for Nonlinear Transmission Line Model

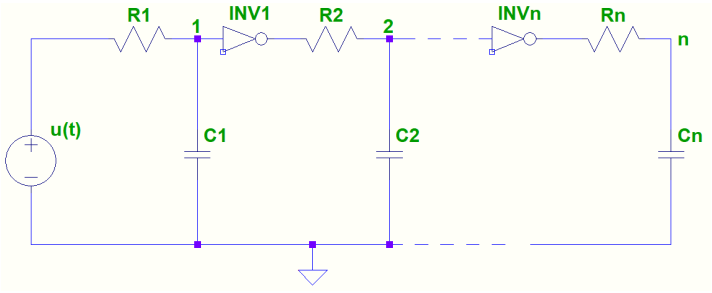


Fig. 7. Inverter Chain Circuit

resistors, and inverters whose input-output relation is

$$V_{out} = f(V_{in}) = V_{dd}tanh(AV_{in}) \tag{13}$$

where V_{dd} is the supply voltage and A is a parameter. We choose $V_{dd} = 1$ and $A = 5$. Input is the voltage source $u(t)$ and output is the voltage at node 1, $v_1(t)$. The states are the corresponding voltage at each node. The parameter values are selected as $R_1 = R_2 = ... = R_n = 1\,\Omega, C_1 = C_2 = ... = C_n = 1\,\text{F}$. With the above given values the circuit has input affine form as given by equations 1 and 2. This higher order nonlinear circuit is reduced using TPWL method with adaptive selection of LPs and the results are compared with the uniform trajectory division case. In both the cases the reduced order and the number of LPs are kept same. In this case, the value of δ is 6 for uniform trajectory division and while using adaptive case the trajectory is first roughly sampled with δ=9 and then used the adaptive algorithm keeping maximum error tolerance angle, θ_{max} as 10°. As a result, the number of LPs came out to be 4 in both the cases. By using the proposed method we are able to reduce the percentage error at the output from 11.5% to 9% without increasing the number of LPs. Hence we can conclude that, in this case also the adaptive sampling with $E_a = 36.2522$ V gives a less erroneous ROM as compared to uniform division with $E_u = 54.8893$ V.

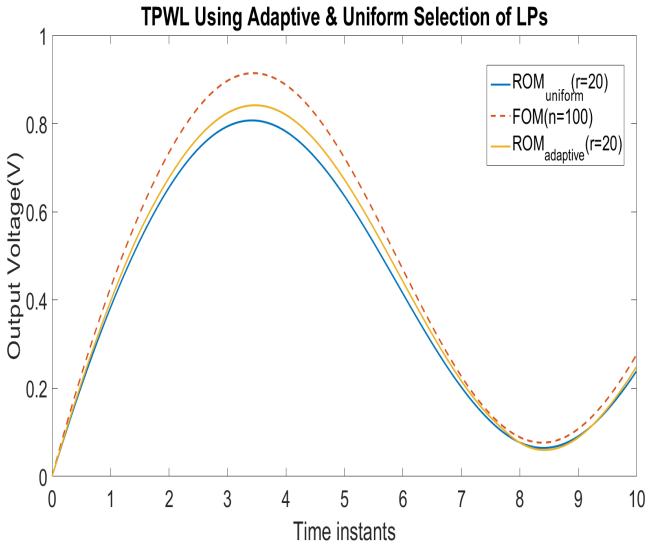


Fig. 8. TPWL using Adaptive Selection of LPs

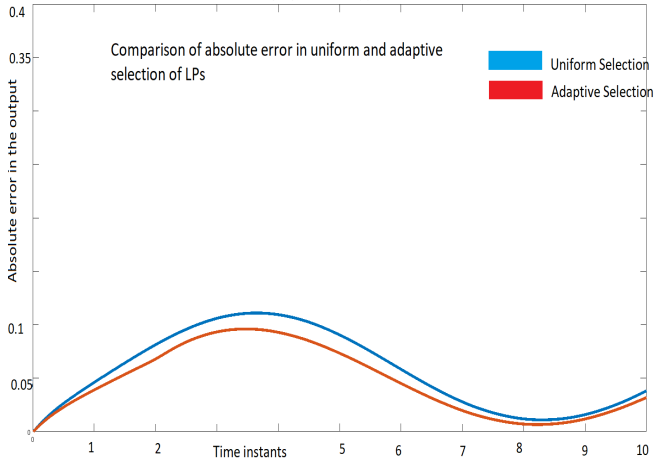


Fig. 9. Absolute error in the output voltage between the FOM and ROM for adaptive LP selection method and uniform trajectory division method

	NL transmission line		Chain of inverters	
	Adaptive	Uniform	Adaptive	Uniform
Full Order	100	100	100	100
Reduced Order	10	10	20	20
% Error in Output	4.2	6.74	9	11.5
No of LPs	5	5	4	4
Online Simulation Time	0.1s	0.1s	0.1s	0.1s

Table 1: Comparison of proposed method with uniform selection of LPs for nonlinear transmission line and chain of inverter circuit.

CONCLUSION

This paper presents an improved TPWL method for nonlinear MOR in which the LPs are placed more accurately. The adaptive LP selection algorithm takes into account of the 'sensitivity' of the nonlinear system towards state removing the uncertainities associated with uniform division methods. Hence it is able to reduce the percentage error in the output between FOM and ROM compared to unform trajectory division by judiciously locating the LPs on the trajectory.The new method is thus able to develop more efficient reduced order models without increasing the number of linearisation points as seen from the simulation results of the two nonlinear benchmark problems. The proposed method can be applied to more nonlinear higher order systems and hence fast simulation of such systems can be achieved more efficiently.

REFERENCES

[1] P. Benner, "Solving large-scale control problems," *IEEE Control Systems Magazine*, vol.24, no.1, pp. 44-59, Feb. 2004.

[2] P. Astrid, *Reduction of Process Simulation Models: A Proper Orthogonal Decomposition Approach*, PhD thesis, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, 2004.

[3] S. Chaturantabut, D.C. Sorensen, *Discrete empirical interpolation for nonlinear model reduction*, Report TR09-05, Rice University (2009)

[4] Y. Chen, *Model Order Reduction for Nonlinear Systems*, M.S. dissertation, MIT, Cambridge, MA, 1999.

[5] Z. Bai, "Krylov subspace techniques for reduced-order modeling of largescale dynamical systems", *Applied Numerical Mathematics*, vol. 43, no. 1-2, pp. 9-44, 2002.

[6] J.R. Phillips, "Projection frameworks for model reduction of weakly nonlinear systems", *Proc. Design Automation Conference*, vol. 37, pp.184-189, 2000.

[7] M. Rewienski, *A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems*, PhD Thesis, MIT, Cambridge, MA, June 2003.

[8] M. Rewienski and J. K. White, "A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices", *Proc. of IEEE/ACM International Conference on Computer Aided-Design*, pp. 252-7, San Jose, CA, USA, November 2001.

[9] M. Rewienski and J. White, "A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices", *IEEE Trans. Computer-Aided Design*, vol. 22, no. 2, pp. 155-70, Feb 2003.

[10] Shahkar Ahmad Nahvi, Mashuq-un-Nabi and S. Janardhanan, "Trajectory Based Methods for Nonlinear MOR: Review and Perspectives", *Proc. of the 2012 IEEE International Conference on Signal Processing, Computing and Control*, Shimla, India, March 2012, pp. 1-6.

[11] Maria Cruz Varona, Mashuq-un-Nabi and Boris Lohmann, "Automatic Adaptive Sampling in Parametric Model Order Reduction by Matrix Interpolation", *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, Sheraton Arabella Park Hotel, Munich, Germany, July 3-7, 2017

[12] Shahkar Ahmad Nahvi, *Reduced Order Modelling and fast simulation Strategies for nonlinear Dynamical Systems*, Phd Thesis, IITD

[13] T. Voss, R. Pulch, E. Maten, and A. Guennouni, "Trajectory piecewise linear approach for nonlinear differential-algebraic equations in circuit simulation", *Scientific Computing in Electrical Engineering, Gabriela Ciuprina and Daniel Ioan*, Eds., vol. 11 of Mathematics in Industry, pp. 167173. Springer Berlin Heidelberg, 2007.

[14] Michael Striebel and Joost Rommes, *Model Order Reduction of Nonlinear Systems in Circuit Simulation: Status and Applications*, Chapter in Lecture Notes in Electrical Engineering January 2011

[15] David Gratton, *Reduced-order, trajectory piecewise-linear models for nonlinear computational fluid dynamics*, MS Thesis, MIT

[16] M. Nassar Albunni,Volker Rischmuller,Thomas Fritzsche,and Boris Lohmann, "Model-Order Reduction of Moving Nonlinear Electromagnetic Devices " *IEEE Transactions on Magnetics*, Vol. 44, No. 7, July 2008

[17] Xiaoda Pan, Fan Yang, Xuan Zeng and Yangfeng Su, "An Efficient Transistor-Level Piecewise-Linear Macromodeling Approach for Model Order Reduction of Nonlinear Circuits", *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2010

[18] S.A. Nahvi, M. Nabi and S. Janardhanan, "Adaptive sampling of nonlinear system trajectory for Model Order Reduction", *Proceedings of 2012 International Conference on Modelling, Identification and Control*, Wuhan, China, June 24-26, 2012

AMMU CHATHUKULAM
Ammu Chathukulam is currently pursuing M.Tech in Control and Automation from Indian Institute of Technology Delhi. Her research field is nonlinear model order reduction using trajectory piecewise linearisation.

DEBASHREE SARKAR
Debashree Sarkar got her M.Tech in Control and Automation at Indian Institute of Technology Delhi. Her research interests includes reduced order modelling and its applications in circuit simulation. She is currently working in Airport Authority of India.

SHIFALI KALRA
Shifali Kalra is working as research scholar in Department of Electrical Engineering at Indian Institute of Technology Delhi. Her research interests includes linear and nonlinear model order reduction.

M.NABI
M.Nabi is currently working as Associate Professor in IIT Delhi. His research includes Control Systems, Guidance and Control, Computational methods for Modeling, Simulation and Control, Finite Element Method, Distributed Parametered Systems, Flexible Structures,Electromagnetic and Coupled Systems, Electromagnetic NDT.

Diode Model Generation for Simulation of Harmonic Distortion

Jennifer Schütt
Nexperia Germany GmbH
Stresemannallee 101
22529 Hamburg
Germany
jennifer.schuett@nexperia.com

Jens Werner
Jade University
of Applied Sciences
Friedrich-Paffrath-Str. 101
26389 Wilhelmshaven
Germany
jens.werner@Jade-HS.de

Ayk Hilbrink
Nexperia Germany GmbH
Stresemannallee 101
22529 Hamburg
Germany
ayk.hilbrink@nexperia.com

KEYWORDS
SPICE equivalent device model, electrostatic discharge, ESD protection, harmonic distortion

ABSTRACT

This paper presents the stepwise SPICE model generation for Transient Voltage Suppression (TVS) diodes allowing to represent the 2nd and 3rd harmonic distortion behavior in circuit based simulations. The model is based on semiconductor physics and generated using measurement data of the device. The generation of this model is based on the I/V characteristics, the behavior of the capacitance versus diode voltage variations, and tuning at one single input power level at the fundamental frequency as well as for the 2nd and 3rd harmonics. The simulation tool used for the device modelling is Keysight Avanced Design System (ADS).

INTRODUCTION

Harmonic Distortion (HD) is a well-known phenomenon in radio frequency (RF) applications like amplifiers and receivers. It is part of standard device characterization to look at the frequency spectrum at the device output. Trends in the market for TVS diodes show rising requests for diodes with optimized HD behavior. Frequencies of interest are mainly found in wireless applications like GSM (900 MHz, 1800 MHz) and WiFi (2.4 GHz, 5 GHz), UMTS (3G) or LTE (4G).
In general, TVS Diodes are made for overshoot protection of supply lines as well as for data lines of interfaces, e.g. like USB, DVI or HDMI. The development of TVS diodes for supply lines in the past had only limited attention on the frequency behavior. To predict the result of design improvements of an existing product, the easiest way is to model the device first and simulate potential improvements. In that manner, a new field of modelling emerged. This paper presents the first steps of this modelling and simulation process.

HARMONIC DISTORTION

According to (Pozar 2011) and (Maas 1988) the HD phenomena can be described by an input voltage v_i and an output voltage v_o which results according to the Taylor series in equation (*1*):

$$v_0 = a_0 + a_1 v_i + a_2 v_i^2 + a_2 v_i^3 + \cdots \quad (1)$$

Considering v_i being a single frequency sinusoid:

$$v_i = V_0 \cos(\omega_0 t) \quad (2)$$

The equations (1) and (2) lead after some transformation to:

$$\begin{aligned} v_0 = &\left(a_0 + \frac{1}{2} a_2 V_0^2\right) \\ &+ \left(a_1 V_0 + \frac{3}{4} a_3 V_0^3\right) \cos \omega_0 t \\ &+ \frac{1}{2} a_2 V_0^2 \cos 2\,\omega_0 t \\ &+ \frac{1}{4} a_3 V_0^3 \cos 3\,\omega_0 t + \cdots \end{aligned} \quad (3)$$

The harmonics in general are related to non-linear dependencies, e.g. the quadratic, cubic and higher terms in equation (1). From equation (3) the generation of second harmonic waves caused by a non-linear device can be identified as:

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

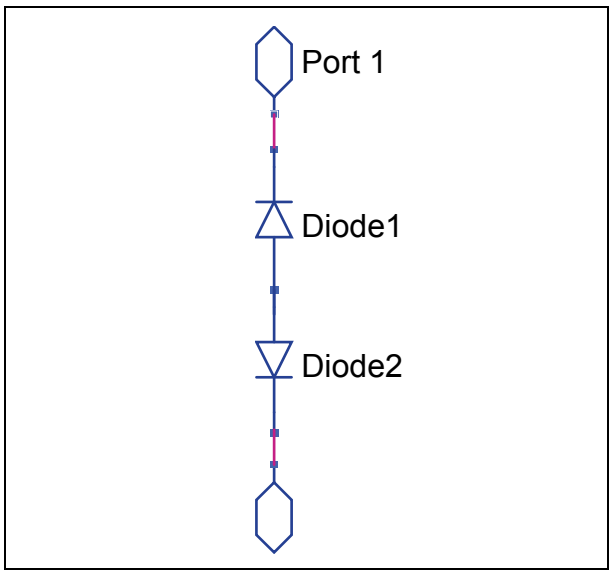


Figure 1: Configuration of a Bi-directional TVS Protection Device.

$$\frac{1}{2}a_2 V_0^2 \cos 2\,\omega_0 t \qquad (4)$$

And the third harmonic wave is

$$\frac{1}{4}a_3 V_0^3 \cos 3\,\omega_0 t \qquad (5)$$

In this paper harmonic distortion of higher orders than three will be neglected. This is reasonable since for higher orders of HD components the magnitude of those components is usually decreasing for increasing order. Nevertheless, an extension of the modelling process to consider higher harmonic orders is straight-forward. In wireless applications the second harmonic and the third harmonic waves are of rather high importance: E.g. for any GSM-850 mobile phone the specified uplink spectrum is 824 – 849 MHz (3GPP, 2005). Therefore any unwanted harmonic signal would appear in the spectrum at 1648 – 1698 MHz for $2 \cdot \omega_0$ and at 2472 – 2547 MHz for $3 \cdot \omega_0$. The latter would partly fall in the 2.4 GHz WiFi spectrum and might cause a degradation. In the same way the 2nd harmonic of a GSM-900 system might interfere with the communication of a DCS-1800 system.

THE PROTECTION DIODE

The purpose of the protection diode (TVS) is to take over the current in case of electrostatic discharge (ESD). In the simplest case, one diode with low ohmic resistance is used. According to the I/V curve it is conductive above a certain voltage. For a silicon diode that just means around 0.7 V. Inside an application, the ESD protecting device is assumed to be inactive until an ESD event occurs. This requirement is usually also in line with the need of low leakage currents. Putting these requirements together it can result in a configuration with two diodes connected in series where the anodes are connected to the same node (Figure 1). In this case one diode is always in reverse direction and makes the combination inactive until the breakdown voltage is reached.

A semiconductor device usually comes along with a capacitive behavior caused by the junction capacitance which is in general voltage dependent (Reisch, 2005). The dependency of a junction capacitance C_J versus the applied voltage V can be seen in equation (6) where C_{J0} is known as the junction capacitance at 0 V, V_J is the junction potential and M is known as grading coefficient. The grading coefficient reflects the abruptness of the PN junction.

$$C_J = \frac{C_{J0}}{\left(1 - \frac{V}{V_J}\right)^M} \qquad (6)$$

This dependency can be measured and used for the model alignment.

THE DIODE MODEL

In general, a SPICE model (Nagel, 1973) of a diode consists of several parameters related to the physical behavior of the diode. An exemplary definition of a diode model can be seen in Figure 2.

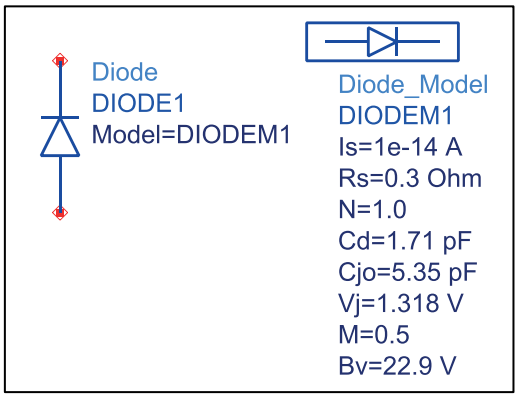


Figure 2: Diode model in ADS with a dedicated set of electrical and physical parameters.

The represented parameters in this model are the saturation current (I_s), the ohmic resistance (R_s), the linear capacitance (C_d) the zero-bias junction

capacitance (C_{jo}), the junction potential (V_j), the reverse breakdown voltage (B_v), the emission coefficient (N), and the grading coefficient (M). As it can be found in (Reisch 2005) the extraction of

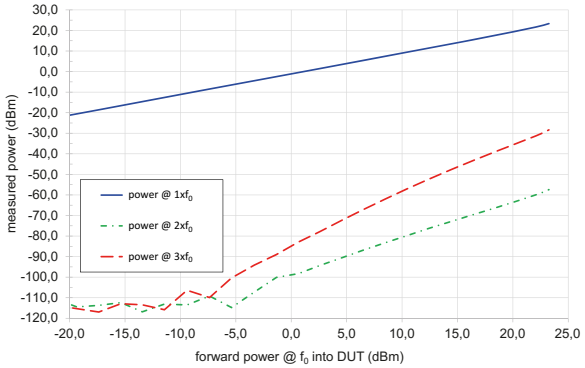


Figure 3: Measured data of Fundamental Wave and Harmonic Distortion versus Input Power.

(Keysight E4980A) which offers precise measurements of the capacitance value with an accuracy of +/- 0.05% and bias level up to 40 V. For the diode to be modelled a range from -4 V to 4 V was used. The test frequency for the capacitance measurement was 1 MHz. The device under test (DUT) was probed with needles.

Harmonic Distortion

The measurement setup for the harmonic distortion itself consists of a signal generator, attenuators and a DC block to protect the input of the spectrum analyzer, a band pass filter for suppression of unwanted signals and a high pass filter to remove the strong fundamental wave from the harmonic measurements in order to avoid saturation of the spectrum analyzer. In particular the 6 dB and 20 dB

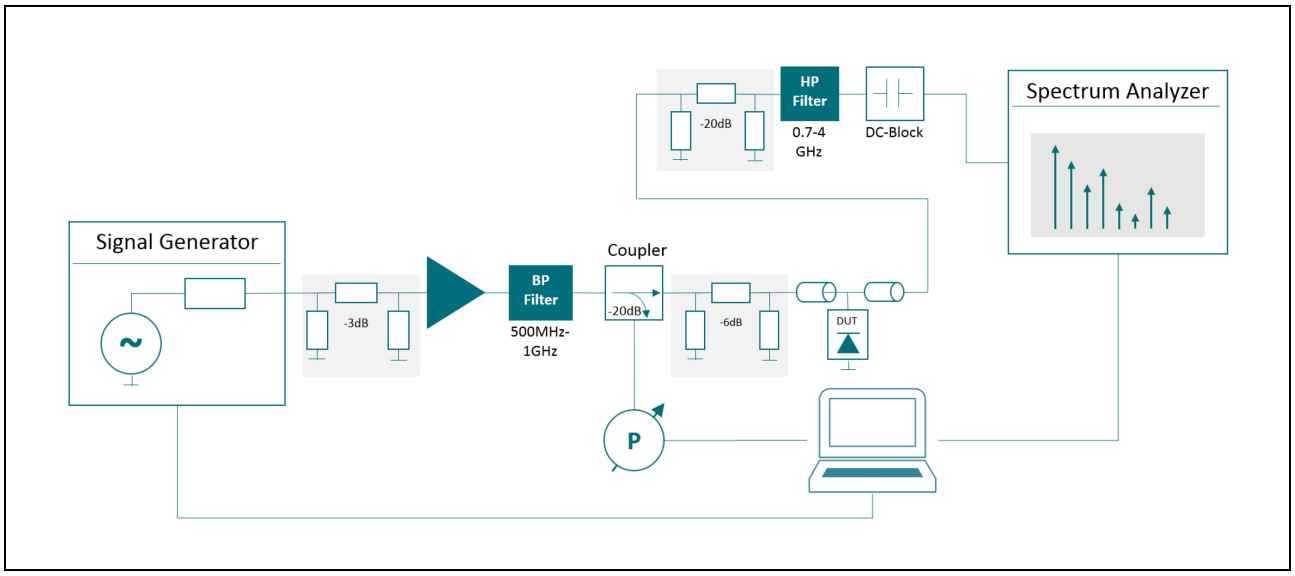


Figure 4: Measurement setup for HD measurement of GSM-900 fundamental frequencies (8xx – 9xx MHz).

the parameters I_s, N and R_s can be done by using three points of an I/V curve. In reverse direction, the breakdown voltage (B_v) and the related current (I_{bv}) can be fitted along the measured I/V curve. To get the values C_d and C_{jo} the model needs to be fitted along an $C(V)$ curve (Reisch 2004, p.75).

MEASUREMENTS

Voltage dependency of capacitance

The capacitance variation versus the applied diode voltage is measured using a precision LCR Meter

attenuators ensure a proper 50 Ohm termination to both sides of the DUT position since the band pass and high pass filter have a reactive behavior outside their passband frequencies.

The forward power applied to the DUT is measured by a directional coupler and a power meter. The spectrum analyzer is finally used to measure the level of the harmonic signals for varying levels of the fundamental wave. The DUT itself was soldered on a co-planar wave guide. The complete measurement setup is depicted in Figure 4. The analyzed fundamental frequency was 837 MHz,

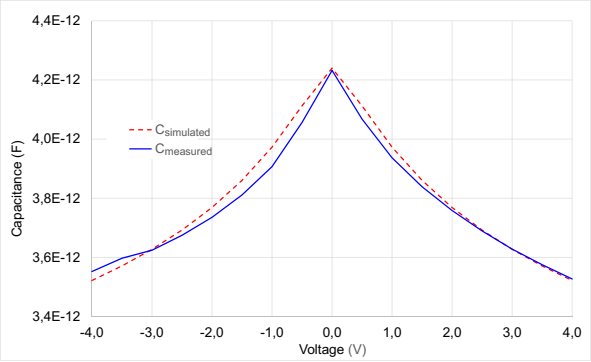


Figure 5: Simulated and Measured Capacitance versus Applied Diode Voltage at 1 MHz.

resulting in the second harmonic wave with 1.674 GHz and the third harmonic at 2.511 GHz. The input power was varied from -26 dBm up to +20 dBm. The measured data is plotted versus the available forward power of the fundamental wave (f_0) at the DUT (Figure 3).

SIMULATION

The simulation was performed in three steps. First a draft fitting of the I/V curve was done. That means the breakdown voltage was adjusted according to the measurement data. This ensures that no high current flows through the diodes. Than the model was tuned in order to achieve the *C(V)* behavior. The third step was the tuning of the harmonic distortion behavior: The model was tuned to match the second and third harmonics for a single level of the fundamental wave of 20 dBm.

Capacitance over voltage

The measured *C(V)* curve was loaded into the simulation setup. For the first step both diodes were modelled with an identical diode model DiodeM1. To get a good matching, a minimum of three diode parameters and an additional capacitance (C_extern) – which models the metal-oxide-silicon capacitance – is needed. The external capacitance was finally tuned to C_extern=0.8 pF. For the diode a linear capacitance of C_d=1.85 pF, a zero-bias junction capacitance of C_{jo}=5.05 pF and a junction potential of V_j=1.308 V was found in order to match the simulation with the measured data. The result can be seen in Figure 5.

Harmonic Distortion

The complete test bench is shown in Figure 6. For the particular analysis of the system the same fundamental frequency of RFfreq=837 MHz was chosen as for the measurements. In order to calibrate the harmonic distortion behavior the ratio between Cd and C_extern was tuned. As this changes also the C over V curve, this step needs to be done iteratively until *C(V)* matches this specific measurement data as well as the measurement and simulation of the power of the fundamental frequency. C_j itself varies with voltage V (Reisch, 2005) as a function of the grading coefficient M. This parameter was used to tune the 3rd harmonic distortion in order to match the measurement data.

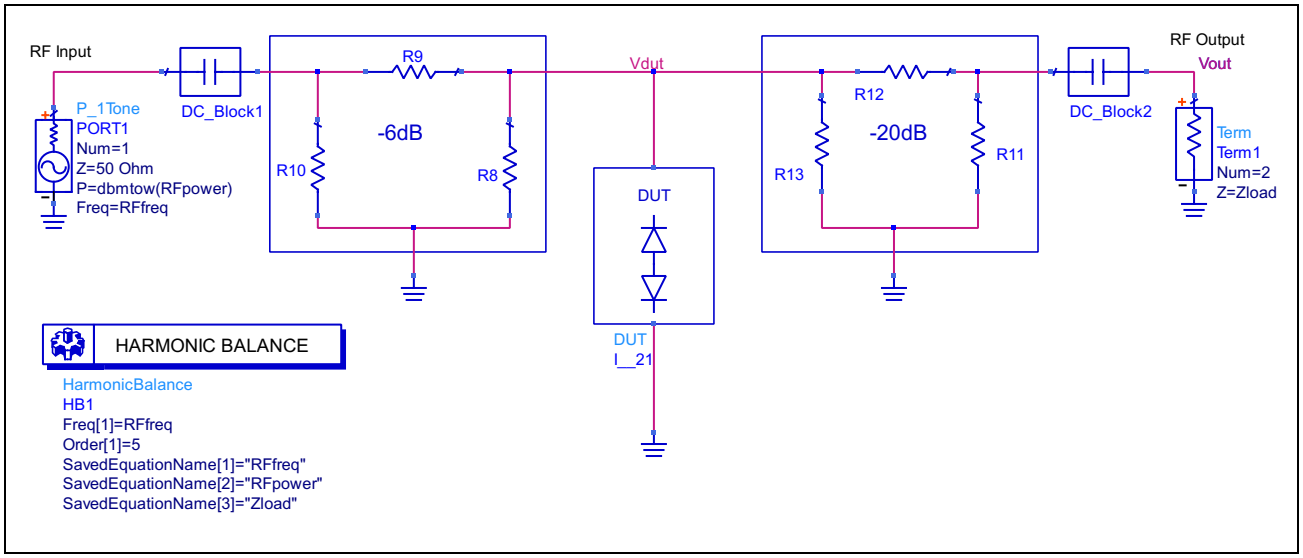


Figure 6: Simulation Setup for HD "Measurements"

To tune the 2nd harmonic distortion a minimal asymmetry of the grading coefficient was useful. Which results in two different diode models. In Figure 8 the "Two-Diode-Model" is shown in detail. Finally, a matching for the HD behavior for an input power level of +20 dBm was achieved. Figure 7 compares the simulation results with the measurement data for a variation of the forward power level at the fundamental frequency. In part a) the matching of the fundamental wave is shown. In

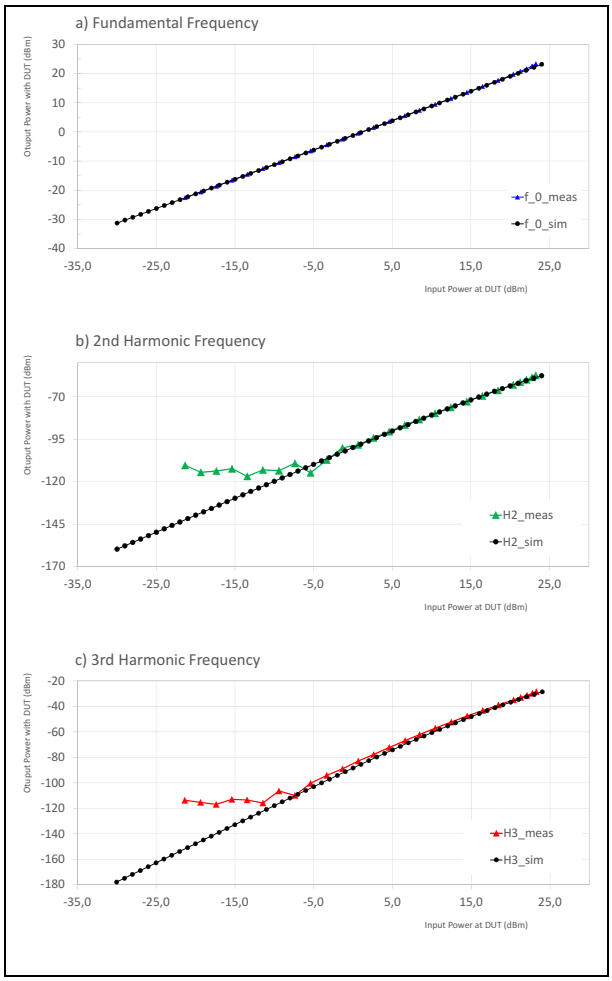


Figure 7: Harmonic Power versus Input Power. Measurements and Simulation in Comparison: a) f0: power of fundamental wave; b) H2: power of 2nd harmonic wave; c) H3: power of 3rd harmonic wave

part b) the second and in part c) the third harmonic wave is shown. For forward power level above -5 dBm a very good matching has been achieved. The deviation for HD2 and HD3 visible for level below -5 dBm is caused by the noise floor of the measurement setup: In fact the low level of the harmonic signals is masked by the noise of the spectrum analyzer.

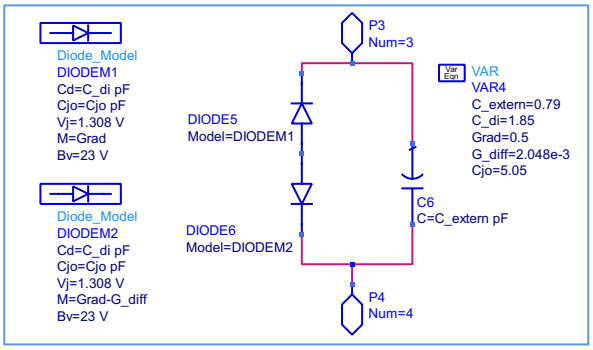


Figure 8: Final Diode Model

CONCLUSION

A diode model was generated, that allows to simulate the harmonic distortion of a device. It can be used for optimization of future devices by changing the semiconductor processing and layout structure towards an optimized design as an outcome of the simulation. It was also shown that even the simplest form of a SPICE based diode model covers the non-linear effects that lead to harmonic distortion. This model can be used to do a simulation in form of a design of experiments (DOE). That means by varying all available component parameters the influence of each parameter variation can be evaluated. This allows to identify layout and process improvements. For example the ratio of C_j, C_extern and C_d can be optimized. Future investigations will lead to modelling of higher harmonics, as well as in more complex semiconductors. As silicon controlled rectifiers (SCR) as well as open base transistors are commonly used instead of TVS diodes, it would be of interest to adapt this SPICE model approach for these devices as well.

REFERENCES

Keysight. (2018) ADS - Advanced Design System. [2018-01-30]. [Online]. Available: http://www.keysight.com/find/eesof-ads

Kories, R., Schmidt-Walter, H. "Taschenbuch der Elektrotechnik", 5. Korrigierte Auflage, 2003, 220.

Pozar, D. M. 2011. "Mircrowave Engineering. 4th" Edition. John Wiley & Sons, Inc. 511-51x

Maas, Stephen A. "Nonlinear microwave and RF circuits", 2nd ed., Artech House microwave library, 1988 (reprinted in 1997)

3GPP (3rd Generation Partnership Project), Technical Specification Group GSM/EDGE, Radio Access Network; Radio transmission and reception, 3GPP TS 05.05 V8.20.0 (2005-11)

L. W. Nagel and D. Pederson, “Spice (simulation program with integrated circuit emphasis),” EECS Department, University of California, Berkeley, Tech. Rep. UCB/ERL M382, Apr 1973. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/1973/22871.html

Reisch, M. „Halbleiterbauelemente“, Springer, 2005, 65-87

Jennifer Schütt was born in Heide, Germany in 1981. She received a degree (Dipl.-Ing.) in electrical engineering from Technical University of Hamburg (TUHH) in 2009. Since then she is with NXP Semiconductors, Hamburg, developing EMI-Filter and ESD protection devices. Since 2017 she started working for Nexperia within the same working area. Main field of expertise is in device modelling, EM simulation, device physics and project management. Currently she is working on common-mode-filter designs with integrated ESD protection for ultra-fast differential data lines. In the field of simulation, especially for ADS related topics, she organizes regular cross team meetings for Nexperia engineers located in Hamburg.

Jens Werner was born in Cologne, Germany in 1969. He received the Dipl.-Ing. and Dr.-Ing. Degrees in electrical engineering from the Technical University of Braunschweig, Braunschweig, in 1996 and 2002, respectively. In 1996 he was working with Aerodata AG as a Flight Inspection Engineer on calibration of airborne antennas. From 1996 to 2001, he was a Research Assistant at the Institute of Electromagnetic Compatibility, TU Braunschweig. His main research interests were measurement techniques and representation of guided and radiated electromagnetic fields. In 2001 he joined the Innovation Centre of Philips Semiconductors Germany GmbH in Hamburg, (since 2006 NXP Semiconductors). In March 2014, he became a Professor at Jade University of Applied Sciences, Wilhelmshaven, Germany. He is responsible for the RF, Wireless and EMC laboratory.

Ayk Hilbrink was born in Stade, Germany in 1992. He studied electrical engineering at Jade University of Applied Sciences, Wilhelmshaven and received the bachelor degree in 2016 and the Master of Science in 2018. His bachelor thesis “Generierung von Spice-Modellen in ADS” was written in cooperation with NXP Semiconductors and covers the field of device modelling in the RF range. His expertise in RF applications led him to his Master thesis at Nexperia about deembedding. Since 2017 Ayk is working for Nexperia as characterization engineer with main focus on RF and EMI measurements.

QUALITY EVALUATION OF MODELS AND POLYMODEL COMPLEXES: SUBJECT-OBJECT APPROACH

Boris Sokolov
[1]St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
[2]ITMO University
49 Kronversky Pr., St.Petersburg, 197101, Russia
sokolov_boris@inbox.ru

Vladislav Sobolevsky
St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
Arguzd@yandex.ru

Stanislav Mikoni
St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
smikoni@mail.ru

Valerii Zakharov
St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
Valeriov@yandex.ru

Ekaterina Rostova
St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
rostovae@mail.ru

KEYWORDS
Models and multiple-model complexes, a situation in progress, qualimetry, complex modeling and proactive control.

ABSTRACT

New approach of models and multiple-models quality evaluation is proposed. This approachis based on twofold ideas. First, when selecting an object for modeling it is reasonable to select not a really existing (designed or abstract) object but a situation in progress, that includes the objects and the subjects of the modeling (people responsible for making decisions (DM), people responsible for the substantiation of a decision (solution), experts, and people responsible for the implementation of solutions. Second, the process of modeling is here assumed as a control process of developing situation under uncertain conditions, caused by absence of information needed for forming the substantiated decisions. The descriptive and formal statement of quality control of models and multiple-model complexesare interpreted as problems of structural-functional synthesis of a model (a multiple-model complex) and also, selection of optimal programs of control and regulation of structural dynamics of a situation in progress (quality control of models and multiple-model complexes).The example of solving the task of poly-model finding distances from a vertex to all other vertices of a graph is proposed.

INTRODUCTION

In the modern world mathematical modeling can be considered a universal tool for study, research and design of objects in various fields of practice. By now the theory, methods and technologies for creation and application of mathematical models have developed to a quite high level. However, unfortunately, the problems of multi-criteria quality evaluation of mathematical models, analysis and classification of different types of models and reasonable selection of models for solving certain practical tasks have not been studied well enough yet. The listed problems are the main objects of study in this paper, which introduces the results of research of a new applied theory being developed by the authors – the qualimetry of mathematical models and multiple-model complexes describing various kinds of complex objects (CO) (Ceany and Raiffa 1981; Ivanov, D.A. et al. 2010; Krishans 2011; Mikoni 2015; Okhtilev 2006).

The paper provides methodological basis for the proposed theory of models' quality evaluation which includes concepts, principles and approaches to solving its main problems. There is also a general formal description and dynamic interpretation of models of a situation in progress, the participants of which are the subjects and objects of modeling, as well as the models used. This description allows to develop the most

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

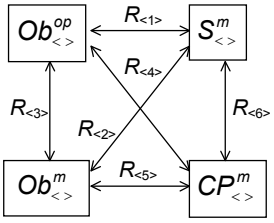
generalized approach to solving the tasks of models' qualimetry based on fundamental and applied results achieved in the modern theory of CO control. There is an example illustrating the achieved results (Azgal'dov 1982; Val'kman 1996).

GENERALIZED DESCRIPTION OF THE PROBLEMS OF SUBJECT-OBJECT MODELING OF A SITUATION IN PROGRESS AND THEIR CONTROL INTERPRETATION

There is a big number of definitions of the word *model*, which is known for its polysemy – the phenomenon of having different meanings depending on a context. Currently there are a few hundred definitions of the concepts of a model and modeling (Aframchuk et al. 1998; Merkuryeva et al. 2011; Mesarovich and Takahara 1978; Rostovtsev and Yusupov 1991). One example of them is the following - a *model* is a multiplace reflection of an original object that, together with its absolutely true content, contains conditionally true and false content, which reveals itself in the process of the object's creation and practical use (Rostovtsev and Yusupov 1991); *modeling* is one of the stages of cognitive activity of a subject that includes the development (selection) of a model, using it for research, obtaining and analyzing the results, developing recommendations on the further activity of the subject and the estimation of the quality of the model itself considering the solved problem and its specific conditions.

The analysis of the mentioned definitions leads to the conclusion that every model that was designed correctly contains the objective truth (that is, correctly reflects the original object in a certain way) (Okhtilev et al. 2006). In addition to this, due to a limited number of elements and relations in designed (applied) models, which describe objects of the unlimitedly diverse reality, and limited resources available for modeling (time, finance and materials), a model always reflects an original object in a simplified and approximate way. However, human practical experience gives enough evidence for these specific features of a model to be quite acceptable for the solution of problems that subjects have to deal with. Applying the principle of accuracies balance, it is always possible to reach a compromise between how detailed the description of an original object is and the pragmatic value of a designed model (Peschel 1981; Sethi and Thompson 2006; Sokolov and Yusupov 2004; Sokolov and Yusupov 2002).

The analysis of definitions given above shows that for modeling of different types of CO, both natural and artificial, it is reasonable to define the following basic elements and relations that characterize a certain process: firstly, a subject or subjects $(S^m_{<>})$, an original object $(Ob^{op}_{<>})$, model-object $(Ob^m_{<>})$, the environment in which the modeling is performed $(CP^m_{<>})$; and, secondly, binary relations between the listed elements $R_{<1>}(Ob^{op}_{<>}, S^m_{<>})$, $R_{<2>}(S^m_{<>}, Ob^m_{<>})$, $R_{<3>}(Ob^{op}_{<>}, Ob^m_{<>})$, $R_{<4>}(CP^m_{<>}, Ob^{op}_{<>})$, $R_{<5>}(CP^m_{<>}, Ob^m_{<>})$, and $R_{<6>}(CP^m_{<>}, S^m_{<>})$. The subscripts "<>" used here stand for personal names of objects (subjects) and relations (Okhtilev 2001; Okhtilev et al. 2006; Sethi and Thompson 2006; Sokolov and Yusupov 2004). It is important to mention that by subjects of modeling we understand the following classes of social subjects – decision-makers (DM); the people who substantiate decisions (PSD); experts; the people who utilize models; the people who design models. Figure 1 represents possible interrelation between the listed elements and relations between them (Okhtilev et al. 2006; Rostovtsev and Yusupov 1991).



Figures 1: All possible interrelations of objects and subjects of modeling of a situation in progress

From the analysis of this figure it can be inferred that the process of modeling is based on the processes of interaction between subjects $(S^m_{<>})$, an original object $(Ob^{op}_{<>})$, a model object $(Ob^m_{<>})$ and an environment $(CP^m_{<>})$, which are set with binary relations between the listed elements $R_{<1>}(Ob^{op}_{<>}, S^m_{<>})$; $R_{<2>}(S^m_{<>}, Ob^m_{<>})$; $R_{<3>}(Ob^{op}_{<>}, Ob^m_{<>})$; $R_{<4>}(CP^m_{<>}, Ob^{op}_{<>})$; $R_{<5>}(CP^m_{<>}, Ob^m_{<>})$; $R_{<6>}(CP^m_{<>}, S^m_{<>})$. It is important to note that all listed elements and relations constantly change with time due to objective-subjective and external-internal reasons. Based on that, we will call any *structure condition* of these four elements in a certain moment *a situation*, and their change with time – *a situation in progress (SP)*. With such description the process of subject-object modeling of CO can be interpreted as a controlled process (as control of a situation in progress).

The purpose of such process will be the constant minimization of a discrepancy between an original object and a model at all stages of their life cycle by constant adaptation of the model to the changes occurring to $Ob^{op}_{<>}$, as well as to $CP^m_{<>}$, $S^m_{<>}$ (for example, if a subject changes the purposes of functioning and modeling of $Ob^{op}_{<>}$).

The given interpretation of the process of object's model development in case of a situation in progress is a very perspective one. Such approach allows to use a quite well-developed set of tools for analysis and synthesis of complex technical systems and their control systems and apply it to such objects of control as models and multiple-model complexes, as well as to a situation in progress as a whole (Ivanov et al. 2010; Krishans et al 2011; Merkuryeva et al. 2011; Okhtilev et al. 2006; Rostovtsev and Yusupov 1991; Sokolov and Yusupov 2004).

So far, a lot of constructive approaches have been developed allowing to describe different kinds of models in general terms, which is necessary for their evaluation and comparative analysis (Laue and Müller 2016; Merkuryeva et al. 2011; Mesarovich and Takahara 1978; Okhtilev et al 2006; Sokolov and Yusupov 2004; Steinburg et al. 1998; Trotsky and Gorodetsky 2009). Now it is necessary to describe possible technologies of subject-object modeling, also in general terms, thoroughly and formally (in general case — subject-object system modeling), applying the proposed control interpretation.

On a descriptive level the problem of modeling of a situation in progress at different stages of its life cycle is reduced to a solution of the following three main types of tasks (Okhtilev et al 2006):

– the task of analysis of structural dynamics of a situation in progress;

– the task of estimation (monitoring) of structural conditions and structural dynamics of a situation in progress;

– the task of structural-functional synthesis of a model (a multiple-model complex) and selection of optimal programs of control and regulation of structural dynamics of a situation in progress (quality control of models and multiple-model complexes) in various environmental conditions.

Let us give an example of a descriptive and formal statement of the task of structural-functional synthesis of a model (a multiple-model complex) and also, selection of optimal programs of control and regulation of structural dynamics of a situation in progress (quality control of models and multiple-model complexes) in various environmental conditions.

The descriptive statement of the task of control of structural dynamics of a situation in progress is reduced to the following: we know the initial structural state of a situation in progress, we know the elements, possible variants of a structure of a situation in progress, we know the space-time, technical and technological constraints of a situation in progress, we know the time interval during which the control over the situation in progress takes place and a certain system of quality indicators for the given control.

It is required to perform multicriteria dynamic structural-functional synthesis of both multiple-model complex itself (based on the purposes of modeling set by a subject and indicators of quality of modeling that are used) and a corresponding technology of system modeling of a situation in progress, so that for each given scenario of changing disturbing actions on a situation in progress the most preferable transition from its current to a required structural state is reached.

Let us provide a formalization of these tasks with the use of the theory of control of structural dynamics of CO, which is being developed by the authors (Ivanov and Sokolov 2010; Ivanov et al. 2010; Okhtilev et al 2006; Sokolov and Yusupov 2002). For a constructive description of relations between above listed subjects and objects which are basic components of a situation in progress we will introduce a dynamic system alternative multigraph (DSAM) with transformable structures that looks the following way:

$$G_{\chi}^{t} =< X_{\chi}^{t}, F_{\chi}^{t}, Z_{\chi}^{t} >, \qquad (1)$$

where χ - index characterizing basic components of a situation in progress, $\chi \in NS = \{1,2,3,4\}$ – the set of indices corresponding to elements $Ob_{<>}^{op}$, $Ob_{<\theta>}^{m}$, $S_{<>}^{m}$, $CP_{<>}^{m}$, $t \in T$ – the set of time moments; $X_{\chi}^{t} = \{x_{\chi l}^{t}, l \in L_{\chi}\}$ – the set of basic components existing in the structure G_{χ}^{t} (the set of DSAM nodes) in the moment of time t; $F_{\chi}^{t} = \{f_{<\chi,l,l'>}^{t}, l, l' \in L_{\chi}\}$ – the set of DSAM edges of the type G_{χ}^{t} reflecting the relations between its basic components in the moment of time t; $Z_{\chi}^{t} = \{f_{<\chi,l,l'>}^{t}, l, l' \in L_{\chi}\}$ – the set of values of parameters that provide quantitative characteristic of the relation between corresponding basic components of DSAM (for example, the parameters of material, energy and information flows that circulate between basic components of a situation in progress).

Graphic control interpretation of the studied tasks of system subject-object modeling of a situation in progress (control of its structural dynamics) in this case is reduced to the search of such structure state $S_{\delta}^{*} \in \{S_1, S_2, ..., S_{K_{\Delta}}\}$ and such sequence (composition) of performing operations of mapping in time $\Pi_{<\delta_1,\delta_2>}^{t_1} \circ \Pi_{<\delta_2,\delta_3>}^{t_2} \circ \Pi_{<\delta',\delta>}^{t_f}$ which enable multi-criteria dynamic structural-functional synthesis of both multiple-model complex itself (based on the purposes of modeling set by a subject and indicators of quality of modeling that are used) and a corresponding technology of system modeling of a situation in progress enabling the transition of dynamic systems (1) from their given to required structural states. Alongside with the graphic interpretation of the studied problem the following *set-theoretic* description can also be proposed — it is necessary to develop principles, approaches, models, methods and algorithms that would help to find such $< U_{*}^{t}, S_{\delta}^{*t_f} >$, so that the following conditions are met

(Ceany and Raiffa 1981; Mikoni 2015; Okhtilev et al. 2006):

$$J_\theta(X^t_\chi,\Gamma^t_\chi,Z^t_\chi,F^t_{<\chi,\chi'>},\Pi^t_{<\tilde{\tilde{\delta}},\tilde{\tilde{\delta}}>},t\in(t_0,t_f])\rightarrow$$
$$\rightarrow \underset{<U^t,S^{*tf}_\delta>\in\Delta_g}{extr}, \quad (2)$$
$$\Delta_g=\{<U^t,S^{*tf}_\delta>\Big|R_\beta(X^t_\chi,\Gamma^t_\chi,Z^t_\chi,F^t_{<\chi,\chi'>},$$
$$\Pi^t_{<\tilde{\tilde{\delta}},\tilde{\tilde{\delta}}>})\le\widetilde{R}_g;\quad, \quad (3)$$
$$U^t=\Pi^{t_1}_{<\delta_1,\delta_2>}\circ\Pi^{t_2}_{<\delta_2,\delta_3>}\circ\Pi^{t_f}_{<\delta',\delta>};\beta\in\mathbf{B}\}$$

where U^t — control actions allowing to synthesize both the most preferable structure and parameters of a multiple-model complex that describes a situation in progress and corresponding technologies of system modeling of a situation in progress; J_θ – cost, time and resources indicators characterizing the quality of selected (in general case) multiple-model complexes and corresponding technologies of system modeling of a situation in progress, $q\in Q=\{1,...,l\}$ – the set of indicators' numbers; Δ_g – the set of dynamic alternatives (the set of functions, structures and parameters of multiple-model complexes and corresponding technologies of system modeling of a situation in progress); **B** – the set of numbers of space-time, technical and technological constraints that define the processes of realization of system modeling of a situation in progress $\widetilde{R}_g$ – given variables; $T=(t_0,t_f]$ – the time interval during which the most preferable structure of a multiple-model complex and corresponding technologies of system modeling of a situation in progress are synthesized.

The analysis of a formal statement of the studied problem shows that it refers to the class of problems of multicriteria selection. We will clarify what it means by giving an example.

THE EXAMPLE OF SOLVING THE TASK OF POLY-MODEL FINDING DISTANCES FROM A VERTEX TO ALL OTHER VERTICES OF A GRAPH

Below is the illustration of the main ideas of the proposed approach to model quality evaluation.

Let us consider the problem of finding distances from a vertex to all other vertices of a graph with unbounded nonnegative edge weights. Let the model to be considered, be a matrix D of edge weights. An example of the matrix D is given below:

$$\mathbf{D}_1=\begin{pmatrix}0&9&6&\infty&3\\9&0&\infty&\infty&\infty\\6&\infty&0&\infty&\infty\\\infty&2&4&0&\infty\\\infty&4&\infty&\infty&0\end{pmatrix}$$

In practice three methods (algorithms) of finding distances from a vertex to all other vertices of a graph are used:

1. Multiplying *i*-th row of the distance matrix by **D** according to the formula:
$$d^s_{ik}=\mathbf{min}\left\{d^{s-1}_{ik},\mathbf{min}_j\left(d^{s-1}_{ij}+d^1_{jk}\right)\right\},\ i,j=\overline{1,n}. \quad (4)$$
2. Dijkstra's algorithm of dynamic programming with stepwise reduction of the shortest route between vertices v_i and v_k:
$$d^s_{ik}=\mathbf{min}\left\{d^{s-1}_{ik},\mathbf{min}_j\, d^{s-1}_{ij}+d_{tk}\right\};$$
$$t=\mathbf{arg}\left(\mathbf{min}_j\, d^{s-1}_{ij}\right), \quad (5)$$
3. Finding all paths from the vertex to all other vertices with the use of the modified adjacency matrix and the following calculation of the minimal distances (Pavlovsky 2000).

Estimation of asymptotic complexity of the methods of solution

1. Matrix multiplication according to (4)

Number of operations: Multiplication of a row by a column: *n*; Determination of the minimal sum: *n*–1; Comparing with the last shortest distance: 1; Total number of row-by-column multiplication operations: $n+n-1+1=2n$; Multiplication of a row by *n* columns: $2\cdot n\cdot n=2\cdot n^2$; Total number of row-by-matrix multiplications for achieving all vertices: *n*–1; The total number of operations is equal to $\mathbf{2\cdot n^2\cdot(n-1)}$.

2. Dijkstra's algorithm of dynamic programming

Number of operations: Determining the vertex nearest to the previous one: *n*–1; Adding the distance between these vertices to the whole distance from the initial vertex: 1; Removal of the labelled vertex from the list of vertices: 1; Total number of searches for the nearest vertex: *n*+1; Removal of labelled vertices reduces the number of searching operations to $n+1-i$; $i=\overline{1,n-1}$. The total number of operation for *n* iterations is equal to $\sum_{i=1}^{n-1}n+1-i=\mathbf{n(n-1)}$.

3. Finding distances through all paths with the use of the modified adjacency matrix

Number of operations: Multiplication of a row by a column: *n*; Determining and removal of cyclic routes: 1; Total number of row-by-column multiplication operations: *n*+ 1; Multiplication of a row by *n* columns: $(n+1)\cdot n$; Total number of row-by-matrix multiplications for achieving all vertices: *n*–1; The total number of

operations for the first phase is equal to $(n+1)\cdot n\cdot(n-1)=n\cdot(n^2-1)$. Calculation of the length for an arbitrary route: $n\cdot(n-1)/2$; The total number of operations for two phases is equal to $\boldsymbol{n\cdot(n^2-1)+n\cdot(n-1)/2}$.

The first two methods have a common model and focus only on the calculation of distances. Therefore, they are compared only by complexity. Unlike the method of finding distances by multiplying row by matrix realizing a parallel search procedure, the dynamic programming algorithm implements a reduced sequential search having the complexity $\boldsymbol{2n}$ times less. Consequently, while giving the same results, it is preferable for solving the problem under consideration. The third method uses a more complex model (edge weights and adjacency matrices), which allows to determine not only the shortest distance, but also all distances from a vertex to all the other vertices. So, it is more universal compared to the first two methods, however it is also more complex than these methods. For the comparison of these methods a two-criteria evaluation of models must be applied (see Table 3).

The last row of Table 3 shows the preferences on the set of parameters. For these preferences Pareto's set is formed in the second and third algorithm. To select one of them, one should aggregate numerical estimates, bringing them to a common scale. The estimates of the third algorithm, that exceed the estimates of other algorithms, are used as normalizing values. The two-criteria estimates of the algorithms depend on the significance of parameters defined by a client (an expert).

Table 1: Capitalize Caption with No-Period

No.	The name of the method (algorithm)	Versatility (number of rotes)	Complexity
1	Multiplying by the edge weights matrix	$(n-1)$	$2\cdot n^2\cdot(n-1)$
2	Dijkstra's algorithm	$(n-1)$	$n\cdot(n-1)$
3	The use of the modified adjacency matrix	$(n-1)\cdot(n-1)$	$n\cdot(n^2-1)$ + $+n\cdot(n-1)/2$
	Preferences	max	min

In conference presentation we will propose example of practical implementation of an interdisciplinary approach that uses broadly the Earth's remote sensing data, service architecture-based forecasting systems, and an intelligent interface to select the type and adjust the parameters of hydrological models, providing the interpretation, user-friendly representation, and accessibility of operational river-flood forecast results as web services.

CONCLUSION

As the result of the conducted research on complex objects' models quality evaluation several conclusions have been made. First, when selecting an object for modeling it is reasonable to select not a really existing (designed or abstract) object but a situation in progress, that includes the objects and the subjects of the modeling (people responsible for making decisions (DM), people responsible for the substantiation of a decision (solution), experts, and people responsible for the implementation of solutions. The main feature of a situation in progress is that the set of states for all its participants varies in time due to different kinds of reasons (objective, subjective, internal, external, etc.). Secondly, modeling of objects is interpreted here as a process of control of structural dynamics of a situation in progress which takes place in uncertain conditions caused by lack of information necessary for subjects to form substantiated decisions. The discussed specifics of the conceptual and formal description of models and multiple-model complexes allow to apply mathematical structures that are being developed in the modern theory of control and engineering knowledge for the purpose of formal representation and study of models and multiple-model complexes.

ACKNOWLEDGEMENTS

The research described in this paper is partially supported by the grant of the Russian Science Foundation #17-11-01254 (section 2 of the paper), by the Russian Foundation for Basic Research (grants 17-08-00797, 17-01-00139, 17-29-07073-ofi-i), grant 074-U01 (ITMO University), state order of the Ministry of Education and Science of the Russian Federation №2.3135.2017/4.6, state research 0073–2018–0003 (section 1 (Introduction), section 3, section 4 (Conclusion) of the paper).

REFERENCES

Aframchuk, E.F., Vavilov, A.A., Emel'yanov S.V., et al., 1998. *Technology of System Modeling*. Emel'yanov S.V. ed. Mashinostroenie, Moscow. [in Russian].

Azgal'dov, G.G., 1982. *The theory and Practice of Estimation of the Quality of Products: Foundations of Qualimetry*. Ekonomika, Moscow. [in Russian].

Ceany, R.L. and Raiffa, H. 1981. *Decision Making under Many Criteria: Preferences and Fillings*. Radio i Svyaz', Moscow. [in Russian].

Ivanov, D.A. and Sokolov B.V. 2010. Adaptive supply chain management. Berlin: Springer.

Ivanov, D.A., Sokolov, B.V., and Kaeschel, J. 2010. "A multi-structural framework for adaptive supply chain planning and operation with structure dynamics considerations". *European Journal of Operation Researc.*, 200(2), 409-420.

Krishans Z., Mutule A., Merkuryev Y. and Oleinikova I. 2011. *Dynamic Management of Sustainable Development. Methods for Large Technical Systems*. Springer-Verlag, London.

Laue, R. & Müller, C. 2016. “The Business Process Simulation Standard (BPSIM): Chances and Limits”. *Proceedings of 30th European Conference on Modelling and Simulation, ECMS*, 413-418.

Merkuryeva, G., Merkuryev, Y. and Vanmaele, H. 2011. “Simulation-Based Planning and Optimization in Multi-Echelon Supply Chains”. *Simulation*. 8(87), 698–713.

Mesarovich, M. and Takahara, I. 1978. *General Systems Theory: Mathematical Foundations*. Mir, Moscow; Academic, New York.

Mikoni, S.V. 2015. “System analysis of multicriteria optimization methods on a finite set of alternatives”. *SPIIRAS Proceedings*. Is. 4(41), 180–199. [in Russian]

Okhtilev, M.Yu. 2001. “Specifics of technology for development of special computer-aided systems analizing information measured in real-time”. *Automatic Control and Computer Science*, Vol.39, No.6.

Okhtilev, M.Y., Sokolov, B.V., Yusupov, R.M. 2006. *Intelligent technologies of complex technical objects monitoring and structure dynamics control*. Nauka, Moscow. [in Russian].

Peschel, M. 1981. *Modellbilding fur Signale und Systeme*. Mir. Berlin: Technik, Moscow.

Rostovtsev, Y.G., Yusupov, R.M. 1991. “The problem of ensuring the adequacy of subject-object modeling”. *Journal of Instrument Engineering*. No.7, 7–14. [in Russian]

Sethi, S.P. and Thompson, G.L. 2006. *Optimal control theory: applications to management science and economics*. Springer, Berlin.

Sokolov, B.V. Yusupov, R.M. 2004. “Conceptual Foundations of Quality Estimation and Analysis for Models and Multiple-Model Systems”. *Journal of Computer and System Sciences International.* No.6, 5-14.

Sokolov, B.V., Yusupov, R.M. 2002. “Complex Simulation of Automated Control System of Navigation Spacecraft Operation”. *Journal of Automation and Information Sciences*. Vol.34, No.10, 19–30.

Steinburg, Alan N., Bowman, Christopher L., White, Franklin E. 1998. “Revisions to the JDL Data Fusion Model”. *Joint NATO/IRIS Conference*, Quebec.

Trotsky, D.V., Gorodetsky, V.I. 2009. “Scenario-based Knowledge Model and Language for Situation Assessment and Prediction”. *SPIIRAS Proceedings*, Iss.8, 94-127. [in Russian]

Val'kman, Yu.R. 1996. “The Problem of "Alienating" Models of Studied Objects from Their Creators in Complex Object Design”. *Izv. Ross. Akad. Nauk, Teor. Sist. Upr.*, 3. [*Comp. Syst. Sci*. 35 (3), 487 (1996)].

AUTHOR BIOGRAPHIES

BORIS SOKOLOV is a deputy director at the Russian Academy of Science, Saint Petersburg Institute of Informatics and Automation. Professor Sokolov is the author of a new scientific lead: optimal control theory for structure dynamics of complex systems. Re-search interests: basic and applied research in mathematical modelling and mathematical methods in scientific research, optimal control theory, mathematical models and methods of support and decision making in complex organization-technical systems under uncertainties and multicriteria. He is the author and co-author of 9 books on systems and control theory and of more than 450 scientific papers. Professor B. Sokolov supervised more over 90 research and engineering projects. His e-mail address is `sokolov_boris@inbox.ru` and his Web-page can be found at `http://litsam.ru`.

STANISLAV MIKONY Ph.D., Dr. Sci., professor, professor of mathematics and modeling department, Petersburg State Transport University, leading researcher of information technologies in the system analysis and modeling laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS. Research interests: system analyses, decision making, intellect technologies. The number of publications — 254. His e-mail adress is `smikoni@mail.ru`.

VLADISLAV SOBOLEVSKY was born in Vitebsk, Republic of Belarus and went to the St. Petersburg State Technological Institute, where he studied systems analysis, control and data processing and obtained his degree in 2017. He worked for a couple of years for the JSC Concern VKO "Almaz-Antey" before moving in 2017 to the SPIIRAS where he is now working in the field of deep learning. His e-mail address is `Arguzd@yandex.ru`.

VALERII ZAKHAROV was born in Saint-Petersburg, Russia. He studied organization of construction operations at the Saint-Petersburg State University of Architecture and Civil Engineering and obtained his master’s degree in 2016. He worked for a couple of years for the building and construction company before moving in 2017 to the SPIIRAS where he is now researcher and PhD student in the Laboratory of Information Technologies in System Analysis and Modeling. His e-mail address is : `Valeriov@yandex.ru`

EKATERINA ROSTOVA graduate student, Laboratory of Information Technology in System Analysis and Integrated Modeling, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: man-machine systems, robotics monitoring and control, methods and models for complex decision support. Her e-mail address is `Rostovae@mail.ru`.

OPTIMAL PLANNING FOR PURCHASE AND STORAGE WITH MULTIPLE TRANSPORTATION TYPES FOR CONCENTRATED LATEX UNDER AGE-DEPENDENT CONSTRAINT

Tuanjai Somboonwiwat[1]
Sutthinee Klomsae[2]
Walailak Atthirawong[3]
[1] King Mongkut's University of Technology Thonburi, Thailand,
[2] Uttaradit Rajabhat University, Thailand,
[3] King Mongkut's Institute of Technology Ladkrabang, Thailand,
E-mail:tuanjai.som@kmutt.ac.th, klomsae_aey@hotmail.com, walailaknoi@gmail.com

KEYWORDS

Age-dependent, Optimal Planning, Perishable Product, Rubber

ABSTRACT

This paper presents the mathematical model to support decision planning over the multi-period for purchasing storage and transportion of concentrated latex for the rubber glove production under age-dependent constraint. The model considers multiple suppliers with different purchasing costs, varying product ages, multi-truck load capacity and costs, and storage times with respect to perishable product. The model is then applied to a rubber glove production case study and the sensitivity analysis is performed. The result reveals that the price of concentrated latex in each period and truck load capacity decision are critical factors for total cost reduction.

INTRODUCTION

Thailand has achieved particularly remarkable in the upstream sector of rubber industry over an extended period. In 2015, the country was the world leader in rubber producer and exporter which produced about 4.5 million tons of output, followed by Indonesia, China, India, and Malaysia, respectively (Thailand Industry Outlook 2016). Nowadays, Thailand has the advantage of expanding production area, especially in the Northeast and North part of the country, which can be divided into 10 regions of plantation areas (http//www.oae.go.th). Rubber is a key input into a number of significant manufacturing sectors, especially rubber gloves and tires. The most suitable raw material for the manufacture of gloves is natural concentrated latex. It was found that rubber gloves had contributed an enormous export values to Thai economy. Figure 1 illustrates the flow of production materials in gloves industry from upstream to downstream.

Nevertheless, currently Malaysia becomes the world leader in developing the downstream part of the rubber

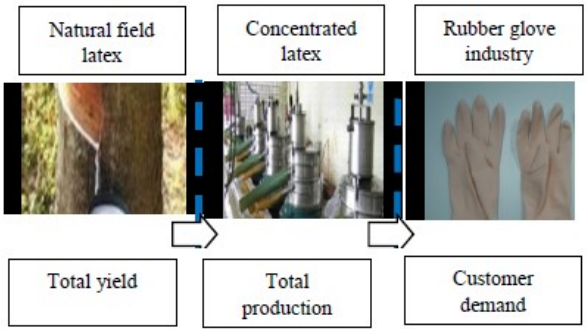


Figure 1. Flow of Production Materials in Rubber Glove Industry from Upstream to Downstream

value chain, particularly medical glove production. As such, the challenge for Thailand to gain competitive, decision planning of this industry is extensively required for purchasing and storage of the concentrated latex throughout the seasonal production cycle for the industry. It was found that the supply volume of field latex varies according to regions and seasons, which is a perishable in nature. As such, it is necessary to pay attention on production and transportation which is a generally complicated task.

According to the statistics report, most of rubber latex manufactures are located in the Middle, Southern and Eastern regions of Thailand account for 40%, 36.7% and 20% respectively (Plastics Institute of Thailand 2014). Due to the proximity of raw materials, the large size of the manufacturers is in the Southern (46%), and Eastern (38%) regions of the country. Generally, the delivery of field latex is made from the plantation to a latex concentration industry within the same region, as it is a perishable product latexes. However, the glove manufacturing factories will require concentrated latex according to their real demand which any shortages of raw material to the factories will result in higher costs. As a result, some concentrated latexes are supplied to the rubber glove companies in the same region when possible, and some of them are shipped across regions according to real demand. After that, the finish products

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

will be exported to customers through the closest seaports from that factory.

Along with rapid change in the marketplace and the expansion of supply chains, highly coordination on purchasing, inventory, and transportation over a multi-echelon supply chain network or a multi-period time frame is vigorous, which has very much impact on customer service and profit margins of the supply chain members. Nonetheless, research related to simultaneous purchase and storage decisions of rubber latex with regard to aging limitations as well as considering a transportation cost at the same time is still quite rare. Most of the literature in the area of the rubber supply chain is focused on the structure of the rubber industry.
The literature of the rubber supply chain is related to the structure of the rubber industry. Haan *et al.* (2003) present the production flow in the rubber supply chain of India, using semi-structured interviews and analysis of the relationships of each part. For research relevant to marketing analysis of the rubber industry, Arifin (2005) studies the supply chain of rubber production in Indonesia and assess the transmission of the price to rubber growers. Fathoni (2009) analyzes rubber market systems and analyze marketing margins and prices for rubber in Jambi Province in Indonesia. Auckara-aree and Boondiskulchok (2010) design a raw material collection system by formulating Mixed Integer Programming for location routing with a step–price policy model to find the optimal solution. Somboonwiwat and Chanclay (2009) create value-added products by coherence of targets in different parts of the supply chain in Thailand. For research related to transportation, Kritchanchai and Chanpuypetch (2009) investigate gateway selections for Thailand rubber exports, using fuzzy analytic hierarchy processes. Klomsae *et al.* (2012) develop a mathematical model to determine the concentrated latex volume to purchase and storage. However, recently the perishable product management has been widely concerned in the logistics planning area, for example in the studies of Wu *et al.* (2015) and Liu *et al.* (2017). Hence, this paper presents the mathematical model for decision planning which simultaneously solve the volume of concentrated latex purchased and storaged in each age including transpotation type in multi-period time frame taking into consideration of product aging and deterioration through each time period. The parameters in the model include purchasing costs of product with varying price and age, transportation costs that vary by truck load capacity and distance between location of suppliers and manufacturers, storage costs, and expiration costs of outdated concentrated latex.

The organization of this paper is as the following. The next section describes the multi-period planning problem of purchasing and storage of concentrated latex from multiple suppliers and transportation types. Then the mathematical model is developed. A numerical example is presented to illustrate an application of the formulation. Finally, summary and concluding remarks are addressed.

PROBLEM DESCRIPTION

The rubber glove industry uses rubber concentrated latex to produce the glove products. The latex requirement is based on dependent demand which the latex can be purchased from different suppliers in differernet regions. The decision planning for purchase, storage and transportion of concentrated latex for rubber glove production begins with the determination of the volume of concentrated latex from different regions and multiple suppliers with number of round trips of different transportation types. The cost of transportation will depend on distance between location of suppliers and manufacturers, as well as size of truck and truck load capacity as well. Since age is significant factor in perishable latex requirement operations, the latex planning includes age of latex for each supplier and transportation type.

Figure 2 delineates the system of purchase, storage and transportation of perishable concentrated latex. There are multiple prices and ages of latex for each supplier. The number of round trips (X_{ijkt}) is decided for each supplier for each age of latex. The volume of purchased concentrated latex for supplier i at age j deliver by truck type k in period t (B_{ijkt}) and the volume of stored concentrated latex at the beginning of period t (I^B_{jt}) must be high enough for the demand of concentrated latex to glove producers (D_t). At the beginning of period t+1 the age of the concentrated latex is increased to j+1 ($I^B_{(j+1)(t+1)}$). The age of concentrated latex continues to increase until it reaches the expiration date, then that concentrated latex will be eliminated.

The primary objective of this paper is to develop a mathematical model for solving the concentrated latex requirement planning under age-dependent and multiple transportation types constrains. The decision criteria used is to minimize the total costs of purchasing, storaging, transportation and expiration.

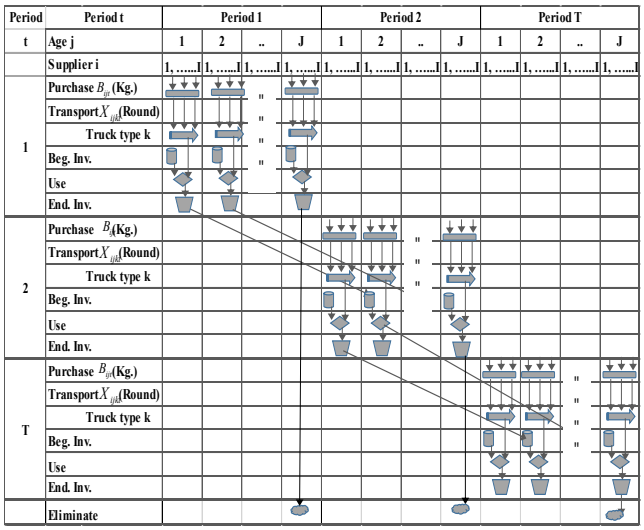


Figure 2. System of Purchase, Storage and Transportation of Concentrated Latex under Age-dependent Constraint

MATHEMATICAL FORMULATION

Indices

i Index of supplier (i = 1, 2, 3, …, I)

j Index of age of concentrated latex (j = 1, 2, 3, …,J)

k Index of transportation type (k = 1, 2, …,K)

t Index of time periods (t = 1, 2, 3, …, T)

Parameters

P_{ijt} Purchasing price of concentrated latex from supplier i for age j in period t

D_t Demand of concentrated latex for production in period t

C_{ijt} Concentrated latex capacity of supplier i for age j in period t

W_k Weight per round of concentrated latex transported by transportation type k

T_{ikt} Transportation cost of concentrated latex transported from supplier i by transportation type k in period t

H_{jt} Holding cost of concentrated latex to age j in period t

E_t Expiration cost of concentrated latex that expires in period t

Decision Variables

B_{ijt} Volume of concentrated latex purchase from supplier i for age j in period t

X_{ijkt} Number of round trips required to transport latex from supplier i for age j by transportation type k in period t

U_{jt} Volume of concentrated latex age j used in production in period t

I^E_{jt} Volume of concentrated latex inventory age j at the ending of period t

I^B_{jt} Volume of concentrated latex inventory age j at the beginning of period t

F_t Volume of concentrated latex that expires in period t

Objective Function

The objective function of the problem is to find the minimum total cost of the purchasing cost, storage cost, transportation cost and expiration cost of concentrated latex as shown in Equation 1.

$$\text{Minimize Total Cost} = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J} P_{ijt}\, B_{ijt} + \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} T_{ikt}\, X_{ijkt} + \sum_{t=1}^{T}\sum_{j=1}^{J} H_{jt}\, I^B_{jt} + \sum_{t=1}^{T} E_t F_t \qquad (1)$$

Constraints

In each period for each supplier the volume of concentrated latex at each age purchased is less than or equal to the available capacity at the same latex age.

$$B_{ijt} \le C_{ijt} \qquad \forall\ i, j, t \qquad (2)$$

In each period for each supplier, the number of round trips for concentrated latex at each age transported by each transportation type is greater than or equal to rate of volume of concentrated latex purchased from supplier i of age j in period t and weight per round of transported concentrated latex from by transportation type k.

$$X_{ijkt} \ge \frac{B_{ijt}}{W_k}; \qquad \forall\ i, j, k, t \qquad (3)$$

At each period, sum of total volume of concentrated latex purchased from supplier i for age j and total volume of concentrated latex inventory for age j-1 at the ending of period t-1 is greater than volume of demand of concentrated latex for production in period t.

$$\sum_{i=1}^{I}\sum_{j=1}^{J} B_{ijt} + \sum_{j=1}^{J} I^E_{(j-1)(t-1)} \ge D_t\ ; \quad \forall\ t \qquad (4)$$

At each period, volume of concentrated latex inventory for age j at the ending of period t equal to the sum of volume of concentrated latex inventory for age j in the beginning of period t and volume of concentrated latex purchased from supplier i for age j in period t, minus volume of concentrated latex for age j used in production.

$$I^E_{jt} = I^B_{jt} + \sum_{i=1}^{I} B_{ijt} - U_{jt} \qquad \forall\ j, t \qquad (5)$$

At period T, volume of expired concentrated latex equal to volume of concentrated latex inventory for age *j* at the ending of period *t*.

$$F_t = I^E_{jt} \qquad \forall\, t \qquad (6)$$

In order to determine the volume of concentrated latex of age *j* used in production in period *t*: if demand for concentrated latex for production in period *t* is greater than or equal to the sum of the volume of concentrated latex inventory for age *j*-1 at the ending of period *t*-1 and the volume of concentrated latex purchased from supplier *i* of age *j* in period *t* then: the volume of concentrated latex for age *j* used in production in period *t* will equal the sum of the volume of concentrated latex for age *j*-1 used in production in period *t*-1 and the volume of concentrated latex purchased from supplier *i* of age *j* in period *t* (Equation 7). But if this condition is not true, then: the volume of concentrated latex age *j* used in production in period *t* will equal to the demand for concentrated latex for the production in period *t* (Equation 8).

$$\text{If} \quad D_t - (I^E_{(j-1)(t-1)} + \sum_{i=1}^{I} B_{ijt}) \geq 0$$

$$\text{True:} \quad U_{jt} = I^E_{(j-1)(t-1)} + \sum_{i=1}^{I} B_{ijt}\ ; \qquad \forall\, t \qquad (7)$$

$$\text{False:} \quad U_{jt} = D_t\ ; \qquad \forall\, t \qquad (8)$$

The constraints address to determine the volume of concentrated latex for age *j* used in production in period *t*. For example, concentrated latex of the oldest age available for production is *j*=4 and determine the volume of concentrated latex available for production of age *j*=3, and determine the volume of concentrated latex available for production in period *t* as demonstrated in Equation 9 and Equation 10.

$$\text{If} \quad (I^E_{(j-2)(t-1)} + \sum_{i=1}^{I} B_{i(j-1)t} - (D_t - U_{jt}) > 0$$

$$\text{True:} \quad U_{(j-1)t} = D_t - U_{jt}\ ; \qquad \forall\, t \qquad (9)$$

$$\text{False:} \quad U_{(j-1)t} = I^E_{(j-2)(t-1)} + \sum_{i=1}^{I} B_{i(j-1)t}\ ; \qquad \forall\ t \qquad (10)$$

Non-negativity constraint:

$$B_{ijt}, X_{ijkt}, I^E_{jt}, I^B_{jt}, F_t, U_{jt}, \geq 0\ ; \qquad \forall\, j, k, t \qquad (11)$$

NUMERICAL EXAMPLE

In this example, the concentrated latex can be purchased from 2 suppliers which are transported from different regions to the production facilities by 2 transportation types. The age of the latex is tracked over varying periods, allowing for a maximum aging limit of 4 months. The problem is to determine the volume of concentrated latex at each age purchased from each supplier, the number of round trips for concentrated latex transported from suppliers to the production facilities, the volume of concentrated latex used in production, the volume of concentrated latex inventory stored in each period, and the volume of expired concentrated latex. The model is applied to an example data set in order to generate planning decisions for the purchase and storage of concentrated latex with multiple prices and ages. The latex volumes must vary in response to production demand for rubber glove production over a 3 period time interval. The prices and volumes of supplied latex are shown in Table 1. Production demand, holding cost and expiration cost are presented in Table 2. Transportation costs and weight per round of concentrated latex for each transportation type are shown in Table 3.

Table 1. Prices and Volumes of Supplied Latex

Supplier i		Period 1				Period 2				Period 3			
	Age j	1	2	3	4	1	2	3	4	1	2	3	4
1	P_{ijt} (baht)	36.99	36.43	35.7	0	35.16	34.81	34.46	0	35.75	35.39	35.04	34.86
	C_{ijt} (Ton)	1,350	1,350	350	0	1,350	1,350	1,350	0	1,350	1,350	1,350	1,350
2	P_{ijt} (baht)	36.74	36.19	35.45	0	34.91	34.56	34.21	33.86	35.5	35.15	34.79	34.44
	C_{ijt} (Ton)	1,420	1,420	1,420	0	1,420	1,420	1,420	1,420	1,420	1,420	1,420	1,420

Table 2. Demand, Holding Cost and Expiration Cost

Period	Period 1				Period 2				Period 3			
Age j	1	2	3	4	1	2	3	4	1	2	3	4
I (kg.)		13,000										
H_{kt} (Baht)	5	5	5	5	5	5	5	5	5	5	5	5
E_t (Baht)	11	11	11	11	11	11	11	11	11	11	11	11
D_t (kg.)	457,000				426,000				442,000			

Table 3. Transportation Costs and Weight per Round of Concentrated Latex for Each Transportation Type

Supplier i	Transportation type	T_{ikt} (Bath)	W_k (kg./Round)
1	10-wheel truck (k=1)	4,694	12,000
	Trailer (k=2)	8,605	25,000
2	10-wheel truck (k=1)	15,679	12,000
	Trailer (k=2)	21,877	25,000

Results

LINGO software is employed to solve this problem and the solution is presented in Table 4. The results show the optimal planning for purchasing, storage and transporting of perishable concentrated latex used in glove production over a 3-period time interval. The concentrated latex is purchased from both suppliers which purchasing cost is more sensitive to decision making than transportation and holding costs. Since the price of concentrated latex decreases when the age of concentrated latex goes up, so the purchasing decision is prioritized from the oldest to the youngest ones. For example, it can be seen from the planned purchase of concentrated latex of age 3 from supplier 1 and 2 in period 1, concentrated latex of age 4 from supplier 2 and of age 3 from supplier 1, and concentrated latex of age 4 from both suppliers in period 3. Then, the sensitivity analysis is performed by changing the purchase price of concentrated latex from the original price ranging from -30% to 30%. The computational results for a 3-period time interval each purchase price in details are presented in Table 5. The results indicate that the volume of concentrated latex purchased from supplier 1 increases and supplier 2 decreases if the purchase price of latex changes by -30%, as shown in Table 6. In the opposite way, if the purchase price of latex changes by +30%, the volume of concentrated latex purchase from supplier 1 decreases and supplier 2 increases, as shown in Table 7. Moreover, the result from sensitivity analysis also shows that the 25-ton trailers are normally selected as the transportation type. Due to the aging of latex and its holding cost, the results also suggest that the rubber glove industry should purchase the concentrated latex with the only required amount for each period.

Table 4. Results

Supplier i		Period 1				Period 2				Period 3			
	Age j	1	2	3	4	1	2	3	4	1	2	3	4
1	Purchase B_{ijt} (Ton)			344				1					442
	Transport X_{ijkt} (Round)												
	- 10-wheel truck (k=1)							1					
	- Trailer (k=2)			14									18
2	Purchase B_{ijt} (Ton)			100					425				
	Transport X_{ijkt} (Round)												
	- 10-wheel truck (k=1)												
	- Trailer (k=2)			4					17				
Use U_{jt} (Ton)			13	444				1	425				442
Ending Inventory I^E_{jt} (Ton)													
Eliminate F_t (Ton)													
Purchase Cost (Bath)		15,825,800				14,424,960				15,408,120			
Transportation Cost (Bath)		207,978				376,603				154,890			
Inventory Cost (Bath)													
Eliminate Cost (Bath)													
Total Cost (Bath)		46,398,351											

Table 5. Sensitivity Analysis for Changing Purchase Price

Supplier i	% Change	-30%	-20%	-10%	0%	10%	20%	30%
1	Purchase B_{ijt} (Ton)	1,212	1,212	787	787	787	787	362
	Transport X_{ijkt} (Round)							
	- 10-wheel truck (k=1)	1	1	1	1	1	1	1
	- Trailer (k=2)	49	49	32	32	32	32	15
2	Purchase B_{ijt} (Ton)	100	100	525	525	525	525	950
	Transport X_{ijkt} (Round)							
	- 10-wheel truck (k=1)							
	- Trailer (k=2)	4	4	21	21	21	21	38
Use U_{jt} (Ton)		1,325	1,325	1,325	1,325	1,325	1,325	1,325
Ending Inventory I^E_{jt} (Ton)								
Eliminate F_t (Ton)								
Purchase Cost (Bath)		32,138,480	36,732,840	41,090,020	45,658,880	50,228,740	54,788,920	59,125,030
Transportation Cost (Bath)		513,847	513,847	739,471	739,471	739,471	739,471	965,095
Inventory Cost (Bath)								
Eliminate Cost (Bath)								
Total Cost (Bath)		32,652,327	37,246,687	41,829,491	46,398,351	50,968,211	55,528,391	60,090,125

Table 6. Volume of Concentrated Latex Purchase for Changing Purchase Price -30%

Supplier i		Period 1				Period 2				Period 3			
	Age j	1	2	3	4	1	2	3	4	1	2	3	4
1	- 10-wheel truck (k=1)							1					
	- Trailer (k=2)			344				425					442
2	- 10-wheel truck (k=1)												
	- Trailer (k=2)			100									

Table 7. Volume of Concentrated Latex Purchase for Changing Purchase Price +30%

Supplier i		Period 1				Period 2				Period 3			
	Age j	1	2	3	4	1	2	3	4	1	2	3	4
1	- 10-wheel truck (k=1)							1					
	- Trailer (k=2)			344									17
2	- 10-wheel truck (k=1)												
	- Trailer (k=2)			100					425				425

CONCLUSION

The main contribution of this paper is the proposed mathematical model to support decision planning over the multi-period interval, as well as planning for purchasing of concentrated latex from different regions, from multiple suppliers with multiple transportation types under age-dependent constaints. Purchase and storage of concentrated latex with constraint of perishability through a multi-period time interval is a complex problem. The supply volume differs in different regions according to seasonal fluctuations, and the concentrated latex is perishable over time if stored for too long. This model can also be extended by including the quality of concentrated latex in each age. In addition, the matching between glove product and concentrated latex prices under uncertainty could be interesting to explore.

REFERENCES

Auckara-aree, K. and Boondiskulchok, R. 2010. "Designing the raw material collection system for profit maximization under a step–price policy", *Songklanakarin Journal of Science Technology*, Vol. 32, 581-588.

Fathoni Z. 2009. "Evaluation of market system and market integration for rubber cultivation in jambi province Indonesia", Wageningen University and Research.

Haan J. de., Groot G. de., Loo E. and Ypenburg M. 2003. "Flows of goods or supply chains; lessons from the natural rubber industry in Kerala, India", *International Journal of Production Economics,* Vol. 81-82, 185-194.

Klomsae, S., Somboonwiwat,T. and Atthirawong,W. 2012. "Optimal multi-period planning for purchase and storage of rubber latex with perishability constraints", *Proceedings of the 7th International Engineering and Management System,* Kitakyushu, Japan, 582-590.

Kritchanchai, D. and Chanpuypetch W. 2009. "Gateway selections for Thailand rubber export", *Proceedings of the Asia Pacific Industrial Congrees on Logistics and SCM Systems*, 244-250.

Liu,H., Zhang, J. , Zhou, C. and Ru, Y. 2017. "Optimal purchase and inventory retrieval policies for perishable seasonal agricultural products", *Omega*, In press, 1–13.

Wu, T., Shen, H. and Zhu, C. 2015, "A multi-period location model with transportation economies-of-scale and perishable inventory", *International Journal of Production Economics,* Vol. 169, 343-349.

AUTHOR BIOGRAPHIES

TUANJAI SOMBOONWIWAT is an Associate Professor in the Industrial Management section, Department of Production Engineering Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand. She received her M.Eng. in Industrial Engineering from Chulalongkhorn University, Thailand and Ph.D. in Industrial Engineering from Corvallis, Oregon State University, USA. Her research interests include green supply chain and logistics, business process and applications of operations research. Her e-mail address is: tuanjai.som@kmutt.ac.th

SUTTHINEE KLOMSAE received M.Eng. in Industrial and Manufacturing Systems Engineering from King Mongkut's University of Technology Thonburi in 2006. She works at Uttaradit Rajabhat University. Her research interests are in area of logistics and supply chain management. Her e-mail address is: klomsae_aey@hotmail.com.

WALAILAK ATTHIRAWONG is an Associate Professor in Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand. She received a Ph.D. in Manufacturing Engineering and Operations Management at the University of Nottingham, England in 2002. Her teaching and research interests include applied operations research, supply chain management, forecasting and location decision. Her e-mail address is: walailaknoi@gmail.com.

Using DEMATEL to explore the relationship of factors affecting consumers' behaviors in buying green products

Walailak Atthirawong*
Department of Statistics
King Mongkut's Institute of Technology Ladkrabang
Bangkok,Thailand
walailaknoi@gmail.com

Wariya Panprung
Management Science Department
Phranakhon Rajabhat University
Bangkok,Thailand
noiwari@gmail.com

Kanogkan Leerojanaprapa
Department of Statistics
King Mongkut's Institute of Technology Ladkrabang
Bangkok,Thailand
kanogkan.le@kmitl.ac.th

KEYWORDS

Buying Behavior, Decision-making Trial and Evaluation Laboratory (DEMATEL), Environmental Issues, Green Products.

ABSTRACT

The main purpose of this paper is to analyze factors influencing consumers' behaviors in buying green products by applying Decision-making Trial and Evaluation Laboratory (DEMATEL) method. Nine criteria i.e. perception on environmental concerns (A), safety and health concerns (B), green packaging (C), convenience to buy (D), environmental attitude (E), subjective norms (F), green product management (G), environmental laws (H) and perceived value (I) were employed from the previous study by Atthirawong and Panprung (2017). In this study, six experts were involved in order to determine the degree of direct influence between two factors through a pairwise comparison. The results revealed that the top three important criteria affecting consumers in buying green products are environmental attitude (E), safety and health concerns (B) and green product management (G), respectively. Furthermore, a cause and effect relation diagram was also constructed to gain a better understanding of the interactive relationship between those criteria**.** It was found that subjective norm (F) has the most influence on other factors, whereas perception on environmental concerns (A) gets the most impact from other factors. Finally, this paper provides some practical suggestions for relevant agencies and policy makers based on the analysis.

INTRODUCTION

Recently, the deterioration of natural resources, pollution problems and global warming has become serious social problems everywhere across the globe. Consequently, the level of consciousness and awareness about environmental problems are increasingly internationally concerned from both society and business. According to UN Global Compact (2010) report, there were about 93% of CEO's around the world fretful about sustainability in their businesses. Those organizations pay greater attention to diminish the harmful impact of business activities in terms of production, consumption and purchasing behavior on the environment. Environmental awareness and concerns would lead to the emergence of sustainable development which minimizes negative impact on the nature, physical environment and society. The key principle of sustainable development is to strengthen the economy and environment in long-term resulting from rigorous laws and regulations concerning the impact of the products during manufacturing, use and end of life (Hartmann and Ibáñez 2006; Hertwich 2005). Consequently, sustainable development has led to eco innovation and green consumption.

On the one hand, eco innovation refers to the development of products and processes via incorporating environmental sustainability practices at every stage of creation (Veleva and Ellenbecker 2001). These so called "green product" refers to product incorporating the strategies in recycling or with recycled content, reduced packaging or using less toxic materials to reduce the impact on the natural environment. On the other hand, green consumption is a concept which ascribes to consumers responsibility or co-responsibility for addressing environmental problems through adoption of environmentally friendly behaviors by purchasing, using and disposing of products (Moisander 2007).

Over the past decade, the issue regarding to green consumerism has been investigated by a various studies. Despite environmental concern, consumers are willing to pay more for environmental benefits (Singh 2011). However, it was argued that, not every

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

consumer takes environmental action in purchasing eco-friendly products into consideration (Chen 2013). The decision in buying may depend on consumers' background and experience or the way they think and behave (Grimmer and Bingham 2013). According to Ward *et al.* (2011), consumer demographics and attitudes affect willingness to purchase green products. Lee (2008) claimed that there are several factors influencing green purchasing behavior of adolescents in Hong Kong; for example, perceived seriousness of environmental problems, perceived environmental responsibility, concern of self-image in environmental protection and so on. A previous study from Padel and Foster (2005) also reported that environmental knowledge has a strong relationship with green purchase behavior. Whereas Elham and Nabsiah (2011) as well as Ottman (1992) indicated that attitude and demand for green products are uneven across different cultures and market segments. In Thailand, however, the studies in this area are still quite rare (Kianpour *et al.* 2014). Recently, Atthirawong and Panprung (2017) had investigated factors influencing consumers' behavior on buying green products in Bangkok, the capital of Thailand. The results of this study via using Factor Analysis revealed that there were nine factors i.e. perception on environmental concerns, safety and health concerns, green packaging, convenience to buy, environmental attitude, subjective norms, green product management, environmental laws and perceived value influencing consumers' behaviors on such issues.

Along with those mentioned studies, it is quite clear that there are several factors which contribute to persuade consumers to purchase green products. Nevertheless, those factors may have interrelationships with each other in nature and some of them could also be impacted by others. According to Gabus and Fontela (1972, 1976), DEMATEL is a sophisticated method for extracting interrelationship among multiple complex factors and determine how a particular factor influences other ones (Abbasi *et al.* 2013). The methodology will help to develop a holistic policy of organizations to explain cause and effect relationship of consumers' decision in buying green products**.** As such, this paper has adopted DEMATEL technique to further investigate the significance of various criteria found in the aforementioned study by Atthirawong and Panprung (2017).

The remainder of this paper is organized as follows. The latest related studies on green products and DEMATEL method are briefly reviewed in Section 2. Next, an example for application in applying DEMATEL method to evaluate the criteria is illustrated in Section 3. Finally, conclusions are drawn and recommendations for future study are mentioned in the final section.

RESEARCH METHODOLOGIES

DEMATEL Method

Decision-making Trial and Evaluation Laboratory (DEMATEL) method was originally developed by the Science and Human Affairs Program (SHAP) of the Battelle Memorial Institute of Geneva between 1972 and 1976 (Wu, 2008; Tzeng *et al.* 2007). The method, which is a graph theory based technique, has been employed to gather knowledge of experts and visualize the causal relationship between complex factors by using a graphical diagram (Shieh *et al.,* 2010). Through analysis of visual relationship among entities and their groups, DEMATEL approach can help to prioritize factors based on type of relationship as well as identify severity of their effect on other factors.

In recent years, DEMATEL has been successfully and extensively applied in many areas; for instance, knowledge management, marketing strategies, control systems, safety problems (Wu and Lee 2007), tourism (Chen 2012) and emergency management (Zhou *et al.* 2011). Hessami and Yousefi (2013) identified factors which influence on consumers' green purchase behavior in Iran using DEMATEL and Fuzzy-Delhi methods. Seven factors i.e. individual's ecological beliefs, environmental factors, consumer values, awareness of green products, attitudes toward green purchases, green purchasing intention, green purchasing behavior were identified from their study.

Briefly, the steps of DEMATEL method based on Gabus and Fontela (1972) are as follows (Wu and Lee 2007; Wu *et al.* 2011):

Step 1: Define the Evaluation Scale. Suppose in a problem that composes n factors to be considered, binary relations and the degree of influence of criteria i to criteria j are investigated. All pairwise comparisons between the i^{th} criteria and the j^{th} criteria is denoted as a_{ij} are performed which takes an integer score ranging between 0 (no influence), 1 (low influence), 2 (middle influence), 3 (high influence), and 4 (very high influence).

Step 2: Establish an Initial Direct-relation Matrix. Each expert would produce an n x n direct matrix. Each value in the matrix represents the size of an interactive influence between factors. When i=j, the diagonal values in the matrix are set as 0.

Suppose m is the number of experts participated in the study. An average matrix Z i.e., $Z = [z_{ij}]_{nxn}$ is then derived through the mean of the same factors in the different direct matrices of the experts as follows:

$$z_{ij} = \frac{(a_{ij(1)} + a_{ij(2)} + \ldots + a_{ij(k)} + \ldots + a_{ij(m)})}{m} \quad (1)$$

Step 3: Calculate the Normalized Direct-relation Matrix. The normalized direct-relation matrix X, i.e. $X = [x_{ij}]_{nxn}$ and $0 \leq x_{ij} \leq 1$ can be acquired from the

equations (2) and (3). All diagonal elements are equal to zero.

$$X = S*Z \tag{2}$$

$$S = \frac{1}{\max_{1\le i\le n}\sum_{j=1}^{n} z_{ij}}; i, j = 1, 2, ..., n \tag{3}$$

Step 4: Derive the total-relation matrix T. The total-relation matrix T can be obtained by using the equation (4), where I is denoted as the identity matrix.

$$T = X(I-X)^{-1} \tag{4}$$

The sum of rows and the sum of columns are separately denoted as R and C within the total-relation matrix T through the equations (5)-(7):

$$T = \left[t_{ij} \right]_{\text{nxn}} \quad i,j = 1,2,\ldots,\text{n} \tag{5}$$

$$R = \sum_{j=1}^{n} t_{ij} \tag{6}$$

$$C = \sum_{i=1}^{n} t_{ij} \tag{7}$$

where R and C represent the sum of rows and the sum of columns respectively.

Step 5: Produce a Causal Diagram. A causal diagram can be acquired by mapping the dataset of (R+C, R–C), where the horizontal axis (R+C) named "Prominence" is made by adding R to C, and the vertical axis (R–C) named "Relation" is made by subtracting C from R. Additionally, when $i = j$ (i.e., the sum of the row and column aggregates) (R_i+C_i) provides an index of the strength of influences given and received, that is, (R_i+C_i) demonstrates the degree of the central role that factor i plays in the problem. If ($R_i–C_i$) is positive, then factor i is affecting other factors and if ($R_i–C_i$) is negative, it implies that factor i is being influenced by other factors (Tzeng *et al.* 2007).

CASE ILLUSTRATION AND RESULTS

To identify the relationship of factors affecting consumers' behaviors in buying green products, the steps of applying DEMATEL approach are employed as follows.

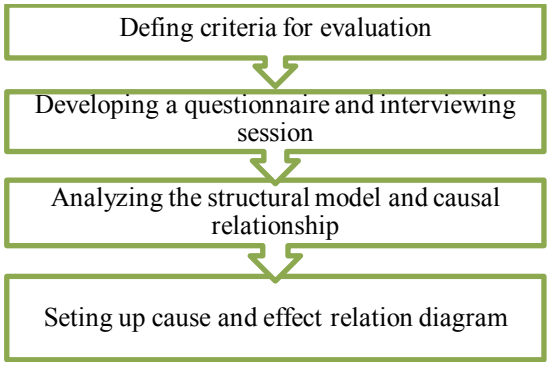


Figure 1: Steps of Applying DEMATEL Approach

Defining Criteria for Evaluation

Based on the previous research of the authors, factors that have an influence on buying of green products were explored via questionnaires. Data were gathered from 288 Bangkok Metropolitan respondents in Thailand and were analyzed using Factor Analysis (Atthirawong and Panprung 2017). Grouping of variables influencing on decision to purchase green products were extracted from the study i.e. perception on environmental concerns (A), safety and health concerns (B), green packaging (C), convenience to buy (D), environmental attitude (E), subjective norms (F), green product management (G), environmental laws (H), and perceived value (I). This previous study was extended by utilizing these nine factors for evaluation of the relationship and prioritization of them.

Developing a Questionnaire and Interviewing Session

A questionnaire of DEMATEL was developed using nine factors. According to Teng (2002), a group decision-making is more appropriate with 5 to 15 experts. In the numerical case example of Gandhi *et al.* (2015), five decision-makers were engaged to determine the degree of direct influence between two factors through a pairwise comparison, while TÜRKER *et al.* (2016) employed eight experts in their evaluation process. Therefore, in this paper, questionnaires were distributed to six experts to determine the relative importance of all criteria. Three of them were targeting consumers, another one is a professor in the university and the rest are managers who have worked in green product companies for several years. Each of them was provided a direct-relation 9x9 matrix by evaluating the influence degree of a factor to the others through a pairwise comparison. To do this comparison, the experts had expressed their opinions through variables ranging from "no influence" to "very high influence" and the verbal variables were converted to absolute numbers as displayed in Table 1 (Hessami and Yousefi 2013).

Table 1: Absolute Numbers of Verbal Variables

Linguistic terms	Influence score
No influence (NO)	0
Low influence (VL)	1
Middle influence high (L)	2
High Influence (H)	3
Very high influence (VH)	4

Analyzing the Structural Model and Causal Relationship

The average matrix *Z* was constructed in accordance with equation (1). Next, the normalized direct-relation matrix was generated by using equations (2) and (3) as displayed in Table 2. Table 3 illustrates the total-relation matrix (*T*) which was computed by using equation (4). Then, the sum of row i of matrix *T* , which is denoted as R_i and the sum of column j of matrix *T*, which is denoted as C_i are computed by using equations (6) and (7). Consequently, a data set of (R_i+C_i, R_i–C_i) is then obtained to identify the degree of prominence and net effects, as shown in Table 4.

Table 2: The Normalized Direct-relation Matrix

	A	B	C	D	E	F	G	H	I
A	0	0.140	0.127	0.047	0.153	0.053	0.100	0.107	0.127
B	0.140	0	0.120	0.080	0.133	0.073	0.140	0.133	0.140
C	0.113	0.140	0	0.053	0.127	0.060	0.140	0.060	0.087
D	0.047	0.080	0.067	0	0.08	0.08	0.067	0.053	0.06
E	0.147	0.133	0.147	0.027	0	0.087	0.147	0.133	0.147
F	0.100	0.127	0.053	0.047	0.113	0	0.06	0.060	0.093
G	0.133	0.127	0.147	0.087	0.140	0.053	0	0.107	0.140
H	0.147	0.133	0.093	0.047	0.120	0.027	0.140	0	0.127
I	0.153	0.120	0.120	0.047	0.113	0.067	0.133	0.113	0

Table 3: The Total-relation Matrix

	A	B	C	D	E	F	G	H	I
A	0.658	0.775	0.713	0.343	0.777	0.387	0.72	0.627	0.732
B	0.842	0.713	0.763	0.400	0.822	0.435	0.809	0.698	0.800
C	0.705	0.722	0.553	0.327	0.705	0.366	0.699	0.546	0.650
D	0.452	0.478	0.433	0.186	0.473	0.285	0.450	0.379	0.441
E	0.863	0.846	0.798	0.362	0.719	0.452	0.828	0.709	0.820
F	0.599	0.616	0.515	0.276	0.600	0.261	0.541	0.469	0.565
G	0.817	0.807	0.768	0.397	0.809	0.410	0.669	0.661	0.782
H	0.777	0.759	0.679	0.340	0.742	0.358	0.742	0.524	0.724
I	0.796	0.764	0.712	0.346	0.752	0.399	0.749	0.636	0.624

Setting Up Cause and Effect Relation Diagram

Lastly, the cause and effect relation diagram was constructed with the horizontal axis (R+C) and the vertical axis (R–C). According to Figure 2 and Table 4, it is clear that the most important factor affecting consumers' decision in buying green products is environmental attitude (E). It is followed by safety and health concerns (B), green product management (G) and perception on environmental concerns (A), respectively. Whereas, convenience to buy (D) is the least important among those nine criteria. In addition, the digraph also separates nine criteria into two groups according to whether their value of (R_i–C_i) is positive or negative. The cause group includes subjective norms (F), convenience to buy (D) and environmental laws (H), while the effect group comprises of perception on environmental concerns (A), green packaging (C), safety and health concerns (B), perceived value (I), green product management (G) and environmental attitude (E).

Table 4: The Degree of Prominence and Net Cause/Effects

Criteria	Details	R_i	C_i	R_i+C_i	Rank	R_i–C_i
A	perception on environmental concerns	5.732	6.509	12.241	4	-0.777
B	safety and health concerns	6.282	6.480	12.762	2	-0.198
C	green packaging	5.273	5.934	11.207	6	-0.661
D	convenience to buy	3.577	2.977	6.554	9	0.600
E	environmental attitude	6.397	6.399	12.796	1	-0.002
F	subjective norms	4.442	3.353	7.795	8	1.089
G	green product management	6.120	6.207	12.327	3	-0.087
H	environmental laws	5.645	5.249	10.894	7	0.396
I	perceived value	5.778	6.138	11.916	5	-0.360

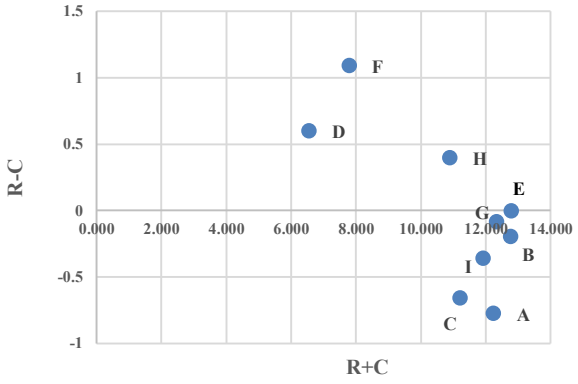


Figure 2: The Causal Diagram

CONCLUSION AND FURTHER RESEARCH

This study applied the DEMATEL method to analyze and identify the most important criteria in buying green products. The top three important criteria are environmental attitude (E), safety and health concerns (B) and green product management (G), respectively. It is in line with the study of Cornelissen *et al.* (2008) which states that environmental attitude has a significant impact on green purchasing behavior

decision. In contract, convenience to buy (D) is found to be the least important for consumers in making their decisions in buying green products.

The results of the study also portray the cause and effect visualization of that subjective norm (F), convenience to buy (D) and environmental laws (H) have the most influence on other factors. Therefore, these factors should be considered for effective criteria that will satisfy customers' decision in buying green products. In opposite, perception on environmental concerns (A) gets the most impact from other factors. The study could help public policies and marketing managers in focusing on factors which have more impact towards consumers' decision in buying green products. Specifically, in order to increase the number of green consumers, it must be paid attention on how convenient or easy it is for people to reach and buy those products. The study also reveals that via using social impact, it could stimulus Bangkokian people in green purchasing behavior (Finisterrado and Raposo 2004). Besides, environmental laws and regulation about green products should also be urgently launched by authorities in order to enforce people and companies to be aware of environmental impact and turn into green consumers or manufactures. The finding is supported by Kainpour *et al*. (2014). It is claimed that laws and regulations on environment could shape people attitude in ecological products as well as helping consumers to have better confidence in products they purchase taking as well.

The proposed model of this study could be extended by taking fuzzy environments into consideration in order to deal with the imprecise judgments, and the ambiguity of human being's judgment. This study provides a static relationship of the relationships among criteria. Nevertheless, these relationships may change over time. As such, what-if analysis based on the dynamic situation should be conducted. Moreover, there are other multi attribute decision-making methods such as TOPSIS and PROMETHEE which could be applied for ranking those criteria or comparing the results of the study. In addition, the test of the model in a different part of Thailand could also help to decrease research gaps in this area.

ACKNOWLEDGEMENT

The authors of this study would like to acknowledge the National Research Council of Thailand (NRCT), the sponsor of this research funding. The authors are also grateful to all experts who were kindly involved and provided valuable and fruitful information for this study.

REFERENCES

Abbasi, M.; Hosnavi R.; and Tabrizi, B. H. 2013. "Application of Fuzzy DEMATEL in Risks Evaluation of Knowledge-Based Networks". *Journal of Optimization Volume,* available at: http://dx.doi.org/10.1155/2013/913467 [accessed 19 December 2017].

Atthirawong, W. and Panprung, W. 2017. "Applying Two-Step Cluster Analysis for Explaining Consumers' Behavior in Buying Green Products". *The Journal Scientific Social Studies,* Vol. 1, No. 1, 83-90.

Chen, C. 2012. "Using DEMATEL Method for Medical Tourism Development in Taiwan". *American Journal of Tourism Research,* Vol. 1, No. 1, 26-32.

Chen, L. 2013. "A Study of Green Purchase Intention Comparing with Collectivistic (Chinese) and Individualistic (American) Consumers in Shanghai, China". *Information Management and Business Review,* Vol. 5, No. 7, 342-346.

Cornelissen, G.; Pandelaere, M.; Warlop, L.; and Dewitte, S. 2008. "Positive cueing: Promoting sustainable consumer bahaviour buying cueing common environmental behavior as environmental". *International Journal of Research in Marketing,* Vol. 25, No. 1, 46-55.

Elham, R. and Nabsiah, A. W. 2011. "Investigation of green marketing tool's effect on consumers' purchase behavior". *Business Strategy Series*, Vol. 12, No. 2, 73-83.

Finisterrado, P. A. M. and Raposo, M. L. B. 2004. "Determining the characteristics to profile the green consumer: an exploratory approach". *International Review on Public Non-profit Marketing,* 211-240.

Fontela, E. and Gabus, A. 1976. The DEMATEL Observer. Report. Geneva, Switzerland Geneva: Battelle Geneva Research Center.

Gabus, A. and Fontela, E. 1972. *World Problems an Invitation to Further Thought within the Framework of DEMATEL.* Switzerland Geneva: Battelle Geneva Research Centre.

Gandhi, S.; Manglab, S.; Kumarc, P.; and Kumard, D. 2015. "Evaluating factors in implementation of successful green supply chain management using DEMATEL: A case study". *International Strategic Management Review,* Vol. 3, 96–109.

Grimmer, M. and Bingham, T. 2013. "Company environmental performance and consumer purchase intentions". *Journal of Business Research*, Vol. 66, No. 10, 1945-1953.

Hartmann, P. and Ibáñez, V. A. 2006. "Green value added". *Marketing Intelligence & Planning,* Vol. 24, No. 7, 673-680.

Hertwich, E. 2005. "Life cycle approaches to sustainable consumption: a critical review*". Environment Science and Technology,* Vol. 39, No. 13, 46-73.

Hessami, H. Z. and Yousefi, P. 2013. "Investigation of major factors influencing green purchasing behavior: Interactive approach". *European Online Journal of Natural and Social Sciences,* Vol. 2, No. 4, 584-596.

Hessami, H. Z.; Yousefi, P.; and Goudarzi, G. 2013. "The conceptual model of effective factors on consumers green purchasing intentions". *International Journal*

of Engineering and Innovative Technology, Vol. 2, No. 7, 10-17.

Kianpour, K.; Anvari, R.; Jusoh, A,; and Othman, M.F. 2014. "Important Motivators for Buying Green Products". *Intangible Capital*, Vol. 10, No. 5, 873-896. available at: http://creativecommons.org/licenses/by-nc/3.0/ [accessed 19 December 2017].

Lee, K. 2008. "Opportunities for green marketing: young consumers". *Marketing Intelligence & Planning*, Vol. 6, No. 6, 573-586.

Moisander, J. 2007. "Motivational complexity of green consumerism". *International Journal of Consumer Study,* Vol. 31, No. 4, 404-409.

Opricovic, S. and Tzeng, G. H. 2003. "Defuzzification within a multicriteria decision model". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.11, No. 5, 635-652.

Ottman, J. 1992. "Sometimes consumers will pay more to go green". *Marketing News,* Vol. 26, 6-16.

Padel, S. and Foster, C. 2005. "Exploring the gap between attitudes and behaviour: Understanding why consumers buy or do not buy organic food". *British Food Journal,* Vol. 107, No. 8, 606-625. available at: DOI=https://doi.org/10.1108/00070700510611002 [accessed 19 December 2017].

Shieh, J. I.; Wub, H. H.; and Huang, K. K. 2010. "A DEMATEL method in identifying key success factors of hospital service quality". *Knowledge-Based Systems*, Vol. 23, 277-282. available at: http://dx.doi.org/10.1016/j.knosys.2010.01.013 [accessed 8 December 2017].

Singh, D. P. 2011. "Indian ecological consumer market profile". *Global Business Review*, Vol. 12, No. 3, 447-457.

Tzeng, G. H.; Chiang, C. H.; and Li, C. W. 2007. "Evaluating intertwined effects in learning programs: a novel hybrid MCDM model based on factor analysis and DEMATEL". *Expert System Applications*, Vol. 32, 1028-1044.

TÜRKER, T., ETÖZ, T. M.;and TÜRKER, Y. A. 2016. "Determination of Effective Critical Factors in Successful Efficiency Measurement of University Departments by using Fuzzy Dematel Method". *The Journal of Operations Research, Statistics, Econometrics and Management Information Systems,* Vol. 4, No. 1, 57-72.

UN Global Compact. 2010. "Accenture release findings of largest CEO research study on corporate sustainability". available at: www.unglobalcompact.org/news/42-06-22-2010 [accessed 12 January 2018].

Veleva, V. and Ellenbecker, M. 2001. "Indicators of sustainable production: framework and methodology". *Journal of Cleaner Production,* Vol. 9, No. 6, 519-549.

Ward, D.O.; Clark, C. D.; Jensen, K.; and Russell, C.S. 2011. "Factors Influencing Willingness-to -pay for the Energy Stats Level". *Energy Policy*, Vol. 39, 1450-1458.

Wu, W. W. 2008. "Choosing knowledge management strategies by using a combined ANP and DEMATEL approach". *Expert System Applications,* Vol. 35, 828-835.

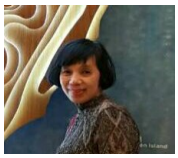
Wu, W. and Lee, Y. 2007. "Developing global managers' competencies using the fuzzy DEMATEL method". *Expert Systems with Applications*, Vol. 32, No. 2, 499-507.

Wu, H. Y.; Lin, Y. K.; and Chang, C. H. 2011. "Performance evaluation of extension education centers in universities based on the balanced scorecard". *Evaluation and Program Planning,* Vol. 34, No. 1, 37-50.

Zhou, Q.; Huang, W.; and Zhang, Y. 2011. "Identifying critical success factors in emergency management using a fuzzy DEMATEL method". *Safety Science*, Vol. 49, No. 2, 243-252.

AUTHOR BIOGRAPHIES

Walailak Atthirawong is Associate Professor of Operations Research at Faculty of Science, King Mongkut's Institute of Technology Ladkrabang (KMITL) in Thailand. She had received doctoral degree from the University of Nottingham in Manufacturing Engineering and Operations Management. She is actively engaged research in logistics and supply chain management, simulation, multi-criteria decision making, applied statistics and optimization. Her e-mail address is: walailaknoi@gmail.com.



Wariya Panprung was born in Chonburi, Thailand. She had got her Master degree from Srinakarinwirot University in Accounting and Ramkhamheang University in Business Administrator. Now she is an Assistant Professor in Management Science at Phranakhon Rajabhat University. She is actively engaged research in accounting, applied statistics, logistics and supply chain management. Her e-mail address is: noiwari@gmail.com.

Kanogkan Leerojanaprapa is an Assistant Professor in Applied Statistics at Faculty of Science, Mongkut's Institute of Technology Ladkrabang (KMITL). She obtained her PhD in Management Science from the University of Strathclyde, UK in 2014. Her research focus is in the area of supply chain risk and statistical analysis. Her current e-mail address is: kanogkan.le@kmitl.ac.th

SOLVING LOCATION PROBLEM FOR VEHICLE IDENTIFICATION SENSORS TO OBSERVE AND ESTIMATE PATH FLOWS IN LARGE-SCALE NETWORKS

Pegah T. Yazdi and Yousef Shafahi
Department of Civil engineering
Sharif University of Technology
Azadi St., Tehran 11365-9313, Iran
E-mail: shafahi@sharif.edu

KEYWORDS
Location Problem, Vehicle Identification Sensor, Large-Scale Networks, Heuristic and Meta-Heuristic

ABSTRACT

Origin-Destination (OD) demand is one of the important requirements in transportation planning. Estimating OD demand could be an expensive and time consuming procedure. These days using vehicle identification sensors for OD estimation has become very common because of its low cost and high accuracy. In this paper, we focus on solving two location problems of these sensors: one to observe and one to estimate path flows. These problems have only been solved for small-scale networks until recently due to being computationally expensive. Therefore, we try to present a method to solve these models for large-scale networks. Due to resemblance of these models and set covering problem, we used heuristic and meta-heuristic methods based on set covering problem. For this purpose, we defined our new set covering matrix based on prime matrix. In order to determine which method is more appropriate, we chose a large-scale and six medium-scale networks. The results represent that through heuristic methods and meta-heuristic methods a greedy algorithm and a Tabu search are more appropriate respectively.

INTRODUCTION

One of the most important requirements in transportation planning is demand data of using networks. OD and path flows estimation is one of the approaches in this regard. Common methods of these estimation are in the form of household surveys, on-board vehicle surveys, using trip distribution, assignment models etc. which are mainly economically expensive and not always reliable. Therefore, these days using output information from ITS is also taken into consideration. This information can estimate OD and path flows with high accuracy and low cost. One type of ITS that is used for this purpose is the use of sensors. However due to high price of these sensors, solving location problem and reducing their number in network is very important.

Gentili and Mirchandani (2012) represented two major categories for sensor location models:

- Flow-observability model: These models locate and determine the optimum number of sensors in the network so that the target flows can be obtained uniquely by solving a linear system of equations associated with the optimum number of located sensors.

- Flow-estimation model: These models locate and determine limited number of sensors in the network due to budget constraints so that the best estimation of target flows can be obtained.

Here, target flow is a term for three type of flows: OD flow, path flow and arc flow. Also four categories of sensor types have been represented: counting sensors, path-ID sensors, image sensors and vehicle identification sensors. Gentili and Mirchandani (2012) In this paper, we focused on path flow observability and estimation based on vehicle identification sensors.

Through literature review, two location models have been selected for studying in this paper. First, a vehicle identification sensor location model for path flow observability represented by Castillo et al. (2008). This model finds the minimum number of arcs to locate vehicle identification sensors so that path flows can be observed uniquely. Second, a vehicle identification sensor location model for path flow estimation represented by Minguez et al. (2010). This model finds limited number of arcs for locating vehicle identification sensors (because of budget constraints) so that the quality of path flow estimation can be maximized with limited number of sensors. In both models the assumption is that the arc number where vehicle has been seen and the vehicle unique ID, from extracted information of vehicle identification sensors is available.

Up to now, these location problems were only solved for small-scale networks. But in practice it is in the large-scale networks that location and installation of sensors are more necessary. Due to volume of calculations and lack of solution we have to present a method to solve these models for large-scale networks.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Literature review shows that Castillo et al. (2008)'s model is an instance of set covering problem. Due to resemblance of our two models we used set covering solutions for both of them. For this purpose, we adjust our model with set covering matrix. Finally, we solve our models for six medium-scale networks and a large-scale network. Our results present a method for solving location models for large-scale networks.

The observability model

In this model, Castillo et al. (2008) allocated the minimum number of vehicle identification sensors in the network required for unique observation of all path flows. For this purpose, they appointed a set of arcs to a vehicle. Each set shows the arcs which vehicle has been seen in them. Thus, The path that each vehicle has taken becomes known. They formulated the model as follows:

A: Minimize $\quad (1)\ N_s = \sum_{a \in A} J_a$

Subject to: $(2)\ J_a \in \{0,1\}\ ,\ \forall a \in A$

$(3)\ \sum_{a \in A} J_a \cdot d(r_1, r_2, a) \geq 1$

$,\ \forall r_1, r_2 \in R | r_1 \neq r_2$

$(4)\ \sum_{a \in A} J_a \cdot \delta_{ra} \geq 1\ ,\ \forall r \in R$

Where J_a is a binary variable and it takes value (of) "1" if a sensor is located on link "a" and it's "0" otherwise. A is the set of all links and "R" is the set of all paths of the network. Parameter δ_{ra} equals "1" if link "a" exists in path "r" otherwise it equals "0". $d(r_1, r_2, a)$ is a binary parameter too. It takes value of "1" if link "a" exists in only one of the paths "r_1" and "r_2" and it's "0" otherwise.

In this model, with the objective function (1) of minimizing number of sensors which should install in the network, constraint (4) ensures that each path gets at least one sensor and constraint (3) guarantees that there is at least one uncommon link between every tow paths that get a sensor. With these two constraints we ensure that path flow could be observed uniquely.

The estimation model

Minguez et al. (2010) developed an estimation model based on previous model. Due to consideration of budget constraints, this model is more practical. Their model with the assumption that no prior OD matrix information is available, was written as follows:

B: Maximize $\quad (5)\ N_r = \sum_{r \in R} x_r$

Subject to: $(6)\ J_a \in \{0,1\}\ ,\ \forall a \in A$

$(7)\ x_r \in \{0,1\}\ ,\ \forall r \in R$

$(8)\ \sum_{a \in A} J_a \cdot d(r, r_1, a) \geq x_r$

$,\ \forall r, r_1 \in R | r \neq r_1$

$(9)\ \sum_{a \in A} J_a \cdot \delta_{ra} \geq x_r\ ,\ \forall r \in R$

$(10)\ \sum_{a \in A} C \cdot J_a \leq B\ ,\ \forall r \in R$

Where C is cost of buying and installing one sensor and B is available financial resource. x_r is a binary variable and when flow of path "r" could be observed it is equal to "1", otherwise it is "0".

In this model, with the objective function (5) of maximizing number of paths that their flow could be observed uniquely, constraint (10) controls the number of sensors that could be afforded with available budget. Other constraints and parameters are same as Castillo et al. (2008)'s model but the right part of the constraints (8) and (9) are x_r. It means that this model tries to maximize the number of paths that get at least one sensor and have at least one uncommon sensored arc with each other paths (flow observable path).

Set covering problem

Zangui et al. (2015) proved that model A is an instance of set covering problem. Thus, solving model A with the solutions of such problem can represent acceptable answers. Due to resemblance of model A and B we use same solutions for model B.

Set covering problem is a NP-complete problem of covering a $m \times n$ zero-one matrix by selecting columns at minimum cost. Let $A = (a_{ij})$ be a $m \times n$ zero-one matrix with $M = \{1.2. \ldots . m\} . N = \{1.2. \ldots . n\}$ and $C = (c_1 . c_2 . \ldots . c_n)$ denoting respectively the sets of rows and columns and cost of each column of A. This is said that column $j \in N$ covers row $i \in M$ if $a_{ij} = 1$. The target in this problem is to find a set of columns $(S \subseteq N)$ with the minimum cost where each row $i \in M$ is covered by at least one column $j \in S$. The classical mathematical formulation for Set covering problem is:

Minimize $\quad (11)\ V = \sum_{j \in N} c_j x_j$

Subject to: $\quad (12)\ \sum_{j \in N} a_{ij} . x_j \geq 1\ ,\ \forall i \in M$

$(13)\ x_j \in \{0,1\}\ ,\ \forall j \in N$

Where x_j is the binary decision variable whose value is "1" if column j is in the solution(S) and "0" otherwise. Constraint (12) ensures that every row is covered by at least one column and constraint (13) is written for definition of x_j.

As previously mentioned $C = (c_j)$ is a vector with positive value as the cost of columns. In a version of set covering problem known as unicost set covering

problem all c_j for each column $j \in N$ are the same. Thus, we can put the value "1" instead of c_j in objective function. In this version, the target is to find the minimum number of columns where each row is covered by at least one column. It could be demonstrated that model A is an unicost set covering problem.

We use four algorithms to solve model A and B:

- Greedy Algorithm – Heuristic method (Johnson 1974)
- Greedy algorithm (four selected row knowledge function)– Heuristic method (Vasco et al. 2016)
  (1) $1/\sqrt{L_i}$ (L_i defined as sum of each row)
  (2) $1/\sqrt{(L_i-1)}$
  (3) $1/\sqrt{(L_i)^2}$
  (4) $1/\sqrt{(L_i-1)^2}$
- Tabu Search Algorithm – Meta-Heuristic method (Cerrone et al. 2015)
- Meta-RAPS Algorithm – Meta-Heuristic method (Lan et al. 2007)

We chose two heuristic and two meta-heuristic methods. Heuristics for lower run time and meta-heuristic for better answers. Johnson (1974)'s greedy algorithm is the basis for most of algorithms after that. Vasco (2016)'s greedy algorithm considers rows priority in selecting best column in each iteration. Tabu search and Meta-RAPS algorithms both are based on Johnson (1974)'s greedy algorithm.
We chose Johnson (1974)'s greedy algorithm to compare our methods with the basis. And we chose Vasco (2016)'s greedy algorithm because in our problem number of rows is more than columns and it seemed that considering rows priority could help reaching better answers. Also reason of choosing Cerrone et al. (2015)'s Tabu search was using it in solving some similar models by Cerrone et al. (2015). We chose Meta-RAPS algorithm because of high randomness in choosing columns to see the effect of that in this type of models.
Tabu search and Meta-RAPS algorithms have some parameter which need to be determined. We determined these parameter by examining different values of parameters for solving our medium-scaled networks which their exact answers are in hand. The value of parameters which reached the closest to the exact answers had been chosen.

DEFINING OUR NEW SET COVERING MATRIX

In this section, we describe how the models adjust with the set covering problem. Firstly, we should define our rows, columns and elements of set covering problem matrix. Then determining the problem for each model is necessary.

We set our columns as arcs of the network that we should select them as the place of installation of each sensor. Every row of set covering matrix stands for one of these two subjects: paths of network and couple-paths of network. The definition of elements which are equal to 1 is different for rows that stand for paths of network and rows that stand for couple-paths of network. The element a_{ij} for paths of network equals 1 if arc j would exist in the related row and the element a_{ij} for couple-paths of network equals 1 if arc j would exist in just one of the related couple-rows.

Let assume that $R_1 = (a_1, a_4)$ and $R_2 = (a_1, a_3)$ are two paths of a network and $A = (a_1, a_2, a_3, a_4)$ is the set of all arcs in network. Set covering matrix of this network is written as below:

	a_1	a_2	a_3	a_4
R_1	1	0	0	1
R_2	1	0	1	0
R_1-R_2	0	0	1	1

Figure 1 An Example of a Set Covering Matrix

Observability models should find the minimum number of arcs for installation of sensors, therefore the target in solving this problem is to find the minimum number of columns that cover all rows.

Estimation models should find exact number of arcs for installation of sensors with which target flows could be estimated with the highest quality. The definition of quality in model B is the number of paths that their flow can be observed uniquely. We changed this definition with the aim of simplification in order to solve this model with set covering solutions. We define the quality of flow estimation as the number of paths that at least have one sensor plus the number of couple paths that have at least one uncommon sensor. The objective function of this new definition has been written as below:

$$\text{Maximize:} \quad N_{rr'} = \sum_{r,r' \in R} x_{rr'}$$

Therefore, the target in solving this problem is to find exact number of columns that cover maximum number of rows.

PROBLEM REDUCTION

As previously mentioned, the goal of this paper is to solve models for large-scale networks. Due to volume of calculations, we looked for approaches to reduce this

volume and time of calculations. There are various approaches that reduce size of set covering matrix. These reduction approaches are either related to the columns (variables) or to the rows (constraints). As we saw in our models, the number of constraints is larger than the number of variables, thus we choose row reduction to diminish set covering matrix.

Beasley (1987) reviewed the literature and presented an effective row reduction procedure. In this procedure redundant rows are being deleted. Beasley defines a redundant row as: "If there is a subset for row i in set covering matrix, row i is redundant."

It is worth mentioning that we can't use row reduction for our estimation model (model B with new objective function) because the target is covering maximum number of rows so existence of every row would be necessary.

RESULTS

To determine how well our selected set covering solutions is doing, we choose six path collections of Sioux Falls medium-scale networks and one path collection of Mashhad's large-scale network. The characteristics of these networks and their set covering matrix are given in table 1. Percent of remaining rows show that row reduction for this problem could delete 85% to 99% of set covering matrix rows.

On each instance, we compare the results returned by Johnson, Vasco et.al (four different row knowledge function), Cerrone et al.(2015)'s Tabu search and Meta-RAPS algorithms for each of our two problems. All algorithms were coded in MATLAB on a 2.5 GHz Intel i5 processor and 6.00 GB RAM.

Table 1: Characteristics of Instances

Network	nodes	OD	arcs	paths	Paths + Couple-Paths (rows)	Rows after reduction	Percent of remaining rows
Sioux Falls 1	24	12	76	140	9'870	1'490	15.963 %
Sioux Falls 2	24	12	76	225	25'425	3'282	12.908 %
Sioux Falls 3	24	12	76	305	46'665	4'269	9.148 %
Sioux Falls 4	24	30	76	245	30'135	91	0.302 %
Sioux Falls 5	24	30	76	448	100'576	216	0.215 %
Sioux Falls 6	24	30	76	837	350'703	399	0.114 %
Mashhad	917	7'157	2'092	44'210	977'284'155	1'468'803	0.150 %

Table 2 report the results of solving model A with four selected algorithms. Values in this table are values of objective function (number of necessary sensors to observe all path flows in network). The exact answers for Sioux falls networks were obtained by GAMS. As it's clear in table 2, Tabu search could reach to better answers in 100% of instances in average of 32.47 seconds for Sioux falls networks and 1'690'824 seconds for Mashhad. Vasco4 (row knowledge function: $1/\sqrt{(L_i-1)^2}$) is the best algorithm among heuristics and it reached better answers in 16'765.23 seconds for our large-scale network.

Table 2: The Value of Objective Function after Solving Model A

Network	Exact answer	Johnson	Vasco1	Vasco2	Vasco3	Vasco4	Tabu search	Meta-RAPS
Sioux Falls 1	24	27	27	28	26	25	24	24
Sioux Falls 2	26	31	30	29	27	29	26	26
Sioux Falls 3	27	31	31	30	29	29	27	27
Sioux Falls 4	31	31	31	31	31	31	31	31
Sioux Falls 5	35	36	36	36	35	36	35	35
Sioux Falls 6	35	39	38	38	35	36	35	35
Mashhad	UK	679	678	671	668	659	651	655

Table 3-5 show the results of solving model B with three levels of budget constraint. We define budget constraint as a percent of observability model's answer (exact number of sensors). For number of sensors we chose 25%, 50% and 75% of exact answers for Sioux falls and 25%, 50% and 75% of Tabu search's answer for Mashhad. Values in these tables are number of uncovered rows, meaning number of rows minus value of objective function.

Thus, bigger values of objective function cause lower number of uncovered rows. As previous, Tabu search could reach to better answers in 100% of instances in three levels of budget constraint. But among heuristic methods that their run time is fewer, Vasco1 (row knowledge function: $1/\sqrt{L_i}$) could reach to better answers.

Table 3: Number of Uncovered Rows after Solving Model B with New Objective Function (Budget Constraint: 25%)

Network	Johnson	Vasco1	Vasco2	Vasco3	Vasco4	Tabu search	Meta-RAPS
Sioux Falls 1	912	859	859	1'034	1'159	841	841
Sioux Falls 2	1'443	1'482	1'482	1'833	2'246	1'384	1'384
Sioux Falls 3	2'127	2'115	2'115	2'468	3'082	2'115	2'115
Sioux Falls 4	2'029	2'029	3'289	2'639	3'289	2'023	2'023
Sioux Falls 5	4'580	4'795	8'897	7'212	8'897	4'432	4'555
Sioux Falls 6	11'774	11'784	24'912	15'104	24'912	10'891	10'891
Mashhad	471'907	387'980	951'983	464'745	977'329	329'500	342'074

Table 4: Number of Uncovered Rows after Solving Model B with New Objective Function (Budget Constraint: 50%)

Network	Johnson	Vasco1	Vasco2	Vasco3	Vasco4	Tabu search	Meta-RAPS
Sioux Falls 1	125	116	116	158	158	107	107
Sioux Falls 2	253	242	242	291	286	214	221
Sioux Falls 3	364	401	389	511	413	294	320
Sioux Falls 4	240	251	317	317	317	216	233
Sioux Falls 5	437	481	724	735	749	414	422
Sioux Falls 6	1'073	981	1'742	1'462	1'911	909	971
Mashhad	32'000	25'596	32'533	36'321	45'660	20'287	21'895

Table 5: Number of Uncovered Rows after Solving Model B with New Objective Function (Budget Constraint: 75%)

Network	Johnson	Vasco1	Vasco2	Vasco3	Vasco4	Tabu search	Meta-RAPS
Sioux Falls 1	24	31	25	33	30	17	20
Sioux Falls 2	43	39	39	45	45	25	30
Sioux Falls 3	62	57	61	60	77	36	36
Sioux Falls 4	50	43	36	50	66	31	31
Sioux Falls 5	71	90	67	62	72	52	60
Sioux Falls 6	103	124	139	175	192	97	103
Mashhad	1'596	1'592	1'535	3'006	2'868	1'026	1'339

CONCLUSION

In this paper, we adjusted two allocation models with set covering problem in order to solve our models for large-scale networks. Two heuristic algorithms from Johnson (1973) and Vasco et al. (2016) and two meta-heuristic algorithms from Cerrone et al. (2015) and Lan et al. (2007) were used. The results of solving the models with these algorithms in the medium-scale and large-scale networks show that Tabu search algorithm which was presented by Cerrone et al. is better than the others. Even though the calculation time of location models are not an important constraint (because these models are only solved once), Tabu search's run-time is much higher than heuristic algorithms. Nevertheless, the greedy algorithm of Vasco et al. (2016) solves our models in shorter time and it's more suited for heuristic methods.

REFERENCES

Beasley, J.E. 1987 “An algorithm for set covering problem”. *European Journal of Operational Research* 31, No.1 85-93.

Castillo, E.; J.M. Menéndez; and P. Jiménez. 2008. “Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations”. *Transportation Research Part B: Methodological* 42, No.5, 455-481.

Cerrone, C.; R. Cerulli; and M. Gentili. 2015. “Vehicle-id sensor location for route flow recognition: Models and algorithms”. *European Journal of Operational Research* 247, No.2, 618-629.

Gentili, M.; and P.B. Mirchandani. 2012. “Locating sensors on traffic networks: Models, challenges and research opportunities”. *Transportation research part C: emerging technologies* 24, 227-255.

Johnson, D.S. 1974. “Approximation algorithms for combinatorial problems”. *Journal of computer and system sciences* 9, No.3, 256-278.

Lan, G.; G.W. DePuy; and G.E. Whitehouse. 2007. “An effective and simple heuristic for the set covering problem”. *European journal of operational research* 176, No.3, 1387-1403.

Mínguez, R.; S. Sánchez-Cambronero; E. Castillo; and P. Jiménez. 2010. “Optimal traffic plate scanning location for OD trip matrix and route estimation in road networks”. *Transportation Research Part B: Methodological* 44, No.2, 282-298.

Vasko, F.J.; Y. Lu; and K. Zyma. 2016. “What is the best greedy-like heuristic for the weighted set covering problem?”. *Operations Research Letters* 44, No.3, 366-369.

Zangui, M.; Y. Yin; and S. Lawphongpanich. 2015. “Sensor location problems in path-differentiated congestion pricing”. *Transportation Research Part C: Emerging Technologies* 55, 217-230.

AUTHOR BIOGRAPHIES

PEGAH T. YAZDI was born in Tehran, Iran in 1992. She went to the University of Tehran, where she studied civil engineering and obtained her bachelor degree in 2015. After that she received her master of science in transportation planning and engineering in 2017 from Sharif university of technology. Her e-mail address is: pghtlbn@gmail.com

YOUSEF SHAFAHI received his Ph.D. in Civil Engineering from University of Maryland, USA in 1997. He joined the Department of Civil Engineering in Sharif University of Technology as an Assistant Professor in 1998 and became an Associated Professor in 2009. His main research interest is operation research in Transportation Planning. His email address is shafahi@sharif.edu and his web-page can be found at http://sharif.edu/~shafahi

Europäische Union
Europa fördert Sachsen.
ESF
Europäischer Sozialfonds

MASTER PRODUCTION SCHEDULING WITH INTEGRATED ASPECTS OF PERSONNEL PLANNING AND CONSIDERATION OF EMPLOYEE UTILIZATION SPECIFIC PROCESSING TIMES

Marco Trost
International Institute (IHI) Zittau
Technical University of Dresden
Markt 23, 02763 Zittau, Germany
E-mail: marco.trost@mailbox.tu-dresden.de

KEYWORDS
Sustainability, Production Planning and Control, Master Production Scheduling, Linear Optimization, Exhaustion

ABSTRACT

This article deals with the integration of social criterias into production planning and control. Firstly, it is emphasized, that working conditions for employees have barely improved and that production planning and control can achieve considerable progress in this field. The presented literature demonstrate, that only a few number of papers consider the social dimension at the context of production planning and control. Therefore this paper presents a linear optimization model for the Master Production Scheduling, which can be considered as a long-term employee utilization management system in order to reduce employee burdens. The special feature of this model is the link between aspects of personnel planning und production planning and control. The system consideres different employee utilization intervals and employee utilization specific processing times and it includes flexible capacities, so it is allowed to build and to reduce the available capacity. This model is used to examine a case study, with the aim to verify the need for the control of employee utilization. The results confirm previous results from short-term planning horizons, which underline, that maximizing the utilization of employees does not necessarily lead to optimal results. Finaly, further research questions can be derived from the analysis of the results.

INTRODUCTION

The development of sustainable models for production planning and control has received considerable attention in recent years. However, many papers ignore the social dimension. This is confirmed by the findings from Schmucker et al. (2014), which shows, that working conditions have hardly improved at all. In addition, the existing work, which takes the social dimension into account, concentrates primarily on optimizing the distribution of burdens. In this way, the overall burden on employees is not reduced, but only redistributed. Long-term load management is not known.

For this reason, the present work combines aspects of personnel planning with the Master Production Scheduling. In addition, different employee utilizations are assumed in order to take into account of employee utilization specific processing times, so that the exhaustion or recovery (hereinafter called exhaustion) of employees is also taken into account. In the area of personnel planning, the hiring and dismissal of personnel resources is integrated, whereby different qualifications and experiences as well as effects on shift models and labour market conditions are taken into account. Thus, the burden of employees can be reduced without restricting production capacity, because available capacity and capacity requirements are flexible.

The model presented here is an extension of the original model from Trost et al. (2017a and 2017b), which underlines the relevance of employee utilization specific processing times and demonstrate, that the most cost-effective production programme is not accompanied by maximum employee utilization. The current paper reviews these results for a long-term planning horizon and develops the model accordingly. The paper is divided into a brief review of the literature, an introduction to the linear optimization model, following by a case study and the results. A conclusion completes the article.

LITERATURE REVIEW

Considering the planning stages of the hierarchical production planning (Master Production Scheduling, Lot Sizing and Scheduling) presented by Drexl et al. (1993), it can be stated, that only a few papers exist, which take the social dimension into account. As already indicated in Trost et al. (2016), the social dimension is often overlooked in the development of sustainable models. Exceptions in the area of lot-sizing are, for example, the work of Arslan and Turkay (2013), in which personnel hours are to be minimized, and the work of Jaber and Bonney (2007), which integrates 2-phase learning and forgetting effects into a classic economic manufacture quantity model. In the scheduling area, for example, the

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

work of Boysen and Fliedner (2011) is worth mentioning, in which the burden on the ground staff at an airport is minimised, which can be transferred to a production environment. Another work is that of Lai and Lee (2013), which integrates learning and forgetting effects into a scheduling model for a single-machine environment. It becomes clear, that many papers in this field mainly takes learning and forgetting effects into account and that no work is known, that takes social aspects into account in long-term planning models.

These findings are supported by the review of Grosse et al. (2017), who are looking for corresponding work and are setting up their own framework. The work classifies relevant papers in the areas of "inventory management and lot-sizing" (IM&LS), "production and assembly management" (P&AM) and "intra-logistics and warehouse management" (I&W). However, this does not apply to work that integrates learning effects into lot-sizing problems, since these have been researched more thoroughly, as mentioned before. As the work shows, although the number of corresponding papers has increased in recent years, there is still considerable potential for research. Relevant papers, mentioned by Grosse et al. (2017) for the IM&LS are, for example, the paper by Khan et al. (2014), which take cognitive human factors into account in lot-sizing and the paper by Andriolo et al. (2016), which determine an ergonomic lot size. In the area of P&AM, the paper by Otto and Scholl (2011) should be mentioned, which integrates a performance comparison with consideration of ergonomic risk factors in constraints. However, it could generally observed, that balancing problems are predominant in the foreground.

Especially physical effects have been less researched. The existing literature mainly deals with the improvement of social conditions by an optimized distribution of burden on different planning levels or tries to exploit learning and forgetting effects. However, the overall burden is not reduced, but merely redistributed. This paper therefore controls the overall burden in terms of keeping the employee utilization within a specific corridor, whereby the employee utilization reflects the quotient of available capacity and capacity requirement. Due to the possibility of building up and reducing capacities, there is also no impairment in the satisfaction of customer orders.

MATHEMATICAL NOTATIONS AND EQUATIONS

In order to reduce the previously outlined gaps in the integration of social criteria into production planning and control, the linear optimization model for Master Production Scheduling introduced in Trost et al. (2017a) and Trost et al. (2017b) was developed. To investigate a long planning horizon, this model was adapted with regard to the relevant requirements. Because of the resulting complexity, solutions cannot easily be achieved within an acceptable period of time. Therefore, the complete optimization problem has been broken down into a main problem and a subproblem. At the main problem, the decision variables are set to floating point numbers. After solving the main problem, the determined shift model is transferred to the subproblem. The subproblem determines the optimal results, whereby the number of employees is determined as an integer and the other variables remain floating point numbers.

Parameter

$CAPA_{ma}$	Available capacity per worker of a worker class *ma*
$d_{k,t}$	Product requirement per product *k* and time period *t*
$f_{z,j,k}$	Capacity requirement per forerun period *z*, production segment *j*, and product *k*
h_k	Cost rate for storage per product *k*
I_k^{Init}	Initial inventory per product *k*
J	Number of production segments ($j = 1, 2, \ldots, J$)
K	Number of products ($k = 1, 2, \ldots, K$)
MA	Number of worker classes ($ma = 1, 2, \ldots, MA$)
$Mit_{ma,j}^{Cost}$	Cost rate per worker of class *ma* and production segment *j*
$Mit_{ma,j,s}^{Init}$	Initial number of worker per worker class *ma*, production segment *j* and shift *s*
$Mit_{ma,j}^{Max}$	Maximum number of worker per worker class *ma* and production segment *j*
$Mit_{ma,j}^{Min}$	Minimum number of worker per worker class *ma* and production segment *j*
$Mit_j^{TotalMax}$	Maximum number of worker per production segment *j*
$Mit_j^{TotalMin}$	Minimum number of worker per production segment *j*
m_{ma}^{Cost}	Cost rate for building capacity per worker class *ma*
n_{ma}^{Cost}	Cost rate for reducing capacity per worker class *ma*
$p_{j,s}^{Init}$	Initial shift model per production segment *j* and shift *s*
$Q_{j,s}^{Cost}$	Cost rate for shift model change per production segment *j* and shift *s*
R_j^{Max}	Maximum worker utilization per production segment *j*
R_j^{Min}	Minimum worker utilization per production segment *j*
S	Number of shifts ($s = 1, 2, \ldots, S$)

$S_{j,s}^{Above}$ Maximum limit for number of worker per production segment j and shift s

$S_{j,s}^{Bottom}$ Minimum limit for number of worker per production segment j and shift s

S_{s}^{Cost} Cost factor for calculating shift bonuses per shift s

T Planning horizon in time periods ($t = 1, 2, ..., T$)

W Number of forerun periods for modifying available capacity ($w = 1, 2, ..., W$)

Z Number of forerun periods for production ($z = 1, 2, ..., Z$)

Decision Variables

$a_{j,t}$ Available capacity per production segment j and time period t

$b_{j,t}$ Capacity requirement per production segment j and time period t

$I_{k,t}$ Inventory per product k and time period t

$m_{ma,j,t}$ Number of worker recruitments per worker class ma, production segment j and time period t

$Mit_{ma,j,s,t}$ Number of worker per worker class ma, production segment j, shift s and time period t

$n_{ma,j,t}$ Number of worker redundancies per worker class ma, production segment j and time period t

$p_{j,s,t}$ Boolean-Variable for calculating the number of shifts per production segment j, shift s and time period t

$q_{j,s,t}^{a}$ Boolean-Variable to determine shift changes per production segment j, shift s and time period t (no change)

$q_{j,s,t}^{b}$ Boolean-Variable to determine shift changes per production segment j, shift s and time period t (shift model becomes active)

$q_{j,s,t}^{ac}$ Boolean-Variable to determine shift changes per production segment j, shift s and time period t (shift model becomes inactive)

$x_{k,t}$ Produced quantity per product k and time period t

Main Problem

In the following the main problem is described, that determines the optimal shift model for each time period, wich is transferred to the subproblem.

The objective function (Equation 1) minimizes the total costs (Equation 2). These comprise the storage costs (Equation 3), employee costs (Equation 4), shift allowances (Equation 5), one-time costs for a possible shift model change (Equation 6), the costs for capacity building (Equation 7) and the costs for capacity reduction (Equation 8).

$$ObjectiveFunction=Minimize(TotalCost) \quad (1)$$

$$TotalCost = StorageCost + MitCost + ShiftCost + Shiftmodell Cost + BuildingCost + ReductionCost \quad (2)$$

$$StorageCost = \sum_t^T \sum_k^K h_k \bullet I_{k,t} \quad (3)$$

$$MitCost = \sum_t^T \sum_s^S \sum_j^J \sum_{ma}^{MA} Mit_{ma,j}^{Cost} \bullet Mit_{ma,j,s,t} \quad (4)$$

$$ShiftCost = \sum_t^T \sum_s^S \sum_j^J Mit_{ma,j}^{Cost} \bullet Mit_{ma,j,s,t} \bullet S_s^{Cost} \quad (5)$$

$$ShiftmodelCost = \sum_t^T \sum_s^S \sum_j^J Q_{j,s}^{Cost} \bullet q_{j,s,t}^{b} \quad (6)$$

$$BuildingCost = \sum_t^T \sum_j^J \sum_{ma}^{MA} m_{ma}^{Cost} \bullet m_{ma,j,t} \quad (7)$$

$$ReductionCost = \sum_t^T \sum_j^J \sum_{ma}^{MA} n_{ma}^{Cost} \bullet n_{ma,j,t} \quad (8)$$

The constraints are shown next. First, the warehouse balance sheet (Equation 9) and the employee balance sheet (Equation 10) are determined. In addition, the warehouse and employee initial quantities (Equations 11 and 12) are determined. Further, the determination of capacity requirements (Equation 13) and available capacity (Equation 14) as well as restrictions on the minimum and maximum utilization of employees (Equations 15 and 16) are specified.

$$x_{k,t} + I_{k,t-1} - I_{k,t} = d_{k,t} \quad (9)$$

$$\sum_s^S Mit_{ma,j,s,t-1} + m_{ma,j,t-w} - n_{ma,j,t-w} = \sum_s^S Mit_{ma,j,s,t} \quad (10)$$

$$I_{k,0} = I_k^{Init} \quad (11)$$

$$Mit_{ma,j,s,0} = Mit_{ma,j,s}^{Init} \quad (12)$$

$$\sum_z^Z \sum_k^K f_{z,j,k} \bullet x_{k,t+z} = b_{j,t} \quad (13)$$

$$\sum_{ma}^{MA} \sum_s^S Mit_{ma,j,s,t} \bullet CAPA_{ma} = a_{j,t} \quad (14)$$

$$R_j^{Min} \bullet a_{j,t} \leq b_{j,t} \quad (15)$$

$$R_j^{Max} \bullet a_{j,t} \geq b_{j,t} \quad (16)$$

The number of employees is limited below. To this end, the minimum and maximum number of employees per production segment (Equations 17 and 18) are

determined in order to guarantee a permanent staff and not to exceed technical requirements (number of workplaces). In addition, the minimum and maximum number of employees per production segment and employee class (Equations 19 and 20) are defined as well, in order to reflect the structure of the core workforce and the supply of skilled workers available on the labour market.

$$\sum_{ma}^{MA}\sum_{s}^{S} Mit_{ma,j,s,t} \geq Mit_{j}^{TotalMin} \quad (17)$$

$$\sum_{ma}^{MA}\sum_{s}^{S} Mit_{ma,j,s,t} \leq Mit_{j}^{TotalMax} \quad (18)$$

$$\sum_{s}^{S} Mit_{ma,j,s,t} \geq Mit_{ma,j}^{Min} \quad (19)$$

$$\sum_{s}^{S} Mit_{ma,j,s,t} \leq Mit_{ma,j}^{Max} \quad (20)$$

The shift model is determined in the following constraints. For this purpose, lower and upper limits are determined for the number of employees per shift model (Equations 21 and 22) and the initial shift model (Equation 23) is specified. In addition, the Boolean variable for shift model determination (Equation 24) is limited. The change of the shift model (Equation 25) is also determined. It is possible that no change takes place (*qa=1*), a shift model becomes active (*qb=1*) or a shift model becomes inactive (*qc=1*). In addition, the Boolean variables have to be restricted (Equation 26).

$$\sum_{ma}^{MA} Mit_{ma,j,s,t} \geq p_{j,s,t} \bullet S_{j,s}^{Bottom} \quad (21)$$

$$\sum_{ma}^{MA} Mit_{ma,j,s,t} \leq p_{j,s,t} \bullet S_{j,s}^{Above} \quad (22)$$

$$p_{j,s,0} = p_{j,s}^{Init} \quad (23)$$

$$\sum_{s}^{S} p_{j,s,t} = 1 \quad (24)$$

$$p_{j,s,t} - p_{j,s,t-1} = 0 \bullet q_{j,s,t}^{a} + 1 \bullet q_{j,s,t}^{b} - 1 \bullet q_{j,s,t}^{c} \quad (25)$$

$$q_{j,s,t}^{a} + q_{j,s,t}^{b} + q_{j,s,t}^{c} = 1 \quad (26)$$

Subproblem

At the subproblem, the shift model from the main problem is used as a parameter. Compared to the main problem, the parameters p^{Init}, Q^{Cost}, S, S^{Cost}, S^{Above} and S^{Bottom} as well as the decision variables p, q^a, q^b and q^c are omitted. In addition, Equations 21 to 26 are no longer used. It should also be noted, that the omission of the shift parameter in all variables and equations means, that the s index is omitted.

The objective function (Equation 27) minimizes the total costs (Equation 28), which are composed of storage costs (Equation 29), employee costs (Equation 30) as well as the costs for building and reducing capacity (Equations 31 and 32). In addition, a *10 % GAP* to the upper bound is allowed for the optimal solution of the subproblem.

$$ObjectiveFunction = Minimize(TotalCostSub) \quad (27)$$

$$TotalCostSub = StorageCostSub + MitCostSub + BuildingCostSub + ReductionCostSub \quad (28)$$

$$StorageCostSub = \sum_{t}^{T}\sum_{k}^{K} h_k \bullet I_{k,t} \quad (29)$$

$$MitCostSub = \sum_{t}^{T}\sum_{j}^{J}\sum_{ma}^{MA} Mit_{ma,j}^{Cost} \bullet Mit_{ma,j,t} \quad (30)$$

$$BuildingCostSub = \sum_{t}^{T}\sum_{j}^{J}\sum_{ma}^{MA} m_{ma}^{Cost} \bullet m_{ma,j,t} \quad (31)$$

$$ReductionCostSub = \sum_{t}^{T}\sum_{j}^{J}\sum_{ma}^{MA} n_{ma}^{Cost} \bullet n_{ma,j,t} \quad (32)$$

EXAMINATION SCENARIO AND CASE STUDY

The 15 examination scenarios differentiate between 5 demand scenarios and 3 exhaustion courses (see Figure 1). It is assumed, that, because of the exhaustion, with lower employee utilization the processing times are also reduced. Therefore, *4* different employee utilization intervals are considered for each production segment: *80-85 %*, *85-90 %*, *90-95 %* and *95-100 %* ($R^{Min}(j)$ *to* $R^{Max}(j)$). Therefore (and due to the assumption of *2* production segments) each examination scenario has to been solved *16* times and the most cost-effective utilization interval is considered as the optimal solution for the examination scenario.

Figure 1: Worker utilization specific weighting factors for processing times per human exhaustion courses

The third exhaustion course corresponds to not-considering human exhaustion, so that the processing times for all employee utilization intervals are identical. All in all, the different exhaustion courses illustrate a wide range of possibilities that can occur in practice.
To determine the 5 different demand scenarios, the results from the Aggregated Production Planning are used as a basis. Thus, an average demand of *234 PCs* (pieces) per period is assumed for the product *k=1* and an average demand of *196 PCs* per period for the product *k=2*. Based on this mean values, the 5 demand scenarios are calculated under the assumption of a normal distribution with a standard deviation of *5 %* for *k=1* and *10 %* for *k=2*. This procedure ensures, that the demand scenarios are generated independently of the optimization model. In summary, under the given assumptions, a representative analysis is achieved for this problem class. The parameters of the case study are presented below, initially with the general parameters in Table 1.

Table 1: General parameters

Parameter	Value
J	*2*
K	*2*
MA	*2*
S	*3*
T	*90*
W	*1*
Z	*1*

Additional, there is no initial stock $I^{Init}(k) = 0$. The storage costs *h(k)* per period and piece amount to *160 MU* (monetary unit) for product *k=1* and *100 MU* for product *k=2*. The original processing times (*f(z)(j)(k)*), which are evaluated with the exhaustion factors mentioned above, are *2,794 TU/PC* (time units per piece) for product *k=1* in production segment *j=1*, *2,329 TU/PC* for product *k=2* in production segment *j=1*, *8,900 TU/PC* for product *k=1* in production segment *j=2* and *6,779 TU/PC* for product *k=2* in the production segment *j=2*, and it is assumed that the requirements must be available at the beginning of each period. Therefore, the processing times are already incurred in the preprocessing period *z=1*. However, it is taken into account, that the complete processing time is not carried out manually and is therefore influenced by exhaustion effects. It is assumed that *60 %* of the activities are carried out manually in production segment *j=1* and *50 %* of the activities in production segment *j=2*. The further parameters define the assumptions for planning personnel requirements. The following tables 2 to 5 show the relevant parameters. Each employee class represents different levels of experience and qualifications. In the case study, the employee class *ma=1* is interpreted as highly qualified and experienced internal personnel. Employee class *ma=2* represents external employees that are procured temporarily. The procurement and release for employee class *ma=2* is carried out by a service company, which is compensated by the employee costs. The initial number of employees ($Mit^{Init}(ma)(j)(s)$) is assumed to be *2* employees in the employee class *ma=1* and the production segment *j=1* as well as *7* employees in employee class *ma=1* and the production segment *j=2*. In the employee class *ma=2*, a number of *0* employees are accepted. Further in Table 5 the respective limits for the determination of the corresponding shift models are presented.

Table 2: Employee parameters for each production segment (j)

Parameter	j=1	j=2
$Mit^{TotalMax}(j)$	*9*	*15*
$Mit^{TotalMin}(j)$	*1*	*1*

Table 3: Employee parameters for each worker class (ma)

Parameter	ma=1	ma=2
CAPA(ma)	*576,000 TU*	*432,000 TU*
$m^{Cost}(ma)$	*6,000 MU*	*250 MU*
$n^{Cost}(ma)$	*20,000 MU*	*150 MU*

Table 4: Employee parameters for each worker class (ma) and production segment (j)

Parameter		j=1	j=2
$Mit^{Cost}(ma)(j)$	ma=1	*3,000 MU*	*3,000 MU*
	ma=2	*3,800 MU*	*3,800 MU*
$Mit^{Max}(ma)(j)$	ma=1	*3*	*10*
	ma=2	*9*	*15*
$Mit^{Min}(ma)(j)$	ma=1	*1*	*1*
	ma=2	*0*	*0*

Table 5: Limits for determination of the correct shift model for each production segment (j) and shift model (s)

Parameter		s=1	s=2	s=3
$S^{Above}(j)(s)$	j=1	*3*	*6*	*9*
	j=2	*5*	*10*	*15*
$S^{Bottom}(j)(s)$	j=1	*1*	*4*	*7*
	j=2	*1*	*6*	*11*

So the shift models *s=1* in the production segment *j=1* and *s=2* in the production segment *j=2* are active ($p^{Init}(j)(s)$). Shift allowances $S^{Cost}(s)$ are assumed to be *0 %* for *s=1*, *1.5 %* for *s=2* and *15 %* for *s=3* of the employee costs per period and in addition, one-time costs $Q^{Cost}(j)(s)$ for a shift model change of *1,500 MU* to activate shift model *s=1* and *2,500 MU* to activate shift model *s=2* are assumed for both production segments.

To activate shift model *s=3*, *5,000 MU* in the production segment *j=1* and *7,500 MU* in the production segment *j=2* are assumed.

RESULTS

The results of the examination scenarios are presented below. Table 6 shows the optimal results for each examination scenario. In comparison, Table 7 shows the results derived from a non-consideration of employee utilization. For this purpose, the results of the employee utilization interval of 95-100 % were used, since it is generally assumed, that maximizing utilization leads to optimal results.

Table 6: Optimal Solutions for each demand scenario and exhaustion course [MU]

Demand Scenarios	Exhaustion Course 1	Exhaustion Course 2	Exhaustion Course 3
Scenario 1	*2,477,865*	*2,529,101*	*2,649,707*
Scenario 2	*2,549,378*	*2,510,901*	*2,551,692*
Scenario 3	*2,493,051*	*2,520,922*	*2,637,919*
Scenario 4	*2,657,297*	*2,728,688*	*2,780,999*
Scenario 5	*2,525,665*	*2,392,793*	*2,538,531*

Table 7: Solutions without consideration of worker utilization [MU]

Demand Scenarios	Solution without worker utilization
Scenario 1	*2,649,707*
Scenario 2	*2,551,692*
Scenario 3	*2,637,919*
Scenario 4	*2,893,396*
Scenario 5	*2,538,531*

As can be seen, the costs resulting from Table 7 are higher than those for the first (by 4.45 %) and second (by 4.63 %) exhaustion course from Table 6 and they are identical for the third exhaustion course, with the exception of the demand scenario 4. As Table 6 shows the true optimal results per exhaustion course and Table 7 the results of a maximum employee utilization, it becomes clear, that irrespective of the extent of the exhaustion or a non-consideration of exhaustion, maximizing employee utilization does not necessarily lead to optimal results, so planning and controlling employee utilization is necessary. This finding is enhanced by Figure 2, which shows the optimal employee utilization interval for each exhaustion course and demand scenario for the production segment *j=1*. The upper interval values are displayed as legends. For example, the legend value *85 %* stands for the utilization interval *80-85 %*. Next to the findings before, it becomes clear, that in order to achieve optimal results, employee utilization should be lower, the stronger the exhaustion effects are. In addition to the cost advantages due to planning and controlling the employee utilization, it is also important, to emphasize the reduction of the burden on employees by taking exhaustion effects and utilization specific processing times into account. It can be stated, that the findings from the investigation of short-term planning horizons (see Trost et al. 2017a and 2017b) can also be confirmed for long-term planning horizons.

In addition, the results for the first and second exhaustion course from Table 6 are compared with the results of the third exhaustion course from Table 6. This illustrates the variations between the consideration and non-consideration of exhaustion effects. It becomes clear, that even with flatter exhaustion curves, considerable deviations of up to *7 %* occur. This suggests a considerable potential for improvement through the integration of exhaustion effects, whereby there is a need for concrete quantification of exhaustion effects, which Grosse et al. (2017) and Trost et al. (2016) also mentioned.

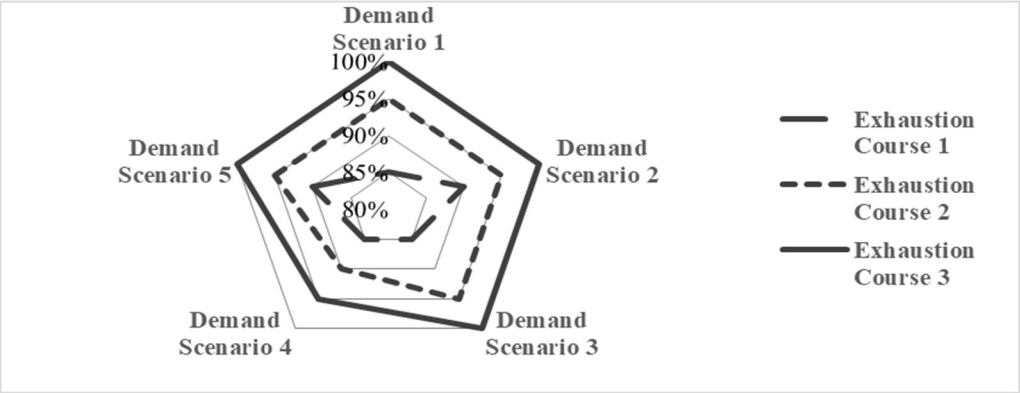


Figure 2: Optimal worker utilization for each demand scenario and exhaustion course

In addition, further research questions can be derived from a more detailed analysis of the results. The analysis of the shift models showed that no shift model changes were necessary, which can be explained by the demand scenarios. For further research, greater fluctuation in demand should also be investigated. These include, for example, the integration of seasonal effects and various scenarios of slumps in demand.

Another research question arises from the analysis of the number of employees per employee class (ma). It becomes clear, that, despite the rather stable demand scenarios, the flexibility of the available capacity is a decisive factor. For example, partly despite an increase in the total number of employees, some employees in employee class *ma=1*, which stands for highly qualified and experienced internal employees, are replaced by employees in employee class *ma=2*, which stands for less qualified and experienced external employees. This finding does not come as a surprise at first, but on the one hand, there is the research question of how this need for flexibility changes with increasing fluctuations in demand and which level of flexibility leads to optimal results. On the other hand, there is the question of how the need for flexibility in connection with the expected shortage of skilled workers will change. In view of the demographic change and the increased demand for skilled workers (for example due to job restructuring in the context of Industry 4.0), it could become problematic or more expensive to always procure the necessary number of employees (*ma=2*), so that an increase in internal personnel (*ma=1*) may be preferable.

Finally, it can be summarized that it has been proven, that the consideration of exhaustion effects and the planning and control of employee utilization in connection with production planning are necessary, since a maximum employee utilization, independent of the exhaustion courses, does not necessarily lead to optimal results. Further studies, which can be carried out with the presented optimization model, should deal with the effects of increasing demand fluctuations in connection with decreasing numbers of available employees.

CONCLUSION

This paper underlines the need for a link between production planning and personnel planning. The basis of this approach is the inadequate improvement of working conditions for employees and the potential of production planning and control confirmed by the literature. However, there are few papers in the context of hierarchical production planning, that integrate the social dimension. Especially in the area of long-term planning, no papers are known.

On this basis, the model presented above was developed, which controls employee utilization in a long-term planning environment without significantly restricting production capacities by integrating aspects of personnel planning into the Master Production Scheduling. In addition, social aspects in the form of employee utilization specific processing times are also taken into account. This makes it possible to reduce the burden of employees without endangering the fulfillment of customer orders.

On the one hand, the findings show, that the recognized results from short-term planning horizons are also reflected in long-term planning environments. Especialy, the finding, that maximizing employee utilization does not necessarily lead to optimal results should be emphasized. Consequently, taking into account employee utilization and the resulting processing times, a cost advantage can be achieved and the burden on employees can be reduced. On the other hand, the results show, that there is further research potential. Continued investigations should examine the effects of increasing demand fluctuations in connection with decreasing numbers of available employees.

REFERENCES

Andriolo, A.; D. Battini; A. Persona and F. Sgarbossa. 2016. ”A new bi-objective approach for including ergonomic principles into EOQ model”. In *International Journal of Production Research*. 54 (9), S. 2610-2627.

Arslan, M. C. and M. Turkay. 2013. “EOQ revisited with sustainability considerations”. In *Foundations Of Computing And Decision Sciences*, 38 (4), S. 223–249.

Boysen, N. and M. Fliedner. 2011. “Scheduling aircraft landings to balance workload of ground staff”. In *Computers & Industrial Engineering*, 60 (2), S. 206-217.

Drexl, A.; B. Fleischmann; H.-O. Günther; H. Stadtler and H. Tempelmeier. 1994. “Konzeptionelle Grundlagen kapazitätsorientierter PPS-Systeme”. In *Zeitschrift für betriebswirtschaftliche Forschung,* 46 (12), S. 1022-1045.

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”. In *IFAC-PapersOnLine*. 50(1), S. 6900-6905.

Jaber, Y. M. and M. Bonney. 2007. “Economic manufacture quantity (EMQ) model with lot-size dependent learning and forgetting rates”. In *International Journal of Production Economics*, 108 (1), S. 359-367.

Khan, M.; M. Y. Jaber and A. R. Ahmad. 2014. “An integrated supply chain model with errors in quality inspection and learning in production”. In *Omega*. 42(1), S. 16-24.

Lai, P. J. and W. C. Lee. 2013. ”Single-machine scheduling with learning and forgetting effects”. In *Applied Mathematical Modelling*. 37(6), S. 4509-4516.

Otto, A. and A. Scholl. 2011. ”Incorporating ergonomic risks into assembly line balancing”. In *European Journal of Operational Research*. 212 (2), S. 277-286.

Schmucker, R. 2014. DGB-Index Gute Arbeit – Der Report 2013. Institut DGB-Index Gute Arbeit, Berlin.

Trost, M.; T. Claus; F. Herrmann; E. Teich and M. Selmair. 2016. ”Social and Ecological Capabilities for a Sustainable Hierarchical Production Planning. In *Proceedings of European Conference on Modelling and Simulation*. 30, S. 432-438.

Trost, M.; E. Teich; T. Claus and F. Herrmann. 2017a. Ein lineares Optimierungsmodell zur Hauptproduktionsprogrammplanung mit Berücksichtigung sozialer Größen. *uwf UmweltWirtschaftsForum*. 25(1-2), S. 71-80.

Trost, M.; T. Claus and F. Herrmann. 2017b. ”Master Production Scheduling and the Relevance of Included Social Criteria”. In *ACC Journal*. 23(2), S. 146-154.

AUTHOR BIOGRAPHIE

MARCO TROST is a doctoral student at the Department of Business Science at the Technical University of Dresden and he is sponsored by the European Social Fund (ESF). His e-mail address is: marco.trost@mailbox.tu-dresden.de.

ASSESSING CROP ROTATION SUSTAINABILITY USING ANALYTICAL HIERARCHY PROCESS

Saturnina Fabian Nisperos and Frederic D. McKenzie
Department of Modeling, Simulation and Visualization Engineering
Old Dominion University
Norfolk, Virginia
E-mail: snisp001@odu.edu, rdmckenz@odu.edu

KEYWORDS
Crop rotation, sustainability assessment model, multi-criteria decision analysis, AHP.

ABSTRACT

With the food security challenge faced by nations globally, agriculture sustainability has been a significant consideration for concerned agencies. Sustainability assessments are significant tools in providing support to stakeholders in their crop production planning. Agricultural sustainability assessment, however, is complex and it involves numerous criteria that can be conflicting. In this study we investigated the use Analytical Hierarchy Process, a multi-criteria decision analysis method, in assessing the sustainability of crop rotation alternatives and its applicability to address the multiple criteria of sustainability and the diverse preferences of stakeholders. The comparable results of the model with a sustainability assessment of cropping systems reported in the literature, validates AHP as an apt method for sustainability assessment of crop rotation alternatives and handling the complex criteria of sustainability and preferences of stakeholders. The model results, when well presented, can be utilized to support stakeholders in their decision making and in evaluating their crop rotation choices.

INTRODUCTION

Sustainable agriculture involves selection of crops appropriate to the location and conditions of the farm, crops diversity, proper soil management and efficient use of farm resources. It promotes crop production practices that enhances productivity and profitability (economic) without compromising the health of natural resources (environment) and the quality of life of the society (social). With the food security challenge faced by nations globally, agriculture sustainability has been a significant consideration for concerned agencies like the Food and Agriculture Organization (FAO) and United Nations (UN). Diverse innovative practices have been explored to improve sustainability. Among the crop production practices endorsed by research agencies is crop rotation, which is the planned successions of crops over time on the same field. Crop rotation has been proven to increase yield, reduce the need for synthetic inputs (i.e. fertilizer and pesticides) and enhance resilience (Stanger and Lauer 2008, Carter, et al. 2009, Lin 2011).

Numerous research methods have been exploited to advance and assess crop rotation sustainability. Crop growth simulation models have been developed to evaluate the impact of climate, water, soil, agricultural inputs and management practices on crops. Model-driven decision support system (DSS), a type of DSS that utilizes complex models, is among the approaches explored to provide support to stakeholders in agriculture in their decision making. DSS tools developed to promote crop rotation have diverse and genuine objectives, but the majority are mainly for experimental simulations, for experts use and not aimed for smallholder farmers use. Limitations on crop rotation sustainability assessment methods include: non-dynamic assessment, lack of regard to the individual crop production preferences and goals of smallholder farmers, and focused only on single years and single crops rotation.

Agricultural sustainability assessment is complex, and it involves numerous criteria that can be conflicting and stakeholders may also have different needs and priorities. One approach to address the complex criteria of sustainability is by alternatives evaluation (rather than just selecting one solution) based on indicators with the aid of multi-criteria decision methods (Dury, et al. 2012). In the critical review of Multi-Criteria Decision Analysis (MCDA) techniques in (Diaz-Balteiro, González-Pachón and Romero 2017), the results indicate that there is a proliferation on the utilization of MCDA techniques in aggregating sustainability criteria which signifies the importance of the method in this context. Furthermore, MCDA techniques have been regarded as an apt framework for assessing agricultural sustainability because of its capacity to evaluate diverse criteria and priorities (Talukder, et al. 2017).

Our research aims to investigate the integration of crop growth simulation model and multi-criteria decision analysis as an approach for a dynamic and multi-criteria sustainability assessment model which can be used to support stakeholders in their decision making. In this paper, we study the use of Analytical Hierarchy Process, an MCDA method, in assessing the sustainability of crop rotation alternatives and its applicability to address the multiple criteria of sustainability and the diverse preferences of stakeholders.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

BACKGROUND

Multi-Criteria Decision Analysis (MCDA)

The MCDA deals with the evaluation of alternatives relating to multiple and conflicting decision criteria. Alternatives are the set of options that a decision maker needs to assess, and the criteria are the factors that are being considered to attain the goal of the decision making (e.g. cost, quality). MCDA is composed of non-linear recursive process which involves structuring the decision problem, articulating and modelling the preferences, aggregation of the alternative evaluations and providing recommendations (Guitouni and Martel 1998).

MCDA methods can be classified as deterministic, stochastic or fuzzy for single or group decision making. They have been regarded as apt methods to perform sustainability assessments. In the "Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment" study by Cinelli et al. (Cinelli, Coles and Kirwan 2014), the authors reviewed the performance of MAUT (Multi attribute utility theory), ELECTRE (Elimination and choice expressing the reality), AHP (Analytical hierarchy process), PROMETHEE (Preference ranking organization method for enrichment of evaluations) and DRSA (Dominance-based rough set approach) with respect to 10 criteria under the domain of scientific soundness, feasibility, and utility. Their result indicates that most of the requirements are satisfied by the MCDA methods but with different extents. MAUT and AHP are for utility-based theory, ELECTRE and PROMETHEE are for outranking relation theory and DRSA is for the sets of decision rules theory. These methods have been the most widely employed MCDA tools in sustainability related research and the selection of which method to employ should be grounded on the basics of the approach and the type of assessment to be performed (Cinelli, Coles and Kirwan 2014).

Analytical Hierarchy Process (AHP)

The AHP method, developed by Dr. Thomas Saaty, is a theory of measurement by pairwise comparisons which derives priority scales through the experts' judgements. AHP decomposes a complex MCDA problem into a system of hierarchies, combines both qualitative input with quantitative data and supports dimensionless analysis. It has been used in different settings for decision making in various projects. The standard procedure for AHP is outlined by (Saaty 2008) as:

1. Define the problem and determine the kind of knowledge sought.
2. Structure the decision hierarchy, starting from the top to the bottom level (i.e. goal, criteria and alternatives, respectively)
3. Construct the set of pairwise comparison matrices using the fundamental scale of absolute numbers (Table 1)
4. Compute priority values and consistency ratio

The consistency ratio (CR) estimates the consistency of the pairwise comparisons and allows checking of reliability.

$$CR = \frac{\text{Consistency Index (CI)}}{\text{Random Index (RI)}}$$

The calculation of the consistency ratio is further explained in (Mu and Pereyra-Roxas 2017). An acceptable consistency ratio value should be less than 10%. The priority value is used to rank the alternatives. The alternative with the highest priority value can be regarded as the best by the decision maker.

Table 1: The Fundamental Scale of Absolute Numbers (Saaty 2008)

Intensity	Definition
1	Equal Importance
3	Moderate importance
5	Strong importance
7	Very strong or demonstrated importance
9	Extreme importance
2, 4, 6, 8	Weak or slight, Moderate plus, Strong plus, and Very, very strong (respectively)

If activity i has one of the above non-zero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i

Sustainability Assessment and Indicators

Sustainability assessment advocates agriculture sustainability by aiding stakeholders in evaluating the sustainability impact of their crop production choices. An increasing number of sustainability assessment tools have been developed to support stakeholders, like farmers and policymakers (Olde, Bokkers and Boer 2017). Sustainability assessment approaches vary on how and what (economic, environmental, and social sustainability) indicators are measured and evaluated.

In their sustainability assessment study, Castoldi and Bechini (2010) aggregated 15 economic and environmental indicator values to come up with a global sustainability index which they used to assess the cropping systems at field level. The indicators were selected from extensive literature review based on the ability to quantify the effects of cropping systems management on the environment and on economic profitability, and data obtainability. The average and standard deviation of the indicators were calculated using a large data set of cropping systems management for 131 fields in Northern Italy, which were obtained through a 2-year periodic interviews with farmers. Figure 1 lists the 15 economic and environmental indicators which are mainly classified as economic, nutrient management, energy management, pesticide management and soil management indicators.

METHODS

With the analysis goal of evaluating the agricultural sustainability of crop rotation alternatives to support stakeholders in their decision making, the AHP method was employed and its standard procedure was followed. The following subsections give further details on these steps.

Decision Hierarchy

The sustainability indicators and alternatives identified by (Castoldi and Bechini 2010) were used in structuring the decision hierarchy. Figure 1 shows the criteria, and subcriteria to evaluate the alternatives and provide solution to the analysis goal. The crop rotation alternatives to be evaluated are continuous maize (*Mc*), maize and other crops (*Mo*), continuous rice (*Rc*), rice and other crops (*Ro*), and winter cereals (*Ce*). The permanent meadows, which was originally part of the assessment in the benchmark study, was not included due to the lack of available model parameters to simulate its impact.

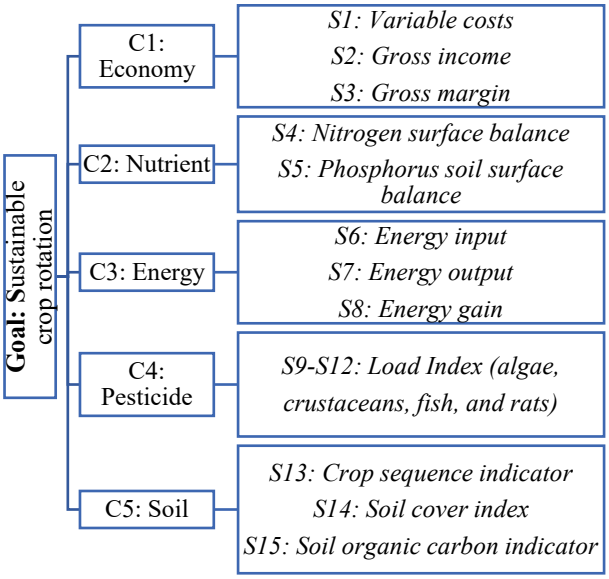


Figure 1: Goal and decision criteria based from the indicators identified by (Castoldi and Bechini 2010)

Indicator Values and Pairwise Comparison

To facilitate comparison of the goal analysis result of AHP with the sustainability assessment of (Castoldi and Bechini 2010), the same sustainability function, parameters and the average indicator values (*x*) from the study were used to compute the subcriteria values *(s)* of the 5 alternatives (*Mc, Mo, Rc, Ro, Ce*).

The subcriteria values of the alternatives were derived using the sustainability function:

$$f(s_i)\begin{cases}\left(\frac{x_i - S_{min}}{S_{opt1} - S_{min}}\right)^k, left\ side\ of\ the\ curve \\ \left(\frac{x_i - S_{max}}{S_{opt2} - S_{max}}\right)^k, right\ side\ of\ the\ curve\end{cases}$$

where x_i is the mean subcriteria value of alternative *i;* S_{opt1} and S_{opt2} are the lower and upper threshold values of the subcriteria, respectively; S_{min} and S_{max} are the thresholds used to define the minimum and maximum sustainable range of the indicators; k sets the linear or non-linear relationship; and, $s_i \in \mathbb{R} \,|0 \geq s_i \leq 1$. Table 2 shows the mean indicator values and the computed subcriteria values of the alternatives.

The alternatives are then compared using the derived subcriteria values (or the sustainability index) and the pairwise comparison matrices are constructed using the fundamental scale of absolute numbers. To automate the pairwise comparison process, the following pairwise function was used:

$$f(P_{ij})\begin{cases}8 * (v_i - v_j) + 1 & v_i \geq v_j \\ \frac{1}{8 * (v_j - v_i) + 1}, otherwise\end{cases}$$

where v_i and v_j are the corresponding subcriteria values of alternatives i and j; and, $P_{ij} \in \mathbb{R} \,|\frac{1}{9} \geq P_{ij} \leq 9$.

Table 2: Mean and Subcriteria Values of Alternatives

		Mc		Mo		Rc		Ro		Ce	
		x	*s*	*x*	*s*	*x*	*s*	*x*	*s*	*x*	*s*
C1	S1	583	0.6	445	1	692	0	466	1	188	1
	S2	1616	1	1284	0.5	2052	1	1736	1	951	0
	S3	1033	1	840	0.5	1360	1	1270	1	763	0.2
C2	S4	182	0	72	0.8	75	0.7	55	1	-18	0.3
	S5	38	0.8	0	1	-5	1	-15	1	-12	1
C3	S6	27.8	0	22	1	22.6	1	18.8	1	10.7	1
	S7	364.5	1	257.3	1	192.6	0.2	204.6	0.4	127.4	0
	S8	336.7	1	235.3	1	169.9	0.4	185.8	0.6	116.7	0
C4	S9	108.2	0.8	106.5	0.8	259.4	0	144.5	0.7	0.3	1
	S10	1.4	0.6	15.5	0	7.6	0	4.1	0	0	1
	S11	2.2	0.5	2.4	0.3	8.5	0	7.6	0	0	1
	S12	1.5	0.9	0.8	1	8.5	0	3.6	0	0.5	1
C5	S13	2	0.3	4.6	0.7	1	0.1	4.1	0.6	3.5	0.5
	S14	0.35	0.3	0.5	1.0	0.33	0	0.4	1.0	0.45	1.0
	S15	6.3	0.8	4.6	0.4	4.3	0.4	2.1	0.1	1.4	0

RESULTS AND DISCUSSION

Multicriteria Sustanability Assessment of Alternatives

Using equal weights (*w*) on the multiple criteria sustainability, the priority values of the alternatives were computed and is shown in Table 3. Each five criteria (C1-C5) are equally assigned a weight of 20, totaling to 100 and this weight is equally divided to the respective sub-criteria.

$$\sum_i^n C_i = 100\ and\ C_i = \sum_j^n S_j$$

The results denote that the best crop alternative, with respect to the set goal criteria, is maize with other crops (*Mo*, 24%) and the least is continuous rice (*Rc,* 13.6%).

Mo outperforms the other alternatives in the energy and soil management criteria (*C3 and C5*). The priority values suggest, however, that rice and other crops (*Ro*) is more favored when it comes to the economic nutrient management criteria (*C1 and C2*) while winter cereals (*Ce*) tops the alternatives on pesticide toxicity. These results are consistent with the findings of the benchmark study. As to the reliability of the pairwise comparisons, the average consistency ratio (CR) value is 2.4% and all are within the acceptable consistency ratio value (i.e. < 10%).

Table 3: Priority Values Result (Equal Criteria Weights)

	w	Mc	Mo	Rc	Ro	Ce	CR
C1	20	4.2	3	4.3	6.1	2.4	0
S1	6.67	0.6	2	0.2	2	2	2.8
S2	6.67	1.7	0.5	2.1	2.1	0.2	2.2
S3	6.67	2	0.5	2	2	0.2	2.9
C2	20	1.2	4.7	4.4	6.6	3.1	0
S4	10	0.3	2.4	2	4.6	0.7	4
S5	10	0.9	2.4	2.4	2.1	2.4	0
C3	20	5.5	6.8	2.5	3.1	2	0
S6	6.67	0.2	1.6	1.6	1.6	1.6	0
S7	6.67	2.6	2.5	0.5	0.9	0.2	4.6
S8	6.67	2.8	2.7	0.4	0.6	0.2	3.7
C4	20	4.6	3.6	0.8	1.2	9.8	0
S9	5	1.1	1.1	0.1	0.6	2.1	2.9
S10	5	1.3	0.3	0.3	0.3	3	2.7
S11	5	0.9	0.6	0.2	0.2	3	4.2
S12	5	1.3	1.7	0.2	0.2	1.7	0.4
C5	20	4.6	5.8	1.6	4.3	3.7	0
S13	6.67	0.6	2.5	0.4	1.9	1.3	1.5
S14	6.67	3.6	1.3	1	0.4	0.3	2.7
S15	6.67	0.4	2	0.2	2	2	2.5
Priority	***100***	***20.2***	***24***	***13.6***	***21.4***	***20.9***	

Addresing Diverse Preferences

To evaluate the applicability of AHP in addressing the diverse preferences of stakeholders, the crop rotation alternatives were assessed using the different criteria and sub-criteria preferences (weights) of the stakeholders (farmer, researcher, agronomist, decision maker and environmentalist) in (Castoldi and Bechini 2010). Figure 2 shows the comparison of the results of AHP with the the rankings of the said study. The rankings are labeled as numbers 1 to 5, with 1 as the best. The permanent meadows were mainly considered as most sustainable system (rank 1) in the benchmark study. However, since it was not included in the AHP ranking, the alternatives ranking in the benchmark study were subsequently adjusted (i.e. rank 2 to rank 1, rank 3 to rank 2, and so on) to facilitate comparison.

In the AHP ranking, the top 1 and 2 crop rotation alternatives among stakeholders vary between *Mo* and *Ro* while the least (5) is mainly *Rc*, with the exeption of the farmer ranking in (*b*) where the lowest rank is *Ce*. For the rank results of the benchmark study., generally, the top 1 and 2 crop rotation are also a switch between *Mo* and *Ro*, with the exeption again of the farmer ranking in (*b*) where *Mc* lands the second. *Rc* is consistently in their lowest in rank.

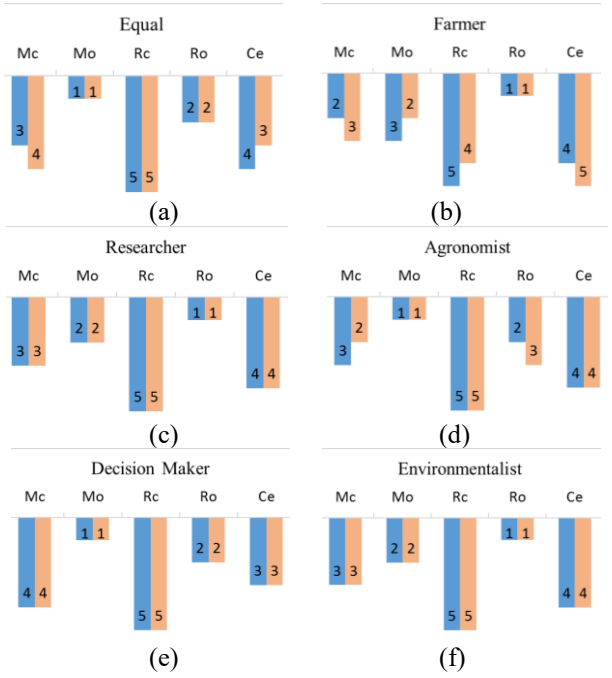


Figure 2: Comparison of Rankings per Stakeholder

Overall, the AHP ranked the same top (1) crop rotation alternative as the benchmark study's result for all stakeholder cases. This demonstrates the capability of AHP to find the best alternative. Both have corresponding rankings in *c*, *e* and *f* but with some variations in *a*, *b*, and *d*. In *a* (equal), *Mc* and *Ce* were switched as rank 2 and 3; in *b* (farmer), there is an interchange in ranks between *Mc* and *Mo*, and *Rc* and *Ce;* and in *d* (Agronomist), *Mc* and *Ro* swapped as 2nd and 3rd ranks. The priority values of the alternatives related to these swapped ranks were examined and the average priority value difference between these swaps is 0.005 (0.5%) which can be considered as negligible and hence, rationalizes the switch in ranks. The overall priority values of the stakeholder groups with switch in ranks were scaled relative to the maximum priority and were plotted as radar graphs in Figure 3. It can be noted in the chart that the alternatives switched in ranks generally falls on a contiguous radial grid or distance. These observations support the validity of the AHP method in evaluating the sustainability of crop rotation alternatives.

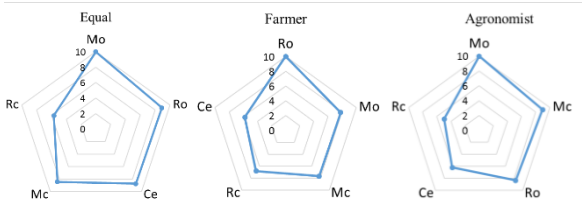


Figure 3. Scaled Priority Values of Equal, Farmer and Agronomist Rankings

CONCLUSION

In this paper, we used and investigated the applicability of Analytical Hierarchy Process as an approach to assess the agricultural sustainability of crop rotation alternatives and to address the diverse sustainability criteria and preferences of stakeholders. The output of the model was compared to the integrated sustainability assessment of the benchmark study. and the resulting ranking of the evaluated crop rotation alternatives are comparable regardless of the different inclinations of the stakeholder groups. This validates AHP as an apt method in handling the multiple and complex criteria of sustainable agriculture and the diverse preferences of stakeholders, and in assessing the sustainability of crop rotation alternatives. Moreover, the resulting priority values of AHP, when well presented, can be utilized to support stakeholders in their decision making and in evaluating their choices.

FUTURE WORK

For a dynamic agriculture and multi-criteria sustainability assessment, we plan to investigate the integration of a crop growth simulation model and the MCDA method.

REFERENCES

Carter, MR, C Noronha, RD Peters, and J Kimpinski. 2009. "Influence of conservation tillage and crop rotation on the resilience of an intensive long-term potato cropping system: Restoration of soil biological properties after the potato phase." *Elsievier (Agriculture, Ecosystems & Environment)* 133 (1-2): 32-39.

Castoldi, Nicola, and Luca Bechini. 2010. "Integrated sustainability assessment of cropping systems with agro-ecological and economic indicators in northern Italy." *European Journal of Agronomy (Elsevier)* 59-72.

Cinelli, Marco, Stuart R. Coles, and Kerry Kirwan. 2014. "Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment." *Ecological Indicators (Elsivier)* 46: 138-148.

Diaz-Balteiro, L, J. González-Pachón, and C. Romero. 2017. "Measuring systems sustainability with multi-criteria methods: A critical review." *European Journal of Operational Research (Elsevier)* 258: 607-616.

Dury, Jérôme, Noémie Schaller, Frédérick Garcia, Arnaud Reynaud, and Jacques Eric Bergez. 2012. "Models to support cropping plan and crop rotation decisions. A review." *Agronomy for Sustainable Development* 32 (2): 567-580.

Guitouni, Adel, and Jean-Marc Martel. 1998. "Tentative guidelines to help choosing an appropriate MCDA method." *European Journal of Operational Research* 109 (2): 501-521.

Lin, Brenda B. 2011. "Resilience in Agriculture through Crop Diversification: Adaptive Management for Environmental Change." *BioScience* 61 (3): 183-193.

Mu, E., and M. Pereyra-Roxas. 2017. "Practical Decision Making." SpringerBriefs in Operations Research.

Olde, Evelien M. de, Eddie A.M. Bokkers, and Imke J.M. de Boer. 2017. "The Choice of the Sustainability Assessment Tool Matters: Differences in Thematic Scope and Assessment Results." *Ecological Economics (Elsevier)* 77-85.

Saaty, Thomas L. 2008. "Decision making with the analytic hierarchy process." *Int. J. Services Sciences* 1 (1): 83-98.

Stanger, Trenton F., and Joseph G. Lauer. 2008. "Corn Grain Yield Response to Crop Rotation and Nitrogen over 35 Years." *American Society of Agronomy Agronomy Journal* 100 (3): 643-650.

Talukder, Byomkesh, Alison Blay-Palmer, Keith W. Hipeland, and Gary W. vanLoon. 2017. "Elimination Method of Multi-Criteria Decision Analysis (MCDA): A Simple Methodological Approach for Assessing Agricultural Sustainability." *Sustainability* 9 (287).

AUTHOR BIOGRAPHIES

SATURNINA F. NISPEROS is a Ph.D. student in Modeling and Simulation at the MSVE department of Old Dominion University. She received her BS in Computer Science and MS in Information Technology from Saint Louis University in Baguio City. She is a Fulbright grantee from the Philippines and affiliated with Mariano Marcos State University as assistant professor of computer science. Her current research interest is on developing decision support systems employing optimization, discrete event and agent-based modeling techniques.

FREDERIC D. MCKENZIE is a professor and department chair of the MSVE Department at ODU where he currently serves as Principal Investigator (PI) and Co-PI on projects involving software architectures for simulation, behavior representation in simulations, and medical modeling and simulation. To date, his projects in these areas have led to several publications relating research in modeling human-like intelligent agents including crowds, formal descriptions of distributed simulation architectures, objective measures of successful prostate surgery, and augmenting standardized patients. Dr. McKenzie received his Ph.D. in computer engineering from the University of Central Florida in 1994. Both his M.S. and Ph.D. work have been in artificial intelligence - focusing on knowledge representation and model-based diagnostic reasoning.

GROUND VEHICLE LOCALIZATION WITH PARTICLE FILTER BASED ON SIMULATED ROAD MARKING IMAGE

Oleg Shipitko and Anton Grigoryev
Institute for Information Transmission Problems – IITP RAS,
Bol'shoy Karetnyy Pereulok 19,
Moscow, Russia, 127051
E-mail: shipitko@visillect.com

KEYWORDS

Model-based localization, vehicle localization, particle filter, lane markings, autonomous vehicle, pose estimation.

ABSTRACT

Precise localization is a prerequisite and a cornerstone for successful operation of any autonomous vehicle. In this paper, consideration is given to a lane feature-based approach to a self-driving vehicle localization. Proposed map-relative localization method is built upon a combination of vision-based lane markings detection and odometry data. Detected lane markings are aligned with a reference map in order to derive global pose estimate while odometry provides path consistency. To combine heterogeneous sensory data use is made of particle filter method. It allows for non-Gaussian noise common for vision-based detectors as well as a further extension of data sources. The approach described in this work was tested on a real vehicle in urban environment and proved itself to be precise and reliable enough for real-world applications. It was able to provide lateral and longitudinal map-relative localization with a precision of 0.2 m.

INTRODUCTION

Recent advances in autonomous vehicle development have drawn attention of a wide audience including scientists, engineers, and the general public. Although nowadays there are already ongoing experiments on autonomous driving in the natural environment, the increasing safety requirements lead to the high demand for improvement of all vehicle subsystems. Precise and reliable localization is a prerequisite for any complex task such as path planning, maneuvering, and navigation in general.

It has been long understood that lane markings in particular and all types of road markings in general contain essential information used by human drivers for decision making and road situation analysis. The idea to provide autonomous vehicles with the same capabilities is not new as well. Lane detection is widely used in both partially and fully autonomous vehicle projects for various purposes (Krokhina et al. 2015; Lu et al. 2014). Hillel et al. (Hillel et al. 2014) provides an extensive overview of lane detection techniques, their limitations, and application. One of the applications where lane detection has proved itself to be an efficient foundation is localization (Ziegler et al. 2014; Jo et al. 2015; Du and Tan 2016). Many systems presented in literature exploit LIDAR as a major source of information for road understanding and lane detection (Huang et al. 2009; Hernández and Marcotegui 2009). One of the subtasks of lane marking detection where LIDAR is superior to other sensors is curb detection. There are many studies of localization systems build upon detector of this type exclusively. Thus in (Hata et al. 2014), authors propose a system using a method named least trimmed squares to fit a curve model to detected curb points and validate the proposed detector in a localization task. However, curbs on their own are prone to occlusion and cannot guarantee high precision in many real-world scenarios, while lane features (which may as well include detection of curbs as shown in (Ziegler et al. 2014)) provide more solid localization foundation. In general, the high (though constantly decreasing) cost of LIDAR systems prevents such kind of systems from becoming wide-spread. On the other hand, vision-based approaches require only relatively inexpensive camera systems and are able to provide a comparable level of detection precision.

The majority of known autonomous ground vehicles use a predefined map as a reference and derive their global pose by associating incoming sensory data with mapped data corresponding to a particular reference pose (Chausse et al. 2005; Tao et al. 2013; Volkov et al. 2017). As shown in (Li et al. 2017), size is an important property of a digital map. For instance, maps built from multidimensional features as (e.g. 3D Point Cloud maps used in (Sheehan et al. 2013)) may require significant storage capacity and relatively high amount of computational resources. At the same time curve-based vector maps require much less storage and are specifically efficient for representing 2D structures like lane markings.

A crucial issue to solve in any localization pipeline is combining together more than one sensor or several concurrent algorithms of vehicle pose estimation. There are two major approaches to data fusion in localization applications, namely Kalman filter, and particle filter (also known as Sequential Monte Carlo localization technique). Both approaches have been used widely in robotics as emphasized in (Thrun 2002) and have their own advantages and drawbacks. Kalman filters provide exact, optimal solution and have fast computational speed, but can handle only linear (or linearized) Gaussian systems (Kalman 1960). Particle filters, on the other hand, provide only approximate so-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

lution and in general, require much more computational resources (since probability evaluation is performed for each particle), but allow to accommodate arbitrary noise models. The latter is crucial for a vision-based systems, where detection noise model often cannot be straightforwardly derived from the architecture. Another attractive feature of particle filters is that they are time-independent and thus are applicable in cases with strongly varying sensor sampling time (Thrun 2002).

In this work, we propose a map-relative vision-based localization approach which matches detected lane markings to a reference vector map. A particle filter algorithm is exploited in order to combine sensory data from heterogeneous sources such as lane marking detectors (including a pedestrian crossings detectors), an a priori occupancy map and odometry.

The rest of the paper is structured as follows: section II explains the details of digital map preparation, section III describes generic particle filter algorithm and emphasizes the implementation details of our particular method, and section IV presents and discusses obtained results.

MAP PREPARATION

Predefined map is a crucial component of the proposed localization system. It allows storing information such as lane marking features and occupancy data in a compact format of a vector image. Such format allows for a natural representation of 2D geometric primitives such as lane markings and makes it easy to access and edit an existing map markup.

Although automatic mapping technique can be exploited, current implementation of map generation process requires manual specification of road marking in the form of polylines as shown in Fig. 1. There are no additional properties associated with line segments which significantly simplifies and speeds up the map preparation process. Apart from lanes detected pedestrian crossings are also used as an input to the localization. They are distinguished by the separate detector and approximated by rectangles in the road plane. As well as lines pedestrian crossings are manually marked on a map preparation step. Another important component of the map is the occupancy data marking places with zero probability of the vehicle presence. Accommodation of this information leads to lower requirements to computational speed by reducing the number of considered pose hypotheses.

Once prepared in vector graphic form the map is transformed to a raster image which allows to access any point on the map in $O(1)$ time. Another important property of the raster representation is an ability to explicitly accommodate the measurement noise model as will be shown below. Although a simple geometric representation of mapped lane markings is used in this work, other properties (e.g. line orientation) may be computed offline and encoded in a multi-channel raster image.

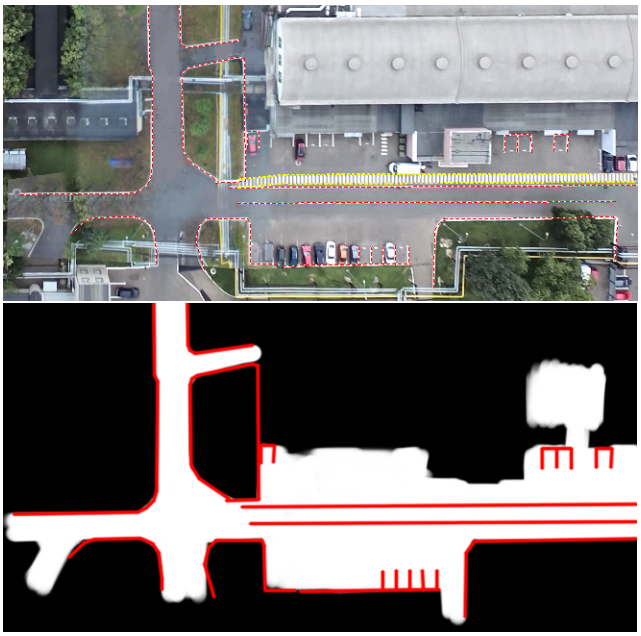


Figure 1. A fragment of a manually prepared map (on the top) and the resulting raster map used by the vehicle (on the bottom). Preprocessed lane markings are shown as red dashed line. Pedestrian crossings are marked as yellow rectangle. The resulting raster markup is shown as solid red lines. Black areas denote occupied space.

PARTICLE FILTERS

In the context of localization, particle filters are also known as Monte Carlo Localization (MCL) technique. They allow to derive the posterior distribution of state variables in a partially observable Markov chains. Let's denote a Markov chain state at time t as $\mathbf{x_t}$. The state $\mathbf{x_t}$ depends on the previous state $\mathbf{x_{t-1}}$ according to the probabilistic law $p(\mathbf{x_t}|\mathbf{u_t}, \mathbf{x_{t-1}})$, where $\mathbf{u_t}$ is the vector of control signals applied at time $t-1$. In the field of robotics, $p(\mathbf{x_t}|\mathbf{u_t}, \mathbf{x_{t-1}})$ is called an actuation model. The state of partially observable Markov chain can be described by measurement vector $\mathbf{z_t}$, generated via probabilistic law $p(\mathbf{z_t}|\mathbf{x_t})$, called a measurement model. The problem being solved by particle filters can be formulated as follows: given all consecutive sensors measurements $\mathbf{z^t} = \mathbf{z_0}, ..., \mathbf{z_t}$, and control vectors $\mathbf{u^t} = \mathbf{u_0}, ..., \mathbf{u_t}$ recover posterior distribution of the state $\mathbf{x_t}$ at any given time t (Thrun 2002). The idea behind MCL is to approximate unknown posterior by the set of particles $\{\mathbf{x_t^n}\}$ with associated weight, where $n = 1, ..., N$ – number of particles used. In our application each particle represents a hypothesis of the whole state vector of the vehicle pose. As discussed in (Thrun 2002), the original particle filter algorithm adopted to field of robotics has the form presented below.

Subsequent sections describe in details implementation specifics of each step of presented general particle filter algorithm.

Algorithm 1 Generic particle filter algorithm

procedure PARTICLE FILTER($\mathbf{x_{t-1}}, \mathbf{u_t}, \mathbf{z_t}$)
  $\{\mathbf{x_t^n}\} = \widetilde{\{\mathbf{x_t^n}\}} = \emptyset$
  for $n = 1$ to N **do**
    sample $x_t^n \sim p(\mathbf{x_t}|\mathbf{u_t}, \mathbf{x_{t-1}^n})$
    $w_t^n = p(\mathbf{z_t}|\mathbf{x_t^n})$
    $\widetilde{\{\mathbf{x_t^n}\}} = \widetilde{\{\mathbf{x_t^n}\}} + (x_t^n, w_t^n)$
  end for
  for $n = 1$ to N **do**
    draw i with probability $\propto \widetilde{w_t^i}$
    $\{\mathbf{x_t^n}\} = \{\mathbf{x_t^n}\} + (\widetilde{\mathbf{x_t^i}}, \widetilde{\mathbf{w_t^i}})$
  end for
  return $\{\mathbf{x_t^n}\}$
end procedure

Actuation model

The actuation model used in this work relies upon odometry data. Thus, on each iteration of particle filter each particle is evolved according to the relative odometry measurements. Gaussian nose with zero mean is applied to the motion of particles in order to accommodate measurement errors. The vehicle state is described by the state vector $\mathbf{x_t} = (\mathbf{x_t}, \mathbf{y_t}, \theta_t)^\mathbf{T}$, where x_t and y_t - vehicle 2D map-relative coordinates at the time t and θ_t is a corresponding heading angle. Given the difference between two consecutive odometry measurements:

$$\begin{cases} \Delta x_t = x_t^{odom} - x_{t-1}^{odom}, \\ \Delta y_t = y_t^{odom} - y_{t-1}^{odom}, \\ \Delta \theta_t = \theta_t^{odom} - \theta_{t-1}^{odom}, \end{cases} \quad (1)$$

the actuation applied to particles is expressed as:

$$\begin{cases} x_t^n = x_{t-1}^n + d_t^n \sin(\theta_{t-1}^n + \Delta\theta^n + \delta_t^n), \\ y_t^n = y_{t-1}^n + d_t^n \cos(\theta_{t-1}^n + \Delta\theta^n + \delta_t^n), \\ \theta_t^n = \theta_{t-1}^n + \Delta\theta^n + \delta_t^n, \end{cases} \quad (2)$$

where $d_t^n = \sqrt{(\Delta x_t)^2 + (\Delta y_t)^2} + \eta_t^n$, δ_t^n and η_t^n represent Gaussian white noise.

The set of sensors used in this work to support pose estimation includes wheel speed sensor and yaw-rate sensor. It is important to note that the presented set does not provide acceptable pose estimation on its own, and prone to significant drift over time. Due to this reason, no absolute pose estimation derived from odometry data is used for localization and only relative measurements $(\Delta x_t, \Delta y_t, \Delta\theta_t)$ are fed to the actuation model. The trajectory derived from odometry is shown in Fig. 2.

Measurement model

In order to compute the weight of each sample, the measurement model expressed as a likelihood function is applied as a next filtering step. The function indicates how well each particle corresponds to sensor measurements obtained at the current moment of time. This section describes how the likelihood function is derived for each type of measurements used in this work.

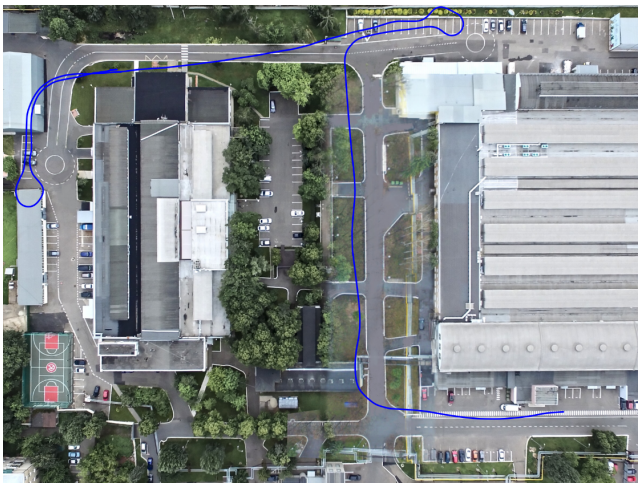


Figure 2. Vehicle trajectory estimated from odometry data.

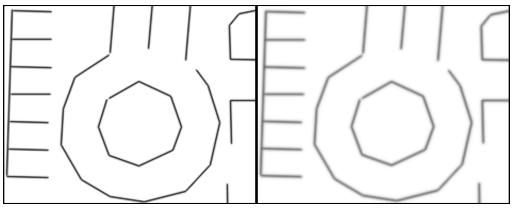


Figure 3. A fragment of the map image before (left) and after (right) Gaussian smoothing.

The main data source for the proposed localization method is the lane marking detector, which works on images from a single monocular camera. Lane markings detected on each incoming image are approximated by polylines. Then each polyline is transformed from the reference frame associated with camera origin to the map reference frame. Particle weight based on lane markings may be expressed via log-likelihood function as:

$$w_{m_t}{}^n = \sum_{i=1}^{I} \sum_{j=1}^{J} \ln f(x_{i,j}, y_{i,j}|\mathbf{x_t^n}), \quad (3)$$

where I – number of line segments detected on image, J – number of points corresponding to ith line segment and $f(x_{i,j}, y_{i,j}|\mathbf{x_t^n})$ is a function indicating the probability of point $(x_{i,j}, y_{i,j})$ to be detected given the state $\mathbf{x_t^n}$.

As it was mentioned previously measurement error model associated with lane markings detection is embedded into the map on offline map preparation stage. We assume that the error follows a normal distribution with zero mean. In order to account for it, Gaussian smoothing is applied to the raster map image as shown in Fig. 3.

The resulting likelihood value is computed as the sum values of the map pixels matched to the detected set of polylines. We propose an efficient approximation of equation (3) presented below:

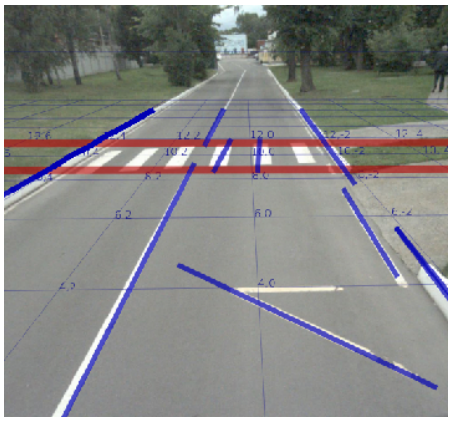


Figure 4. Example of lane markings (blue) and pedestrian crossing (red) recognition.

$$w_{m_t}^{\ n} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{I(x_{i,j}^n, y_{i,j}^n)}{\sigma}\right)^2}, \qquad (4)$$

where $x_{i,j}^n, y_{i,j}^n$ are coordinates of jth point of ith detected line segment in the map reference frame, $I(x_{i,j}^n, y_{i,j}^n)$ is a function which accepts coordinates in a map reference frame and returns intensity value of the pixel corresponding to this coordinates on a digital map image, σ is a standard deviation of lane markings detection error.

In addition to lane markings, other types of features and prior information can be used to improve the localization precision. As an example, a detector of pedestrian crossings similar to one proposed in (Povolotskiy et al. 2017) is used in this work to reduce longitudinal error. It becomes possible due to the higher specificity of the detector in comparison to the more generic lane marking detector. Fig. 4 shows the example of pedestrian crossing as well as lane marking detection. As it was discussed previously, pedestrian crossings are represented by rectangles on the map. The likelihood function for detected crossings is represented by a shifted sigmoid function as indicated below:

$$w_{ct}^{\ n} = \begin{cases} \sum_{i=1}^{I} \frac{1}{1+e^{(-C_n+S_i)}}, & \text{if crossing was detected,} \\ 1, & \text{otherwise.} \end{cases} \qquad (5)$$

In equation (5) C_n is an intersection area of a crossing marked on the reference map with the rectangle approximating the detected crossing transformed to the map reference frame according to a position of nth particle; S_i corresponds to a bounding rectangle area of ith crosswalk on the map.

Another type of information used to increase accuracy and computational speed is the occupancy map. It constrains the area where particle presence is possible (utilizing so-called "vehicle on road" assumption) and therefore discards unlikely particles as early as possible and improves the posterior distribution accuracy. The likelihood function for occupancy map is expressed as:

$$w_{om_t}^{\ n} = \begin{cases} 1, & \text{if pose is not occupied,} \\ 0, & \text{otherwise.} \end{cases} \qquad (6)$$

The resulting measurement model combining likelihood functions (4), (5), and (6) is presented below:

$$p(\boldsymbol{z_t}|\boldsymbol{x_t^n}) \propto w_t^n = \frac{w_{m_t}^{\ n} w_{ct}^{\ n} w_{om_t}^{\ n}}{\sum_{n=1}^{N} w_t^n} \qquad (7)$$

EVALUATION AND DISCUSSION

The proposed method was experimentally evaluated on a real ground vehicle of a small bus kind.

The analysis and comparison of localization methods are complicated by the fact that there is no straightforward way to obtain the reference trajectory from the sensors. In the scope of this work, in order to generate reference trajectories, the described approach described below was utilized. During the test run of a localization algorithm, on each iteration of particle filter the posterior particle distribution is saved with the corresponding indices relating each particle to one from the prior distribution from which it was derived. Together these measurements form the data structure referred to within the scope of this work as a particle transition graph. To obtain the reference trajectory the graph is traversed backward in time and the reference pose is computed on each step by averaging poses of particles which are known to survive on the subsequent steps.

While many systems proposed in the literature had been tested in relatively simple scenarios, we have deliberately chosen a challenging test bed with two 180 degree turns on the roundabouts to evaluate the proposed localization technique. The comparison of trajectory recovered from online localization versus reference trajectory is presented in the Fig. 5 along with quantitative localization results presented in table 1. The number of particles used in this experiment was set to 1000, which can be easily computed on any modern low-end computer. The x and y coordinates of starting position are assumed to be known in all the presented experiments unless otherwise indicated. Meanwhile the initial heading direction is unknown and has a uniform distribution.

Table 1: Summary of localization precision in experimental run.

	x, m	y, m	θ, degrees	Euclidean distance, m
Max abs. error	1.310	1.207	0.246	1.346
Mean abs. error	0.159	0.127	0.029	0.235
Standard deviation	0.246	0.237	0.044	0.251

The obtained results demonstrate that the proposed method is able to perform precise localization in com-

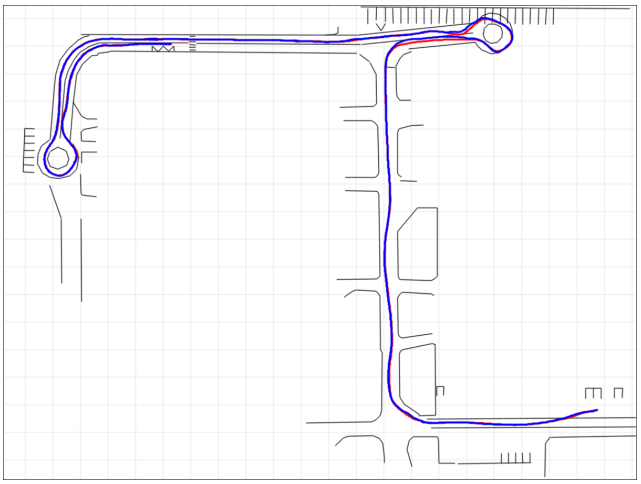


Figure 5. Reference trajectory (red) vs. trajectory recovered by particle filter algorithm (blue). Cell size is 10 m along both directions.

Table 2: Summary of localization precision in experimental run without lane marking detection.

	x, m	y, m	θ, degrees	Euclidean distance, m
Max abs. error	3.958	6.724	1.002	6.725
Mean abs. error	0.793	1.395	0.122	1.862
Standard deviation	1.107	1.826	0.221	1.159

plex real-world scenarios. Even though the maximum absolute error is quite high the average precision and standard deviation suggest that most of the time the localization precision stays within 0.5 m tolerance. It is important to note that most of the previous research is focused on regular cars where only limited use of lane markings can be made for localization purposes due to the relatively low camera position (and consequently narrow field of view), while for autonomous buses it is possible to install the camera at around 2 m height above the road surface and therefore to obtain a more informative picture of road marking.

In order to analyze the precision gain introduced by using lane feature-based approach trajectory obtained by using all the available data (i.e. odometry, occupancy map, crosswalk detection) except lane markings was compared with the reference one. The experimental results are presented in Fig. 6 and in table 2.

The results show that the use of lane markings yields about fivefold increase in localization precision.

It is worth mentioning that the developed approach is capable of solving the global localization problem, i.e. localization under global uncertainty. Fig. 7 shows the convergence of particles distribution over time in the scenario when all starting positions on the map have equal probability. The number of particles used in this experiment was increased to 30000 in order to populate the whole map uniformly.

CONCLUSION

A vision-based localization method for autonomous ground vehicles was proposed in this work. It is based on a combination of visual detectors (i.e. lane marking detector and pedestrian crossing detector) and odometry data. Particle filter technique was applied in order to fuse heterogeneous data sources. The proposed method has a straightforward yet computationally efficient implementation which allows it to be used even in applications with limited computational recourses. Another advantage of this method is an ease of map preparation and use. The required vector map contains lane markings and pedestrian crossings marked as simple geometric primitives. Once transformed into a raster image the map can also embed extra precomputed properties as it was demonstrated with lane detection error represented by Gaussian blur. Such map representation allows to reduce the number of operations in the online localization algorithm and guarantees constant data access time.

As it was demonstrated, another possible application of the proposed method is to be used as a preliminary reference trajectory generation approach for further map refinement and enrichment with other data, especially in cases where precise external localization is impossible, and implementing a full-featured simultaneous localization and mapping (SLAM) approach is impractical due to the algorithm complexity.

Tests on a real vehicle have demonstrated the applicability of the proposed method to the real-world scenarios. In considering the applicability of this method one should take into account that the tests were conducted on a bus-like platform which has a camera system placed at a considerable height, which may be unachievable for regular cars. Therefore a degradation of localization precision may be expected when used on conventional platforms.

Further improvement may include the use of additional lane marking features in the likelihood estimation such as line orientation as well as an extension of the set of visual detectors, e.g. stable local image features or stereo vision-based detection of buildings and curbs.

ACKNOWLEDGMENT

This work was supported by the Russian Science Foundation, project no 14-50-00150.

REFERENCES

Chausse, Frederic; Jean Laneurit; and Roland Chapuis. 2005. "Vehicle localization on a digital map using particles filtering." In *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*, 243–248.

Du, Xinxin and Kok Kiong Tan. 2016. "Vision-based approach towards lane line detection and vehicle localization." *Machine Vision and Applications*, 27(2):175–191.

Hata, Alberto Y; Fernando S Osorio; and Denis F Wolf. 2014. "Robust curb detection and vehicle localization in

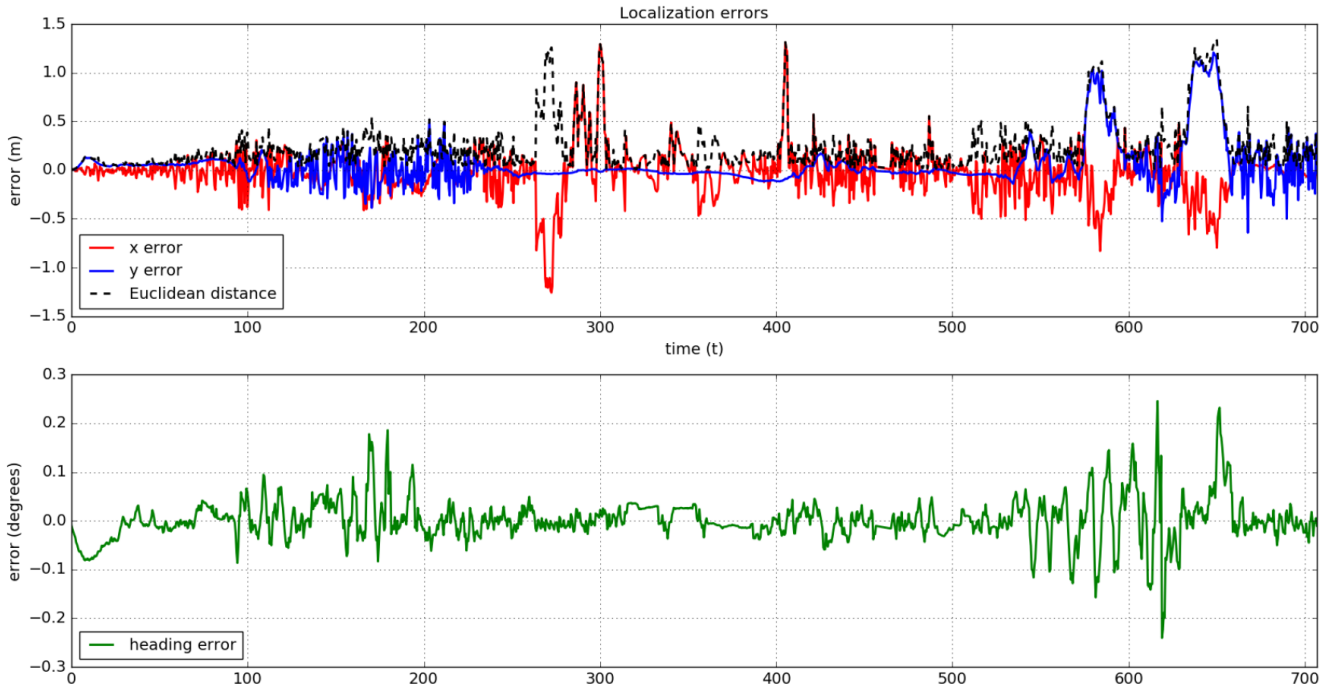


Figure 6. Localization errors in a testing run.

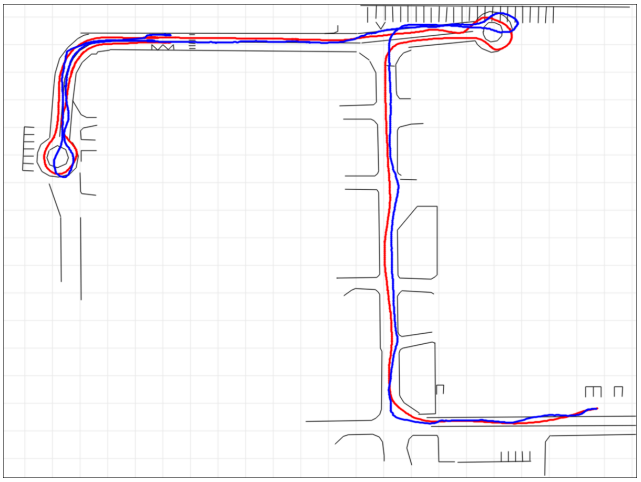


Figure 7. Reference trajectory (red) vs. trajectory recovered by particle filter algorithm in the absence of lane markings detector (blue). Cell size is 10 m along both directions.

urban environments." In *Intelligent vehicles symposium proceedings, 2014 IEEE*, 1257–1262.

Hernández, Jorge and Beatriz Marcotegui. 2009. "Filtering of artifacts and pavement segmentation from mobile lidar data." In *ISPRS Workshop Laserscanning 2009*.

Hillel, Aharon Bar; Ronen Lerner; Dan Levi; and Guy Raz. 2014. "Recent progress in road and lane detection: a survey." *Machine vision and applications*, 25(3):727–745.

Huang, Albert S; David Moore; Matthew Antone; Edwin Olson; and Seth Teller. 2009. "Finding multiple lanes in urban road networks with vision and lidar." *Autonomous Robots*, 26(2-3):103–122.

Jo, Kichun; Yongwoo Jo; Jae Kyu Suhr; Ho Gi Jung; and Myoungho Sunwoo. 2015. "Precise localization of an autonomous car based on probabilistic noise models of road surface marker features using multiple cameras." *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3377–3392.

Kalman, Rudolph Emil. 1960. "A new approach to linear filtering and prediction problems." *Journal of basic Engineering*, 82(1):35–45.

Krokhina, D.; V. Blinov; S. Gladilin; I. Tarxanov; and V. Postnikov. 2015. "Fast roadway detection using car cabin video camera." In *ICMV 2015*, volume 9875, 98751F–1–98751F–5 (SPIE, 2015).

Li, Liang; Ming Yang; Bing Wang; and Chunxiang Wang. 2017. "An overview on sensor map based localization for automated driving." In *Urban Remote Sensing Event (JURSE), 2017 Joint*, 1–4.

Lu, Wenjie; Emmanuel Seignez; F Sergio A Rodriguez; and Roger Reynaud. 2014. "Lane marking based vehicle localization using particle filter and multi-kernel estimation." In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, 601–606.

Povolotskiy, Mikhail A; Elena G Kuznetsova; and Timur Khanipov. 2017. "Russian license plate segmentation based on dynamic time warping."

Sheehan, Mark; Alastair Harrison; and Paul Newman. 2013. "Continuous vehicle localisation using sparse 3d sensing, kernelised rényi distance and fast gauss transforms." In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 398–405.

Tao, Zui; Ph Bonnifait; Vincent Fremont; and Javier Ibanez-Guzman. 2013. "Lane marking aided vehicle localization." In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, 1509–1515.

Thrun, Sebastian. 2002. "Particle filters in robotics." In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, 511–518.

Volkov, Alexey; Egor Ershov; Sergey Gladilin; and Dmitry Nikolaev. 2017. "Stereo-based visual localization without triangulation for unmanned robotics platform." In *2016 International Conference on Robotics and Machine Vision*, volume 10253, page 102530D.

Ziegler, Julius; Henning Lategahn; Markus Schreiber;

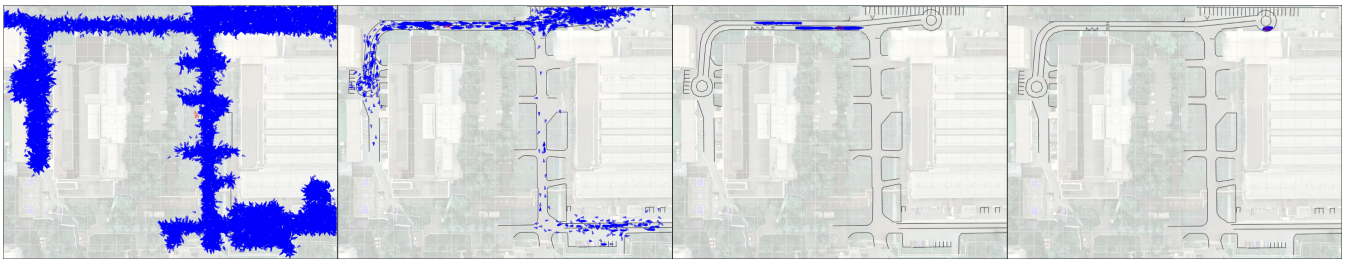


Figure 8. Particles evolution under global uncertainty. Blue markers denote particles positions. Red marker shows the estimated vehicle pose.

Christoph G Keller; Carsten Knoppel; Jochen Hipp; Martin Haueis; and Christoph Stiller. 2014. "Video based localization for bertha." In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, 1231–1238.

AUTHOR BIOGRAPHIES

OLEG SHIPITKO

was born in Yeisk, Russia. He obtained his bachelor degree in Electrical Engineering from Bauman Moscow State Technical University in 2015 and masters degree in Computer Science from Skolkovo Institute of Science and Technology in 2017. Since then he has works in the Laboratory of Vision Systems at the Institute for Information Transmission Problems. His research is focused on vision-based localization algorithms for robotics. Send mail to oleg.shipitko@iitp.ru.

ANTON GRIGORYEV

was born in Petropavlovsk-Kamchatskiy, Russia. Having graduated from Moscow Institute of Physics and Technology, he has been developing industrial computer vision systems with the Laboratory of Vision Systems at the Institute for Information Transmission Problems since 2010. Now he is heading the Autonomous Machine Vision Systems Research Group developing visual navigation solutions for robotic vehicles. Send mail to me@ansgri.com.

Thermistor Problem: Multi-Dimensional Modelling, Optimization, and Approximation

Ciro D'Apice
Dipartimento di Science Aziendali-
Management e Innovation Systems,
University of Salerno,
Via Giovanni Paolo II,
132, Fisciano (SA), Italy
Email: cdapice@unisa.it

Umberto De Maio
Department of Mathmatics
University of Naples Federico II,
Complesso Monte S. Angelo, via Cintia,
Napoli, 80126, Italy
Email: udemaio@unina.it

Peter I. Kogut
Department of Differential Equations
Oles Honchar Dnipro National University,
Gagarin av., 72,
Dnipro, 49010, Ukraine
Email: p.kogut@i.ua

KEYWORDS

Nonlinear elliptic equations, control in coefficients, $p(x)$-Laplacian, approximation approach, thermistor problem.

ABSTRACT

We consider a problem of an optimal control in coefficients for the system of two coupled elliptic equations also known as thermistor problem which provides a simultaneous description of the electric field $u = u(x)$ and temperature $\theta(x)$. The coefficients of operator div $(A(x)\nabla\theta(x))$ are used as the controls in $L^\infty(\Omega)$. The optimal control problem is to minimize the discrepancy between a given distribution $\theta_d \in L^r(\Omega)$ and the temperature of thermistor $\theta \in W_0^{1,\gamma}(\Omega)$ by choosing an appropriate anisotropic heat conductivity matrix B. Basing on the perturbation theory of extremal problems and the concept of fictitious controls, we propose an "approximation approach" and discuss the existence of the so-called quasi-optimal and optimal solutions to the given problem.

Introduction

Thermistor is a generic name for a device made from materials whose electrical conductivity is highly dependent on temperature. The advantages of thermistors as temperature measurement devices are low cost, high resolution, and flexibility in size and shape. The applications of thermistors can be summarized as follows:

- temperature sensing and control: thermistors provide inexpensive and reliable temperature sensing for a wide temperature range;
- thermal relay and switch: voltage regulation, surge protection;
- indirect measurement of other parameters: when a thermistor is heated its rate of change of temperature depends on its surroundings. This property can be used to monitor other quantities such as liquid level and fluid flow.

In a bounded open domain $\Omega \subset \mathbb{R}^N$, $N \geq 2$, we consider the following steady-state thermistor

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

problem

$$\operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right) = \operatorname{div} g \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0, \tag{1}$$

$$-\operatorname{div}\left(B\nabla\theta\right) = |\nabla u|^{p} \quad \text{in } \Omega, \quad \theta|_{\partial\Omega} = 0, \tag{2}$$

$$p(\cdot) = \sigma(\theta(\cdot)) \quad \text{a.e. in } \Omega, \tag{3}$$

where $B \in BV(\Omega)^{N\times N}$ is a given squared matrix.

System (1)–(3) describes the coupling between the electric field with potential u and the temperature θ in an anisotropic thermistor, where its anisotropic heat conductivity is given by a matrix of positive coefficients $B = [b_{i\,j}(x)]_{i,\,j=1,\ldots,N}$. This model is based on rational mechanics of electrorheological fluids, that takes into account the complex interactions between the electromagnetic fields and the moving liquid. In particular, the electrorheological fluids have the interesting property that their viscosity depends on the electric field in the fluid. A great deal of attention has been paid by many authors in the study of the thermistor problem during the last two decades. The search of the least assumptions on $\sigma(\theta)$, ensuring the (weak) solvability of the system (1)–(3), has been in the agenda of experts for decades. Earlier, existence theorems were proved only under some smallness conditions, e.g., in the case of a sufficiently small Lipschitz constant for the function $\sigma(\theta)$. However, the most essential progress in the study of existence and qualitative properties of solutions to the boundary value problem (1)–(3) was achieved by Zhikov [Zhikov 2011]. It has been shown that the solvability of these systems can be obtained in the multi-dimensional case without any smallness requirements on the function $\sigma(\theta)$ via a regularization approach and further passing to the limit over the parameter of regularization.

Setting of Optimization Problem

Our main goal is two-fold. The first one is to prove an existence result for the thermistor optimal control problem in coefficients with nonlinear state equations containing the p-Laplacian with variable exponent $p = p(x)$. The second one is to provide the asymptotic analysis of a special class of well-defined parametrized optimal control problems with fictitious controls and show that the original problem can be considered as a variational "limit"of the corresponding constrained minimization problems.

In view of this, in a bounded open domain $\Omega \subset \mathbb{R}^N$, $N \geq 2$, with sufficiently smooth boundary $\partial\Omega$, we deal with the following optimization problem:

$$\text{Minimize} \quad \left\{ J(B,u,\theta) = \int_\Omega |\theta(x) - \theta_d(x)|^r \, dx \right\} \tag{4}$$

subject to the constraints (1)–(3) and, in addition, $B \in \mathfrak{B}_{ad}$, where

$$\mathfrak{B}_{ad} = \left\{ \begin{array}{c} B \in BV(\Omega)^{N\times N}, \\ m_1 I \leq B(\cdot) \leq m_2 I, \\ \int_\Omega |Db_{ij}| \leq \mu \ \forall\, i,j = \overline{1,N}, \end{array} \right\} \tag{5}$$

$r \in \left(1, \frac{N}{N-2}\right)$ if $N > 2$ and $r \in (1, +\infty)$ for $N = 2$ is a given value, m_1 and m_2 are constants such that $0 < m_1 \leq m_2 < +\infty$, I is the identity matrix in $\mathbb{R}^{N\times N}$, the inequalities (5) are in the sense of the quadratic forms defined by $(B\xi,\xi)_{\mathbb{R}^N}$ for $\xi \in \mathbb{R}^N$, $\theta_d \in L^r(\Omega)$ and $g \in L^\infty(\Omega)^N$ are given distributions, σ is a continuous function such that $\alpha \leq \sigma(y) \leq \beta$ for all $y \in \mathbb{R}$ and the constants α and β satisfy the condition

$$1 < \alpha \leq \beta < \alpha^* = \left\{ \begin{array}{ll} +\infty, & \text{if } \alpha \geq N, \\ \frac{\alpha N}{N-\alpha}, & \text{if } \alpha < N, \end{array} \right. \tag{6}$$

$\mathfrak{B}_{ad}$ stands for the class of admissible controls, and μ is a given positive value. For motivation of BV-choice for the set of admissible controls, we refer to [D'Apice et al. 2008, D'Apice et al. 2010, D'Apice et al. 2012, D'Apice et al. 2014].

As for the optimal control problem (4)–(6), to the best of the authors knowledge, the existence of optimal solutions for the above thermistor problem remains an open question. Only very few articles deal with optimal control for the thermistor problem in two dimensional case (see [Hömberg et al. 2010], [Hrynkiv 2009] and references therein). There are several reasons for this:

- it is unknown whether the set of feasible points to the problem (4)–(6) is weakly closed in the corresponding functional space;
- we have no a priori estimates for the weak solutions to the boundary value problem (1)–(3) under conditions (6);

- the asymptotic behaviour of a minimizing sequence to the cost functional (4) is unclear in general;

- the optimal control problem (4)–(6) is ill-posed and relations (1)–(3) require some relaxation (see, for instance, [Durante et al. 2017, Kupenko and Manzo 2016, Kupenko and Manzo in press]).

To circumvent the problems listed above, we propose an "indirect approach" to the solvability of the optimal control thermistor problem in coefficients. Basing on the perturbation theory of extremal problems and the concept of fictitious controls (see, for instance, [Casas et al. 2016, Kogut and Leugering 2011, Kogut et al. 2016]), we prove the existence of so-called quasi-optimal and optimal solutions to the problem (4)–(6) and show that they can be attained by the optimal solutions of some appropriate approximations for the original optimal control problem. The main idea of our approach is based on the fact that weak solutions to the Dirichlet problem (1)–(3) can be attained through a special regularization of the exponent $p = p(x)$ and an approximation of the operator $\mathcal{A}(u) = \operatorname{div}\big(|\nabla u|^{p-2}\nabla u\big)$, using its perturbation by the $\varepsilon\Delta_\beta$-Laplacian, and the right-hand side of (2) by its transformation to $\operatorname{div}\big[\big(|\nabla u|^{\sigma(\theta)-2}\nabla u - g\big)\,u\big] + (g, \nabla u)_{\mathbb{R}^N}$. Here, by attainability of a weak solution (u, θ), we mean the existence of a sequence $\{(u_\varepsilon, \theta_\varepsilon)\}_{\varepsilon>0}$, where (u_ε, θ) are the solutions of "more regular" boundary value problems, such that $(u_\varepsilon, \theta_\varepsilon) \to (u, \theta)$ in some appropriate topology as ε tends to zero.

Preliminaries

We recall the well-known facts for nonlinear elliptic problems with variable exponent and discuss how each of the equations in system (1)–(3) can be interpreted. Assuming that the temperature $\theta = \theta(x)$ is known for some admissible control $B(x)$, we introduce the Sobolev-Orlicz space

$$W_0^{1,p(\cdot)}(\Omega) := \left\{u \in W_0^{1,1}(\Omega) \ : \ \int_\Omega |\nabla u(x)|^{p(x)}\,dx < +\infty\right\}$$

and equip it with the norm $\|u\|_{W_0^{1,p(\cdot)}(\Omega)} = \|\nabla u\|_{L^{p(\cdot)}(\Omega)^N}$, where $p(x) = \sigma(\theta(x))$. Here, $|\cdot|$ denotes the Euclidean norm $|\cdot|_{\mathbb{R}^N}$ in $\mathbb{R}^N$, and $L^{p(\cdot)}(\Omega)^N$ stands for the set of all measurable functions $f : \Omega \to \mathbb{R}^N$ such that $\int_\Omega |f(x)|^{p(x)}\,dx < +\infty$. It is well-known that, unlike in classical Sobolev spaces, smooth functions are not necessarily dense in $W = W_0^{1,p(\cdot)}(\Omega)$. Hence, with variable exponent $p = p(x)$ $(1 < \alpha \le p(\cdot) \le \beta)$ it can be associated another Sobolev space, $H = H_0^{1,p(\cdot)}(\Omega)$ as the closure of the set $C_0^\infty(\Omega)$ in $W_0^{1,p(\cdot)}(\Omega)$-norm. Since we can lose the density of the set $C_0^\infty(\Omega)$ in $W_0^{1,p(\cdot)}(\Omega)$ for some (irregular) variable exponents $p(x)$, it follows that a weak solution to the problem (1) is not unique, in general.

Definition 1. *We say that a function $u \in W_0^{1,p(\cdot)}(\Omega)$ is a weak solution of the problem* (1) *if*

$$\int_\Omega \big(|\nabla u|^{p-2}\nabla u, \nabla\varphi\big)_{\mathbb{R}^N}\,dx = \int_\Omega (g, \nabla\varphi)_{\mathbb{R}^N}\,dx, \tag{7}$$

for all $\varphi \in C_0^\infty(\Omega)$, and we say that u is the H-solution of problem (1)*, if $u \in H_0^{1,p(\cdot)}(\Omega)$ and the integral identity* (7) *holds for any test function $\varphi \in H_0^{1,p(\cdot)}(\Omega)$.*

As for the second equation (2), its right-hand side $|\nabla u|^p$ with $p(\cdot) = \sigma(\theta(\cdot))$, a priori belongs to the space $L^1(\Omega)$. In this case, following the L^1-theory of the Dirichlet problem for the Laplace operator, the solution to the boundary value problem (2) can be defined as solution obtained as a limit of approximations (SOLA).

Definition 2. *A function $\theta : \Omega \to \mathbb{R}$ is the SOLA to* (2) *if the following two conditions holds:*

- *$u \in L^1(\Omega)$ is a duality solution of* (2) *in the sense of Stampacchia, i.e.*

$$\int_\Omega \theta\varphi\,dx = \int_\Omega |\nabla u|^p v\,dx, \quad \forall\,\varphi \in L^\infty(\Omega),$$

where $v \in H_0^1(\Omega) \cap L^\infty(\Omega)$ is the weak solution of

$$-\operatorname{div}\big(B^t\nabla v\big) = \varphi \quad \text{in } \Omega \qquad v = 0 \text{ on } \partial\Omega;$$

- *For any sequence $\{f_n\}_{n\in\mathbb{N}} \subset L^\infty(\Omega)$ such that $f_n \to |\nabla u|^p$ strongly in $L^1(\Omega)$ and $\|f_n\|_{L^1(\Omega)} \le$*

$\||\nabla u|^p\|_{L^1(\Omega)} = \|u\|^p_{W_0^{1,p(\cdot)}(\Omega)}$ *for all* $n \in \mathbb{N}$, *we have*

$\theta_n \to \theta$ *strongly in* $L^1(\Omega)$, *weakly in* $W_0^{1,\gamma}(\Omega)$

for all $\gamma \in [1, \frac{N}{N-1})$, *where* $\theta_n \in H_0^1(\Omega) \cap L^\infty(\Omega)$ *is the weak solution of*

$$-\operatorname{div}(B\nabla\theta_n) = f_n \quad \text{in } \Omega \quad \theta_n = 0 \text{ on } \partial\Omega.$$

The main result concerning the existence of a SOLA to the problem (2), can be stated as follows: if Ω is a bounded domain with sufficiently smooth boundary and $|\nabla u|^{p(\cdot)} \in L^1(\Omega)$, then the Dirichlet problem (2) has the unique SOLA $\theta \in W_0^{1,\gamma}(\Omega)$ with $\gamma \in [1, \frac{N}{N-1})$; moreover, there exists a constant $C = C(\gamma)$ independent of $f = |\nabla u|^{p(\cdot)}$ such that

$$\|\theta\|_{W_0^{1,\gamma}(\Omega)} \le C(\gamma) \int_\Omega |\nabla u(x)|^{p(x)}|\, dx. \qquad (8)$$

In fact, if the datum $f = |\nabla u|^p$ is more regular, say $f \in L^{1+\delta}(\Omega)$ for some $\delta > 0$, we have the following result: if $|\nabla u|^p \in L^{1+\delta}(\Omega)$, $0 < \delta < \frac{N-2}{N+2}$ then the unique SOLA of (2) belongs to $W_0^{1,q}(\Omega)$ with $q = \frac{N(1+\delta)}{N-1-\delta} = 1 + \delta + \frac{(1+\delta)^2}{N-1-\delta}$.

The optimal control problem we consider in this paper is to minimize the discrepancy between a given distribution $\theta_d \in L^r(\Omega)$ and the temperature of thermistor $\theta \in W_0^{1,\gamma}(\Omega)$ by choosing an appropriate anisotropic heat conductivity matrix $B \in \mathfrak{B}_{ad}$. It is assumed here that $r \in \left(1, \frac{N}{N-2}\right)$ where the choice of such range is motivated by Sobolev Embedding Theorem. Namely, in view of the fact that the embedding $W_0^{1,\gamma}(\Omega) \hookrightarrow L^q(\Omega)$ is compact for all $q \in [1, \frac{N}{N-2})$, the exponents γ and r can be related as follows $\gamma = \frac{Nr}{N+r}$. As a result, for a given $r \in \left(1, \frac{N}{N-2}\right)$ we have $\gamma \in \left[1, \frac{N}{N-1}\right)$.

Characteristic features of the OCP (4)–(6)

Since for a "typical" measurable or even continuous function $\sigma(\theta)$ with properties (6), the set $C_0^\infty(\Omega)$ is not dense in $W_0^{1,p(\cdot)}(\Omega)$, and, hence, no uniqueness of weak solutions to (1)–(3) can be expected, the mapping $B \mapsto (u, \theta)$, where (u, θ) is a weak solution to the boundary value problem (1)–(3), can be multi-valued in general. In view of this, we introduce the set of feasible solutions to the OCP (4)–(6) as follows: $(B, u, \theta, p) \in \Xi_0$ if and only if

$$\left\{ \begin{array}{c} B \in \mathfrak{B}_{ad},\ u \in H_0^{1,p(\cdot)}(\Omega),\ \theta \in W_0^{1,\gamma}(\Omega), \\ p \in L^\infty(\Omega), \quad \gamma = \frac{Nr}{N+r}, \\ p(\cdot) = \sigma(\theta(\cdot)) \ \text{ a.e. in } \Omega, \\ u \text{ is the } H\text{-solution of (1)}, \\ \theta \text{ is the SOLA to (2)}. \end{array} \right\} \qquad (9)$$

It is clear that $J(B, u, \theta, p) < +\infty$ for all $(B, u, \theta, p) \in \Xi_0$.

So, the characteristic feature of the OCP (4)–(6) is the fact that a priori it is unknown whether the set Ξ_0 is nonempty. Using the assumption (6) and basing on a special technique of the weak convergence of fluxes to a flux, it was established in [Zhikov 2011] that the thermistor problem (1)–(3) for $B = \xi I$, with $\xi \in [m_1, m_2]$, and for any measurable function $\sigma(\theta)$ admits a weak solution $u \in W_0^{1,p(\cdot)}(\Omega)$. However, in this case the inclusion $u \in H_0^{1,p(\cdot)}(\Omega)$ is by no means obvious even in the case of diagonal constant matrix $B \in \mathfrak{B}_{ad}$. Hence, the OCP (4)–(6) requires some relaxation. With that in mind we propose to consider the function $p(\cdot)$ as a fictitious control with some more regular properties and interpret the fulfilment of equality $p(\cdot) = \sigma(\theta(\cdot))$ with some accuracy.

Relaxation of the original OCP

We consider the following extension of the set of feasible solutions to the original OCP. Let $k_0 > 0$ and $\tau \ge 0$ be given constants.

Definition 3. *We say that a tuple* (B, u, θ, p) *is quasi-feasible to the OCP* (4)–(6) *if* $(B, u, \theta, p) \in \widehat{\Xi}_0(\tau)$, *where*

$$\left\{ \begin{array}{c} B \in \mathfrak{B}_{ad},\ u \in H_0^{1,p(\cdot)}(\Omega), \\ \theta \in W_0^{1,\gamma}(\Omega), p \in \mathfrak{S}_{ad}, \\ \|p - \sigma(\theta)\|_{L^2(\Omega)} \le \tau, \quad \gamma = \frac{Nr}{N+r}, \\ u \textit{ is the } H\textit{-solution of } (1), \\ \theta \textit{ is the weak solution to } (2). \end{array} \right\} \qquad (10)$$

$$\mathfrak{S}_{ad} = \big\{ q \in C(\overline{\Omega}) \text{ such that } \quad (11)$$
$$|q(x) - q(y)| \le \omega(|x-y|),$$
$$\forall\, x, y, \in \Omega,\ |x-y| \le 1/2,$$
$$\omega(t) = k_0/\log(|t|^{-1}),\ 1 < \alpha \le q(\cdot) \le \beta \text{ in } \overline{\Omega}\big\}. \quad (12)$$

We also say that $(B^0, u^0, \theta^0, p^0) \in BV(\Omega)^{N\times N} \times H_0^{1,p(\cdot)}(\Omega) \times W_0^{1,\gamma}(\Omega) \times C(\overline{\Omega})$ *is a quasi-optimal solution to the problem* (4)–(6) *if*

$$(B^0, u^0, \theta^0, p^0) \in \widehat{\Xi}_0(\tau) \ \text{ and}$$
$$J(B^0, u^0, \theta^0, p^0) = \inf_{(B,u,\theta,p)\in\widehat{\Xi}_0(\tau)} J(B, u, \theta, p),$$

and this tuple is called to be optimal if $p^0(\cdot) = \sigma(\theta^0(\cdot))$ *a.e. in* Ω. It is clear that $\widehat{\Xi}_0(\tau) \subset \Xi_0$ for $\tau = 0$ and, moreover, as we will see later on, the set $\widehat{\Xi}_0(\tau)$ is nonempty if only $\tau \ge \sqrt{|\Omega|}(\beta - \alpha)$. It is also worth to emphasize that the condition $p \in \mathfrak{S}_{ad}$ implies that $p(\cdot)$ has some additional regularity. Moreover, in view of the obvious relation $\lim_{t\to 0} |t|^\delta \log(|t|) = 0$ with $\delta \in (0,1)$, it is clear that $p \in C^{0,\delta}(\Omega)$ implies $p \in \mathfrak{S}_{ad}$. Because of this $p \in \mathfrak{S}_{ad}$ is often called a locally log-Hölder continuous exponent. Another point about benefit of the choice of the subset $\mathfrak{S}_{ad}$ is related with the following properties: (i) $\mathfrak{S}_{ad}$ is a compact subset in $C(\overline{\Omega})$ and thus provides uniformly convergent subsequences; (ii) Every cluster point p of a sequence $\{p_k\}_{k\in\mathbb{N}} \subset \mathfrak{S}_{ad}$ is a regular exponent (i.e. in this case the set $C_0^\infty(\Omega)$ is dense in $W_0^{1,p(\cdot)}(\Omega)$), which plays a key role in our further study; (iii) Because of the log-Hölder continuity of an exponent $p \in \mathfrak{S}_{ad}$, the corresponding weak solution $u \in W_0^{1,p(\cdot)}(\Omega)$ to the variational problem (7) is such that $|\nabla u|^{(1+\delta)p(\cdot)} \in L^1(\Omega)$ for some $\delta > 0$ and satisfies the estimate

$$\int_\Omega |\nabla u(x)|^{(1+\delta)p(x)}\, dx \le C \int_\Omega |g(x)|^{(1+\delta)p'(x)}\, dx + C, \quad (13)$$

where $\delta > 0$ and $C > 0$ depend only on Ω, α, N, k_0, and $\int_\Omega |g|^{p'}\, dx$. The property (13) is crucial for the proof of existence of quasi-optimal solutions to the problem (4)–(6). It is easy to show that if $u \in W^{1,p(\cdot)}(\Omega)$ is a solution to $\operatorname{div}(A(u)\nabla u) = \operatorname{div} g$ in $\mathcal{D}'(\Omega)$, then

$$\big(A(u)\nabla u, \nabla u\big) = \operatorname{div}\big((A(u)\nabla u - g)u\big) + g \cdot \nabla u,$$

also in $\mathcal{D}'(\Omega)$, where

$$A(u) = |\nabla u(x)|^{p(x)-2}$$
$$\text{or} \quad A(u) = |\nabla u|^{p(x)-2} + \varepsilon |\nabla u|^{\beta-2}.$$

As a result, it allows to deduce the existence of the unique weak solution to the variational problem

$$-\operatorname{div}(B\nabla\theta) = \operatorname{div}\big((A(u)\nabla u - g)u\big) +$$
$$(g, \nabla u) \quad \text{in } \mathcal{D}'(\Omega)$$

which is also the SOLA to the Dirichlet BVP

$$-\operatorname{div}(B\nabla\theta) = |\big(A(u)\nabla u, \nabla u\big)| \quad \text{in } \Omega, \quad \theta|_{\partial\Omega} = 0.$$

Our main goal in this paper is to present the "approximation approach", based on the perturbation theory of extremal problems and the concept of fictitious controls. With that in mind, we make use of the following family of approximated problems: Minimize $J_{\varepsilon,\tau}(B, u, \theta, p)$, where

$$J_{\varepsilon,\tau}(B, u, \theta, p) = \int_\Omega |\theta - \theta_d|^r\, dx$$
$$+ \frac{1}{\varepsilon}\mu_\tau\left(\int_\Omega |p - \sigma(\theta)|^2\, dx\right) \quad (14)$$

subject to the constraints

$$\operatorname{div}\Big(|\nabla u|^{p(x)-2}\nabla u + \varepsilon|\nabla u|^{\beta-2}\nabla u\Big) = \operatorname{div} g \quad \text{in } \Omega, \quad (15)$$
$$u|_{\partial\Omega} = 0, \quad (16)$$
$$-\operatorname{div}(B\nabla\theta) =$$
$$= \operatorname{div}\Big[\Big(|\nabla u|^{p(x)-2}\nabla u + \varepsilon|\nabla u|^{\beta-2}\nabla u - g\Big)u\Big] \quad (17)$$
$$+ (g, \nabla u) \quad \text{in } \Omega, \quad \theta|_{\partial\Omega} = 0, \quad (18)$$
$$B \in \mathfrak{B}_{ad}, \quad p \in \mathfrak{S}_{ad}. \quad (19)$$

Here, the function $\mu_\tau : \mathbb{R}_+ \to \mathbb{R}_+$ is defined as follows

$$\mu_\tau(s) = 0 \ \text{ if } \ 0 \le s \le \tau^2$$

and

$$\mu_\tau(s) = s - \tau^2 \ \text{ if } \ s > \tau^2.$$

There are several principle points in the statement of approximated problem (14)–(19) that should be emphasized. The first one is related with $\varepsilon\Delta_\beta$-regularization of $p(\cdot)$-Laplacian. Though this is a

standard trick in order to establish the existence of H-solution to the Dirichlet problem (15) with a given exponent $p(\cdot)$, however, this approach does not allow to arrive at the existence of a weak solution $(u,\theta) \in H_0^{1,p(\cdot)}(\Omega) \times W_0^{1,\gamma}(\Omega)$ to the thermistor problem (1)–(3). This can be done if only the exponent $p(\cdot) = \sigma(\theta(\cdot))$ is regular, i.e. if the set $C_0^\infty(\Omega)$ is dense in $W_0^{1,p(\cdot)}(\Omega)$, and the energy density $|\nabla u(\cdot)|^{p(\cdot)}$ belongs to the space $L^{1+\delta}(\Omega)$ for some $\delta > 0$ so that the equation (18) holds in the sense of the distributions. With that in mind we consider the condition $p \in \mathfrak{S}_{ad}$ as an additional option for the regularization of the original OCP. The another point that should be indicated, is related with some relaxation of the equation (2). Namely, it is easy to see that after the formal transformations, the equation (2) can be transformed to the following one

$$-\operatorname{div}(B\nabla\theta) = \operatorname{div}\left[\left(|\nabla u|^{\sigma(\theta)-2}\nabla u - g\right)u\right] + (g, \nabla u) \tag{20}$$

in $\mathcal{D}'(\Omega)$. The benefit of such representation and condition $p \in \mathfrak{S}_{ad}$ is the fact that, due to the estimate (13), the expression $\left(|\nabla u|^{\sigma(\theta)-2}\nabla u - g\right)u$ under the divergence sign in (20) is integrable with degree greater than 1. As follows from our further analysis, this property plays an important role in the study of OCP (4)–(6) and we consider the representation (20) as some relaxation of the relation (2).

Some Auxiliary Results

Orlicz spaces

Let $p(\cdot)$ be a measurable exponent function on Ω such that $1 < \alpha \le p(x) \le \beta < \infty$ a.e. in Ω, where α and β are given constants. Let $p'(\cdot) = \frac{p(\cdot)}{p(\cdot)-1}$ be the corresponding conjugate exponent. It is clear that $\beta' \le p'(\cdot) \le \alpha'$ a.e. in Ω, where β' and α' stand for the conjugates of constant exponents. Denote by $L^{p(\cdot)}(\Omega)^N$ the set of all measurable functions $f(x)$ on Ω such that $\int_\Omega |f(x)|^{p(x)}\,dx < \infty$. Then $L^{p(\cdot)}(\Omega)^N$ is a reflexive separable Banach space with respect to the Luxemburg norm

$$\|f\|_{L^{p(\cdot)}(\Omega)^N} = \inf\left\{\lambda > 0 \ : \ \rho_p(\lambda^{-1}f) \le 1\right\}, \tag{21}$$

where

$$\rho_p(f) := \int_\Omega |f(x)|^{p(x)}\,dx.$$

The dual of $L^{p(\cdot)}(\Omega)^N$ with respect to $L^2(\Omega)$-inner product will be denoted by $L^{p'(\cdot)}(\Omega)^N$. The following estimates are well-known: if $f \in L^{p(\cdot)}(\Omega)^N$ then

$$\|f\|^\alpha_{L^{p(\cdot)}(\Omega)^N} \le \int_\Omega |f(x)|^{p(x)}\,dx \le \|f\|^\beta_{L^{p(\cdot)}(\Omega)^N}, \tag{22}$$

$$\text{if } \|f\|_{L^{p(\cdot)}(\Omega)^N} > 1,$$

$$\|f\|^\beta_{L^{p(\cdot)}(\Omega)^N} \le \int_\Omega |f(x)|^{p(x)}\,dx \le \|f\|^\alpha_{L^{p(\cdot)}(\Omega)^N}, \tag{23}$$

$$\text{if } \|f\|_{L^{p(\cdot)}(\Omega)^N} < 1,$$

$$\|f\|_{L^{p(\cdot)}(\Omega)^N} = \int_\Omega |f(x)|^{p(x)}\,dx, \tag{24}$$

$$\text{if } \|f\|_{L^{p(\cdot)}(\Omega)^N} = 1,$$

$$\|f\|^\alpha_{L^{p(\cdot)}(\Omega)^N} - 1 \le \int_\Omega |f(x)|^{p(x)}\,dx \le \|f\|^\beta_{L^{p(\cdot)}(\Omega)^N} + 1, \tag{25}$$

$$\|f\|_{L^\alpha(\Omega)^N} \le (1+|\Omega|)^{1/\alpha}\,\|f\|_{L^{p(\cdot)}(\Omega)^N}. \tag{26}$$

Moreover, due to the duality method, it can be shown that

$$\|f\|_{L^{p(\cdot)}(\Omega)^N} \le (1+|\Omega|)^{1/\beta'}\,\|f\|_{L^\beta(\Omega)^N}, \tag{27}$$

$$\beta' = \frac{\beta}{\beta-1}, \quad \forall\, f \in L^\beta(\Omega)^N.$$

We make use of the following results. **Lemma 1.** [[Zhikov 2011], Lemma 13.3] *If a sequence $\{f_k\}_{k\in\mathbb{N}}$ is bounded in $L^{p(\cdot)}(\Omega)$ and $f_k \rightharpoonup f$ in $L^\alpha(\Omega)$ as $k \to \infty$, then $f \in L^{p(\cdot)}(\Omega)$ and $f_k \rightharpoonup f$ in $L^{p(\cdot)}(\Omega)$, i.e.*

$$\lim_{k\to\infty}\int_\Omega f_k\varphi\,dx = \int_\Omega f_k\varphi\,dx, \quad \forall\,\varphi \in L^{p'(\cdot)}(\Omega).$$

Lemma 2. *Let $\{p_k\}_{k\in\mathbb{N}} \subset \mathfrak{S}_{ad}$ and $p \in \mathfrak{S}_{ad}$ be such that $p_k(\cdot) \to p(\cdot)$ uniformly in $\overline{\Omega}$ as $k \to \infty$. If a sequence $\left\{\|f_k\|_{L^{p_k(\cdot)}(\Omega)}\right\}_{k\in\mathbb{N}}$ is bounded and $f_k \rightharpoonup f$ in $L^\alpha(\Omega)$ as $k \to \infty$, then $f \in L^{p(\cdot)}(\Omega)$.*

On the weak convergence of fluxes to flux

A typical situation arising in the study of most optimization problems and which is of fundamental importance in many other areas of nonlinear

analysis, can be stated as follows: we have the weak convergence $u_k \rightharpoonup u$ in some Sobolev space $W^{1,\alpha}(\Omega)$ with $\alpha > 1$ and we have the weak convergence of fluxes $A_k(\cdot, \nabla u_k) \rightharpoonup z$ in the Lebesgue space $L^{\delta}(\Omega)$, $\delta > 1$, where by flux we mean the vector under the divergence sign in an elliptic equation (in our case it is $A_k(\cdot, \nabla u_k) = |\nabla u_k|^{p_k(\cdot)-2}\nabla u_k$ or $A_k(\cdot, \nabla \theta_k) = B_k(\cdot)\nabla \theta_k$). Then the problem is to show that $z = A(\cdot, \nabla u)$, although the validity of this equality is by no means obvious at this stage.

Assume that the fluxes $A_k(x, \xi)$ satisfy the following conditions:

$$A_k : \Omega \times \mathbb{R}^N \to \mathbb{R}^N \tag{28}$$

are Carathéodory vector-valued functions,

$$(A_k(x, \xi) - A_k(x, \eta), \xi - \eta)_{\mathbb{R}^N} \geq 0, \tag{29}$$

$$A_k(x, 0) = 0, \text{ for a.e. } x \in \Omega \text{ and } \forall \xi, \eta \in \mathbb{R}^N, \tag{30}$$

$$|A_k(x, \xi)|^{\beta'} \leq C_1|\xi|^{\beta} + C_2, \tag{31}$$

$$\lim_{k\to\infty} A_k(x, \xi) = A(x, \xi)$$

$$\text{for a.e. } x \in \Omega \text{ and } \forall \xi \in \mathbb{R}^N. \tag{32}$$

Let $\{v_k\}_{k\in\mathbb{N}}$ and $\{A_k(\cdot, v_k)\}_{k\in\mathbb{N}}$ be weakly convergent sequences in $L^1(\Omega)^N$, and let v and z be their weak L^1-limits, respectively. In order to clarify the conditions under which the equality $z = A(x, v)$ holds and the fluxes $A_k(\cdot, v_k)$ weakly converge to the flux $A(\cdot, v)$, we cite the following result.

Lemma 3.[[Zhikov 2011], Theorem 4.6] *Assume that $\{u_k\}_{k\in\mathbb{N}}$ and $\{A_k(\cdot, \nabla u_k)\}_{k\in\mathbb{N}}$ are the sequences such that conditions* (29)–(31) *hold true and*

(i) $u_k \rightharpoonup u$ in $W^{1,\alpha}(\Omega)$ and $u_k \in W^{1,\beta}(\Omega)$ for all $k \in \mathbb{N}$;

(ii) $\sup_{k\in\mathbb{N}} \|A_k(\cdot, \nabla u_k)\|_{L^{\beta'}(\Omega)^N} < +\infty$;

(iii) $\sup_{k\in\mathbb{N}} \| (A_k(\cdot, \nabla u_k), \nabla u_k)_{\mathbb{R}^N} \|_{L^1(\Omega)} < +\infty$;

(iv) the exponents α and β are related by the condition

$$1 < \alpha \leq \beta < \begin{cases} +\infty, & \text{if } \alpha \geq N - 1, \\ \frac{\alpha(N-1)}{N-1-\alpha}, & \text{if } \alpha < N - 1. \end{cases} \tag{33}$$

Then, up to a subsequence, the fluxes weakly converge to the flux

$$A_k(\cdot, \nabla u_k) \rightharpoonup A(\cdot, \nabla u) \quad \text{in} \quad L^{\beta'}(\Omega)^N.$$

It is worth to note that in the case of equality $\alpha = \beta$, Lemma 3 becomes the well-known result of Tartar and Murat also known as the div-curl Lemma.

On Consistency of Approximated Optimal Control Problems in Coefficients

Let ε be a small parameter. Assume that the parameter ε varies within a strictly decreasing sequence of positive real numbers which converges to 0. Let $\tau \geq 0$ be a given constant. We consider the collection of approximated optimal control problems in coefficients for nonlinear elliptic equations (14)–(19). For every $\varepsilon > 0$ we denote by $\widehat{\Xi}_{\varepsilon}$ the set of all feasible points to the problem (14)–(19).

Definition 4. *We say that (B, u, θ, p) is a feasible point to the problem* (14)–(19) *if $B \in \mathfrak{B}_{ad}$, $p \in \mathfrak{S}_{ad}$, and $u \in W_0^{1,\beta}(\Omega)$ and $\theta \in W_0^{1,\gamma}(\Omega)$ are the solutions to the following variational problems*

$$\operatorname{div}\left(|\nabla u|^{p(x)-2}\nabla u + \varepsilon|\nabla u|^{\beta-2}\nabla u\right) = \operatorname{div} g \tag{34}$$

$$-\operatorname{div}(B\nabla\theta) = (g, \nabla u)$$

$$+\operatorname{div}\left[\left(|\nabla u|^{p(x)-2}\nabla u + \varepsilon|\nabla u|^{\beta-2}\nabla u - g\right)u\right] +_{\mathbb{R}^N}, \tag{35}$$

where each of these relations we consider as equalities in $\mathcal{D}'(\Omega)$

The following lemma reflecting the consistency of approximated optimal control problem (14)–(19) [Kupenko and Manzo 2015].

Lemma 4. Let $\theta_d \in L^r(\Omega)$ and $g \in L^{\infty}(\Omega)^N$ be given distributions with $r \in \left(1, \frac{N}{N-2}\right)$, let $\sigma \in C(\mathbb{R})$ be a given function satisfying the conditions (6), and let τ be an arbitrary non-negative value. Then the approximated optimal control problem (14)–(19) is consistent for each $\varepsilon > 0$, i.e. $\widehat{\Xi}_{\varepsilon} \neq \emptyset$.

As an obvious consequence of the reasoning given in proof of Lemma 4, we can draw the following inference.

Corollary. *For given $\tau \geq 0$ and $\varepsilon > 0$, let $(B, u, \theta, p) \in \widehat{\Xi}_{\varepsilon}$ be a feasible point to the problem* (14)–(19). *Then θ the unique SOLA to the Dirichlet problem*

$$-\xi \operatorname{div}(\nabla\theta_{\varepsilon,k}) = |\nabla\widehat{u}|^{\beta} \quad \text{in} \quad \Omega, \quad \widehat{\theta}\Big|_{\partial\Omega} = 0. \tag{36}$$

The next results are crucial for our analysis.

Lemma 5. *The set of fictitious controls $\mathfrak{S}_{ad}$ is convex, bounded and compact with respect to the strong topology of $C(\overline{\Omega})$.*

Lemma 6. *Let $\{(B_{\varepsilon,k}, u_{\varepsilon,k}, \theta_{\varepsilon,k}, p_{\varepsilon,k})\}_{k\in\mathbb{N}} \subset \widehat{\Xi}_\varepsilon$ be an arbitrary sequence. Then there exist a distribution $u_\varepsilon \in W_0^{1,\beta}(\Omega)$, an exponent $p_\varepsilon \in \mathfrak{S}_{ad}$, and a subsequence of $\{u_{\varepsilon,k}\}_{k\in\mathbb{N}}$, still denoted by the suffix (ε, k), such that*

$$\varepsilon \|u_\varepsilon\|^\beta_{W_0^{1,\beta}(\Omega)} \le 2^{\alpha'+1}\left(\int_\Omega |g|^{\alpha'}\,dx + |\Omega|\right), \quad (37)$$

$$u_{\varepsilon,k} \rightharpoonup u_\varepsilon \ \text{ in } \ W_0^{1,\beta}(\Omega) \ \text{ as } \ k\to\infty, \quad (38)$$

$$u_\varepsilon \in W_0^{1,p_\varepsilon(\cdot)}(\Omega). \quad (39)$$

Lemma 7. *Let $\{(B_{\varepsilon,k}, u_{\varepsilon,k}, \theta_{\varepsilon,k}, p_{\varepsilon,k})\}_{k\in\mathbb{N}} \subset \widehat{\Xi}_\varepsilon$ be an arbitrary sequence, and let $p_\varepsilon \in \mathfrak{S}_{ad}$ and $u_\varepsilon \in W_0^{1,\beta}(\Omega)$ be such that $p_{\varepsilon,k}(\cdot) \to p_\varepsilon(\cdot)$ uniformly in $\overline{\Omega}$ and $u_{\varepsilon,k} \rightharpoonup u_\varepsilon$ in $W_0^{1,\beta}(\Omega)$ as $k\to\infty$. Then, up to a subsequence, we have the weak convergence of fluxes to a flux:*

$$|\nabla u_{\varepsilon,k}|^{p_{\varepsilon,k}-2}\nabla u_{\varepsilon,k} + \varepsilon|\nabla u_{\varepsilon,k}|^{\beta-2}\nabla u_{\varepsilon,k} \rightharpoonup |\nabla u_\varepsilon|^{p_\varepsilon-2}\nabla u_\varepsilon + \varepsilon|\nabla u_\varepsilon|^{\beta-2}\nabla u_\varepsilon \quad \text{in} \quad L^{\beta'}(\Omega)^N. \quad (40)$$

Lemma 8. *Let $p_\varepsilon \in \mathfrak{S}_{ad}$ and $u_\varepsilon \in W_0^{1,\beta}(\Omega)$ be as in Lemma 8. Then u_ε is the unique weak solution to the Dirichlet problem*

$$\operatorname{div}\left(|\nabla u|^{p_\varepsilon(x)-2}\nabla u + \varepsilon|\nabla u|^{\beta-2}\nabla u\right) = \operatorname{div} g \quad \text{in} \quad \Omega,$$
$$u|_{\partial\Omega} = 0.$$

Lemma 9. *Let $\{(B_{\varepsilon,k}, u_{\varepsilon,k}, \theta_{\varepsilon,k}, p_{\varepsilon,k})\}_{k\in\mathbb{N}} \subset \widehat{\Xi}_\varepsilon$ be a sequence such that $B_{\varepsilon,k} \overset{*}{\rightharpoonup} B_\varepsilon$ in $BV(\Omega)^{N\times N}$ and $\theta_{\varepsilon,k} \rightharpoonup \theta_\varepsilon$ in $W_0^{1,\gamma}(\Omega)$ for some $\gamma \in [1, \frac{N}{N-1})$. Then we have*

$$\lim_{k\to\infty}\int_\Omega (B_{\varepsilon,k}\nabla\theta_{\varepsilon,k}, \nabla\varphi)_{\mathbb{R}^N}\,dx = \int_\Omega (B_\varepsilon\nabla\theta_\varepsilon, \nabla\varphi)_{\mathbb{R}^N}\,dx, \quad (41)$$

for all $\varphi \in C_0^\infty(\Omega)$.

To conclude this section, we give the existence result for the approximated OCP (14)–(19).

Theorem 1. Let $\theta_d \in L^r(\Omega)$ and $g \in L^\infty(\Omega)^N$ be given distributions with $r \in \left(1, \frac{N}{N-2}\right)$, let $\sigma \in C(\mathbb{R})$ be a given function satisfying the conditions (6), and let τ be an arbitrary non-negative value. Then the optimal control problem (14)–(19) admits at least one solution for each $\varepsilon > 0$.

Main results

The main result of this paper is the following theorem, where we claim that if the OCP (4)–(6) has a sufficiently regular feasible point, then there exist optimal solutions to the OCP and some of them are the limit as $\varepsilon \searrow 0$ of optimal solutions to (14)–(19).

Theorem 2. *Let Ω be an open bounded domain in $\mathbb{R}^N$ with a sufficiently smooth boundary. Assume that $\widehat{\Xi}_0(\tau) \neq \emptyset$ for $\tau = 0$, i.e. there exist a matrix $\widehat{B} \in \mathfrak{B}_{ad}$, an exponent $\widehat{p} \in \mathfrak{S}_{ad}$, and a weak solution to the thermistor problem (1)–(3) $(\widehat{u}, \widehat{\theta}) \in W_0^{1,\sigma(\widehat{\theta}(\cdot))}(\Omega) \times W_0^{1,\gamma}(\Omega)$ with $\gamma = \frac{Nr}{N+r}$ and $B(\cdot) = \widehat{B}(\cdot)$ such that $\widehat{p} = \sigma(\widehat{\theta})$ almost everywhere in Ω. Then OCP (4)–(6) has a non-empty set of optimal solutions and some of them can be attained in the following way*

$$B_\varepsilon^0 \overset{*}{\rightharpoonup} B^0 \ \text{ in } \ BV(\Omega)^{N\times N}, \quad u_\varepsilon^0 \rightharpoonup u^0 \ \text{ in } \ W_0^{1,\alpha}(\Omega), \quad (42)$$

$$\theta_\varepsilon^0 \rightharpoonup \theta^0 \ \text{ in } \ W_0^{1,\gamma}(\Omega), \quad p_\varepsilon^0 \to p^0 \ \text{ uniformly on } \ \overline{\Omega}, \quad (43)$$

as $\varepsilon \to 0$, where $(B_\varepsilon^0, u_\varepsilon^0, \theta_\varepsilon^0, p_\varepsilon^0)$ are the solutions to the approximated problems (14)–(19) with $\tau = \varepsilon$ in (14).

It is clear that the condition $\widehat{p} = \sigma(\widehat{\theta})$ in the statement of Theorem 1, where $\widehat{p}$ has logarithmic modulus of continuity, imposes some additional and rather special constraint on the function $\sigma \in C(\mathbb{R})$. The principle point here is the fact that this relation has to be valid for a particular function $\widehat{\theta}$ and it is not required that the function $\sigma(\theta(\cdot))$ must be at least continuous for every solution $\theta \in W_0^{1,\gamma}(\Omega)$ of (2). It is rather delicate problem to guarantee the fulfilment of the equality $\widehat{p} = \sigma(\widehat{\theta})$ by the direct description of function $\sigma \in C(\mathbb{R})$ even if we make use of the "typical" assumption: σ is a Lipschitz continuous function.

Since it is unknown whether OCP (4)–(6) is solvable or the main assumptions of Theorem 1 are satisfied, it is reasonable to show that this problem

admits the quasi-optimal solutions and they can be attained (in some sense) by optimal solutions to special approximated problems. We prove the following result.

Theorem 3. *Let* $\left\{(B_\varepsilon^0, u_\varepsilon^0, \theta_\varepsilon^0, p_\varepsilon^0)\right\}_{\varepsilon>0}$ *be an arbitrary sequence of optimal solutions to the approximated problems (14)–(19). Assume that either there exists a constant* $C^* > 0$ *satisfying condition*

$$\limsup_{\varepsilon\to 0} \inf_{(B,u,\theta,p)\in\widehat{\Xi}_\varepsilon} J_{\varepsilon,\tau}(B,u,\theta,p) \le C^* < +\infty$$

or $\tau \ge \sqrt{|\Omega|}(\beta-\alpha)$*, where* $\widehat{\Xi}_\varepsilon$ *stands for the set of feasible solutions to the problem (14)–(19). Then any cluster tuple* $\left(B^0, u^0, \theta^0, p^0\right)$ *in the sense of convergence* (42)–(43) *is a quasi-optimal solution of the OCP* (4)–(6)*. Moreover, in this case the following variational property holds*

$$\lim_{\varepsilon\to 0} \inf_{(B,u,\theta,p)\in\widehat{\Xi}_\varepsilon} J_{\varepsilon,\tau}(B,u,\theta,p)$$
$$= J\left(B^0, u^0, \theta^0, p^0\right) = \inf_{(B,u,\theta,p)\in\widehat{\Xi}_0(\tau)} J(B,u,\theta,p).$$

REFERENCES

E. Casas, P.I. Kogut and G. Leugering, *Approximation of Optimal Control Problems in the Coefficient for the p-Laplace Equation. I. Convergence Result*, SIAM J. Control Optim., 54(3)(2016), 1406–1422.

C. D'Apice, U. De Maio, P.I. Kogut, *Gap phenomenon in the homogenization of parabolic optimal control problems*, Journal of Mathematical Control and Information, 25(4) (2008), 461–489;

C. D'Apice, U. De Maio, O. Kogut, *On shape stability of Dirichlet optimal control problems in coefficients for nonlinear elliptic equations*, Advances in Differential Equations, 15(7-8) (2010), 689–720;

C. D'Apice, U. De Maio, O. Kogut, *Optimal Control Problems in Coefficients for Degenerate Equations of Monotone Type: Shape Stability and Attainability Problems*, SIAM Journal on Control and Optimization, 50(3) (2012), 1174–1199;

C. D'Apice, U. De Maio, P.I. Kogut, R. Manzo, *Solvability of an Optimal Control Problem in Coefficients for Ill-Posed Elliptic Boundary Value Problems*, Electronic Journal of Differential Equations, 2014(166) (2014), 1–23;

T. Durante, O. P. Kupenko and R. Manzo, *On attainability of optimal controls in coefficients for system of Hammerstein type with anisotropic p-Laplacian*, Ricerche di Matematica, 66(2) (2017), 259-292.

D. Hömberg, C. Meyer, J. Rehberg and W. Ring, *Optimal control for the thermistor problem*, SIAM Journal on Control and Optimization, 49(5)(2010), 3449-3481.

V. Hrynkiv, *Optimal boundary control for a time dependent thermistor problem*, Electr. Journal of Diff. Equations., 2009(83)(2009), 1–22.

P. Kogut and G. Leugering, *Optimal control problems for partial differential equations on reticulated domains. Approximation and Asymptotic Analysis*, Series: Systems and Control, Birkhäuser Verlag, Boston, 2011.

P. I. Kogut, R. Manzo and A. O. Putchenko, emphOn approximate solutions to the Neumann elliptic boundary value problem with non-linearity of exponential type, Boundary Value Problems, 2016(1)(2016), 1–32.

O. P. Kupenko and R. Manzo, *Approximation of an optimal control problem in the coefficients for variational inequality with anisotropic p-Laplacian*, Nonlinear Differential Equations and Applications, NoDEA, (2016) 23:35. https://doi.org/10.1007/s00030-016-0387-9.

O. P. Kupenko and R. Manzo, *On optimal controls in coefficients for ill-posed non-linear elliptic Dirichlet boundary value problems*, Discrete and Continuous Dynamic Systems. Series B, (to appear).

O. P. Kupenko and R. Manzo, *Shape stability of optimal control problems in coefficients for coupled system of Hammerstein type*, Discrete and Continuous Dynamical Systems Journal, Series B, 20(9)(2015), pp. 2967-2992.

V. V. Zhikov, *On variational problems and nonlinear elliptic equations with nonstandard growth conditions*, Journal of Mathematical Sciences, 173(5) (2011), pp. 463-570.

Simulation of Fluid-Mechanically Effective Microstructures and Combustion Processes

Towards Immersed Boundary Methods for Complex Roughness Structures in Scale-Resolving Simulations

Konrad M. Hartung
Institute of Energy and Process Engineering
Jade University of Applied Sciences
26389 Wilhelmshaven, Germany

Philipp Gilge
Florian Herbst
Institute of Turbomachinery and Fluid Dynamics
Leibniz Universität Hannover
30167 Hannover, Germany

KEYWORDS

Direct Numerical Simulations, Immersed Boundary Method, Surface Roughness

ABSTRACT

In many technical applications the effect of surface roughness on the local flow as well as on integral characteristics is significant. Understanding and modeling their effect is an ongoing challenge as there are plenty of surface structures caused by intention, manufacturing, or wear which have different or even contrary effects on the boundary layer flow. Scale-resolving simulations like direct numerical simulations are a valuable tool in this context as they provide highly-resolved view of the local effect of roughness on the flow. However, complex surface structures pose challenges to the generation of commonly used body-fitted structured computational grids. Immersed boundary methods (IBM) are a promising tool for bypassing this challenge. In this paper the IBM implemented in the CFD-solver OpenFOAM is qualified for scale-resolving simulations of turbulent channel flows over rough surfaces by introducing an additional mass-flow controller. By means of three characteristic test-cases the direct numerical simulations with IBM are verified against corresponding simulations with body-fitted grids. The excellent quantitative prediction of average flow quantities as well as turbulent statistics demonstrate the suitability of the method for the simulation of turbulent flows over arbitrary complex rough surfaces.

INTRODUCTION

Surface roughnesses and their effect on the boundary layer flow was first investigated by Nikuradse (1933) and Schlichting (1936). They used sand grain surface structures to quantify the effect of these roughnesses in pipe flows. To correlate the roughness height with the effects in the boundary layer, they established the sand grain diameter k_s as the quantification of the roughness height. Since then, k_s is a standard parameter to characterize and to describe surface roughness for model based low order fluid simulation.

Real surface roughnesses resulting from tribological effects in real gas turbine machines, like heavy duty or instationary machines like aircraft engines, usually form non-equivalent sand grain structures (Bons 2010; Bons et al. 2001; Achary et al. 1986; Taylor 1990). Because of particle impacts, erosion, and wear a mix of different undirected (isotropic) and directed (anisotropic) structures are formed which are inhomogeneously distributed along a blade surface (Fig. 1). Also surface roughness structures resulting from manufacturing or refurbishment are forming individual structures combining isotropic and anisotropic elements (Denkena et al. 2015; Nespor et al. 2016). Thus, real surface topologies are too complex to describe them with a single parameter and model-based roughness investigations for near-wall boundary layer flows correlated to this single parameter are not always expedient. To investigate the effect of roughness on the flow, many investigations apply scale-resolving simulations like Large Eddy Simulations (LES) or Direct Numerical Simulations (DNS) to avoid the use of model-based interpretation of roughness effects on the flow (Hohenstein and Seume 2013; Flack and Schultz 2010; Cui et al. 2003a,b; Ikeda and Durbin 2007; Singh et al. 2007; Iacone et al. 2008). However, DNS and LES are very sensitive to the spatial discretization of the fluid. High quality standards are necessary, to get a proper discretization, especially of the near wall boundary layer region. For very complex structures, these high standards of the discretization in the near wall grid generation may not be met.

One solution is the use of unstructured grids with e.g. tetrahedral cells. However, because of the their geometric shape, they are susceptible to higher numerical dissipation (Bensow and Liefvendahl 2008) than hexahedral cells and consequently they are not expedient to study the effects of roughness on the near wall flow. The hexahedral discretization reduces the dissipative effects, but is difficult or even impossible to apply on complex structures (e.g. with undercuts). Thus, only simplified (viz. filtered) roughness structures can be

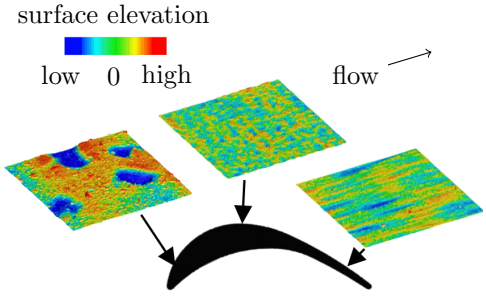


Fig. 1. Real worn measured surface roughness on a high pressure turbine blade from a aircraft engine

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

investigated by LES or DNS with the necessary grid quality (Hohenstein and Seume 2013).

An alternative method to discretize the near wall region is the Immersed Boundary Method (IBM). The IBM is based on a decoupling of the discretized fluid region and the solid body structures. The effect of a wall is implemented by an analytic force term which dependes on the relative position of the wall and neighbouring fluid cells (Mittal and Iaccarino 2005). This method is usually used in LES or DNS investigation with moving objects in a fluid domain (Kempe and Fröhlich 2012). A first attempt to use the IBM for roughness investigations showed that the IBM can also be adopted for near wall flow investigations (Busse et al. 2015). However, a direct comparison of surface roughness with different stages of complexity (deterministic model structures and real surface structures) has not been given so far. The objective of the present work is a validation of the IBM for complex surface structures in channel flows. For this reason, plane channel flow, the flow over a riblet structure, and an operationally stressed surface structure measured on a real high-pressure turbine blade from an aircraft engine are investigated and compared to literature results.

SETUP

Immersed Boundary Method

The DNS in the current work are conducted within the open-source simulation environment foam-extend-3.1, which is part of the open source library OpenFoam including various extensions and additional features, e.g an implementation of the IBM. The IBM makes it possible to perform CFD simulations of arbitrary complex and moving geometries with preferably cartesian, non body-fitted grids. Hence, the time-consuming task of conformal grid generation, often associated with additional problems regarding the mesh quality or interpolation errors in inter-grid transfer can be greatly simplified. Since the method was introduced by Peskin (1972) a number of variants have been developed, which differ in the approach used to impose the effect of the immersed boundary (IB) on the flow. The implementation of the IBM in foam-extend-3.1, initially introduced by Tuković and Jasak (2012), is based on a discrete forcing approach with direct imposition of boundary conditions. As the IBM grid's boundary does not reflect the contour of the geometry, the effect of viscous walls on the flow is modeled by a force term F_i added to the momentum equation for cells near the IB

$$\frac{\partial\left(\rho u_i\right)}{\partial t}+\frac{\partial\left(\rho u_i u_j-\tau_{ij}\right)}{\partial x_j}=F_i \qquad (1)$$

with the density ρ and the shear stress τ.

The approach requires a classification of cells according to Fig. 2 into fluid cells, solid cells, and IB cells. Please note that the distinguishing criterion is the position of the cell's center relative to the IB. In order to transfer

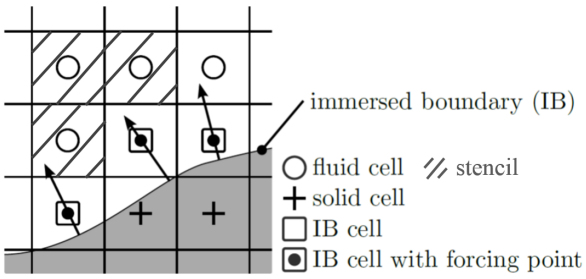


Fig. 2. Schematic of the direct forcing immersed boundary method

the physical boundary condition to the cartesian grid, forcing points at the center of wall layer cells and wall-normal vectors through each forcing point are defined. Flow quantities at the forcing point are reconstructed by interpolation using neighboring cells on an extended stencil and the Dirichlet or the Neumann boundary conditions on the IB directly imposed by specified values. The unknown coefficients for the quadratic polynomial interpolation on the stencil are determined using a weighted least square method. The computed flow quantities in the forcing point makes it possible to determine the additional force term at time instance $n+1$ for fluid cells interacting with IB-cells according to Eq. (2):

$$F_i=\rho\frac{u_{\mathrm{IB}}^{n+1}-u_i^n}{\Delta t}-\frac{\partial\left(\rho u_i u_j-\tau_{ij}\right)}{\partial x_j}. \qquad (2)$$

For a detailed overview of the mathematical formulation of the IBM the reader may refer to Peskin (2002) and Mittal and Iaccarino (2005).

Implemented velocity/pressure correction

Presently foam-extend-3.1 does not include solvers suitable for solving the incompressible laminar Navier-Stokes equations for channel flows with constant mass flow rates, that additionally support the IBM. For this purpose a solver was developed based on the IBM-compatible finite volume flow solver icoIbFoam, which solves the incompressible laminar Navier-Stokes equations using the PISO (Pressure-Implicit with Splitting of Operators) algorithm. The time marching scheme is an implicit second-order backward scheme, and spatial discretization is based on bounded second order central schemes.

To maintain a constant mass flow rate and to counterbalance energy dissipation induced by friction losses driving forces were corrected by imposing an additional pressure gradient calculated by means of comparison of the current mean flow velocity in the channel and a specified value. The mean velocity in the channel is estimated by the sum of the local velocity of each cell weighted with its corresponding volume (V_i) in relation to the overall volume of the computational domain

$$u_{\text{mean}} = \frac{\sum_{i=1}^{N} u_i V_i}{\sum_{i=1}^{N} V_i}. \quad (3)$$

In terms of using a cartesian non body-fitted background grid, which includes solid cells, these cells are identified and neglected for the estimation of the mean velocity by setting their volume to zero in Eq. (3).

Computational Setups

In order to qualify the IBM for the simulation of complex surface structures, initianally a smooth wall configuration and two different roughness geometries providing different characteristic challenges to the IBM are investigated systematically. The investigation includes the flow over a smooth surface, an operationally stressed surface structure measured on a high-pressure turbine blade, and a riblet structured surface with trapezoid cross-section. The simulations were conducted at the friction Reynolds number

$$\text{Re}_\tau = \frac{u_\tau \delta}{\nu}; u_\tau = \sqrt{\frac{\tau_\text{w}}{\rho}} \quad (4)$$

of $\text{Re}_\tau = 180$ and verified with the literature data of Moser et al. (1999), Hohenstein (2014) and data provided by Koepplin et al. (2017b), respectively. The comparative data was obtained by simulations, which used the traditional approach to impose the presence of solid obstacles by applying the no-slip boundary condition on body-fitted grids.

In accordance with the studies of Ashrafian et al. (2004) the dimensions of the channel were set to $L_x \times L_y \times L_z = 6\delta \times 2\delta \times 3\delta$ in streamwise, spanwise, and wall-normal direction with $\delta = 0.5$ m representing half the height of the channel. Depending on the the applied surface structure the height of the computational domain can differ slightly in order to maintain a constant channel volume (Fig. 4). For the cartesian background grid a uniform grid spacing is selected in streamwise and spanwise direction. The grid resolutions in the two homogeneous directions are set equally to $\Delta x^+ = \Delta z^+ \approx 3.6$ ($\vec{x}^+ = (\vec{x}\ u_\tau)/\nu$). In wall-normal direction equally spaced cells are applied within the distance $y = 2\delta_{\text{IB}+10}$ with a grid resolution of $\Delta y^+ \approx 0.6$. Therefore, approximately six nodes are embedded within the viscous sublayer. Table I gives an overview of the channel dimensions and total number of cells used for the numerical investigations in reference to literature data.

In streamwise and spanwise directions periodic boundary conditions were applied representing a setup with indefinitely extended channel dimensions. The time step was chosen such that the maximal CFL number remains below 0.3. In order to obtain converged turbulent statistics, a non-dimensional through flow time of at least

$$t^+ = \frac{t\ u_\text{b}}{\delta} = 1000 \quad (5)$$

TABLE I: Geometry and grid parameters of the computational domains

	$L_x \times L_y \times L_z$	N_x	N_y	N_z
IBM	$6\delta \times 2\delta \times 3\delta$	300	208	150
Moser et al.	$4\pi\delta \times 2\delta \times \frac{4}{3}\pi\delta$	128	129	128
Hohenstein	$6\delta \times 2\delta \times 3\delta$	518	216	258
Koepplin et al.	$6\delta \times 2\delta \times 2\delta$	128	181	620

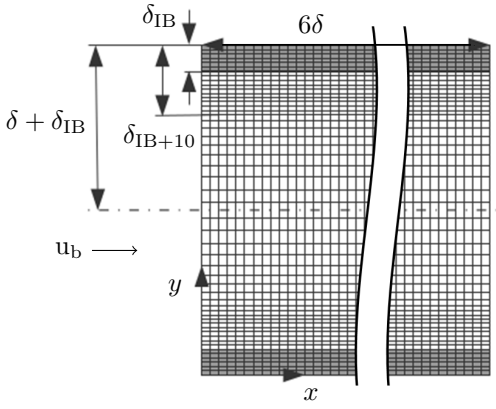


Fig. 3. Computational grid for IBM investigations

is ensured for all investigated test-cases, whereby t is the computed time, u_b the bulk velocity, and δ the channel half height.

RESULTS

The numerical results for the three test-cases are compared to literature data in terms of the spatial and temporal averaged velocity profiles in wall normal direction normalized by the friction velocity u_τ. Furthermore, the Reynolds stresses $\overline{u'u'}$ and $\overline{u'v'}$ are compared as they are known to be highly sensitive to fluctuations in the turbulent flows.

Plane Channel Flow

Initially, the IBM implementation is verified by a smooth wall configuration of a turbulent channel flow. In order to provide the required classification of cells, it has been ensured, that near wall cells are cut by the immersed boundary (Fig. 6). Figure 7 shows the wall-normal profile of the mean velocity in wall units compared to the DNS results of Moser et al. (1999). The mean velocity profiles show a remarkable agreement over the entire channel height. Special attention should be paid to the first two cells, because they are directly affected by the direct forcing approach. Furthermore, the distributions of the Reynolds stresses illustrated in Figure 10 also show good correlation with the data of Moser et al. (1999). The deviation of the Reynolds normal stress $\overline{u'u'}$ in the range of $10 < y^+ > 20$ can be attributed to the slightly mismatching friction Reynolds number. The simulation of Moser et al. (1999) was conducted at a friction Reynolds number of 178.13, wheras for the simulation based on the IBM a friction Reynolds number of 176.6 was attained. The numerical studies of

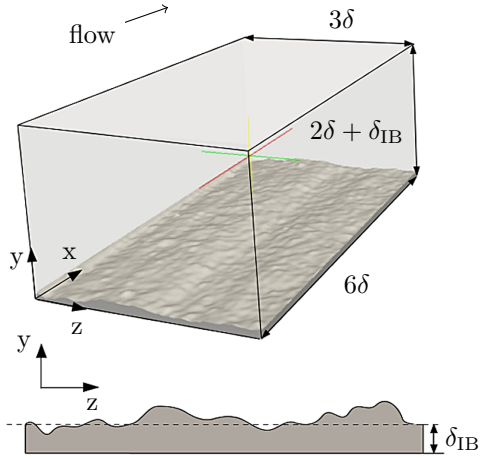


Fig. 4. Schematic illustration of the flow configuration for the operationally stressed surface structure

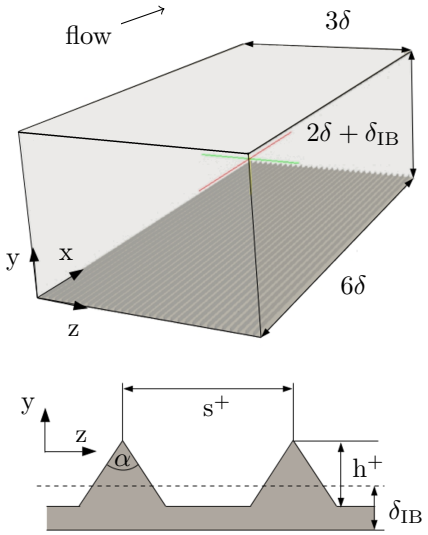


Fig. 5. Schematic illustration of the flow configuration for the riblet surface structure

Moser et al. (1999) for various friction Reynolds numbers clearly state that an increased friction Reynolds number leads to an increasing maximum Reynolds normal stress.

Operationally Stressed Surface

Figure 8 shows the wall normal profile of the mean velocity in wall units for the operationally stressed surface structure. Overall the results are in good agreement with the existing DNS by Hohenstein (2014). In the logarithmic region, however, the velocity is slightly overestimated. Furthermore, the velocity especially in the first two cells shows a slight deviation compared to the results of Hohenstein (2014). This possibly results from an insufficiently precise reconstruction of the flow quantities in the IB-cells, which are used to determine the force term for neighboring fluid cells. As these near-wall cells are additionally used to compute the friction velocity u_τ, more precisely the mean wall shear stress τ_W a deviation of the mean velocity profile, the normalized Reynolds stresses and the computed friction Reynolds number was to be expected (cf. Equation 4). For the DNS of Hohenstein (2014) and the results based on the IBM, a friction Reynolds number of 197.1 and 188.5 are computed, respectively. The obviously underestimated friction velocity results in the overestimation of the normalized maximum Reynolds stresses (Fig. 11) with simultaneously underestimating the friction Reynolds number.

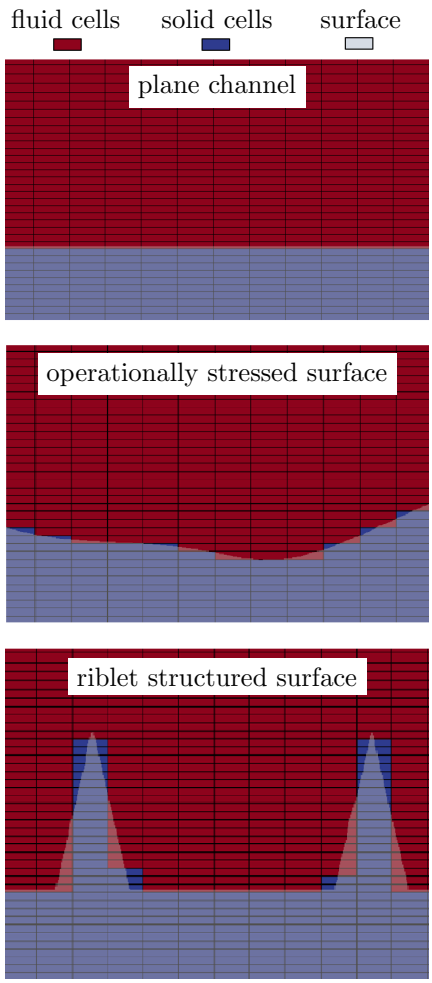


Fig. 6. Comparison of the three test cases with IBM, surface roughness geometry, and definition of fluid cells and solid cells

Riblet Structured Surface

The turbulent channel flow with a riblet structured surface shows the overall highest discrepancies between the numerical results. Both, the mean velocity profile and the Reynolds stresses exhibit a significant deviation over the entire channel height. Therefore, comparable to the results of the operationally stressed surface structure the friction velocity is obviously underestimated, which coincides with the computed friction Reynolds numbers (cf. Eq. 4) and the normalized Reynolds stresses (Fig. 12). The frictional Reynolds number for data provided by Koepplin et al. (2017b) and simulation based on the IBM were computed to 179.0 and 164.1, respectively.

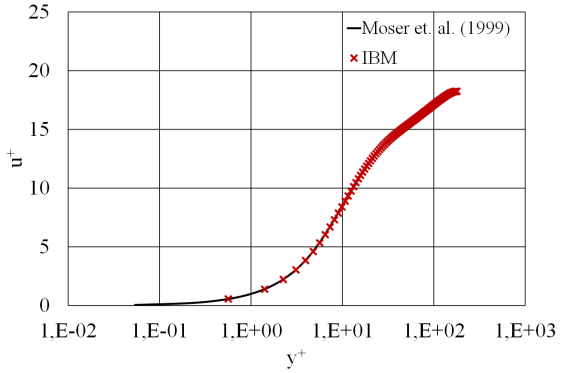


Fig. 7. Mean velocity profile in wall units for the plane channel test-case

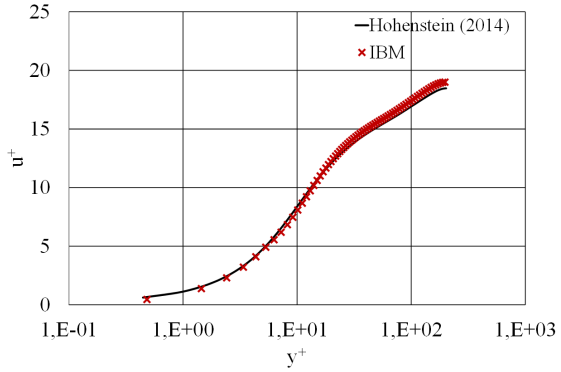


Fig. 8. Mean velocity profile in wall units for the operationally stressed surface

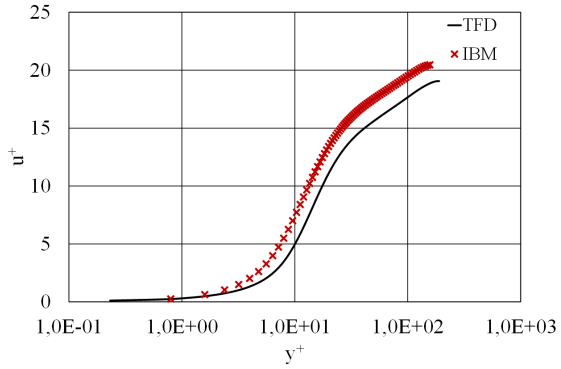


Fig. 9. Mean velocity profile in wall units for the riblet structured surface

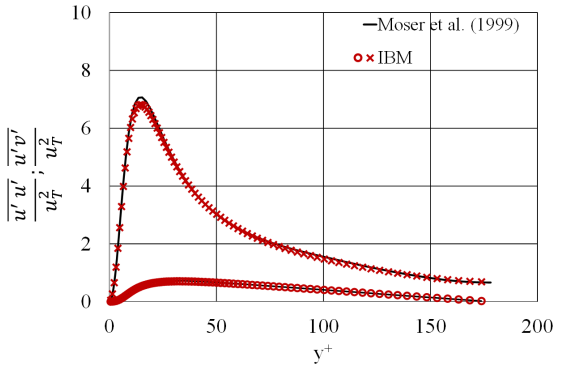


Fig. 10. Normalized Reynolds stresses $\overline{u'u'}$(×) and $\overline{u'v'}$(○) for the plane channel test-case

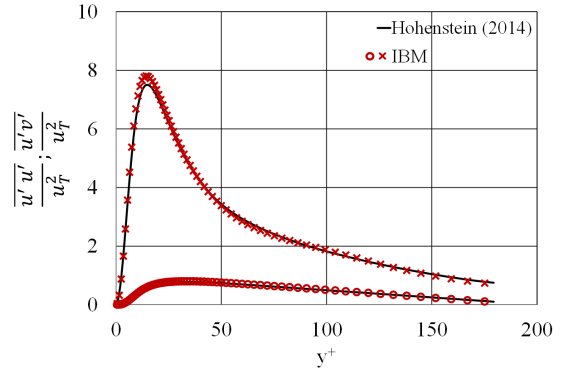


Fig. 11. Normalized Reynolds stresses $\overline{u'u'}$(×) and $\overline{u'v'}$(○) for the operationally stressed surface

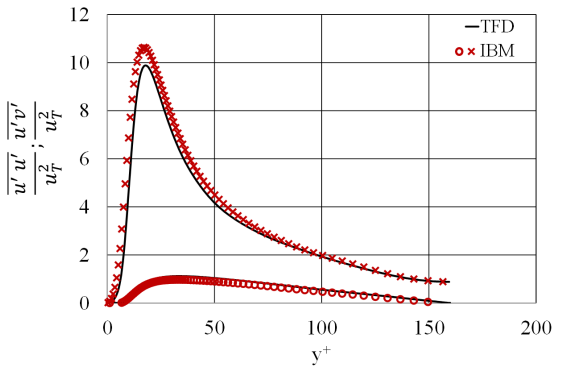


Fig. 12. Normalized Reynolds stresses $\overline{u'u'}$(×) and $\overline{u'v'}$(○) for the riblet structured surface

There is reason to presume that the IBM is not able to provide a sharp representation of the riblet structure with the applied background grid resolution (Fig. 6). Thus, considering the fact that it is well known that the riblet effect strongly depends on small geometric details like the overall riblet shape, height, spacing, and tip angle (Koepplin et al. 2017a) the observed deviation of the mean velocity profile is not surprising. Moreover, the positive shift of u^+ for a given y^+ indicating less frictional losses in the IBM appears physically plausible. Whereas trapezoidal riblets are discretized by the body-fitted grid of Koepplin et al. (2017b) the IBM represents idealized blade-shaped riblets which are known to be more effective (Hage 2005) in drag reduction.

A higher spatial resolution of the IBM background grid would probably resolve the differences to the reference data. However, in the author's experience the required number of cells would be higher or at least in the same range compared to the corresponding body-fitted grid. Thus, for deterministic geometries with sharp edges and corners like riblets, body-fitted grids seem to be the convenient choice regarding computational efficiency.

CONCLUSIONS

An extension of OpenFOAM for investigating the effect of complex surface roughness on turbulent boundary layers by means of the IBM is presented. In order to

allow for simulations of turbulent equilibrium boundary layers by streamwise infinite computational setups of channel flows a massflow controller is developed and implemented. DNS of one smooth wall configuration and two roughness configurations are conducted and verified by previous simulations with body-fitted computational grids. In all cases mean flow quantities as well as Reynolds stresses are predicted in reasonably good accordance with the reference data. Thus, it is concluded that the improved IBM implementation in OpenFOAM can be applied to the investigation of the effect of arbitrary complex roughness on turbulent boundary layers. However, based on the findings it can be concluded that for deterministic sharp-edged surface structures like Riblets body-fitted grids are the convenient choice regarding computational efficiency.

ACKNOWLEDGEMENTS

This work has been conducted within the framework of subproject B3 entitled *Influence of complex surface structures on the aerodynamic loss behaviour of blades* of the Collaborative Research Centre (CRC)871 *Regeneration of Complex Capital Goods* funded by the German Research Foundation (DFG). The authors thank the Leibniz University IT Services (LUIS) for their provision of computational resources. Furthermore, the authors thank Viktor Köpplin and Oleg Schmunk for their valuable input.

REFERENCES

Achary, M., Bornstein, J., and Escudier, M. P. (1986). Turbulent boundary layers on rough surfaces. *Experiments in Fluids*, (4):33–47.

Ashrafian, A., Andersson, H. I., and Manhart, M. (2004). Dns of turbulent flow in a rod-roughened channel. *International Journal of Heat and Fluid Flow*, 25(3):373–383.

Bensow, R. E. and Liefvendahl, M. (2008). Implicit and Explicit Subgrid Modeling in LES Applied to a Marine Propeller.

Bons, J. P. (2010). A Review of Surface˜Roughness Effects in Gas Turbines. *ASME˜Journal of Turbomachinery*, (132).

Bons, J. P., Taylor, R. P., McClain, S. T., and Rivir, R. B. (2001). The Many Faces of Turbine Surface Roughness. *ASME˜Journal of Turbomachinery*, (123):739–748.

Busse, A., Lützner, M., and Sandham, N. D. (2015). Direct numerical simulation of turbulent flow over a rough surface based on a surface scan. *Computers & Fluids*, 116:129–147.

Cui, J., Lin, C.-L., and Patel, V. C. (2003a). Use of Large-Eddy-Simulation to Characterize Roughness Effect of Turbulent Flow Over a Wavy Wall. *Journal of Fluids Engineering*, (125):1075–1077.

Cui, J., Patel, V. C., and Lin, C.-L. (2003b). Large-eddy simulation of turbulent flow in a channel with rib roughness. *International Journal of Heat and Fluid Flow*, (24(3)):372–388.

Denkena, B., Böß, V., Nespor, D., Gilge, P., Hohenstein, S., and Seume, J. (2015). Prediction of the 3D Surface Topography after Ball End Milling and its Influence on Aerodynamics. In *Procedia CIRP*, volume 31, pages 221–227.

Flack, K. A. and Schultz, M. P. (2010). Review of Hydraulic Roughness Scales in the Fully Rough Regime. *Journal of Fluids Engineering*, (132).

Hage, W. (2005). Zur Widerstandsverminderung von dreidimensionalen Riblet-Strukturen und anderen Oberflchen. *Dissertation, Technische Universität Berlin*.

Hohenstein, S. (2014). Einfluss von Oberflächenstrukturen auf das aerodynamische Verlustverhalten von Turbinenbeschaufelungen. *Dissertation, Institute of Turbomachinery and Fluid Dynamics of the Leibniz Universtät Hannover*.

Hohenstein, S. and Seume, J. (2013). Numerical Investigation on the Influence of Anisotropic Surface Roughness on the Skin Friction. *Proceedings of the European Turbomachinery Conference*.

Iacone, G. L., Reynolds, A. M., and Tucker, P. G. (2008). Particle Deposition Onto˜Rough Surfaces. *Journal of Fluids Engineering*, (130).

Ikeda, T. and Durbin, P. A. (2007). Direct simulations of a rough-wall channel flow. *Journal of Fluid Mechanics*, (571):235–263.

Kempe, T. and Fröhlich, J. (2012). Collision modelling for the interface-resolved simulation of spherical particles in viscous fluids. *Journal of Fluid Mechanics*, 709:445–489.

Koepplin, V., Herbst, F., and Seume, J. (2017a). Correlation-based riblet model for turbomachinery applications. *Journal of Turbumachinery*, 139(7):071006.

Koepplin, V., Seume, J., and Herbst, F. (2017b). Investigation of Local Flow Phenomena of Misaligned Riblets in a Turbulent Boundary Layer. In *European Drag Reduction and Flow Control Meeting*, 6 April 2017, Frascati, Rome, Italy.

Mittal, R. and Iaccarino, G. (2005). Immersed boundary methods. *Annu. Rev. Fluid Mech.*, 37:239–261.

Moser, R. D., Kim, J., and Mansour, N. N. (1999). Direct numerical simulation of turbulent channel flow up to re τ= 590. *Physics of fluids*, 11(4):943–945.

Nespor, D., Denkena, B., Grove, T., and Pape, O. (2016). Surface topography after re-contouring of welded ti-6al-4v parts by means of 5-axis ball nose end milling. *The International Journal of Advanced Manufacturing Technology*, 85(5-8):1585–1602.

Nikuradse, J. (1933). Strömungsgesetze in rauhen Rohren. *VDI-Forschungsheft 361*, (361).

Peskin, C. S. (1972). Flow patterns around heart valves: a numerical method. *Journal of computational physics*, 10(2):252–271.

Peskin, C. S. (2002). The immersed boundary method. *Acta numerica*, 11:479–517.

Schlichting, H. (1936). Experimental Investigation of the Problem of Surface Roughness. *Ingenieur-Archiv*, (1(3)).

Singh, K. M., Sandham, N. D., and Williams, J. J. R.

(2007). Numerical Simulation of Flow over a Rough Bed. *Journal of Hydraulic Engineering*, (133):386–398.

Taylor, R. P. (1990). Surface Roughness Measurements on Gas Turbine Blades. *ASME Journal of Turbomachinery*, (112):175–180.

Tuković, Z. and Jasak, H. (2012). A moving mesh finite volume interface tracking method for surface tension dominated interfacial fluid flow. *Computers & fluids*, 55:70–84.

KONRAD HARTUNG received his Master of Science in mechanical engineering in 2016 from the Leibniz Universität Hannover. Since 2017 he is working as scientific assistant in the area of computational fluid dynamics at the Jade University of Applied Sciences in Wilhelmshaven, Germany. He is a contributor in the research project OstrALas funded by the Federal Ministry of Education and Research. His email address is: konrad.hartung@jade-hs.de.

PHILIPP GILGE received his Diploma in mechanical engineering in 2014 from the Leibniz Universität Hannover. Since 2014 he is working as scientific assistant in the area of complex surface structures and their influence on the aerodynamic loss behavior of blades at the Institute of Turbomachinery and Fluid Dynamics of the Leibniz Universität Hannover, Germany. The project is funded by the German Research Foundation in the framework of the Collaborative Research Centre (CRC)871 *Influence of complex surface structures on the aerodynamic loss behaviour of blades*. His email address is: gilge@tfd.uni-hannover.de.

FLORIAN HERBST received his PhD in mechanical engineering from Leibniz Universität Hannover in 2013. He is Head of the Junior Research Group Multiphysics of Turbulent Flows at the Institute of Turbomachinery and Fluid Dynamics. The group focuses on the investigation of the physics of turbulence using scale-resolving numerical methods (DNS, LES) and on the application of this knowledge to the development of physical RANS models for turbomachinery CFD design-codes. His email address is: herbst@tfd.uni-hannover.de.

Numerical Supported Design of Continuously Adapted Riblets for Viscous Drag Reduction on a NREL Wind Turbine Airfoil

Karsten Oehlert
Jan H. Haake
Konrad M. Hartung
Institute of Energy and Process Engineering
Jade University of Applied Sciences
26389 Wilhelmshaven, Germany

KEYWORDS

Riblets, Drag reduction, Transition modeling, Wind turbines, Laser structuring.

ABSTRACT

Viscous drag is a significant factor contributing to the efficiency of aerodynamic application. During the last decades several studies have shown that functional surface structures provide the potential to reduce the viscous drag in the turbulent boundary layer and thus increase efficiency. Due to the rather challenging manufacturing process of micro-scale riblets most studies were conducted using constant riblet dimensions causing losses of the possible drag reduction. With the rapid development in micro structuring technologies it is now possible to manufacture continuously adapted riblets in almost industrial processing scales. For the design of continuously adapted riblets an algorithm is developed considering the effect of a deviation of the riblet spacing from its theoretical optimal value in order to maximize the cumulative drag reduction additionally effected by a misalignment of the riblet structure. The algorithm is applied for the design of riblets with a trapezoid cross-section for the NREL wind turbine airfoil S809 at a Reynolds number of $Re = 2 \cdot 10^6$. The shear stress data required to determine an appropriated region for the riblet application and optimal riblet dimensions are provided by 2D numerical studies using an empirical three equation transition model which shows good consistency of the predicted location of the transition onset with experimental results. With the outlined design approach an increased drag reduction in rage of 25.7% compared to riblets with constant dimensions can be expected.

INTRODUCTION

In most engineering and aerospace applications, such as airfoils, wind turbines or turbomachinery a transitional boundary layer flow occurs. The initially laminar boundary layer controlled by viscous forces becomes turbulent triggered by various types of sufficient flow disturbances. Within the turbulent boundary layer vortices naturally increase the momentum transfer and hence the viscous drag which is considered a major barrier to further optimization of aerodynamic performance. In the last three decades there has been extensive research on drag reduction techniques, mainly focusing on delaying the laminar-turbulent transition and modifying the turbulent structures in the boundary layer. As a passive drag reducing technique inspired by designs found throughout living nature, the dermal denticles of the skin of fast swimming sharks were found to produce low drag. In order to study the drag reducing effect in experimental fluid flow, most important characteristics have been identified and transcribed to simplified riblet geometries, including various types of blade, sawtooth, scalloped, and trapezoid cross-sections arranged periodically in wall flow direction. The extensive investigation on these micro-scale structures since the late 1970's e.g. by (Walsh 1980; Choi 1989; Walsh and Lindemann 1984; Bechert et al. 1997) have shown that riblets provide a drag reduction up to 10%. Furthermore, the experimental results of (Bechert et al. 1997) and (Bruse 1998) have shown that an adaption of the riblet dimensions to the local flow conditions is a basic requirement for an effective drag reduction by riblets. However due to the complex manufacturing process of micro-scale surface structures almost all experimental measurements were conducted using riblets with constant dimensions, such as provided by the 3M Corp. Therefore, the drag reducing potential of riblet structure could yet not be fully exploited. Thanks to the rapid development in micro structuring technology, especially in the field of high-rate laser micro processing (Loeschner et al. 2015; Schille et al. 2017) it is now possible to manufacture continuously adapted riblet in almost industrial processing scales. Aim of this paper is to present a numerical supported design process for continuously adapted riblets for the National Renewable Energy Laboratory (NREL) airfoil S809 for subsequent experimental and numerical investigations.

GENERAL INFORMATION ON RIBLETS

Drag Reducing Mechanism

In the near wall region, stripe-like turbulent structures arise whose middle axis of rotation are oriented in streamwise direction. These coherent vortices in

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

the viscous sublayer were proven both by experimental (Kline et al. 1967; Clark 1990) and numerical results (Kim et al. 1987; Choi et al. 1993; Goldstein et al. 1995) . A flow visualization of the streamwise vortices was carried out by (Lee and Lee 2001) using a synchronized smoke-wire technique over a flat plate and a riblet structured surface with semicircular grooves (Figure 1). (Bechert et al. 1986) as well as (Bechert and Bartenwerfer 1989) assumed that the drag reducing effect of riblets is based on the blocking of the streamwise vortices in the near-wall region, which leads to a decreased momentum exchange in the vicinity of the wall. (Bhushan 2012) summarized the drag reducing effect of riblets as a coupled mechanism of blocking streamwise vortices in the viscous sublayer and an additional elevating effect towards the dominate vortices. The theory was later confirmed by direct numerical simulation of the boundary layer flow over riblet-structured surfaces e.g. by (Choi et al. 1993; Goldstein et al. 1995) as well as by experimental results (Vukoslavevi et al. 1992; Suzuki and Kasagi 1994) . Essential for an effective drag reduction is the correct dimensioning especially of the riblet spacing. In case of streamwise vortices in the near wall region with a middle diameter greater than the local riblet spacing, vortices are impeded and elevated resulting in a decreased momentum exchange in the near wall region. As the vortices only interact with the riblet tips, the overall wetted surface area in contact with fluid of high momentum is decreased leading to a significant viscous drag reduction (Vukoslavevi et al. 1992). For an oversized riblet spacing the streamwise vortices occur within the riblet valley. The overall wetted surface area in contact with fluid of high momentum is thereby increased resulting in a higher drag (Choi et al. 1993).

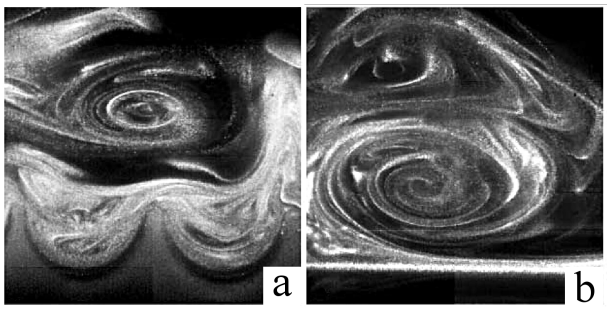


Fig. 1. Flow visualization of streamwise vortices in the turbulent boundary layer (Lee and Lee 2001): a) lateral cross-section of a smooth wall and b) lateral cross-section of riblets with semi-circular grooves

Impact of the Riblet Geometry

The geometrical characterization of riblets can be determined by the riblet spacing s, the riblet height h, the tip width t and the opening angle α. Furthermore, a misalignment angle φ is defined describing the misalignment between the riblet orientation and the freestream direction. For a better comparison of experimental data with various riblet geometries and flow conditions, non dimensional parameters measured in wall units, are introduced. These parameters are denoted by an additional $^+$-symbol as shown in Equation (1). Here, the expression τ_0 relates to the wall shear stress of a smooth reference surface.

$$s^+ = \frac{s}{\nu}\sqrt{\frac{\tau_0}{\rho}}; h^+ = \frac{h}{\nu}\sqrt{\frac{\tau_0}{\rho}}; t^+ = \frac{t}{\nu}\sqrt{\frac{\tau_0}{\rho}} \qquad (1)$$

The investigations of (Walsh 1982) and (Bechert et al. 1997) have shown that certain configurations of the riblet shape, the tip width, the riblet spacing and height maximize the achievable drag reduction. (Walsh 1982) showed in wind tunnel experiments that for riblets with a sawtooth cross-section the maximum drag reduction can be reached for a riblet height-to-spacing-ratio of h/s = 1. Later on (Bechert et al. 1997) discovered that the optimal height-to-spacing-ratio measured by (Walsh 1982) does not possess general validity. The systematic investigations of (Bechert et al. 1997) for various height-to-spacing-ratios showed that blade shaped and trapezoid shaped riblets exhibit a maximum drag reduction for a height-to-spacing-ratio of h/s = 0.5. Contrary riblets with semicircular grooves show a maximum drag reduction for a height-to-spacing-ratio of h/s = 0.7. Further investigations on the impact of the riblet shape and the riblet spacing on the drag reduction were carried out by (Bruse 1998). The results of the studies were later summarized by (Hage 2005) (Fig. 2). For riblets with a sawtooth cross-section, an opening angle of $\alpha = 60°$ and a height-to-spacing-ratio of approximately h/s $\approx$ 0.9 a maximum viscous drag reduction in range of $\triangle\tau/\tau_0 = -5,5\%$ was reached for a non dimensional riblet spacing of s^+ = 16. The overall highest reduction of viscous drag was measured for blade-shaped riblets with a height-to-spacing-ratio of h/s = 0.5 and a tip width of t = 0.02s. With this kind of cross section, a maximum drag reduction of $\triangle\tau/\tau_0 = -9,9\%$ was achieved for a non dimensional riblet spacing of s^+ =17. For riblets with a trapezoid cross-section a maximum drag reduction of $\triangle\tau/\tau_0 = -8,2\%$ was achieved with a height-to-spacing-ratio of h/s = 0.5 and a non dimensionlal riblet spacing of s^+ = 17. Due to the higher mechanical durability of riblets with trapezoid cross-section these should be preferred for practical applications. Furthermore the drag reducing effect of riblets shows a significant sensitivity towards the orientation relative to the flow direction as exemplary illustrated in Figure 3 for riblets with a trapezoid cross-section. The drag reduction reaches a maximum when the riblets are aligned in streamwise direction. In case of misalignment, the drag reduction decreases. Exceeding a certain misalignment angle leads to an increased surface friction compared to a smooth wall.

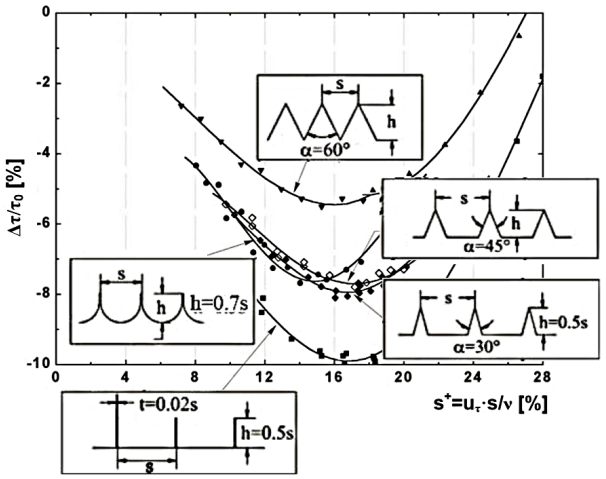


Fig. 2. Impact of the non dimensional riblet spacing on the relative change of the wall shear stress for various riblet shapes (Hage 2005)

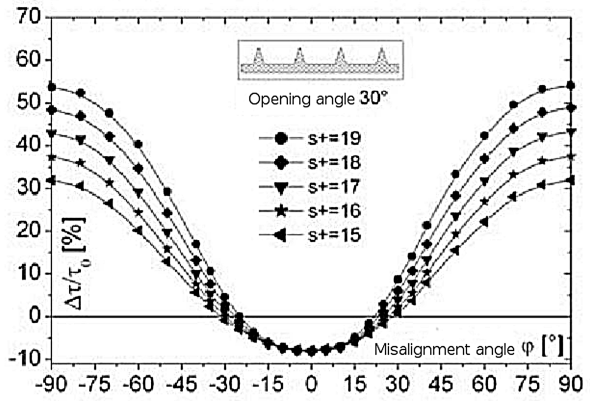


Fig. 3. Impact of the misalignment angle on the relative change of the wall shear stress for riblets with trapezoid cross-section (Hage 2005)

NUMERICAL SUPPORTED DESIGN OF CONTINOUSLY ADPADED RIBLETS

Riblets only lead to a drag reduction within the turbulent boundary layer while there height is within the viscous sublayer. The application of riblets within a laminar boundary layer would have a comparable impact as the application of a surface roughness increasing the overall viscous drag (Indinger 1999). Therefore, a subdivision of laminar, transitional and turbulent boundary layers is necessary to determine an appropriate area for the riblet application.

Transition Modeling

Presently turbulence and transitional phenomenona have still not been fully physically understood. In fact, most knowledge on turbulent flow is empirical. Therefore, a general model able to predict the transitional effects does not exist. To avoid the computational costs associated with solving the full transient Navier-Stokes equations, a number of empirical correlation based turbulence models, such as the $kk_L-\omega$ model and the $\gamma - Re_\Theta$ model have been developed showing promising results for the flat plate test cases as well as for flows over airfoils (Fürst, J. et al. 2013; Keith Walters and Cokljat 2008). The $kk_L-\omega$ turbulence model initially proposed by (Keith Walters and H. Leylek 2004) is based on $k-\omega$ framework which solves the two additional transport equations for the turbulent kinetic energy and the specific dissipation rate ω. Furthermore the $kk_L-\omega$ model solves a third transport equation for the laminar kinetic energy k_L to predict the magnitude of low frequency velocity fluctuations that have been identified triggering transition in the boundary layer (Keith Walters and Cokljat 2008). The model is based on the assumption that velocity fluctuations in the pre-transitional region can be divided into small-scale vortices contributing to turbulence production and large-scale longitudinal vortices contributing to the production of non-turbulent fluctuation. The transport equations for the turbulent kinetic energy, the laminar kinetic energy, and the specific dissipation rate are given by the Equations (2-4) and contain terms representing production, destruction, and transport mechanism. The production is expressed by the first term on the right hand side modeled by a small-scale eddy viscosity concept. The termsand appearing with opposite signs in the transport equations for laminar and turbulent kinetic energy are related to the transition mechanisms for bypass and natural transition, respectively. The terms and represent the non-isotropic part of the near wall dissipation of the turbulent and laminar kinetic energy. In the transport equation for the specific dissipation rate, the second term on the right hand side is the transition production term intended to reduce turbulent length scales during the transition process (Keith Walters and H. Leylek 2004). The fourth term on the right hand side decreases the length scale in the outer region of the turbulent boundary layer necessary to ensure an adequate prediction of the wake region (Keith Walters and H. Leylek 2003). In summary, the transition process is modelled by the effect of energy transfer from the laminar kinetic energy of large-scale longitudinal vortices to the turbulent kinetic energy of small-scale vortices with a concurrent reduction in turbulence length scale. A detailed overview of the model and the model constants is given by (Keith Walters and Cokljat 2008).

$$\frac{\mathrm{Dk_T}}{\mathrm{DT}} = P_{kT} + R_{BP} + R_{NAT} - D_T + \frac{\partial}{\partial x_j}\left[\left(v + \frac{\alpha_T}{\alpha_K}\right)\frac{\partial k_T}{\partial x_j}\right] - \omega k_T \qquad (2)$$

$$\frac{\mathrm{Dk_L}}{\mathrm{Dt}} = P_{kL} - R_{BP} - R_{NAT} - D_L + \frac{\partial}{\partial x_j}\left[v\frac{\partial k_L}{\partial x_j}\right] \qquad (3)$$

$$\frac{\mathrm{D}\omega}{\mathrm{Dt}} = C_{\omega 1}\frac{\omega}{K_T}P_{kT} + \left(\frac{C_{\omega R}}{f_w} - 1\right)\frac{\omega}{k_T}(R_{BP} + R_{NAT}) - C_{\omega 2}\omega^2 + C_{\omega 2}f_w\alpha_T f_w^2\frac{\sqrt{k_T}}{d^2} - \frac{\partial}{\partial x_j}\left[\left(v + \frac{\alpha_T}{\sigma_\omega}\right)\frac{\partial \omega}{\partial x_j}\right] \quad (4)$$

Numerical Setup

In this study the numerical supported design of continuously adapted riblets is conducted for the NREL wind turbine airfoil S809 designed for horizontal-axis wind turbines (HAWT). The coordinates of the airfoil were collected from NREL website and transferred into a 2D model by spline interpolation. The computational domain for the simulations consist of a semicircular section in the front with a radius of 10 chord lengths. The downstream distance of the computational domain from the airfoils leading edge is set to 21 chord lengths as illustrated in Figure 4. The semicircular as well as the bottom boundary were defined as velocity inlet. The top and downstream boundary were defined as pressure outlet. The number of cells and the near wall resolution of the C-type structured mesh were chosen such that the requirement of $y^+ < 1$ for the wall resolved turbulence modeling at a Reynolds number of $Re = 10^6$ and a velocity magnitude of 48m/s is satisfied. The inlet turbulence intensity and the turbulent viscosity ratio were set to 0.2% and 10, respectively. Simulations were conducted for an angle of attack ranging from 2° to 10° in intervals of 2° by adjusting the velocity components at the inlet. Steady state incompressible flow simulations were performed using the $kk_L - \omega$ turbulence model in the software environment Ansys.

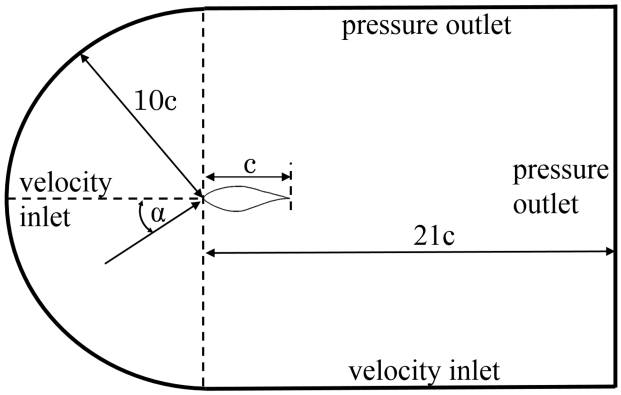


Fig. 4. Computational domain and boundary conditions for the 2D flow over an airfoil

2D Results

Based on the simulation setup discussed in the previous section, the numerical results were compared to the experimental data from (Somers 1997) in terms of the location of transition onset versus the angle of attack. According to the definitions of the $kk_L - \omega$ turbulence models the transition onset location was determined by an initially increasing turbulent kinetic energy in the near wall region. The numerical results show a good agreement of the predicted location and displacement for the transition onset with the experimental data (Figure 5 - 6). On the suction side, the location of transition onset steadily moves towards the leading edge of the airfoil with an accelerated displacement within the range between 5° and 9°. From an angle of approximately 8° on the location of transition onset remains apparently at the vicinity of the leading edge offering high potential for the drag reduction by riblets. On the pressure side the displacement for the location of transition onset shows an opposite behavior. With an increasing

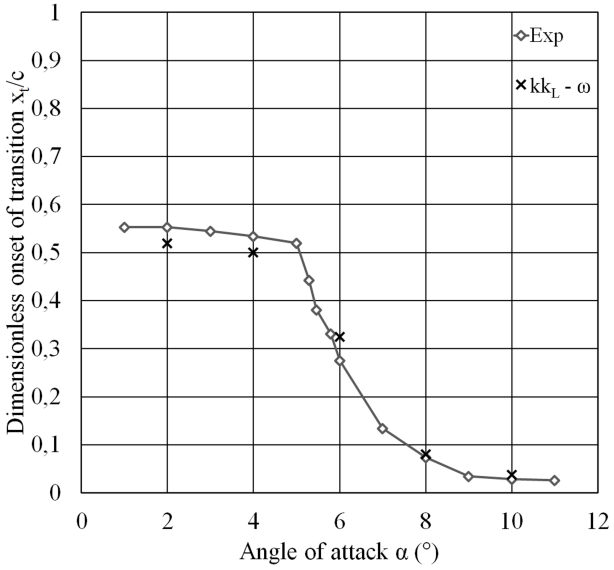


Fig. 5. Predicted and measured location of the transition onset on the suction side of the airfoil S809

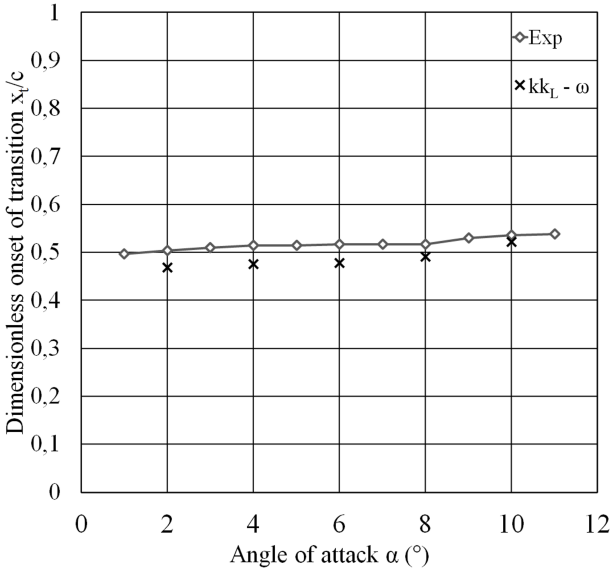


Fig. 6. Predicted and measured location of the transition onset on the pressure side of the airfoil S809

angle of attack the location of the transition onset moves towards the trailing edge of the airfoil showing an overall significant lower sensitivity compared to the suction side. Note that in literature, the location of transition onset is used for comparison of experimental and numerical result. For the selection of an appropriate area for the riblet application, however, information on the onset of the fully developed turbulent boundary layer is necessary. Therefore, the distribution of the wall shear stress as a function of the non dimensional airfoil length is computed and illustrated in Figure 7 together with the resulting classifications of laminar, transitional and turbulent boundary layers exemplary for an angle of attack of 8°.

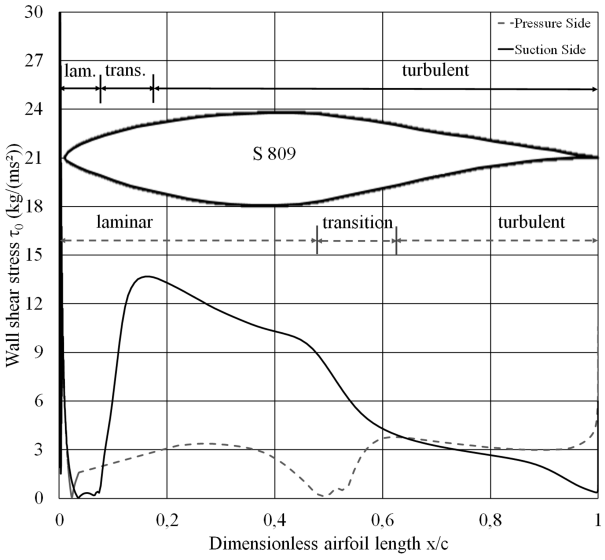


Fig. 7. Computed distribution of the wall shear stress on the smooth airfoil S809 at an angle of attack of 8°

Design of Continously Adapted Riblets

In accordance with the experimental results of (Lietmeyer et al. 2013), which show higher drag reducing potential of riblets on the suction side of an airfoil the design process of continuously adapted riblets is outlined for the suction side of the airfoil S809 at an angle of attack of 8°. Concerning the subsequent manufacturing process, the region between a non dimensional airfoil lengths of 0.2 and 0.95 seems appropriated of the riblet application. For the design process, a trapezoid riblet shape with an opening angle of $\alpha = 30°$, a non dimensional riblet spacing of $s^+ = 17$, and a riblet height that is determined by the local riblet spacing using $h = s/2$ is chosen. The theoretical optimal riblet spacing within the turbulent boundary layer can be determined by rewriting Equation (1) as

$$s = s^+ v \sqrt{\frac{\rho}{\tau_0}} \tag{5}$$

The resulting theoretical optimal riblet spacing based on the computed wall shear stress of the smooth airfoil is illustrated in Figure 8.

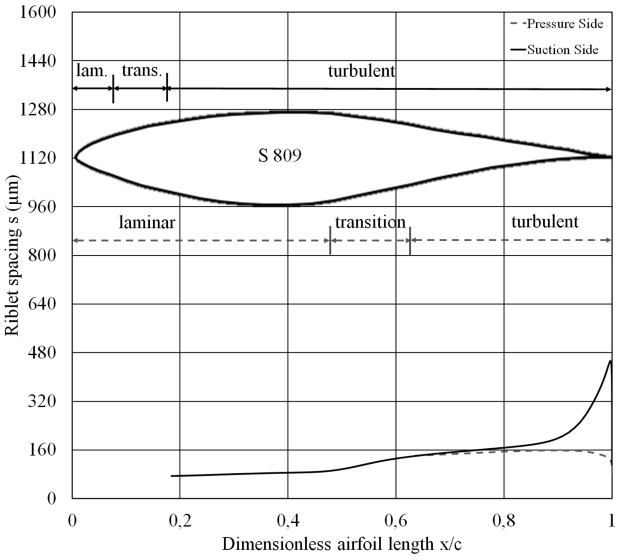


Fig. 8. Computed riblet spacing in the turbulent boundary layer for the smooth airfoil S809 at an angle of attack of 8°

The local varying riblet spacing leads inevitably to a local misalignment of the riblet orientation relative to the flow direction reducing the possible drag reduction as (Hage 2005) summarized (Figure 3). Therefore, a local deviation from the theoretical optimal non dimensional riblet spacing $s^+ = 17$ has to be considered in order to maximize the cumulative drag reduction. For this purpose, an algorithm was designed based on the experimental results of (Bechert et al. 1997) and (Bruse 1998) (cf. Fig. 2 - 3), which utilizes a polynomial approximation for the impact of the non dimensional riblet spacing and the misalignment angle on the drag reduction. To reduce the number of variables the misalignment angle is expressed in term of the local riblet spacing according to

$$\varphi = \arctan\left(\frac{s_{ij}(s^+) - s_{ij-1}}{x_{ij} - x_{i-1j}}\right), \tag{6}$$

whereby the index i is representing the streamwise and j the spanwise direction. By locally varying the non dimensional riblet spacing in Equation (5-6) and computing the corresponding drag reduction by the polynomial approximation the local misalignment angle and riblet spacing which minimize the cumulative drag can be identified. Nevertheless, the possible drag reduction of a riblet-structured surface with a continuously widening cross-section decreases in spanwise direction due to the increasing misalignment. Therefore, a abort criteria had to be considered, which terminates the structuring process in spanwise direction. The abort criteria was chosen in such a manner that an adding of further riblet structures in spanwise direction would lead to

a decreased drag reduction compared to the resulting drag when mirroring the structure in spanwise direction. To avoid additional stagnation points in the mirror plane an offset of the mirrored structures was provided. Figure 9 illustrates a segment of the continuously adapted riblet structure computed by the developed algorithm. For visual simplification a projection of the riblet structure on the curved surface of the airfoil was avoided. With this design approach a drag reduction of -8.02% would be possible. The application of riblets with constant dimensions computed based on the mean wall shear stress would result in a drag reduction of -6.38%. Therefore, through the application of continuously adapted riblets the drag reduction can be increased by 25.7% compared to constant riblet dimensions.

CONCLUSIONS

In this study a numerical supported design process for continuously adapted riblets is introduced. It is demonstrated that the correlation based transition $kk_L - \omega$ model can be successfully used to predict the location of the transition onset for the NREL S809 airfoil at various angle of attacks. However, it is not yet sufficiently investigated, if the $kk_L - \omega$ model provides accurate results for the wall shear stress distribution in the turbulent boundary layer, which is essential for the design process. Furthermore it is shown, that continuously adapted riblets require a deviation from the theoretical optimal riblet spacing in order to maximize the cumulative drag reduction. For the riblet structure generated by the developed algorithm an increased drag reduction of 25.7% in comparison to constant riblet dimensions can be expected. To prove evidence for the superior efficiency of the designed riblet structures the authors are currently exploring the drag reduction by numerical studies and experimental measurements.

ACKNOWLEDGEMENTS

This work has been conducted within the cooperative research project entitled *Fluidmechanical Optimization of Turbomachinery by High-rate Laser Structuring Technologies (OstrALas)* funded and supported by the German Federal Minister of Education and Research.

REFERENCES

Bechert, D. W. and Bartenwerfer, M. (1989). The viscous flow on surfaces with longitudinal ribs. *Journal of Fluid Mechanics*, 206:105129.

Bechert, D. W., Bartenwerfer, M., Hoppe, G., and Reif, W.-E. (1986). Drag reduction mechanisms derived from shark skin. 2:1044–1068.

Bechert, D. W., Bruse, M., Hage, W., van der Hoeven, J. G. T., and Hoppe, G. (1997). Experiments on drag-reducing surfaces and their optimization with an adjustable geometry. *Journal of Fluid Mechanics*, 338:5987.

Bhushan, B. (2012). *Biomimetics Bioinspired Hierarchical Structured Surfaces for Green Science and Technology*. Springer Verlag.

Bruse, M. (1998). *Zur Strömungsmechanik wandreibungsvermindernder riblet-Oberflächen*. Fortschritt-Berichte VDI. Reihe 7, Strömungstechnik. VDI-Verlag.

Choi, H., Moin, P., and Kim, J. (1993). Direct numerical simulation of turbulent flow over riblets. *Journal of Fluid Mechanics*, 255:503539.

Choi, K.-S. (1989). Near-wall structure of a turbulent boundary layer with riblets. *Journal of Fluid Mechanics*, 208:417458.

Clark, D. G. (1990). Boundary layer flow visualisation patterns on a riblet surface. In Coustols, E., editor, *Turbulence Control by Passive Means*, pages 79–96, Dordrecht. Springer Netherlands.

Fürst, J., Straka, P., Phoda, J., and imurda, D. (2013). Comparison of several models of the laminar/turbulent transition. *EPJ Web of Conferences*, 45:01032.

Goldstein, D., Handler, R., and Sirovich, L. (1995). Direct numerical simulation of turbulent flow over a modeled riblet covered surface. *Journal of Fluid Mechanics*, 302:333376.

Hage, W. (2005). Zur widerstandsverminderung von dreidimensionalen riblet-strukturen und anderen oberflchen. (dissertation).

Indinger, T. (1999). Einfluss von riblets auf die natrliche transition von grenzschichten. Technical report, Technical University of Dresden.

Keith Walters, D. and Cokljat, D. (2008). A three-equation eddy-viscosity model for reynolds-averaged navierstokes simulations of transitional flow. *Journal of Fluids Engineering*, 130:121401 121401 14.

Keith Walters, D. and H. Leylek, J. (2003). A cfd study of wake-induced transition on a compressor-like flat plate. 6.

Keith Walters, D. and H. Leylek, J. (2004). A new model for boundary layer transition using a single-point rans approach. *Journal of Turbomachinery*, 126:193 202.

Kim, J., Moin, P., and Moser, R. (1987). Turbulence statistics in fully developed channel flow at low reynolds number. *Journal of Fluid Mechanics*, 177:133166.

Kline, S. J., Reynolds, W. C., Schraub, F. A., and Runstadler, P. W. (1967). The structure of turbulent boundary layers. *Journal of Fluid Mechanics*, 30(4):741773.

Lee, S.-J. and Lee, S.-H. (2001). Flow field analysis of a turbulent boundary layer over a riblet surface. *Experiments in Fluids*, 30(2):153–166.

Lietmeyer, C., Oehlert, K., and Seume, J. R. (2013). Optimal application of riblets on compressor blades and their contamination behavior. *Journal of Turbomachinery*, 135:011036 011036 10.

Loeschner, U., Schille, J., Streek, A., Knebel, T.,

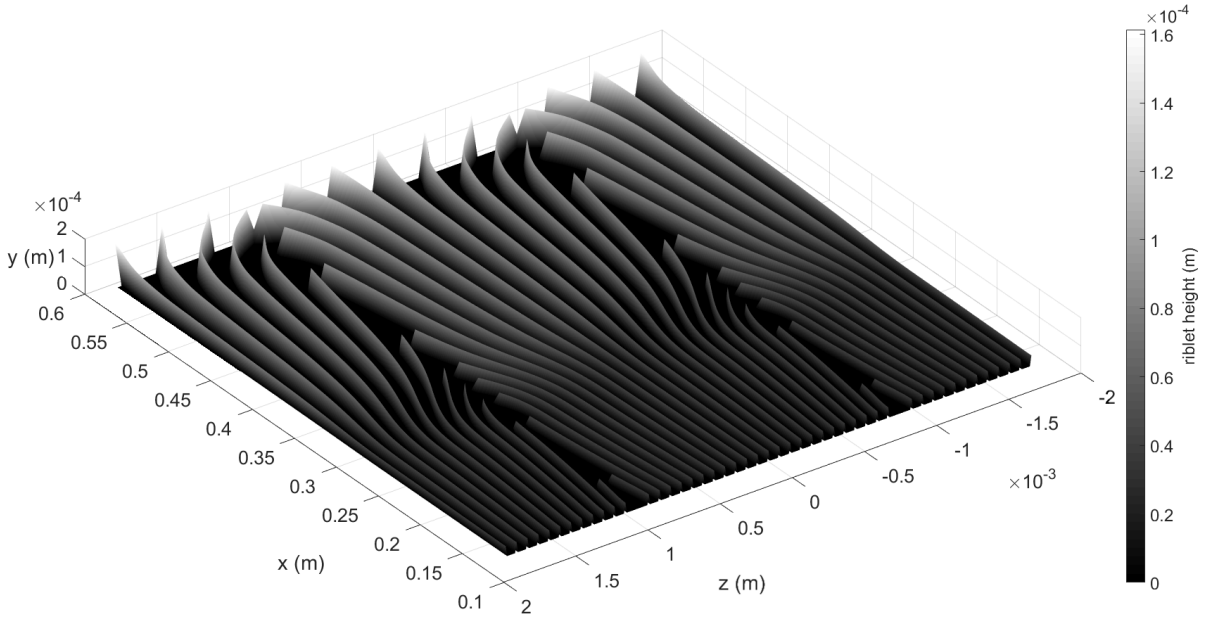


Fig. 9. Segment of the computed continuously adapted riblet structure for the suction side of the S809 airfoil

Hartwig, L., Hillmann, R., and Endisch, C. (2015). High-rate laser micro processing using a polygon scanner system. *Journal of Laser Application*, 27:29303 1 7.

Schille, J., Schneider, L., Hartwig, L., and Loeschner, U. (2017). High rate laser processing of metals using high-average power ultrashort pulse laser. *Journal of Laser Application*, 36:31 50.

Somers, D. (1997). Design and experimental results for the s809 airfoil. (WE711110).

Suzuki, Y. and Kasagi, N. (1994). Turbulent drag reduction mechanism above a riblet surface. 32:1781–1790.

Vukoslavevi, P., Wallace, J., and Balint, J.-L. (1992). Viscous drag reduction using streamwise-aligned riblets. 30:1119–1122.

Walsh, M. (1980). Drag characteristics of v-groove and transverse curvature riblets. 72.

Walsh, M. (1982). Turbulent boundary layer drag reduction using riblets. 6:82–169.

Walsh, M. and Lindemann, M. (1984). Optimization and application of riblets for turbulent drag reduction.

AUTHOR BIOGRAPHIES

KARSTEN OEHLERT graduated at the Leibniz University of Hannover, Germany, in 2005. He received his Dr.-Ing. degree in 2011 for the investigation of drag reducing riblet structures on compressor blades at the Institute of Turbomachinery and Fluid Dynamics. From 2009 to 2013 he worked for the E.ON AG in the *Department of System Technology and Security Analysis*. In the years 2013-2015 he was employed at the Volkswagen Group in the department of R&D in Wolfsburg. Presently he is Professor for Turbomachinery at the Institute of Energy and Process Engineering at the Jade University of Applied Science in Wilhelmshaven, Germany and in this function in the leadership of the cooperative research project *OstrALas* funded by the Federal Ministry of Education and Research. His research interests are mainly in the fields of turbomachinery, fluidmechanics and boundary layer subjects.
His email address is: karsten.oehlert@jade-hs.de.

JAN H. HAAKE received his bachelor in mechanical engineering in 2017. Presently he is studying for a master degree in mechanical engineering.
His email address is: Jan.Haake@student.jade-hs.de.

KONRAD M. HARTUNG received his Master of Science in mechanical engineering in 2016 from the Leibniz University of Hannover. Since 2017 he is working as scientific assistant in the area of computational fluid dynamics at the Jade University of Applied Science in Wilhelmshaven, Germany. He is part of the research project OstrALas funded by the Federal Ministry of Education and Research.
His email address is: konrad.hartung@jade-hs.de.

OPTIMIZATION OF THE PLANT CONTROL SYSTEMS AT WILHELMSHAVEN POWER PLANT BASED ON COAL MILL MODELS AND STATE CONTROLLERS

Dr. Nicolas Mertens*
Uniper Technologies GmbH
Alexander-von-Humboldt-Straße 1
45897 Gelsenkirchen
Nicolas.mertens@uniper.energy
+49 175 5985823

Uwe Krüger
Uniper Kraftwerke GmbH
Zum Kraftwerk
26386 Wilhelmshaven
Uwe.Krueger@uniper.energy

Prof. Dr. Henning Zindler
Uniper Technologies GmbH
Alexander-von-Humboldt-Straße 1
45897 Gelsenkirchen
Henning.Zindler@uniper.energy

Marc-Hendrik Prabucki
Uniper Technologies GmbH
Alexander-von-Humboldt-Straße 1
45897 Gelsenkirchen
Marc-Hendrik.Prabucki@uniper.energy

*Corresponding author

KEYWORDS

Dynamic simulation
Hard-coal fired power plants
Promecon coal dust measurement
Optimization of secondary response capability
Apros simulation software
Kalman filter
State controllers
Observer

ABSTRACT

In the light of increasing renewable feed-in and challenging market conditions, flexibility requirements for hard-coal fired power plants become more important. The focus is on the capability for primary and secondary control response. In this work, a procedure has been developed using Kalman filters to smoothen stochastic oscillations of a Promecon coal dust mass flow measurement while capturing the dynamic behavior during steep load ramps. The filtered signal can be used to improve the performance of the power correction controller in the power plant.

A state controller with observer is proposed for this purpose, where the observer is corrected by the power output measurement and the coal dust mass flow measurement. Test results of the implementation in the control system at Wilhelmshaven power plant show promising behavior of the enhanced controller.

INTRODUCTION

The intermittency in the German electrical grid due to renewable power generation is mainly compensated by hard coal fired power plants. This leads to increasing requirements for primary and secondary control reserve for these power plants (Hentschel 2017). Current control structures are typically based on standard PI controllers, simplified reference models and open-loop control for step load changes. These control concepts typically use constant parameters and cannot adapt automatically to new boundary conditions like new coal types or wear of the coal mill rollers.

Modern state controllers with observers can partly mitigate the time variance of the power plant and reduce the need for optimization of controller parameters, which is cost- and time-intensive. In coal fired power plants, the controlled section from inlet signal at the coal feeder to the output signal at the generator is long. The quality of the observer is therefore reduced. The additional measurement of a state variable in the middle of the controlled section would help to improve the quality of the observer.

Additional measurement of the coal dust mass flow downstream of the coal mill is a promising solution. However, this measurement is not easy and yields lot of stochastic oscillations. More precise coal mill models and the use of Kalman filters can smoothen the measurement signal without influencing the time response of the measurement. This article describes how Kalman filters are successfully applied to

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

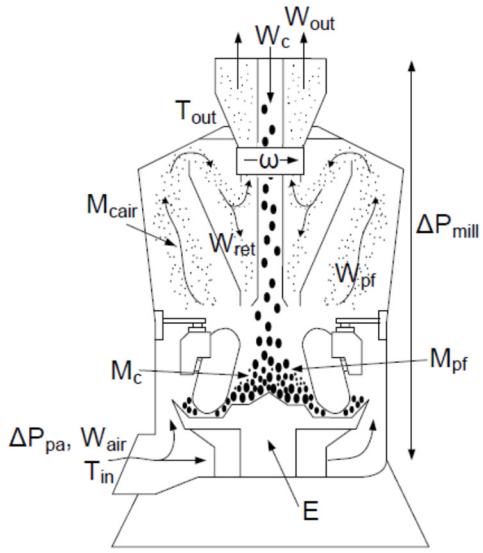


Figure 1: Hard coal mill and its internal coal storages (Niemcyk 2012)

optimize coal dust measurement signals based on Promecon measurement systems. In addition, new control concepts are discussed and the results of the practical implementation at Wilhelmshaven power plant are presented.

COAL MILL MODELS

Figure 1 shows the processing schematic of raw coal and coal dust in a coal mill. The raw coal falls in the middle of the mill plate. The coal is transported towards the outer edge of the rotating plate. In the middle of the plate, coal is grinded by the rollers. The primary air transports all coal upwards that falls over the edge of the plate, where heavy particles are thrown back to the plate. Smaller particles are directed upwards to the classifier. Very small particles are transported to the burners. All other particles are separated in the classifier and thrown back to the center of the plate. The separation grain size in the classifier is a function of the primary air volume flow.

The development of a new coal mill model is based on the complete dynamic power plant simulator of a commercial power plant and the comparison with its process data measurement. The basic idea is that the furnace, heat transfer and fluid flow can be modelled precisely since the plant geometry and process data measurements are available. E.g. heat transfer parameters can be optimized to meet the temperature measurements. The remaining variable to account for the dynamic behaviour of the power plant is the coal mill.

Figure 2: Optimization of the coal mill model

Figure 2 shows the principle of the optimization process. The simulation model has been developed in Apros simulation software, please compare also (Zindler 2016). Measured feeder inlet values are used as transient boundary condition of the model. The model response e.g. generator output is compared to the real measured generator output. The discrepancy is used to optimize the parameters of the mill model in an iterative procedure. Introducing a coal dust measurement downstream of the mill allowed to improve the optimization process. The mill model itself is a grey box model based on transient mass balances. However, the grain size distribution cannot be calculated so that the distributions in the mill are approximated by simplified relations. Details of the process are described in (Hentschel 2016).

FILTERING OF THE COAL DUST MEASUREMENT

Uniper has installed a Promecon coal dust mass flow measurement in Wilhelmshaven power plant to optimize the capacity for secondary control response. Figures 3 and 4 show the measurements of feeder speed and coal dust mass flow during a “double-hump” test. The coal dust measurement signal is clearly oscillating in a stochastic manner. The signal is therefore not suitable for control optimization in its original quality. The reason for the oscillation could not be identified.

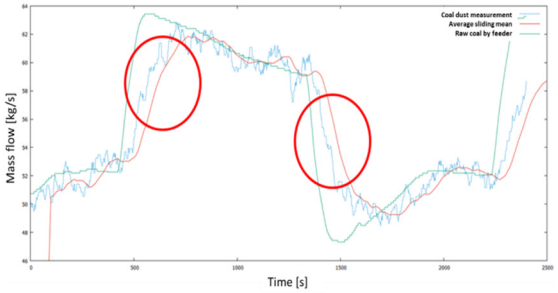


Figure 3: Measurements of a double-hump test and optimization of the oscillating coal dust mass flow measurement signal (blue) based on sliding average mean value (red) and first-order filter (green)

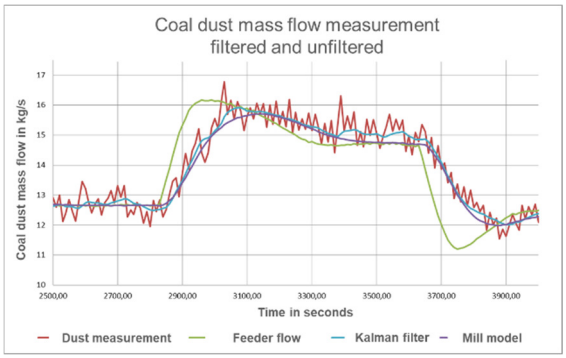


Figure 4: Test of the Kalman filter in Apros simulation software using real measurements

In order to make the coal dust measurement signal available to the control system, different filters have been tested. Figure 3 shows the filtered results based on average sliding mean value and on a first-order filter. In both cases the signal is smoothened, but the dynamic behavior is not accurately captured (particularly ramps). This signal is not suitable for the optimization of large load changes. A filter is required, which does not alter the dynamic behavior of the measurement signal. Uniper has developed a Kalman filter for this purpose.

KALMAN FILTER

The basic principle of a Kalman filter can be explained in two steps. Firstly, the coal dust mass flow is predicted or calculated based on the feeder speed, other boundary conditions and a mathematical model of the mill. The second step is a continuous correction of the model based on the measured coal dust mass flow. The signal influence can be weighted by parameters (Bruns 2016). For in-depth explanation of the Kalman filter, the interested reader is referred to the pertinent literature such as (Föllinger 2013, Lunze 2008).

Figure 4 shows the results of the Kalman filter based on the mill model. The feeder speed and the primary air flow are input signals of the model. The model is corrected continuously by the Promecon coal dust measurement. The signal is well filtered while accurately capturing the dynamic behavior of the measurement signal during a step load change. In this form, the signal is suitable as basis for control optimization.

Figure 5 shows the results of the Kalman filter implementation in the Siemens T3000 control system of Wilhelmshaven power plant, where the green curve is the Promecon coal dust mass flow measurement and the red curve is the filtered signal.

Figure 5: Implementation of the Kalman filter in Wilhelmshaven plant control system

The calculation precision of the control system is sufficiently high to implement the Kalman filter directly into T3000.

USE OF THE FILTERED SIGNAL FOR OPTIMIZED STATE CONTROLLER

The coal dust measurement can now be used for the optimization of the power correction controller. The controller is normally implemented as correction controller to the power set point. The corrected power set point is required in block control to calculate set points for further parameters like fuel mass flow, feed water mass flow or combustion air mass flow.

Figure 6 shows the simplified schematic of a power correction controller for a limited load range (Load-dependency of the parameters is omitted). In the upper section, the real power plant process is shown (orange colour), which can be divided into two parts:

1. Feeder and coal mill

2. Furnace, steam generator, turbine and generator

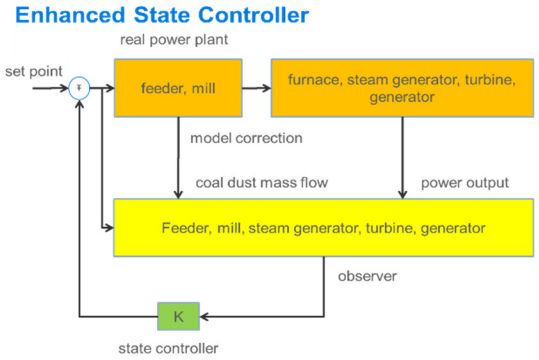


Figure 6: Use of the coal dust measurement to optimize secondary control quality

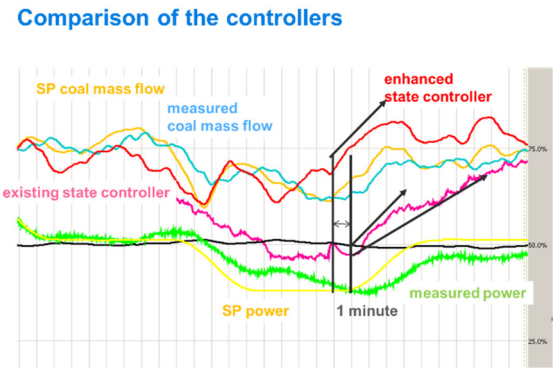


Figure 7: Comparison of a PI controller and the new state controller (Kühling 2017)

In the lower section, the observer is shown (yellow colour). The first part of the observer model describes the feeder and the coal mill, which is corrected by the coal dust measurement. The second part of the observer describes the furnace, the steam generator, the turbine and the generator, which is corrected by the power measurement. The state controller design is based on the merged observer. The main advantage of this approach is the reduced time delay before a discrepancy is identified by the correction controller, particularly in the coal mill (e.g. due to new coal types or mill roller wear).

Figure 7 shows a comparison of the existing state controller of the power plant and the enhanced state controller. Both controllers are implemented in the T3000 Siemens control system, where the existing state controller is in operation. The enhanced state controller is running in parallel. The results show that both state controllers have a qualitatively similar behaviour in operation. In the marked time period, the coal mass flow set point is increased by pilot control for a positive load change. The deviation between this set point and the coal mass flow measurement is readily considered by the enhanced state controller to increase the feeder speed, whereas the existing state controller is limited to the deviation between power set point and measured generator power. The control response time is thus reduced by approximately one minute.

By comparison, for secondary control response the target load must be achieved five minutes after the initial jump of the load set point, with limited overshoot and undershoot not permitted. The enhanced controller is thus a significant contribution to upgrade hard coal-fired power plants for secondary control response and to counteract grid intermittency due to renewable feed-in.

SUMMARY

In this work, a procedure has been developed using Kalman filters to smoothen stochastic oscillations of a Promecon coal dust mass flow measurement while capturing the dynamic behavior during steep load ramps. The filtered signal can be used to optimize the power correction controller of the power plant. The coal dust measurement is suitable to quantify the time variance of a coal mill due to different coal types or mill roller wear.

A state controller with observer is proposed where the observer is corrected by the power output measurement and the coal dust mass flow measurement. Real tests at Wilhelmshaven power plant have shown that the enhanced state controller reacts approximately one minute faster in case of an issue in the coal mill, resulting in increased capability and reliability of plant secondary control response.

LITERATURE

Bruns J. 2016. „Optimierung einer Kohlenstaubmengenmessung mit Hilfe von Kalman - Filtern zur Optimierung der Sekundärregelfähigkeit eines Kohlekraftwerks“, Master Thesis, Hochschule Bremen.

Föllinger O. 2013. „Regelungstechnik“, VDE Verlag, Berlin.

Hentschel J., H. Zindler, C. Wieland and H. Spliethoff. 2016. „Angepasste Sekundärregelcharakteristik durch dynamische Kraftwerkssimulation unter Berücksichtigung des Zeitverhaltens von Steinkohlemühlen“. In: *48. Kraftwerkstechnisches Kolloquium*, Dresden.

Hentschel J., H. Zindler and H. Spliethoff. 2017. „Modelling and transient simulation of a supercritical coal-fired power plant: Dynamic response to extended secondary control power output“. In *Energy*, Volume 137, 927-940.

Kühling J. 2017. „Optimierung eines Zustandsreglers zur Sekundärregelung konventioneller Kraftwerke und Implementierung in einem dynamischen Kraftwerkssimulator“, Bachelor Thesis, Jade-Hochschule Wilhelmshaven

Lunze J. 2008. „Regelungstechnik 2: Mehrgrößensysteme, Digitale Regelung“, Springer-Verlag, Berlin.

Niemczyk P., J.D. Bendtsen, A.P. Ravn, P. Andersen and T.S. Pedersen. 2012. „Derivation and validation of a coal mill model for control”, In *Control Engineering Practice*, Volume 20, Issue 5, 519-530.

Zindler H., U. Krüger, M. Rech and S. Tuuri. 2015. „Dynamischer Kraftwerkssimulator zur leittechnischen Optimierung der Sekundärantwort des E.ON Kraftwerks Wilhelmshaven“ In: *47. Kraftwerkstechnisches Kolloquium,* Dresden.

AUTHOR BIOGRAPHIES

NICOLAS MERTENS holds a diploma in mechanical engineering from RWTH Aachen University and a PhD on transient power plant simulation, which he obtained at TU Darmstadt in 2017. He is now working as process engineer at Uniper Technologies and is the author of several scientific articles in the area of dynamic simulation and power plant flexibility.

HENNING ZINDLER studied process engineering at TU Clausthal and holds a PhD from TU Braunschweig on transient thermal power plant simulation. He worked at E.ON/Uniper as process engineer for ten years until 2017 and is now a professor for energy technologies at the Ostfalia University for Applied Sciences in Wolfenbüttel.

UWE KRÜGER studied mechanical engineering at TH Zittau and started his career at Preussen Elektra in 1990. He was head of the permit department at Brokdorf nuclear power station and is currently head of production for the power plant group Wilhelmshaven.

MARC-HENDRIK PRABUCKI studied marine engineering at the University of Bremerhaven. After several early career positions, he worked nine years as commissioning manager for fossil power plants at Deutsche Babcock Werke and another nine years as lead engineer at Uhde GmbH in various chemical plant projects worldwide. Since 2007, he uses his extensive experience with E.ON/Uniper Technologies and he is currently deputy head of the Plant Engineering department.

Multiphysical Finite Element Simulation

MODELING AND SIMULATION OF BIOHEAT POWERED SUBCUTANEOUS THERMOELECTRIC GENERATOR

Ujjwal Verma, Jakob Bernhardt, Dennis Hohlfeld
Institute of Electronic Appliances and Circuits
University of Rostock, Rostock, Germany
Email: dennis.hohlfeld@uni-rostock.de

KEYWORDS
Thermoelectric generator, Seebeck effect, Bioheat, Medical implants.

ABSTRACT

Electrically active implants are gaining interest for an aging European population. The current generation of implants are powered by batteries that have limited lifetime; once depleted they require surgical reinterventions for their replacement. In this paper, we present a multi-physical model of a thermoelectric generator that utilizes the subcutaneous temperature gradient. The gained electrical power can be used to supply an electrically active implant. Furthermore, this paper studies various parameters that influence the temperature gradient. We implemented a simple human tissue model and a more detailed geometry model based on segmented magnetic resonance imaging (MRI) data.

INTRODUCTION

By 2060 every third person in Europe is expected to be more than 65 years old, the subsequent socio-economic impacts lead to an increase in associated medical treatments. Implantable medical devices, more specific electrically active implants, have found success in clinical trials. These are gaining interest especially for treatments like bone tissue regeneration and treating motion disorders using deep brain simulation (Watkins, Shen and Venkatasubramanian, 2005). All of these implants require electrical power to fulfil their function. Typically, non-rechargeable batteries are used as a source. According to a study (Parsonnet and Cheema, 2003), individuals with pacemakers powered by lithium batteries required a reoperation every 7 or 8 years; most commonly for replacement of the battery.

Many alternative methods have been explored for substituting the lithium batteries, e.g. bio-fuel cells that use glucose as a fuel to power the implant, and nuclear cells for pacemakers, but due to an added risk of radiation poisoning and reliability concerns this is not a viable option (Amar, Kouki and Cao, 2015). While these methods are an improvement in energy autonomy, they still have certain drawbacks such as high cost, possible contamination or inadequate performance, etc. To ensure proper operation, implants need to rely on continuous and sufficient power supply.

Among the various potential energy sources available from the human body, here the thermal gradient between the skin surface and the body core (37 °C) is investigated at different locations. Thermoelectric generators (TEG) present a viable opportunity for tapping into these sources to provide stable and sufficient power.

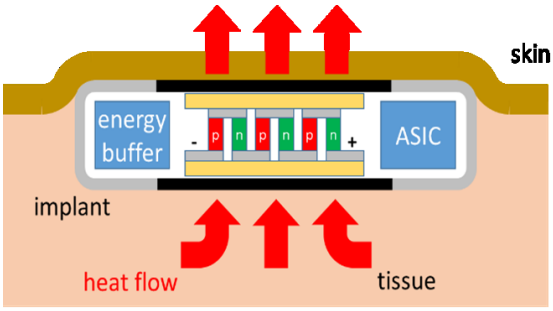


Figure 1: Subcutaneous implant with thermoelectric energy conversion powered by body heat

This paper investigates the energy harvesting potential of a custom TEG compared to commercially available modules. Finally, a possible approach of integrating TEGs into human tissue is presented.

THERMOELECTRIC GENERATOR

Thermoelectric generators are solid-state devices that enable conversion of thermal to electrical energy. Figure 2 shows the setup of a typical TEG. The assembly is made from an array of thermocouples consisting of p-type (hole transporting) and n-type (electron transporting) semiconductor elements. These are connected electrically in series with copper interconnects. The thermocouples are thermally connected in parallel between two ceramic plates.

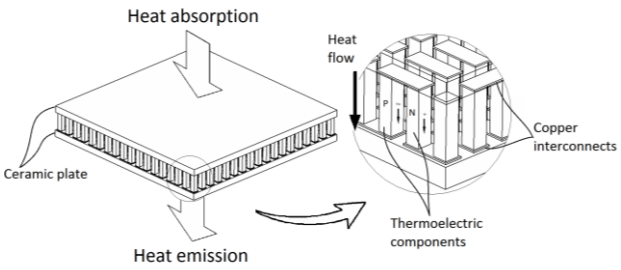


Figure 2: Thermoelectric module with ceramic insulating plates, thermocouples and copper interconnects. A temperature difference generates voltage across the chain of thermocouples

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Thermocouple

The energy conversion in a thermocouple is based on the Seebeck effect, which is illustrated in Figure 3, where a temperature difference drives charge carrier diffusion towards lower temperatures. This results in a potential difference across the thermocouple.

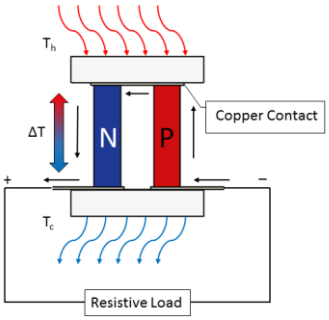


Figure 3: Simplified thermoelectric device with a temperature gradient across the device

The magnitude of the voltage output from equation (1) is proportional to the difference of the Seebeck coefficients α_1 and α_2, the number of semiconductor thermocouples n and the temperature difference ΔT. The maximum power delivered into a load resistor can be calculated using equation (2) with R_{el} as the internal resistance of the TEG (Strasser et al., 2002).

$$V_{out} = n \cdot \Delta T(\alpha_1 - \alpha_2) \tag{1}$$

$$P_{max} = \frac{V_{out}^2}{4\,R_{el}} \tag{2}$$

Figure of Merit

The performance of thermoelectric modules is measured using a dimensionless quantity called figure of merit (ZT) given by equation (3). To obtain a high value, both Seebeck coefficient (α) and electrical conductivity (σ) should be increased, whereas thermal conductivity (κ) shall be minimized. For a thermoelectric device with two semiconductor materials (Tritt, 2002), the figure of merit is calculated using equation (4) where, ρ is the electrical resistivity and the respective material properties of p-type and n-type material are used.

$$ZT = \frac{\sigma \alpha^2 T}{\kappa} \tag{3}$$

$$ZT = \frac{(\alpha_p + \alpha_n)^2 T}{\left[(\rho_n \kappa_n)^{1/2} + (\rho_p \kappa_p)^{1/2}\right]^2} \tag{4}$$

Thermoelectric Material Properties

Most commercially available thermoelectric devices use doped semiconductors with large values for the Seebeck coefficients. For room temperature applications bismuth telluride is used with a typical ZT value ranging from 0.8 to 1.0. To accurately model the thermoelectric behavior, temperature dependent material properties are implemented. Figure 4 illustrates the ZT and Seebeck coefficients' dependence on ambient temperature for p-type and n-type bismuth telluride.

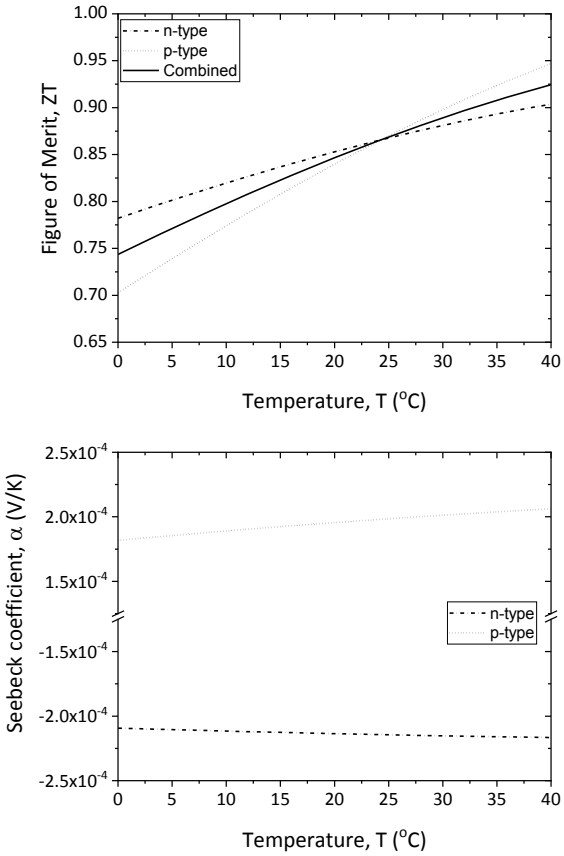


Figure 4: Temperature dependence of figure of merit (top) and Seebeck coefficients (bottom)

For the top and bottom plates of the TEG, ceramic aluminum oxide (96% purity) is used showing high electrical resistivity (10^{14} Ω m) and low thermal conductivity (25 W/(m K)). Interconnects between the p- and n-type semiconductors are made using copper with high thermal conductivity (400 W/(m K)) and low electrical resistivity (1.68 Ω m).

SIMULATION APPROACH AND RESULTS

We use steady state thermal-electric analysis in ANSYS Mechanical 18.2. The geometry model is based on commercially available thermoelectric modules from European thermoelectrics and Laird technologies. The simulations are carried out with top and bottom plates kept at a temperature difference of 1 K. The top face of the terminal thermocouple is electrically grounded. Table 1 comprises voltage output at open circuit conditions. The voltage obtained from the simulation results differ only slightly compared to the rated values. The voltage of a thermoelectric device scales with the number of thermocouples for a given temperature difference, while the power output depends upon the output voltage and the internal electrical resistance of the module.

Table 1: Comparison of manufacturer's data values with ANSYS model performance

					Manufacturer's data	ANSYS model		
Model	Dimensions (mm^3)	n	t_{TEG} (mm)	$A_{Thermoleg}$ (mm^2)	Rated voltage, $V\|_{1K}$ (V)	Voltage, $V\|_{1K}$ (V)	Power, $P\|_{1K}$ (μW)	Power$\|_{1K}$/Total area (μW/mm^2)
A[1]	40×40×6.8	324	3.8	1	0.05454	0.0667	95.85	0.2958
B[2]	30×30×3.7	256	2.0	1	0.04781	0.0524	141.8	0.5538
C[3]	23×23×3.6	144	2.1	0.960	-	0.0290	71.73	0.5186
D[4]	25×25×3.4	256	2.06	0.672	-	0.0524	93.79	0.5448

[1, 2] GM250-161-12-40 and GM250-127-10-15 from European thermoelectrics (thermoelectric generators)
[3, 4] 926-1027-ND and 926-1015-ND from Laird technologies (thermoelectric coolers)

In turn, the electrical resistance of the module depends on the total device area, element length and material properties. To support geometry optimization of the TEG along with the implant assembly, power density provides a viable comparison parameter between the modules.

BIOHEAT MODELING

The thermoelectric modules discussed in the previous section were studied at a constant and uniform temperature difference of 1 K. To find an optimal power output for operation inside the human body and accurate *in silico* results, simulations of temperature distribution in human body are performed so as to identify locations with highest temperature gradient.

The human body is subject to the laws of thermodynamics, the food consumed is converted to bio-chemical energy, which among other things is used to maintain a body core temperature of about 37 °C (Parsons, 1993). The human body dissipates around 100 W at rest. The fat layer in the human body provides a thermal insulation, the largest temperature differences (typically 1-5 K) are found in the highest fat regions of the body (Chen, 2011). We consider the laws of thermodynamics to evaluate the temperature distribution in the human body, the heat-transfer processes are categorized into two categories: internal heat transfer and external heat transfer. Before considering the transient regulation effects, a steady-state simulation model is created in this work.

Internal Heat Transfer

To maintain the body core temperature, the body has to generate energy by processing the food ingested. Metabolism transforms the food into useable energy for the body and perfusion of blood allows the transport of heat throughout the body. Initially, we assume that the body's blood vessels have a fixed temperature of 37 °C and thereby neglect the blood circulation and direction of blood flow in the veins and arteries. The three main heat contributions for internal heat transfer are: conduction, metabolic heat generation and blood perfusion.

Heat conduction

Heat Conduction is the direct translation of heat energy, where heat flows from a hot body to a cold one. This heat flow is defined by the heat flux q_c. From the heat conduction equation (5), the heat flux is directly proportional to the temperature gradient, where the proportionality constant is a material parameter, the thermal conductivity κ.

$$q_c = -\kappa \frac{\partial T}{\partial x} \tag{5}$$

Metabolic Heat Generation

The human body will attempt to preserve or lose sufficient heat to the environment and try to maintain the body core temperature. For cells to perform the metabolic process, glucose and oxygen are required. They are transported to the individual cells through blood. This blood dependency connects the value for the metabolic heat generation directly with the blood perfusion. Since the simulation considers a steady state without blood flow, typical values for the metabolic heat for different tissues are assumed to be constant.

Blood Perfusion

Blood perfusion represents local blood flow through the capillary network and extracellular spaces in the tissue. The main blood vessels are the arteries and the veins. As these blood vessels reach the extremities of the body, many smaller blood vessels branch off to perfuse the organ, muscle, fat and skin tissue. In our model we assume stable one-directional blood flow resulting into constant values for blood perfusion for different types of tissues considered. The heat exchange between the blood and the perfused tissue is dependent on the local temperature T and the blood temperature T_{artery}, leading to a heating of the tissue. To characterize heat transfer in the tissues, the Pennes bioheat equation was used (Pennes, 1948):

$$\nabla \kappa \nabla T + Q_b + Q_m = \rho c \frac{\partial T}{\partial t} \tag{6}$$

The equation describes the influence of homogeneous distributed blood flow on the temperature distribution in

the tissue. where, $Qb = \rho_b c_b \omega \left(T_{artery} - T(x,t)\right)$; ρ, c, κ are the density, specific heat capacity and thermal conductivity of the tissue types ρ_b, c_b and ω are the density, specific heat and the perfusion rate of blood, κ is the thermal conductivity of the tissue and Q_m is the metabolic heat generation.

External Heat Transfer

The fact that the internal temperature is maintained at around 37 °C dictates that there is heat balance between the human body and its environment, the heat generated inside the human body should be balanced by the various heat losses due to convective, radiative and evaporative heat transfer. From a study (Gordon et al., 1976) related to model of temperature regulatory system the boundary condition at skin surface is described as:

$$-Ak\frac{\partial T}{\partial r}\Big|_s = h_c A_s(T_s - T_a) + \sigma\varepsilon FA(T_s^4 - T_r^4) + \dot{E}_s \quad (7)$$

From a computational fluid dynamic (CFD) study of combined simulation of airflow (Murakami, Kato and Zeng, 2000), the major contributions of heat loss were: radiation at about 38.1%, convective loss about 29% and evaporative heat loss of about 24.2%.

Radiation

The process of thermal radiation is described by the Stefan-Boltzmann law or black-body-radiation. As the human body (skin) is not a perfect black body, the emissivity parameter leads to a description of a grey body. The Stefan-Boltzmann law states:

$$Q_r = \sigma\varepsilon A_{skin}(T_{skin}^4 - T_r^4), \quad (8)$$

with the Stefan-Boltzmann constant σ, the skin surface area A_s and emissivity ε.

Convection

The skin tissue heats the local air through a continuous heat loss through convection as long as the ambient air temperature is below the skin surface temperature. The heat flux through convection is described by:

$$Q_b = h_c A_{skin}(T_{skin} - T_{ambient}), \quad (9)$$

where, h_c and A_s are the film coefficient and the skin surface area respectively.

SIMULATION AND RESULTS

The bioheat simulations are done for two different tissue geometries: Three layer simple cylindrical model and a tissue geometry obtained from magnetic resonance imaging (MRI) data.

MRI based Tissue Model – Thermal Simulation

The MRI human tissue model is obtained from the VHP-Female Version 2.2, which has been created using the open-source high-resolution cryosection image dataset from the Visible Human Project®[1] of the U.S. National Library of Medicine. For the steady state thermal analysis in ANSYS Mechanical, we separated the right human forearm from the model and applied realistic tissue parameters available from the IT'IS database[2] such as specific heat, density, perfusion, metabolic heat generation rate (available in Table 2). Blood perfusion along with metabolic heat generation is applied to the muscle, fat and skin layers. The blood vessel bodies are set at a constant temperature of 37 °C. Additionally, the external heat transfer takes place to the environment at 12ºC by means of radiation and convection applied to the skin surface, with a heat transfer coefficient of 3.1 W/(m² K) and emissivity of the body at 0.95.

Table 2: Material properties of various tissue types

Tissue type	Density ρ (kg/m^3)	Specific heat c (J/kgK)	Blood Perfusion ω (1/s)	Metabolic heat Q_m (W/m^3)
Muscle	1090.4	3421.2	6.67×10^{-4}	988.03
Fat	911	2348.3	4.96×10^{-4}	461.48
Skin	1109	3309.5	1.96×10^{-3}	1827.1

Figure 5 illustrates the temperature distribution across the human forearm. As no heat generation is applied to the bone, it remains cooler when compared to other tissues within the arm. Besides metabolic heat generation in muscle, fat and skin tissue heat is also originating from the blood vessels. Minimum temperature of around 26 °C is observed at the fingertips.

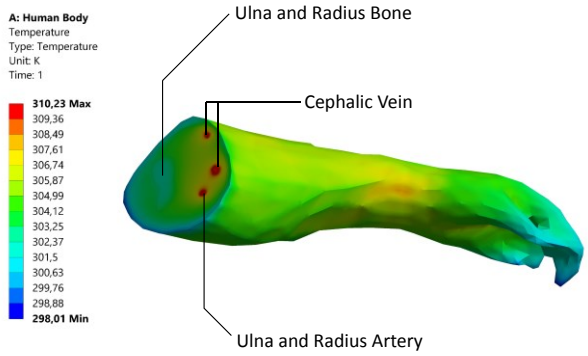


Figure 5: Temperature distribution across the human forearm

The maximum temperature in Figure 6 does not reach the temperature of the arterial blood, because the path does not cross any blood vessels. The temperature

[1]https://www.nlm.nih.gov/research/visible/visible_human.html
[2] https://www.itis.ethz.ch/virtual-population/tissue-properties/database/database-summary/

decreases significantly across the fat tissue. The fat tissue has the lowest heat conduction property compared to other tissues. This leads to an isolating effect creating the largest temperature gradients.

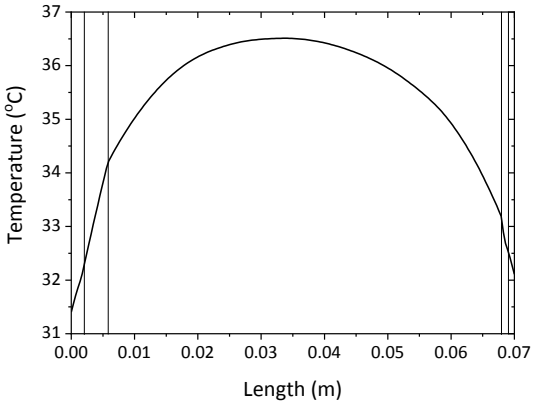


Figure 6: Temperature profile along a path through the forearm

Simplified Model – Fluid Simulation

A fluidic simulation considering heat transport from the fluid to the surrounding tissue overcomes the necessity to consider fixed temperatures of the blood inside the vessel structures. The fluidic model has been implemented in ANSYS FLUENT. The simulation domain is a concentric cylindrical tissue structure, comprised of three layers: muscle, subcutaneous fat and skin. Additionally, arteries and veins along with a simple bone are included to mimic the structure of a human forearm.

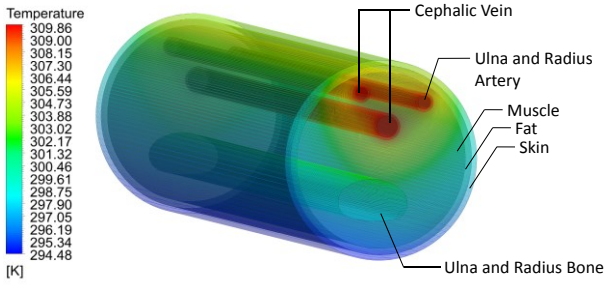


Figure 7: Temperature distribution in simplified geometry

An inlet temperature of 37 °C is applied to the arterial blood flowing with a velocity of 0.4 m/s, while the flow direction in the veins is considered in the opposite direction at 37 °C as there is a negligible drop in blood temperature. Individual metabolic heat generation rates are assigned to the various tissue types. The heat is dissipated by convection and radiation at the skin surface, with a heat transfer coefficient of 3.1 W/(m² K), emissivity of the body is set at 0.95 and the environmental temperature is set to 12 °C.

It can be observed from Figure 8, that the curve follows a similar trend when compared to Figure 6, which confirms that the maximum temperature drop occurs across the fat layer.

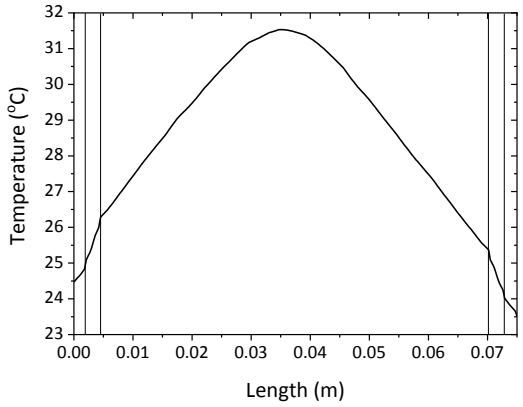


Figure 8: Temperature profile along a path through the simplified geometry

CONCLUSION AND FUTURE WORK

From the results, it is evident that the maximum temperature drop occurs across the fat layer. Thus the main task of further investigation is to integrate the thermoelectric generator into this human tissue region. Furthermore, as the heat transfer from human body is subject to variety of ambient changes throughout the day, we will study the thermoregulation response of the body in different environmental conditions

REFERENCES

Amar, A., Kouki, A. and Cao, H. (2015). *Power Approaches for Implantable Medical Devices.* Sensors, 15(12), pp.28889-28914.

Chen, A. (2011). *Thermal Energy Harvesting with Thermoelectrics for Self - powered Sensors: With Applications to Implantable Medical Devices, Body Sensor Networks and Aging in Place.* Ph.D. University of California, Berkeley.

Gordon, R., Roemer, R. and Horvath, S. (1976). *A Mathematical Model of the Human Temperature Regulatory System - Transient Cold Exposure Response.* IEEE Transactions on Biomedical Engineering, BME-23(6), pp.434-444.

Murakami, S., Kato, S. and Zeng, J. (2000). *Combined simulation of airflow, radiation and moisture transport for heat release from a human body.* Building and Environment, 35(6), pp.489-500.

Parsonnet, V. and Cheema, A. (2003). *The Nature and Frequency of Postimplant Surgical Interventions:* Pacing and Clinical Electrophysiology, 26(12), pp.2308-2312.

Pennes, H. (1948). *Analysis of Tissue and Arterial Blood Temperatures in the Resting Human Forearm.* Journal of Applied Physiology, 1(2), pp.93-122.

Tritt, T. (2002). *Thermoelectric Materials: Principles, Structure, Properties, and Applications.* Encyclopedia of Materials: Science and Technology, pp.1-11.

Watkins, C., Shen, B. and Venkatasubramanian, R. (2005). *Low-grade-heat energy harvesting using superlattice thermoelectrics for applications in implantable medical devices and sensors.* ICT 2005. 24th International Conference on Thermoelectrics, 2005.

MULTIPHYSICS MODELING AND SIMULATION OF A DUAL FREQUENCY ENERGY HARVESTER

Sofiane Bouhedma, Yuhang Zheng, Dennis Hohlfeld
Institute of Electronic Appliances and Circuits
University of Rostock, Rostock, Germany
Email: dennis.hohlfeld@uni-rostock.de

KEYWORDS
Energy harvesting, Mechanical resonator, Finite element analysis, Multiphysics modeling, Piezoelectricity, Frequency tuning, Magnetic force.

ABSTRACT

Vibration based energy harvesting gained a lot of importance in the last two decades, due to the various potential vibration sources in industrial environments, where sensors have to be implemented in inaccessible environments. In this paper we present a concept of a dual frequency piezoelectric energy harvester with magnetic tuning capabilities together with characterization results. We subsequently investigate a potential gain in the electrical power output by implementing separated piezoelectric patches.

INTRODUCTION

Vibration based energy harvesting means converting vibrations present in the environment into electrical energy. It is particularly well suited for industrial condition monitoring applications where sensors need to be implemented in harsh, dark or dirty environments. During the last two decades, several energy harvesters and autonomous systems have been introduced. Further research has been devoted to the development of new resonator designs and to elaborate different strategies for performance optimization of these devices, e.g. by increasing the resonator bandwidth, or tuning the operating frequency of the device.
(Vinod, Prasad, Shi and Fisher, 2008) presented an approach using magnetic forces for tuning the energy harvester's frequency. (Neiss, Goldschmidtboeing, Kroener anf Woias, 2014) proposed a tuning mechanism that allows a compensation of the hysteresis as well as maintaining the optimal working point. A method has been proposed by (Upadrashta and Yang, 2016) to simplify modelling and simulation of the nonlinear magnetic tuning mechanism of a piezoelectric energy harvester using a nonlinear spring. (Hoffmann, Willmann, Hehn, Folkmer and Manoli, 2016) presented a self-adaptive energy harvesting system, which is able to adapt its operating frequency to the dominant vibration frequency of the environment. The energy harvester itself delivers the power required for frequency tuning. Finally, (Wu, Tang, Yang and Soh, 2013) developed a novel compact piezoelectric energy harvester with two resonant modes. (Bouhedma and Hohlfeld, 2017) presented a full experimental characterization of a single and dual frequency resonator with a potential use as an energy harvester.
In this paper, we study the behavior of a dual frequency piezoelectric energy harvester incorporating permanent magnets for frequency tuning. We calculated the magnetic forces to implement them as nonlinear springs. Finally, we investigated the potential gain in the power output of a simple harvester, by using segmented patches connected in series.

FREQUENCY TUNING

We investigate the magnetic tuning of a dual frequency mechanical resonator. A pair of fixed magnets exert forces on another permanent magnet which is mounted on the flexible beam. The net magnetic force superimposes with the restoring force of the deflected beam. Thereby, an effective spring constant arises. In other words, the magnetic forces will soften or harden the structure, depending on the magnets' orientation and their mounting configuration. This will result in a shift in resonance frequency. In order to experimentally investigate the frequency tuning performances, we consider a resonator fabricated from stainless steel. It features a so-called folded beam as shown in figure 1.

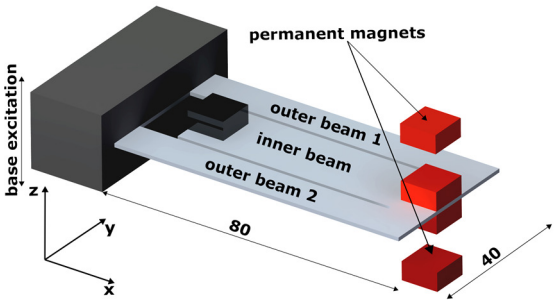


Figure 1: Descriptive scheme of the 1 mm thick folded beam mechanical resonator with the frequency tuning permanent magnets in the vertical configuration

The structure is subjected to a harmonic base acceleration with a magnitude of $a = 0.2$ g generated by a vibration test equipment. The tuning mechanism incorporates permanent magnets in attractive and repulsive orientations, mounted such that the force reaction is either collinear to the axis of the beam (referred as 'axial configuration') or perpendicular to the beam surface (referred as 'vertical configuration').

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

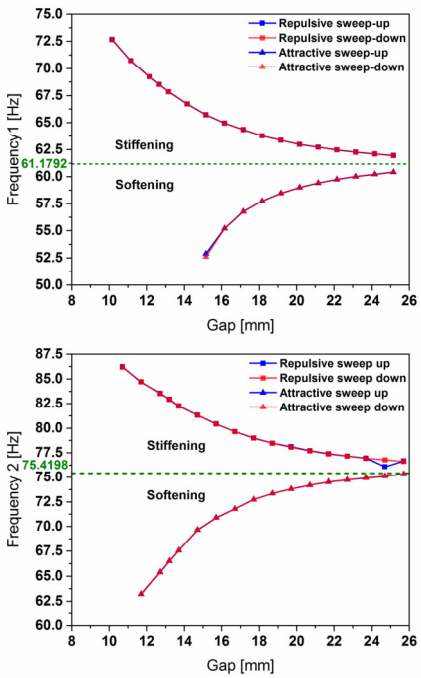


Figure 2: Frequency shift of the dual frequency resonator in the vertical configuration, considering both, outer part tuning (top) and the inner part (bottom). The magnetic forces enable tuning the operating frequencies by approximately 12% in both directions

The experimental results in figure 2 show natural frequency tuning of the dual frequency resonator by up to 20%. The tuning of the two frequencies can be achieved independently of each other; this open up the opportunity to reduce the frequency gap between the two resonances. Furthermore, achieving a frequency overlap is possible. However, we observed a drop in displacement amplitudes during tuning, which constitutes a limitation of using such a strategy (Bouhedma and Hohlfeld, 2017).

SIMULATION APPROACH

In order to understand the behavior and estimate the power output of the proposed energy harvester, we implemented the structure in ANSYS Multiphysics. We considered the mechanical resonator shown in figure 1 with two identical masses m = 7.6 g.

Mode shape comparison

A modal analysis yielded the mode shapes at the two resonance frequencies. We implemented a harmonic response simulation in ANSYS and verified its results via experimental data. We excited the structure with a harmonic base acceleration generated by a vibration test system with an amplitude of a = 0.5 g.

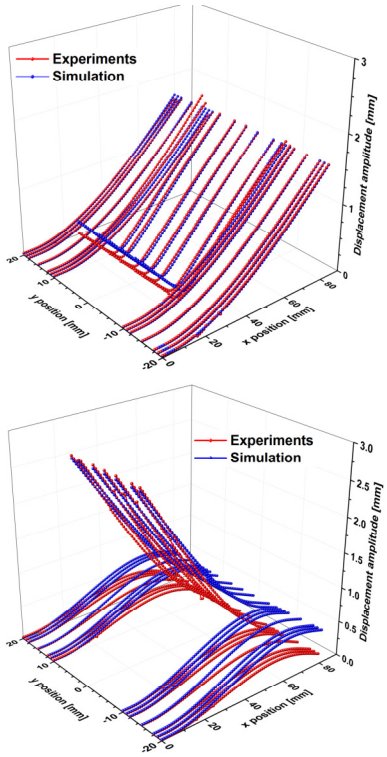


Figure 3: Comparison between the simulation and experimental first (top) and second (bottom) mode shapes of the mechanical resonator

The results on figure 3 show a good qualitative fit between simulation and experimental results. The slight discrepancy in the amplitude values is presumed to stem from measurement errors and the clamping system. Moreover, the positions of the magnets, which in this case are used only as tip masses, can also induce errors.

Piezoelectric model

Furthermore, we implemented the piezoelectric transducer elements in our ANSYS model. Patch dimensions are 60×10×0.2 mm^3 on the outer beam and 48×18×0.2 mm^3 for the inner beam. The material properties of the piezoelectric ceramic originate from PIC-255 supplied by PI Ceramic.

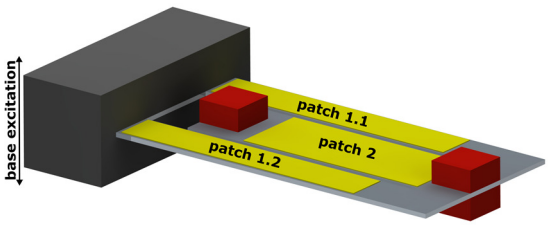


Figure 4: Arrangement of piezoelectric energy transducing patches on the flexible segments of the resonator

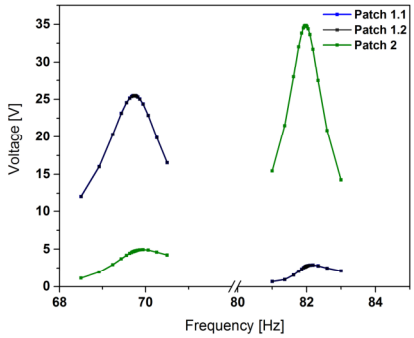


Figure 5: Estimated voltage output of the dual frequency energy harvester

The power output can be given by the following expression:

$$P_{ij} = \frac{1}{4} V_{0\,j}^2 \, 2\pi \, f_j \, C_i \qquad (1)$$

Where: $V_{0\,j}$ is the open-circuit output voltage amplitude at mode j, f_j is the mode frequency and C_i is the capacitance of the piezoelectric patch i. The total power output, as represented in figure 6, is then the sum of all powers delivered by each patch at every mode.

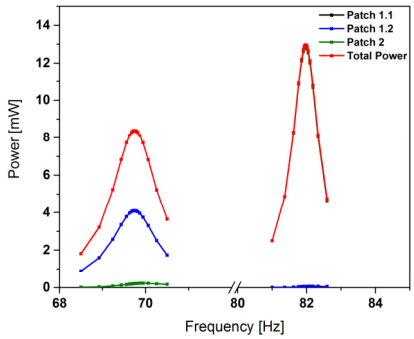


Figure 6: Estimated power output of the dual frequency energy harvester

These results illustrate the possibility of harvesting energy from the first two resonance frequencies. Power generation, as shown on figure 6, occurs mostly at the patches located at the inner respectively outer beams. Only one patch is mechanically stressed at each of the two operational frequencies.

Magnetic forces simulation

We also investigate the magnetic forces for later implementation in our ANSYS model. This will enable us to simulate the frequency tuning effect in different configurations and orientations. We considered a pair of neodymium permanent magnets in COMSOL Multiphysics, with N42 magnetization and with the geometry showed in figure 7. The magnet's dimensions are 10×10×5 mm^3. For simplification purposes, we ignore the rotation of the magnet as the beam deflects.

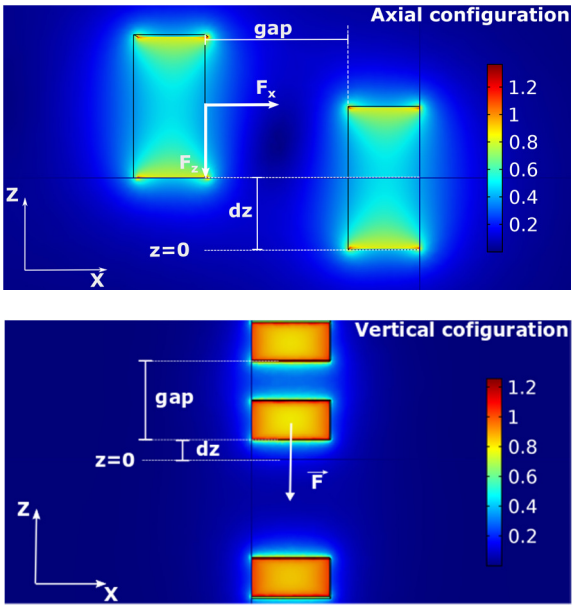


Figure 7: Axial (top) and vertical (bottom) magnet configuration; the color scale corresponds to the magnetic flux density

The simulation considered different orientations (attractive and repulsive) and different configurations (axial and vertical). It has been performed for different gap values and for different vertical displacements as depicted in figures 8 and 9.

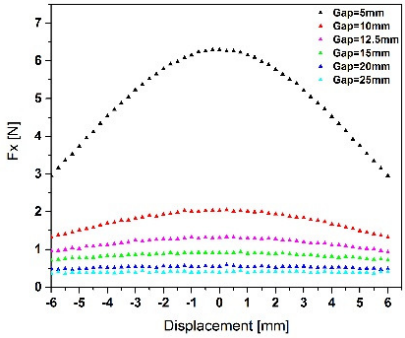


Figure 8: Axial component F_x of the attractive force in axial configuration

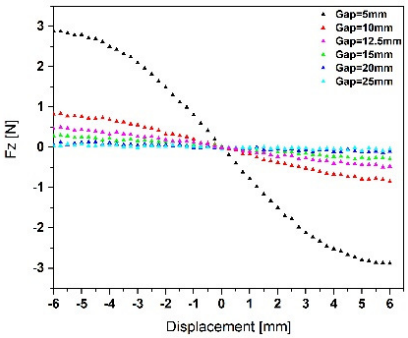


Figure 9: Vertical component F_z of the attractive magnetic force in axial configuration

These results show a nonlinear force-displacement relationship between the force components and the vertical displacement of the magnet, which in our case

represents the beam respective magnet displacement amplitude.

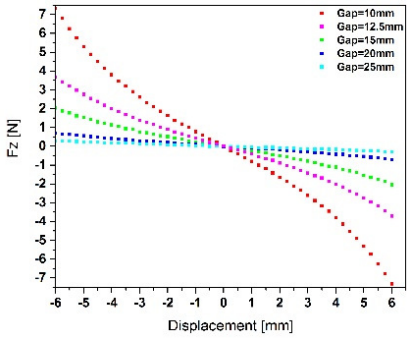


Figure 10: Repulsive magnetic force components in the axial configuration

We intend to implement the effect of the magnetic forces in our dual frequency resonator model by either considering springs elements with nonlinear stiffness or by a position-dependent force. Therefore, we identify fitting functions represented in figures 11 and 12.

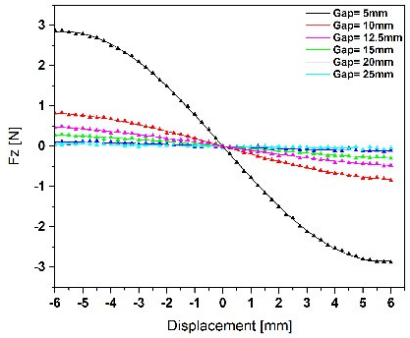


Figure 11: Fitting function for the attractive magnetic force component F_Z in the axial configuration

In both configurations, we mainly focus on F_Z, and this is due to the approach similarities used for both components (Fx and Fz).

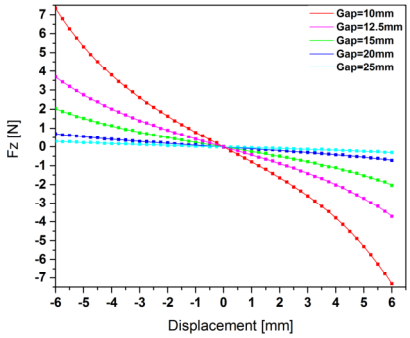


Figure 12: Fitting function for the repulsive magnetic force component F_Z in the vertical configuration

The defined functions in all configurations show an excellent match with the simulation results. As an example, the fitting functions Fz_a and Fz_v corresponding to the minimum gap values (5 mm in the axial configuration and 10 mm in the vertical one) are given as follows:

$$Fz_a = -0.002 + 2.85 \sin\left(\pi \frac{z - 11.46}{11.46}\right) \quad (2)$$

$$Fz_v = \sum_{i=0}^{5} a_i z^i \quad (3)$$

where: a_1 = -0.006, a_2 = -0.782, a_3 = 4.11×10^{-4}, a_4 = -0.009, a_5 = 1.65×10^{-7} and a_6 = 9.02×10^{-5}. These force-displacement functions can be directly implemented into our ANSYS model. Alternatively, we implement their derivatives, which give the nonlinear stiffness of the lumped springs.

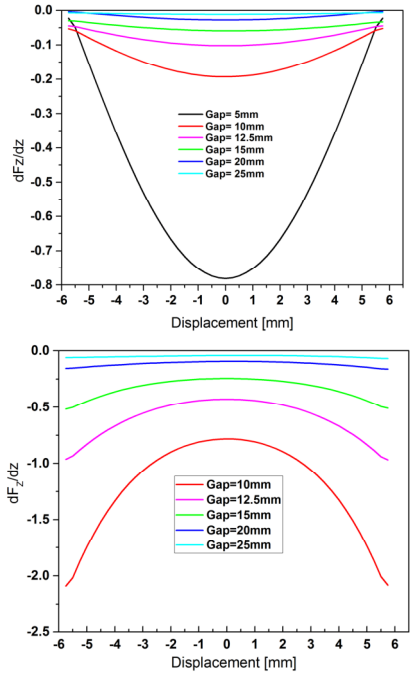


Figure 13: Derivatives of the fitting functions, which give the nonlinear stiffness of the spring elements in the axial configuration (top) and in the vertical configuration (bottom)

The fitting functions are valid for a specific gap. Therefore, we will compose fitting functions with two variables, which will enable us and to parameterize the problem and run the model for different gaps.

Power optimization

Subsequently, a simulation using ANSYS Multiphysics has shown the potential to increase the power output of a piezoelectric harvester by considering two segmented patches electrically connected in series; instead of using a single patch. We considered a simple tip loaded (m = 3.8 g) energy harvester with the geometry shown in figures 14 and 15 subjected to a harmonic base excitation amplitude of d = 6 μm.

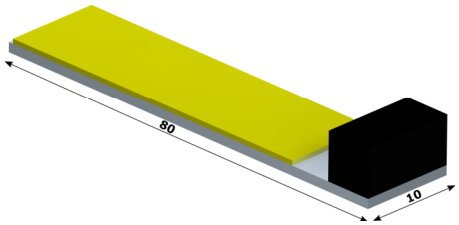


Figure 14: Simple tip loaded energy harvester with a full piezoelectric patch

By implementing PIC-255 as a piezoelectric material we observed that the power can be increased by up to 13% compared to a full patch (see figure 16, the full patch configuration corresponds to L_p = 60 mm). The simulations have been executed with unrealistically low damping coefficients, which explains the high power output values.

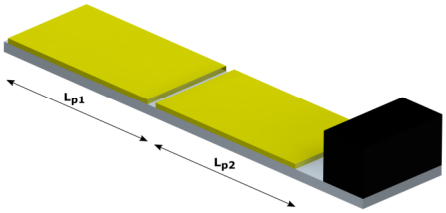


Figure 15: Simple tip loaded energy harvester with two-segmented piezoelectric patches electrically connected in series

The results indicate that optimum power delivery is achieved if the single patch is segmented into two patches of approximately equal sizes. This gain in power is due to the superposition of the voltage generation of the individual patches. Simultaneously, the total capacitance decreases due to the series connection, leading to a higher optimum load. As described in equation (1), the power output scales quadratically with available voltage and only linear with the capacitance. Only a minor gain in total power output is achievable if three or more patches are used.

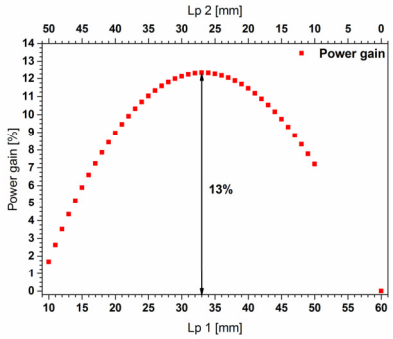


Figure 16: Normalized power output of a tip-loaded simple energy harvester for various piezoelectric patch dimensions, compared to a full patch configuration

The power gain is limited to 13% in the present case (tip masses). Cantilevers without additional masses show a different strain distribution and bear the potential of up to 30% of power gain (Bouhedma and Hohlfeld, 2017).

CONCLUSION AND OUTLOOK

In this work, we experimentally investigated the magnetic frequency tuning of a dual frequency resonator. The mechanical resonator was characterized in two configurations of different magnet arrangements (axial and vertical) with different magnets orientations (attractive and repulsive). It has been shown that the frequency can be tuned by up to 20% for both parts (inner and outer). As the tuning of one mode does not affect the other mode, the approach can also be used for reducing the frequency gap between the two modes.

On the other hand, we simulated the behavior of the mechanical resonator incorporating the piezoelectric energy transducer. We demonstrated the dual frequency operation and that energy can be harvested at two different frequencies. In order to simulate the behavior of the resonator under the influence of the magnetic forces, a static simulation of the magnetic forces between two permanent magnets, in the axial and the vertical configurations has been performed. The results of these simulations enabled us to define fitting functions giving the force-displacement relationships and their derivatives. The latter functions give the nonlinear stiffness of the magnetic spring elements. These data still needs to be thoroughly investigated and extended to define a more general gap-displacement-force relationship, since the magnetic forces are gap-dependent as well.

Subsequently, we investigated the possibility of increasing the power output of a tip-loaded simple piezoelectric energy harvester, by using an array of segmented piezoelectric patches, connected in series, instead of a full patch. The simulation results revealed that the gain in power can reach 13%. These results still need to be validated by an experimental work and will be generalized for the presented energy harvester.

REFERENCES

Vinod, R., C., Prasad, M., G., Shi, Y., and Fisher, F., T., (2008): "A vibration energy harvesting device with bidirectional resonance frequency tenability". *Smart Mater. Struct.* 17(1), p. 15035.

Neiss, S., Goldschmidtboeing, F., Kroener, M., Woias, P., (2014): "Tunable nonlinear piezoelectric vibration harvester". *J. Phys.: Conf. Ser.* 557, p. 12113.

Upadrashta, D., Yang Y., (2015). "Finite element modeling of nonlinear piezoelectric energy harvesters with magnetic interaction". *Smart Mater. Struct.* 24 (4), p. 45042.

Hoffmann, D., Willmann, A., Hehn, T., Folkmer, B., Manoli, Y., (2016). „A self-adaptive energy harvesting system". *Smart Mater. Struct.* 25 (3), p. 35013.

Wu, H., Tang, L., Yang, Y., Soh, C. K., (2013), "A novel two-degrees-of-freedom piezoelectric energy harvester". *J. Int. Mater. Syst. Struct.* 24 (3), pp. 357–368.

Bouhedma, S., Hohlfeld, D., (2017), "Frequency tunable piezoelectric energy harvester with segmented electrodes for improved power generation", In *Proceedings of the 2017 Mikrosystemtechnik Kongress.*

PARAMETRIC MODEL ORDER REDUCTION OF INDUCTION HEATING SYSTEM

Ananya Roy, M. Nabi
Department of Electrical Engineering
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi - 110016, India
Email: roy.ananya2009@gmail.com, mnabi@ee.iitd.ac.in

KEYWORDS

Finite Element Modeling; Parametric Model Order Reduction; Induction Heating System

ABSTRACT

Induction Heating system have been widely used nowadays. Modeling of such system remains a challenge. In this paper, modeling of a axi-symmetrical structured induction heating system is carried out. Few geometrical parameters affect the modeling of the system. Variation of these parameter values leads to multiple simulations, making the design procedure computationally expensive. Original system is produced from finite element method and is reduced to lower order system to make the simulation cost effective. As the size of system matrices are different with different state space models, matrix interpolation is used with two stage model reduction. Simulation results of the models at different parameter values are shown.

Introduction

Induction Heating process has wide application areas from industrial area to household appliances. Induction heating process is a multi-physical system. Uniform current density is applied to the copper coils. It produces a magnetic field in the system. Due to ohmic loss of eddy current induced in the iron body, heat is generated. Experimental design of induction-heated systems is not only tedious but also expensive. Thus, it necessitates the need for a precise mathematical model with help of simulation software[1]-[3].

Most mathematical models of the induction heating system are expressed as distributed parameter models with partial differential equations (PDEs). These PDEs are converted to ordinary differential equations (ODEs) with Finite element (FE) analysis. The generated ODEs are large in number and make the system computationally costly. The original system can be projected to a lower dimensional subspace to produce its lower order approximation using Model Order Reduction (MOR) techniques. There are various methods for model order reduction depending on the need such as Singular value decomposition method, Krylov subspace based method, structure-preserving model order reduction etc[4]-[8].

Sometimes few analytically inexpressible parameters like geometrical parameters in FE model based large systems, play important role in the design process. Changing such parameter values while using standard MOR techniques are computationally costly because of repetitive simulations. In such problems Parametric MOR (pMOR) techniques are used. It tries to preserve the parameter dependency of the original systems in the reduced models so that multiple simulations for different values of the parameters can be carried out in the reduced space. While considering geometric parameters regular pMOR cannot be applied, although the geometric parameters does effect the FE model and solution implicitly. Matrix Interpolation (MI) based pMOR framework [9] has been proposed for such scenarios. In [9], system matrices of reduced order models using MOR techniques, were transformed as the reduced states do not have same physical interpretation. Then weighted interpolation is performed to get a parametric reduced model. In the present induction heating model, it is not directly applicable as the higher order FE models are of different sizes for different parameter values. Hence some modifications are proposed with some results.

Modeling of Induction Heating System

Here we are considering an induction heating system as shown in Fig. 1. The model consists of a iron cylinder and copper wires around it. The

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

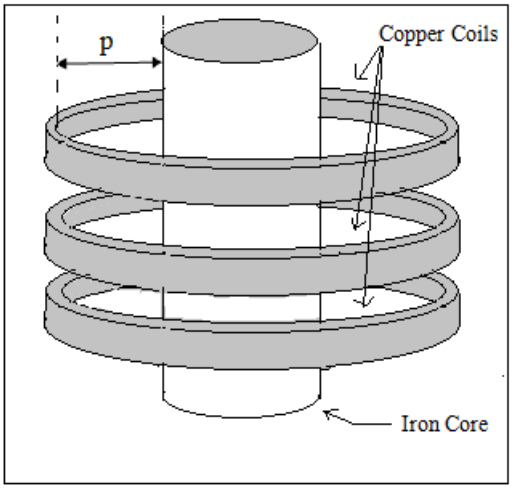


Fig. 1: Three dimensional model of Induction Heating System

copper coils carry alternating current producing electromagnetic effect. Due to ohmic losses of eddy current in the iron cylinder, heat is generated. The distance p, between copper coil and iron core, is considered as a parameter.

The governing equation describing the magnetic vector potential A is given as

$$\frac{1}{\mu}\nabla^2 A - j\nu A + J_s = 0 \tag{1}$$

where J_s is the current density and it varies sinusoidally in time with a single frequency, ω. As it is time harmonic in nature, it is sinusoidal function of time. So it can be represented in terms of vector phasors that depend on the space coordinates but not on time. The permeability is represented with μ and ν is a scalar quantity with $\nu = \omega\sigma$ with σ being electrical conductivity.

The induced heat (Q) can be written as

$$E.J = \sigma A^2 = Q \tag{2}$$

where E and J are electric field intensity and current density respectively. As only magnitude of Magnetic vector potential is considered, Q is a scalar quantity. Due to the ohmic losses of the eddy currents in iron, heat is produced. The time dependent temperature profile $T(r, z, t)$ can be given as

$$c\rho\frac{\partial T(r,z,t)}{\partial t} + \nabla(\kappa\nabla T(r,z,t)) = Q \tag{3}$$

where the model is considered to be thermally insulated and heat is not convected or radiated from the iron body. Here, specific heat and mass density of the materials are represented with c and ρ, whereas the thermal conductivity is denoted by κ.

After discretizing the model, the Finite Element representation of (1) can be written as

$$[\mathbf{K} + j\mathbf{S}]\bar{A} = \bar{F} \tag{4}$$

where $\mathbf{K}$ and $\mathbf{S}$ are the real, sparse, symmetric matrices and $\mathbf{K}$ has full rank.

The heat transfer equation can also be written as

$$\mathbf{C}\frac{d\bar{T}(t)}{dt} + \hat{\mathbf{K}}\bar{T}(t) = \bar{Q} \tag{5}$$

Again, $\mathbf{C}$ and $\hat{\mathbf{K}}$ are sparse, real and symmetric matrices with $\hat{\mathbf{K}}$ having full rank. The input vector is denoted by $\bar{Q}$. Specific heat constant and mass density are consumed in $\mathbf{C}$ and κ in $\hat{\mathbf{K}}$. Hence, (4) & (5) is the coupled model of the induction heating system.

Parametric Model Order Reduction

There are some geometrical parameters which affect the performance of the system. Such a parameter, the distance between the coils and the cylindrical core, is shown in Fig. 1. As the distance increases, the magnetic field or the magnitude of magnetic vector potential in the iron body decreases and heating time increases.

The FE electromagnetic equation where the parameter varies implicitly, can be written as

$$[\mathbf{K}(p) + j\mathbf{S}(p)]\bar{A} = \bar{F}(p) \tag{6}$$

Similarly, the FE heat transfer equation can be written as

$$\mathbf{C}(p)\frac{d\bar{T}(t)}{dt} + \hat{\mathbf{K}}(p)\bar{T}(t) = \bar{Q}(p) \tag{7}$$

where, $p \in \mathbf{R}$.

FE generated ODEs are quite large in number making it complex to handle. These systems can be replaced with reduced order model projecting them to lower dimensional subspaces. The parameters which influences the system modeling, can not be preserved using conventional model order reduction methods. Parametric model order reduction (pMOR) techniques are very useful for preserving the parameter in the reduced models.

All the reduced models from pMOR have equal number of states. But they do not interprete the same physical quantities. This necessitates the need of state transformation. The required state transformation can be achieved employing the matrix interpolation method as described in [9]. Here, all the system matrices are considered to be equal and this leads to equal size of projecting matrices.

While modeling it has been seen that with change in the parameter value, p the size of system matrices ($\mathbf{K}, \mathbf{S}, \mathbf{C}, \hat{\mathbf{K}} \in \mathbf{R^{n_i \times n_i}}$) changes. Therefore, the projecting matrices are of different sizes. Usual matrix interpolation method as described in [9] is inapplicable. So, a two stage model order reduction method is suggested. In the first stage reduction, the original systems are projected to a moderately lower dimensional subspace making all the system matrices of same size. In the second stage, these moderately large systems are reduced to get lower order approximations.

In the first stage of order reduction the original system is projected to a lower order ($\mathbf{R^{r_1}}$) plane using one sided Arnoldi algorithm. The reduced system can be written as

$$[\mathbf{K}_{r_1}(p) + j\mathbf{S}_{r_1}(p)]\bar{A}_{r_1} = \bar{F}_{r_1}(p) \tag{8}$$

$$\mathbf{C}_{r_1}(p)\frac{d\bar{T}_{r_1}(t)}{dt} + \hat{\mathbf{K}}_{r_1}(p)\bar{T}_{r_1}(t) = \bar{Q}_{r_1}(p) \tag{9}$$

Here, system matrices are represented as $\mathbf{K}_{r_1} = \mathbf{V}_{r_1}^T\mathbf{K}\mathbf{V}_{r_1}$, $\mathbf{S}_{r_1} = \mathbf{V}_{r_1}^T\mathbf{S}\mathbf{V}_{r_1}$, $\bar{F}_{r_1} = \mathbf{V}_{r_1}^T\bar{F}$, $\mathbf{C}_{r_1} = \mathbf{V}_{r_1}^T\mathbf{C}\mathbf{V}_{r_1}$, $\hat{\mathbf{K}}_{r_1} = \mathbf{V}_{r_1}^T\hat{\mathbf{K}}\mathbf{V}_{r_1}$ and $\bar{Q}_{r_1} = \mathbf{V}_{r_1}^T\bar{Q}$ with $\mathbf{V}_{r_1}$ as the projection matrix.

As described earlier, (8) & (9) are moderately large systems. So, it can be reduced further by projecting it to a lower order subspace $\mathbf{R^{r_2}}$ with V_{r_2}. The new reduced order model becomes

$$[\mathbf{K}_{r_2}(p) + j\mathbf{S}_{r_2}(p)]\bar{A}_{r_2} = \bar{F}_{r_2}(p) \tag{10}$$

$$\mathbf{C}_{r_2}(p)\frac{d\bar{T}_{r_2}(t)}{dt} + \hat{\mathbf{K}}_{r_2}(p)\bar{T}_{r_2}(t) = \bar{Q}_{r_2}(p) \tag{11}$$

For each discrete value of parameter, large scale model and reduced models are obtained employing projection matrices ($\mathbf{V}_{r_2,i}$) with one-sided Arnoldi algorithm.

The transformed reduced systems can be expressed as

$$[\mathbf{M}_i\mathbf{K}_{r_2,i}\mathbf{T}_i^{-1} + j\mathbf{M}_i\mathbf{S}_{r_2.i}\mathbf{T}_i^{-1}]\bar{A}_{r_2} = \mathbf{M}_i\bar{F}_{r_2,i} \tag{12}$$

$$\mathbf{M}_i\mathbf{C}_{r_2,i}\mathbf{T}_i^{-1}\frac{d\bar{T}_{r_2}(t)}{dt} + \mathbf{M}_i\hat{\mathbf{K}}_{r_2,i}\mathbf{T}_i^{-1}\bar{T}_{r_2}(t) = \mathbf{M}_i\bar{Q}_{r_2,i} \tag{13}$$

Now, following the standard MI procedure, the transformation matrices can be found from

$$\mathbf{T}_i = \mathbf{R}^T\mathbf{V}_{r_2,i} \text{ and } \mathbf{M}_i = (\mathbf{V}_{r_2,i}^T\mathbf{R})^{-1}$$

For obtaining $\mathbf{R}$, whose q columns span the universal subspace of the MOR strategy, the following has been suggested. The projection matrices $\mathbf{V}_{r_2,i}$s form a pool of independent directions and thus are accumulated in

$$\mathbf{V}_{all} = [\mathbf{V}_{r_2,1}\ \mathbf{V}_{r_2,2}\ \cdots \mathbf{V}_{r_2,k}]$$

followed by its SVD, given by $\mathbf{V}_{all} = \mathbf{U}\mathbf{\Sigma}\mathbf{N}^T$. The first q columns of $\mathbf{U}$ span q different directions and serve as $\mathbf{R}$.

Results and Simulations

The actual model of Induction Heating system is three dimensional. But due to its axi-symmetrical structure, two dimensional model is considered as shown in Fig. 2. The dimensions

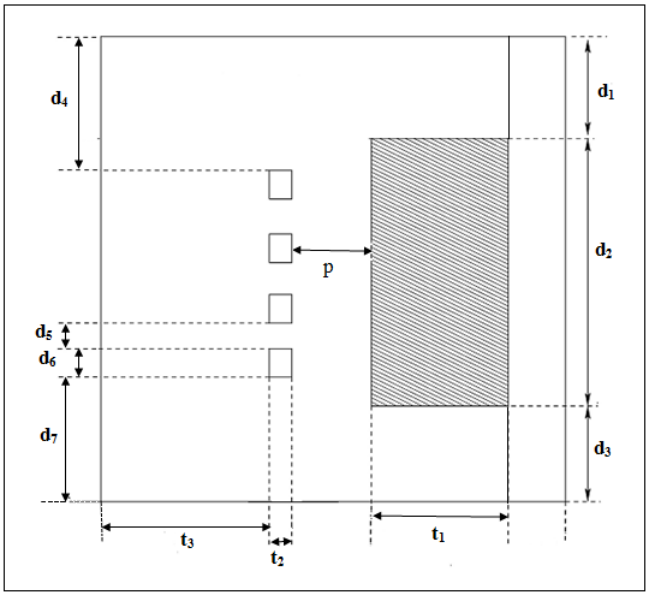


Fig. 2: Parameter affecting the Induction Heating System

of the model considered is given in Table 1. The

TABLE I: Geometric characteristics of the problem domain

p_1, p_2, p_3	$5mm, 10mm, 15mm$
d_1	$70mm$
d_2	$60mm$
d_3	$70mm$
d_4	$71mm$
d_5	$6mm$
d_6	$10mm$
d_7	$71mm$
t_1	$25mm$
t_2	$5mm$
t_3	$55mm - 65mm$

geometric parameter considered here is the distance between coil and the iron core. The parameter value varies from 5 mm to 15 mm. The material properties considered here are given in Table 2. Uniform current density of $8A/m^2$ is applied to the copper coils. Here zero boundary conditions are assumed for (4). As the boundaries are far away from the model, they do not have any effect of the magnetic field. Produced

magnetic vector potential in the system is shown in Fig 3. The temperature profile in iron core is shown in Fig 4. The thermal field is simulated for 40 sec. It is clear from Fig. 4 that the heating time varies with the change in the distance between coil and core. As the FE generated models are of different sizes, in first stage all the models are projected to lower dimensional subspace of order 1000. In the second stage all the models are reduced to an order of 16. With the help of matrix interpolation, interpolated reduced system is generated. Fig. 5 shows the bode plots of the original system, first stage reduced system, second stage reduced system and matrix interpolated reduced system. From the plots it can be suggested that the reduced models and the matrix interpolated reduced model can replace the original system for lower frequencies.

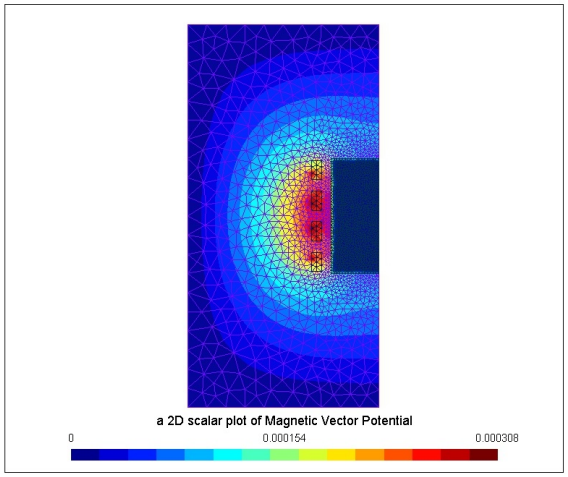


Fig. 3: Magnetic field of Induction Heating System

TABLE II: Material Properties

Parameters	Iron	Copper	Air
Relative Permeability (N/A^2)	1500	1	1
Electrical Conductivity (S/m)	10^7	5.96×10^6	3×10^{-15}
Mass Density (kg/m^3)	7874	8940	1.225
Specific Heat $(J/kg - K)$	450	385	1000
Thermal Conductivity $(W/m - K)$	83.5	401.0	0.001

CONCLUSIONS

In this article, finite element modeling of induction heating system is described. In the design process, variation of the parameter value leads

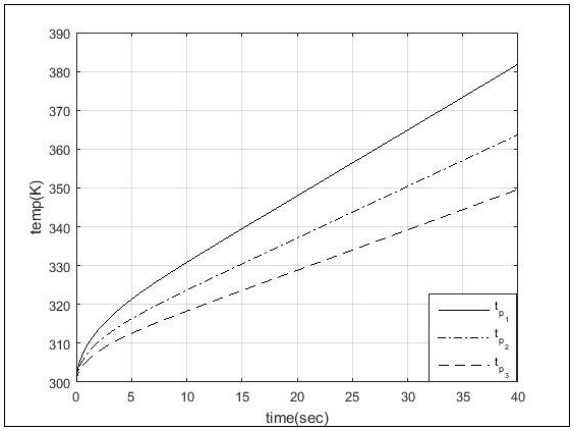


Fig. 4: Temperature plot in Iron core

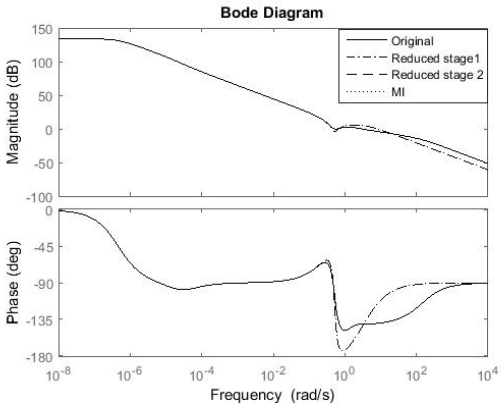


Fig. 5: Bode Plot of Induction Heating System (Thermal Model)

to repetitive simulations which is computationally costly. So, the original system is reduced employing parametric model order reduction. The distance between copper coil and iron core is the parameter. As the size of system matrices changes with the variation of parameter values, standard matrix interpolation method cannot be applied directly. Hence, the original finite element model is reduced to a lower order approximation in two stages. Simulation results show the efficacy of the proposed modification.

References

[1] C. Chaboudez, S. Clain, R. Glardon, D. Mari,J. Rappaz, M. Swierkosz, 'Numerical modeling in induction heating for axisymmetric geometries', IEEE Transactions on Magnetics 33.1 (1997): 739-745

[2] M. H. Tavakoli, H. Karbaschi, F. Samavat,'Computational modeling of induction heating process', Progress in Electromagnetics research letters, 11, 93-102

[3] O. Klein, P. Philip,'Correct voltage distribution for axisymmetric sinusoidal modeling of induction heating with prescribed current, voltage, or power', IEEE transactions on Magnetics, 38(3), 1519-1523.

[4] J. N. Reddy, 'An introduction to the finite element method', Vol. 2. No. 2.2. New York: McGraw-Hill, 1993.
[5] Klaus-Jurgen Bathe, E. L. Wilson, 'Numerical methods in finite element analysis' (1976): 6-12.
[6] B. Salimbahrami, B. Lohmann, 'Order reduction of large scale second-order systems using Krylov subspace methods', Linear Algebra and its Applications 415 (2006) 385 - 405.
[7] Zu-Qing Qu, 'Model Order Reduction Techniques with Applications in Finite Element Analysis', Springer, 2013.
[8] A. C. Antoulas, D. C. Sorensen, 'Approximation Of Large-Scale Dynamical Systems: An Overview', Int. J. Appl. Math. Comput. Sci., 2001,Vol.11,No.5,1093-1121.
[9] H. Panzer, J. Mohring, R. Eid, B. Lohmann, 'Parametric Model Order Reduction by Matrix Interpolation', Automatisierungstechnik Methoden und Anwendungen der Steuerungs- Regelungs- und Informationstechnik, Volume 58, Issue 8, Pages 475 - 484, ISSN (Print) 0178-2312, DOI: 10.1524/auto.2010.0863, August 2010.

Ananya Roy is a Research Scholar in Control And Automation Research Group, in the Department of Electrical Engineering in Indian Institute of Technology, Delhi, India.

M. Nabi is an Associate Professor in Control And Automation Research Group, in the Department of Electrical Engineering in Indian Institute of Technology, Delhi, India.

Finite - Discrete - Element Simulation

COUPLING FINITE AND DISCRETE ELEMENT METHODS USING AN OPEN SOURCE AND A COMMERCIAL SOFTWARE

Ákos Orosz, Kornél Tamás, János P. Rádics, 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
Discrete Element Method, Finite Element Method, Yade, Ansys

ABSTRACT

There are many cases where mechanical engineering structures interact with bulk materials, e.g. cultivating and mining machines.
The tool of the machine has effect on the aggregate and vice versa. The aggregate puts complex forces on the machines. These loads are usually simplified and replaced with a mean, distributed force, which is the input of further computations (e.g. finite element (FEM) simulations). This can be done, because the difference between two different, but statically equivalent loads becomes very small at sufficiently large distances from load. However, near the area, where loads act, the results using the simplification are not real, therefore finding the optimal tool design is difficult.
The discrete element method (DEM) models the materials using particles (elements) with independent translational and rotational degrees of freedom and arising forces (interactions) between them. This allows the simulation and tracking of each particle independently, which makes DEM ideal for modelling bulk materials. The FEM is usually used for modelling continua. If DEM is coupled with FEM, the result is a detailed stress distribution near the loads. In the possession of the detailed stress state, the optimization of the tool can be performed, and a better construction may be created.

INTRODUCTION

In the field of mechanical engineering, there are many types of machines which were designed to handle bulk materials. Good examples are the ones which came into interaction with stone aggregates: stones are yielded, crushed, transported and deposed by massive and complex machines to be finally used in e.g. railway ballasts and road base courses. The crushed rock aggregates are essential additives of concrete and asphalt mixtures. The tools of the machines are affected by static and dynamic loads from the interaction with stones, which cause high level of wear and the risk of fatigue.
Another important field, where machine tools interact with bulk materials is the agricultural machinery. The mechanical characteristics of the inhomogeneous soil are highly nonlinear. The tools of the different machines turn over, rip, compress and mix the soil to raise productivity. The efficiency of these processes is highly affected by the shape of the tools. The effect of the soil on the tool cannot be neglected, as it causes remarkable abrasion.
The various types of sands, similarly to stone aggregates, are needed to be yielded, transported and sorted which is also done by machinery. The solid pharmaceutical industrial products must be treated carefully, which requires the precise design of the geometry and material of the machine tools.
The examples show the wide variety of machinery, which came into interaction with bulk materials. The design and optimization of these structures are traditionally relying on the routine of practice or applying approximation equations. These methods require the manufacturing of several prototype variants and carrying out a large number of benchmark tests to find the best concept.
The numeric simulation of different processes creates the opportunity to virtually test the construction variants in the phase of conception creation, which reduces the number of manufactured prototypes, thus reducing the cost of design and validation process significantly.

SIMULATION METHODS

Discrete Element Method

The discrete element method (DEM) is a numerical technique where the material is made up from discrete elements (particles). The elements have independent motional and rotational degrees of freedom (DoF). The model can track the finite displacements and rotations (possibly deformations) of the particles. Interaction forces can be risen and extinguished between the elements (Bagi 2007, Cundall and Hart 1992).
Because of the definition, the behaviour of the model depends on two factors: the properties of the particles (shape and material) and on the characteristics of interactions (constitution law). The element type is chosen based on the features of the original material, as well as the constitution law, but parameters of interactions between elements have to be calibrated.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

The Finite Element Method

The finite element method (FEM) (Zienkiewicz 1971) is most often used to model continuum materials. Its major application fields are the simulation of mechanical (static, dynamic, buckling, fatigue), thermodynamic and electrodynamic processes.
The finite element method (FEM) also models the material with limited number of elements, however, the adjacent elements have common nodes. These nodes have common DoF (Dill 2012).
So, the similarity between FEM and DEM is that both describes the material with finite number of elements, however, these elements are not independent in case FEM, where the neighbouring elements have common nodes (with same DoF). This makes FEM more suitable for modelling continua and DEM for describing bulk materials.

THE MODELLING POSSIBILITIES OF MACHINE TOOLS

As the main goal of DEM simulations in the field of mechanical engineering is to model the interaction between the machine and the handled material, both the machine tool and the treated substance have to be modelled, which can be done in different ways. The DEM gives a good solution for modelling bulk materials, but there are multiple solutions for describing the tool of the machine depending on the needed information and computational efficiency. There are models that are based on purely the DEM (e.g. using zero thickness elements or cohesive interaction law), but the coupling of DEM and FEM is also a possible way.

The Use of Elements with Zero Thickness

There are two DEM element types, which are ideal for modelling boundary conditions and the geometry of tools.
The *wall* element is an infinite plane with zero thickness, which is usable to separate half-spaces. It has the same constitutional model as the volume elements have, so it is able to come into interaction with them, which arises e.g. repulsive and shear forces.
The *triangular facet* elements are defined by 3 points in 3D space. They have also zero thickness, and the capability to come into interaction with volume elements. In the right constellation, they can represent any surface, so they can be used to model machine tools with complex geometry. Such a geometry can be created by exporting parametric models in STL format using computer aided design software. Figure 1 shows a shaft represented by triangular facets.

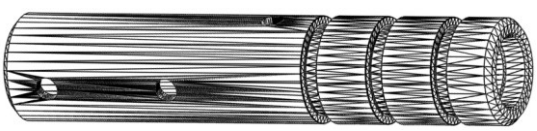


Figure 1: A shaft made up of triangular facet elements

Zero thickness elements model the tool as an ideally rigid body, so they can be used only if the deformations of the tool have so little influence that they can be neglected. Zero thickness elements give information about the load distribution affecting the tool.

The Use of Bonded Interactions

Besides the modelling of granular media, DEM is capable to simulate continua. It can be done by defining the interactions in the way that they have tensional resistance (cohesion). Figure 2 represents the uniaxial tension test of a material with cohesive constitutional model.

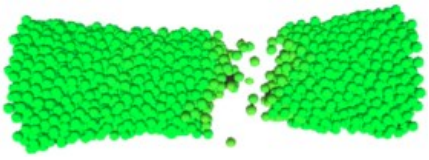


Figure 2: Tension test of a cohesive material made up of sphere elements (Šmilauer et al. 2015)

The advantage of modelling the tool with cohesive DEM elements is the potential to simulate deformations, cracks, breakage and wear. The stress field in the body is also computable by analysis of interaction forces. The disadvantage of such model is in the creation process, as finding the optimal constitutional model and calibrating its parameters takes considerable effort and time.

One-Way Coupling

To get the stress distribution and deformations of the tool, the FEM can be connected (coupled) with DEM. If the deformations of the tool are small, the so-called one-way coupling can be carried out. Where different data (e.g. force, velocity field) are computed via DEM are imported into the FEM simulation as a condition. It can be done with a special interface or a text file in the appropriate format. Such a series of simulations can be seen on Figure 3.

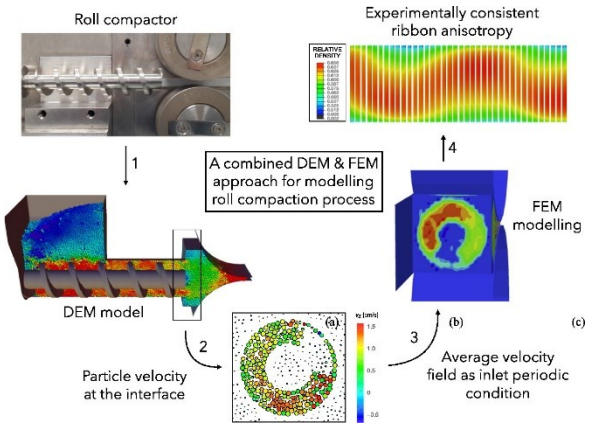


Figure 3: Flow Diagram of a Complex Roll Compactor Analysis (Mazor et al. 2017)

The DEM and FEM simulation are carried out in different software in most cases. The main task is to create the connection between the programmes and create the corresponding geometries. If the connection is established, the one-way DEM-FEM coupling provides an effective solution to find the stress distribution of a machine tool at small deformations with a low a computational demand.

Two Way (Parallel) Coupling

In reality, the interaction of the tool and the bulk material is dual. The moving tool shapes the bulk material, but this substance also deforms the tool. A deformed tool has slightly different effect on the material from the original one, which affects the forces on the tool etc. If the magnitude of this deformation is small, the one-way coupling can be used, else a new model is needed.
The solution is the creation of a two-way (parallel) simulation. In this case, the tool is made of a finite element mesh (Figure 4), which is able to establish contacts with discrete elements.

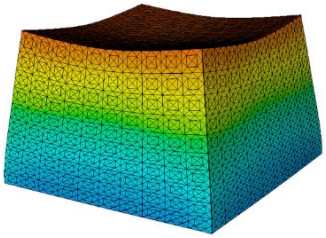


Figure 4: A geometry made of finite element mesh (Xiang et al. 2009)

The two-way coupling technique is also capable to create deformable elements by defining an internal finite element mesh. (Figure 5).

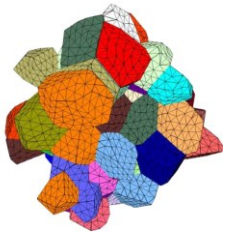


Figure 5: Polyhedron elements which contain internal finite element mesh for modelling crystalline solids (Li et al. 2017)

The method requires the interactive, real-time cooperation of a DEM and an FEM software, or a complex program, which have both DEM and FEM core. Due to the switching between the two models, iterations are needed where the conditions of the convergence are hard to define. For the same reason the simulations are computationally demanding, but in case of the proper definition, the solution will be accurate.

CREATING A ONE-WAY CONNECTION

The interaction of a bulk material (soil) and a tool (tine) was already studied and modelled in our research group with the use of zero-thickness element (Tamás et al. 2013), and the next step was to create a one-way coupled simulation.
There are commercial DEM and FEM software (e.g. EDEM and ANSYS) available that offer an interface to establish a connection, but in our study, the open source Yade (Šmilauer et al. 2015) was chosen as a DEM software to couple with ANSYS Workbench 18.2 FEM program. As this is a novel combination, a new way of connection had to be created.

DEM Model

In DEM simulations, the simplest and oldest element type is the sphere. It is widely used for modelling cohesive granular materials, like soil. However, more complex element shapes exist, for example polyhedra, which are excellent for modelling crushed rock aggregates (Orosz et al. 2017).
As there are several applications for one-way coupling, which uses spheres, the more uncommon polyhedral elements were chosen. They are randomly created based on Voronoi method (Asahina and Bolander 2011) with the desired size and shape (aspect ratio). The material of the particles is ideally rigid, and the stiffness of the real rocks is modelled with the model of the contact forces.
The constitutional model (Eliáš 2014) is cohesionless, (normal) compression and shear force is included. The normal force (F_n [N], Equation (1)) is linearly proportional with the common volume (V_c [m^3]) of the ideally rigid elements that come into contact where the factor k_n [N/m^3] is called the volumetric normal stiffness.

$$F_n = k_n V_c \tag{1}$$

The shear force (F_s [N], Equation (2)) is linearly proportional with the relative rotations and displacements (u_s [m]) of the elements, where the factor is the shear stiffness (k_s [N/m]). The value of the shear force is maximized by the coulomb friction law (Equation (3)) where φ [-] is the inter-particle friction coefficient.

$$F_s = k_s u_s \tag{2}$$

$$F_s \leq F_n tan\varphi \tag{3}$$

The model was tested with the modelling of a uniaxial compression test (Figure 6) in a former study (Orosz et al. 2017).

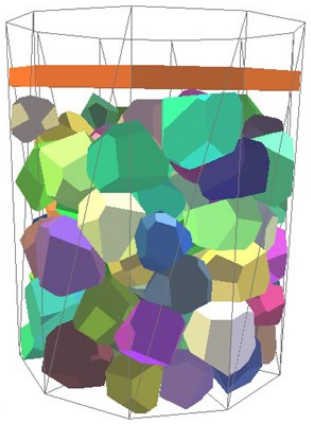


Figure 6: Uniaxial Compression Simulation of a Pack of Polyhedral Elements (Orosz et al. 2017)

Two geometrical models were created with the application of the polyhedral model using one rock: a press and a drop simulation.
In the case of the press test, a random polyhedron was created with a 100x100x100 mm bounding cube and a wall element was placed under it. Then gravity (9,81 m/s^2) was applied and the polyhedron fell onto the wall. The simulation was continued until equilibrium was reached. The wall was replaced with a plate made up from triangular facet elements. The polyhedron was compressed from upwards with another wall element (Figure 7) with a constant velocity until reaching the predetermined maximum normal force (50 kN), that was measured on the wall. At that point the force data were saved.

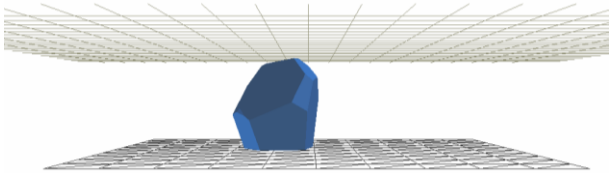


Figure 7: Compression Simulation of a Polyhedron

In the case of the drop test, the same, randomly created polyhedral element was falling onto the same plate under the influence of gravity (9,81 m/s^2) from 500 mm height and with an initial vertical velocity of 10 m/s (Figure 8). The forces were saved when the polyhedron was at its lowest position (maximum penetration into the triangular facets).

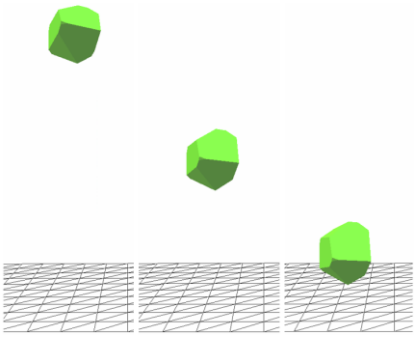


Figure 8: Free-Fall Simulation of a Polyhedron

FEM Model

The elaborated model was created in ANSYS Workbench V18.2. (ANSYS 2017). A simple prismatic body was used to model the ground which was contacted with the rock in the DEM simulation. During the FE analysis the same mesh was produced on the top surface of the body as in the DEM simulation. 10-nodes quadratic tetrahedron elements was used. A fix constrain was applied (which locked all of the 6 DoF) on the bottom surface of the solid body. Figure 9 shows the used geometry and the mesh during the FE simulation.

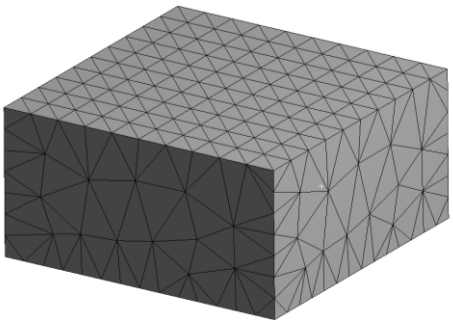


Figure 9: Geometry and the mesh used during the FE analysis

The Connection

A method was created to establish a one-way connection between Yade and ANSYS. The arising forces on the centre of mass of the triangular facets were saved in the appropriate time and were saved in a text file with a special format, containing their magnitude and coordinate of application. 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.

RESULTS

Press Test

The interactions and magnitude of normal forces between the polyhedron-upper plate (upper, red cylinder) and the polyhedron-triangular facets are shown on Figure 10 at the time of the maximum load.

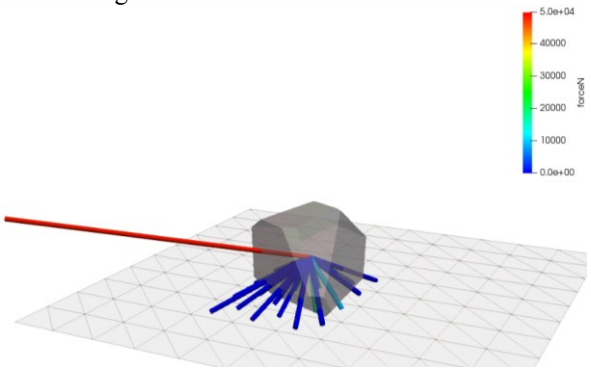


Figure 10: Normal Interactions During the Compression Simulation [N]

With the exporting of the forces acting on the triangular facets and mapping them onto the FE mesh, the Von Mises equivalent stresses and deformations can be computed (Figure 11). The effect of sharp edges and corners can be observed on the stress distribution of the prismatic body.

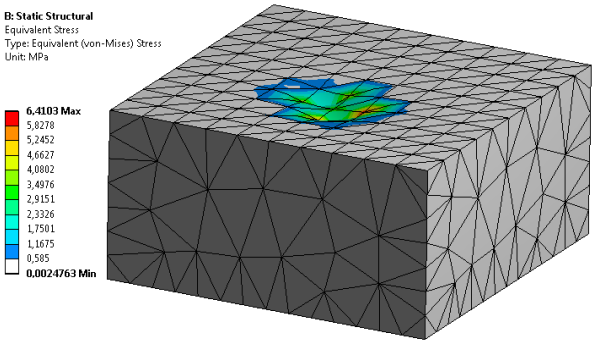


Figure 11: Von Mises Equivalent stresses [MPa] on Deformed Geometry (Deformation Scale 9100:1)

Drop Test

Figure 12 shows the interactions between the polyhedron and the triangular facets at the moment of maximum forces in case of drop test. The difference between magnitudes are easy to see, which is the result of that only one corner touches the surface of the simulated body.

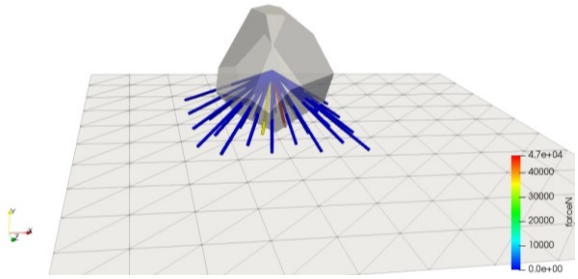


Figure 12: Normal Interactions During the Drop Simulation [N]

The effect of the sharp corner can be seen on the Von Mises stress field of the affected body (Figure 13) where the enlarged deformations also illustrate the influence of the single polyhedron.

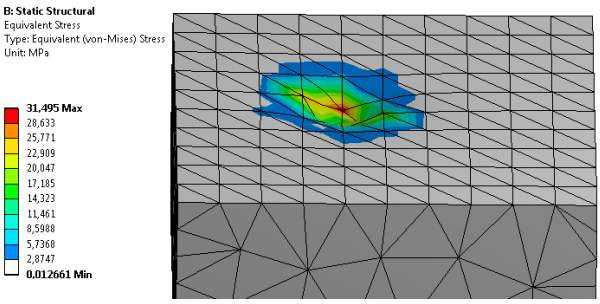


Figure 13: Von Mises Equivalent stresses [MPa] on Deformed Geometry (Deformation Scale 2000:1)

CONCLUSIONS

The introduced examples showed that there are many fields where machines interact with bulk materials and the modelling of these processes can help to improve the design and reduce cost and time.

The following conclusions were made about different techniques to model the tool of the machines:

- Zero thickness elements: fast, but only gives information about the bulk material and forces on the tool
- Cohesive constitutional: has many possibility, but hard to create
- One-way coupled simulation: gives information about stress and deformations. Fast, but only usable in small tool deformation range.
- Two-way coupling: is usable at large deformations also, but hard to define and execute due to the need of iterations.

The paper also showed that it is possible to connect a free source DEM (Yade) and a commercially available FEM software (ANSYS) in order to perform one-way coupled simulations, and a workaround was created.

The polyhedral elements are also capable for coupled simulations and can model particles that have sharp edges and corners effectively.

REFERENCES

ANSYS, Inc. 2017. "*ANSYS V18.2 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. *A diszkrét elemek módszere*. BME Department of Structural Mechanics, Budapest, 5-12.

Cundall, P.A. and D.H. Hart. 1992. "Numerical modelling of discontinua". *Journal of Engineering Computations* 9(2), 101-113.

Dill, E.H. 2012. *The Finite Element Method for Mechanics of Solids with ANSYS Applications*. CRC Press; Boca Raton, ISBN: 978-1-4398-4583-7

Eliáš, J. 2014. "Simulation of railway ballast using crushable polyhedral particles". *Powder Technology* 264, 458–465.

Li, X.F., H.B. Li, and J. Zhao. 2017. "3D polycrystalline discrete element method (3PDEM) for simulation of crack initiation and propagation in granular rock". *Computers and Geotechnics* 90, 96-112.

A. Mazor, L. Oreficeb, A. Michrafya, A. de Rycka and J. G. Khinast. 2017. "A combined DEM & FEM approach for modelling roll compaction process" *Powder Technology*, In press

Orosz, Á., J. P. Rádics and K. Tamás. 2017. "Calibration of railway ballast DEM model" In *Proceeding of the 31th European Conference on Modelling and Simulation 2017* (Budapest, Hungary. May. 23-26.). ECMS, 523-528.

Šmilauer, V. et al. 2015. "Yade Documentation 2nd ed." *The Yade Project* (http://yade-dem.org/doc/)

Tamás. K., I. J. Jóri and A. M. Mouazen. 2013. "Modelling soil–sweep interaction with discrete element method", *Soil & Tillage Research* 134. 223-231.

Xiang, J., Munjiza, A., and Latham, J. -P. 2009. "Finite strain, finite rotation quadratic tetrahedral element for the combined finite-discrete element method". *International Journal for Numerical Methods In Engineering*, 79(8), 946-978.

Zienkiewicz, O.C. 1971. *The Finite Element Method in engineering Science*. McGraw Hill, New York

AUTHOR BIOGRAPHIES

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. He is also a member of a research group in the field of discrete element modelling. His e-mail address is: orosz.akos@gt3.bme.hu and his web-page can be found at http://gt3.bme.hu/oroszakos.

KORNÉL TAMÁS is an assistant professor at Budapest University of Technology and Economics where he received his MSc degree and then completed his PhD degree. His professional field is the modelling of granular materials with the use of discrete element method (DEM). His e-mail address is: tamas.kornel@gt3.bme.hu and his web-page can be found at http://gt3.bme.hu/tamaskornel.

JÁNOS P. RÁDICS is an assistant professor at Budapest University of Technology and Economics where he received his MSc degree. He completed his PhD degree at Szent István University, Gödöllő. His main research is simulation of soil respiration after different tillage methods, and he also takes part in the DEM simulation research group of the department. His e-mail address is: radics.janos@gt3.bme.hu and his web-page can be found at http://gt3.bme.hu/radics.

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/zwierczykpetertamas.

COUPLED DEM-FEM SIMULATION ON MAIZE HARVESTING

Ádám Kovács and Péter Tamás Zwierczyk
Department of Machine and Product Design
Budapest University of Technology and Economics
H-1111, Budapest, Hungary
E-mail: kovacs.adam@gt3.bme.hu

KEYWORDS
Discrete element method, finite element method, maize harvesting, agricultural material.

ABSTRACT

One of the main objectives of today's agriculture is to develop agricultural machineries that guarantee more and better quality agricultural products. However, due to the seasonal characteristics of agricultural products in-situ tests of constructions are limited in time and often prove to be very expensive.
With more than 1000 billion tons of annual production, corn is one of the most essential agricultural crops in the world. Consequently, our study focuses on the modeling of corn harvesting through coupled discrete element and finite element method (DEM-FEM) to provide more input for machine development.
With the use of a complex DEM model of maize and the CAD model of a common corn header unit a large-scale simulation was carried out. The main phenomena were directly analyzed by quantitative and image-based qualitative evaluation methods. For further analysis the stalk rollers of the corn header unit were also investigated by finite element analysis.
The simulation results clearly demonstrate that the discrete element model of maize is capable of modeling the crucial phenomenon during corn harvesting in the future.

INTRODUCTION

The increasing demand for more and better quality agricultural products presents a big challenge for farmers, breeders and developers of agricultural machinery. Most of the agricultural processes are carried out by using different machineries, thus, one of the perpetual goal of precision agriculture is to advance these appliances. There are several different methods to improve agricultural machineries, yet this study discusses how their efficiency and working quality could be increased.
Due to the seasonal characteristics of agricultural products, in situ tests of constructions are limited in time and often prove to be very expensive. In the field of agricultural machine design numerical methods, which could properly replace field tests, are not available.
The utilization of corn plants and crops is remarkable worldwide; the corn production in the world is almost 1000 million tons annually. In 2016 almost 9 million tons of corn were harvested by farmers in Hungary (Hungarian Central Statistical Office 2018), which demonstrates the significance of the plant in the agriculture of the country.
During mature corn processing the first step is harvesting, during which combine harvesters with special corn headers gather the crop. The first time maize gets in contact with the machine is when the corn header collects the corn ears from the stalk. In this process the maize stalk is pulled down by the stalk rollers while it is chopped by the cutting blade of the chopping unit. The corn ear is separated from the stalk by the deck plates and is conveyed into the machine by the gathering chains. To improve working efficiency and quality of corn headers processing of stalk parts under the corn ear is the most crucial factor because these parts of the plant provide the highest resistance against external loads and their dry matter content is higher than in the upper parts of the plant.
Therefore, our study focuses on the modeling of corn harvesting by coupled discrete element and finite element method (DEM-FEM) to provide input for machine development.
DEM is widely used to investigate bulk agricultural materials. The micromechanical parameters of a sunflower DEM model were calibrated based on odometer tests so that the model can sufficiently approach the macro mechanical behavior of the real bulk material (Keppler et al. 2011). In another study the effect of the particle shape of corn kernels on flow was investigated by discrete element method simulation of a rotary batch seed coater (Pasha et al. 2016).
In connection with entire plant and corn ear modeling fewer studies can be found. The interaction among grass stalk and rotation mower was investigated by DEM in Kemper et al. (2014). To develop a sensor for monitoring rice grain sieve losses in a combine harvester a special hollow structure was created about rice stems by Liang et al., (2016). A special solid geometrical structure of DEM was analyzed for corn stalks in quantitative and qualitative ways in Kovács et al. (2015). Several possible DEM geometrical structures for modelling of fibrous agricultural materials were compared in Kovács and Kerényi (2016). A DEM model of corn stalks and corn ears was developed for analysis of losses during corn harvesting in Kovács and Kerényi (2017).
Finite element analysis is a more common method to predict the mechanical behavior of machines, thus, several studies can be found in connection to structural analysis of agricultural machinery. Using finite element method (FEM) model analysis on a V-belt pulley of a fodder crushing machine was carried out by Celik et al. (2010). Stresses and displacements of a coffee harvester

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

structure (engine frame, body right and left sides, front and rear end, main beam, coffee reservoir, wheels and fuel tank) were analyzed by using FEM static simulation in da Silva et al. (2014).
Consequently, based on the literature review there is no numerical method that can provide useful input for machine development. Therefore, this paper analyzes interactions among parts of the machine and maize, external loads on the parts of the machine and working quality by a coupled DEM-FEM simulation method during corn harvesting.

MATERIALS AND METHODS

Discrete element method (DEM) was developed to investigate bulk materials which contain separate parts. The DEM model is defined as follows: it contains separated, discrete particles which have independent degrees of freedom and the model can simulate the finite rotations and translations, connections can break, and new connections can come about in the model (Cundall and Hart 1993). Thus, this method enables the investigation of the mechanical beahviour and breaking phenomenon of solid materials under different loading cases.
Based on the harvesting processes of harvest-ready maize the main loads of the stalk and maize ear were determined. Tassel, leaves and husk of the plant were neglected in our study. First, the physical and physiological properties of the plant parts (mass, length, diameter, shape, position, center of mass) were measured and observed. Laboratorial tests (compression, bending, cutting test on the stalk and ear detachment) were conducted to define the main mechanical parameters and the behavior of maize. The results of the measures weren't directly usable for the modeling method so necessary data and graphs were calculated with mathematical and statistical and image processing methods for the numerical modeling.
In the next step, the DEM physical geometry of maize was created to calibrate its mechanical properties based on the experimental results. With modifications of the micro-mechanical parameters of the contact models, during an iteration process, the right assembly was found. Based on observations of 13 commercial maize heads the CAD geometry of a common maize header unit was designed and imported into the discrete element software (EDEM®, DEM Solutions Ltd., Edinburgh, UK), where its kinematics was also defined.
By using the whole plant DEM model and the model of the maize header unit a large-scale simulation was carried out in EDEM® software. The main phenomena were directly analyzed by quantitative and image-based qualitative evaluation methods. For further analysis the stalk rollers of the maize header unit were also investigated by finite element software (Ansys®, Ansys Inc., Canonsburg, Pennsylvania, US) based on the DEM predicted external loads.

MODEL FORMATION

During the model formation hollow, solid and chain of spheres geometrical structures were used to create each part of the model. Based on the importance and role of maize parts during harvesting, their geometrical structure is more or less detailed in our model, as shown on Figure 1. Section “A” contains the most important parts of the stalk and the maize ear with the shank. Here, the internodes and nodes provide the highest resistance against external loads, their bending behavior plays an important role in losses. Another important part of the stalk is the shank that holds the maize ear thus it has a significant effect on gathering. In section “B”, parts don’t play an important role in losses but they provide significant resistance during processing. In section “C”, stalk parts provide very low resistance against external loads. In the stalk model, all cross sections of nodes and internodes are circular while special traits (groove, wrinkle) and core of internodes were neglected.

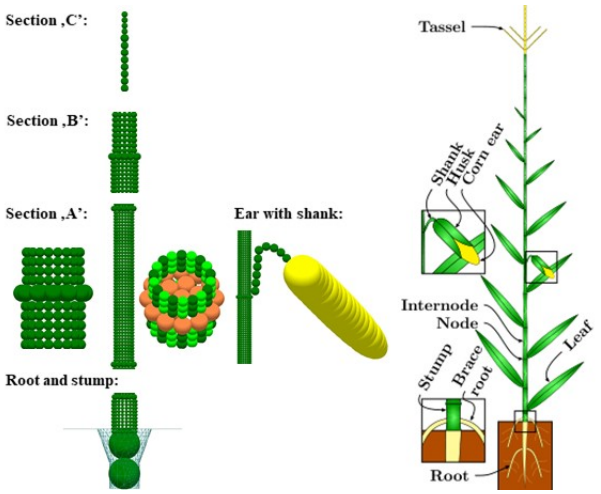


Figure 1: DEM geometrical model for maize

The root and soil-root connection was approached with chain of spheres geometrical model that is situated in a conical geometry to provide relaxation against external bending loads. The stump is modelled with a hollow structure with 18 particles in one cross section.
From the 1st to the 10th, nodes were modelled by a solid geometrical structure in which particles were composed in one circular layer of particles that is made up of 19 particles: 12 in the outer circle, 6 in the middle circle and 1 in the center. This model provides higher resistance against external loading near the nodes. Above the 10th node, nodes were modelled by larger spherical particles.
From the 1st to the 6th, internodes were modelled with hollow geometrical structure. In a cross section of internode, particles were composed in one circular layer that is made up of 18 particles. This model provides a good opportunity to model the typical failure modes (buckling and ovalisation) of internodes. Between the 7th and 10th, internodes were modelled with hollow geometrical structure but there are 12 particles in each

circular layer. Above 10th, internodes were created with chain of spheres geometrical model.
The shape of the shank was formed in such a way that it can carry a hanging ear by using the chain of spheres model. It is impossible to model this broken condition of the shank so a curved shape with unbroken bonds was created.
The geometrical model of the maize ear is one particle that is formed by several sphere surfaces. The previously described ideal shape of the maize ears was approached with 25 sphere surfaces, supposing that the maize ears are axis symmetric. This detailed maize ear model provides a good opportunity to analyze the interaction between the maize ear and parts of the machine. The maize ear was situated in such a way that its center of the mass was near the same as the results from the measurements.
The examination of the available contact models was the first step during the calibration of mechanical behavior of different parts of the model. After that the models were compared and the Timoshenko-Beam-Bonded model (Brown et al. 2014), which is based on the Timoshenko-beam theory, was selected for the study.
In the chain of spheres geometrical structure three different contacts could be defined among the particles (Figure 2, a). The red line marks the connection between particle type one, which formed the node of the stalk, and particle type three, which is the first particle of the internode. Among the particles of the internode (particle type three) the same connections can be observed, on the figure these are marked by pink lines. The blue line marks the connection between the last particle of the internode and the particle of the next node. These types of connections have different mechanical parameters that can model the mechanical characteristic of the stalk in axial direction. In the model of the chain of spheres the mechanical properties of the stalk in tangential direction can be described by the stiffness of the different particle types.
In the hollow geometrical structure, the mechanical properties of the internode were modeled with the connections among the same type of particles (P4:P4; P5:P5) in axial direction and were modeled with the connections among the even-numbered and the odd-numbered type of particles (P4:P5) in tangential direction. The mechanical properties among the particles of the node and the internode were defined through P1:P5; P2:P4; P7:P5; P6:P4 in axial direction and the mechanical properties of the node were defined through the P1:P2; P6:P7 connections in tangential direction (Figure 2, b).
The imported CAD geometry of a maize header unit is shown on Figure 3. Each part was designed based on the parts of a real machine, however, the entire design is not identical with any of them. The kinematics of the parts was also determined based on real operation parameters of a maize header; thus, the pace of the unit is 2 m s^{-1}; the speed of the gathering chains is 4 m s^{-1}; the rotational speed of the stalk rollers and chopping unit are 400 rpm and 3500 rpm, respectively.

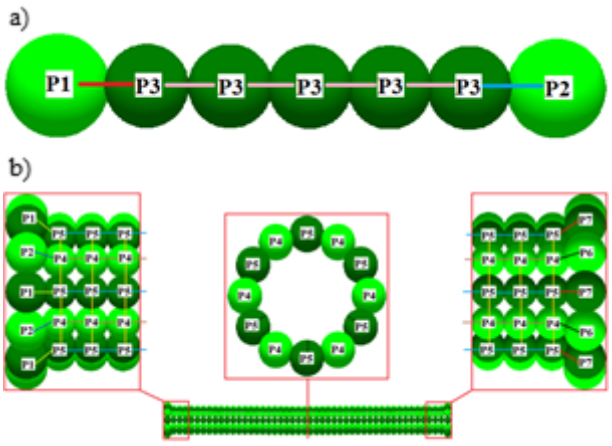


Figure 2: Geometrical and contact structure of the stalk: a) chain of spheres in Section C; b) hollow geometrical structure in Section A and B

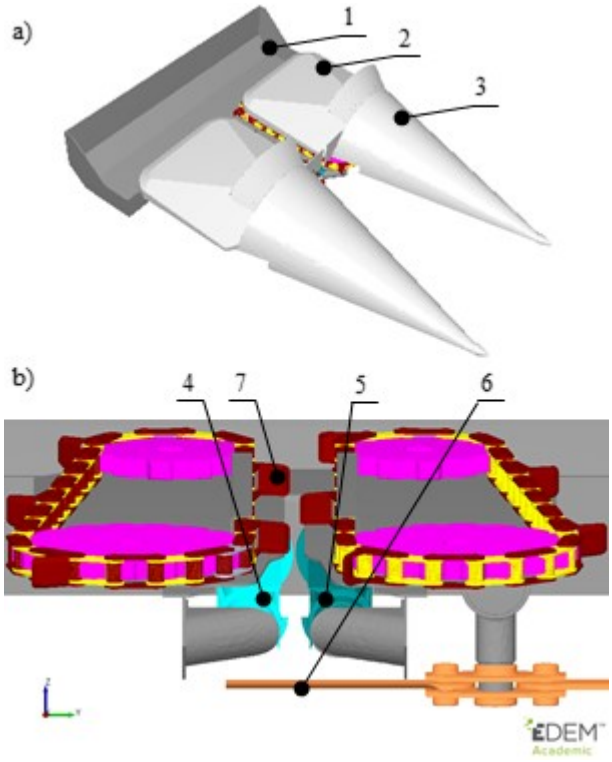


Figure 3: Imported CAD model of a common maize head: 1. gathering box; 2. hood; 3. row divider; 4. right stalk roller; 5. left stalk roller; 6. chopping blade; 7. gathering chains

RESULTS

In this study DEM predicted external loads and working quality of stalk rollers and chopping unit were analyzed in the time interval of the first contact between the header and the stalk and the ear-detachment. For this analysis quantitative and image-based qualitative methods were used.

Quantitative results from DEM analysis

To improve the energy efficiency of a maize header one of the most important aspects is the power requirement of each unit. In case of a maize header the stalk rollers and the chopping unit provide the highest power consumption, thus, results on these parts will be explained here.

While the stalk rollers are spinning in opposite directions their knifes compress the maize stalk locally and pull it down. During the compression the stalk provides high resistance that appears as a torques on the stalk rollers, it is shown on Figure 4. At the beginning of the process, when the stalk is virgin, an extremely high torque-peak (~180 Nm) is observable because of the supporting effect of the roots. In this case the stalk is compressed locally in its radial direction and in its longitudinal direction as well, that causes that special, initial loading case of the stalk rollers. After the first cut the stalk lost its connection to the ground and the peaks are more similar. In the steady state the peaks are between 40 and 70 Nm, while the mean torque on each stalk roller is 23.3 Nm. Except one peak torque on the left stalk roller around its third rotation the characteristics are nearly the same on both stalk rollers.

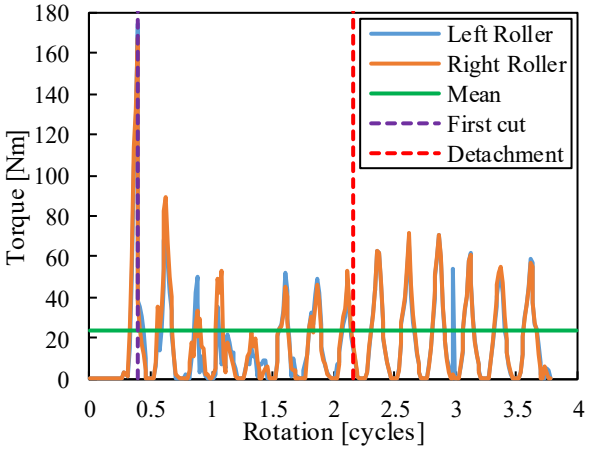


Figure 4: Plot of the external torque on the axes of each stalk roller vs its rotation in cycles with highlights of the first cut and ear detachment during simulation of maize harvesting

Based on the external torque and the rotational speed of the stalk rollers the total power requirement was calculated as shown on Figure 5. In the steady state the peaks are in the range of 2.5-5.5 kW, while the mean required power was 1.9 kW that corresponds to the power requirement of a real maize header unit.

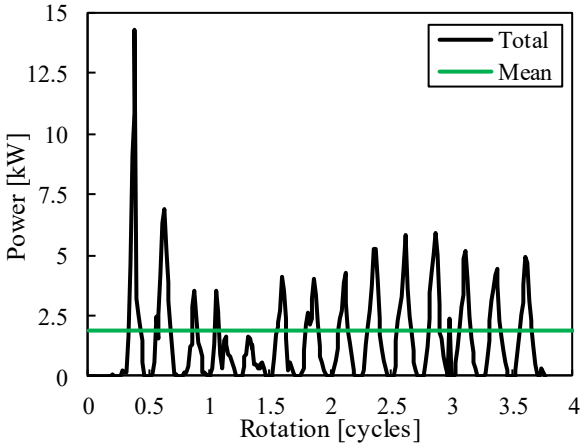


Figure 5: Plot of the total power requirement of stalk rollers vs rotation in cycles during simulation of maize harvesting

While the stalk rollers are pulling down the stalk the chopping unit cuts it into small pieces. During this process the stalk provides resistance against the chopping blade and a resistance torque can be calculated on the axis of the chopping unit, as shown on Figure 6. The range of the appeared peaks is between 8 and 230 Nm. The high peaks can be caused by a numerical error. By decreasing the time-step of the DEM simulation these extremely high peaks will disappear. The calculated mean torque on the chopping unit was 11.8 Nm during the analyzed period.

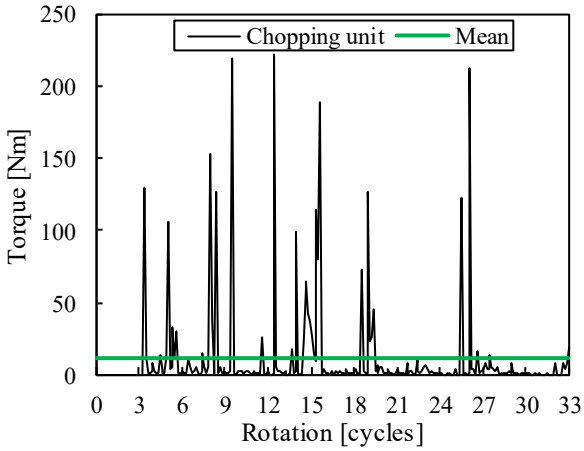


Figure 6: Plot of the external torque on the axis of the chopping blade vs its rotation in cycles during maize harvesting

Similarly to the stalk rollers, the power requirement of the chopping unite was calculated, see on Figure 7. The extremely high peaks caused extremely high power requirement for a short time interval but the calculated mean power requirement was 4.3 kW that corresponds to the power requirement of a real maize header unit.

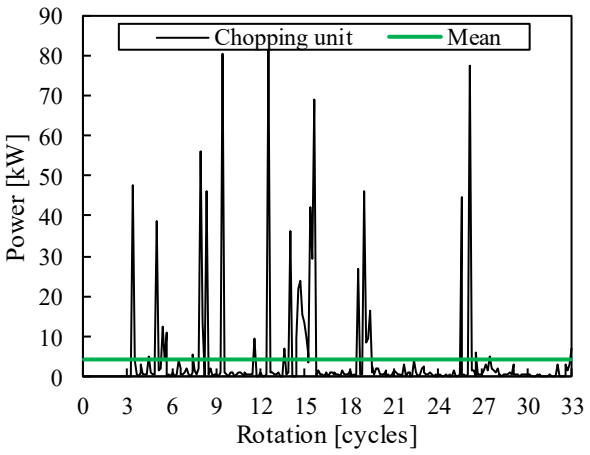


Figure 7: Plot of the power requirement of the chopping unit vs its rotation in cycles during maize harvesting

Image-based qualitative results from DEM analysis

To improve the working quality of a maize header it is important to analyze every phenomenon during maize harvesting. Here, an image-based analysis on the working process of stalk rollers and the chopping unit is shortly explained.

The most important function of the stalk rollers is to fix the position of the stalk during the chopping and ear-detachment. On Figure 8, the main stages of the interaction among the stalk and stalk rollers can be observed: approaching, maximal compression and distancing. There is a short time interval between the distancing and approaching when there is no interaction among the stalk rollers and the stalk. To know what happens in this time interval can be significant for development of the number and shape of the knives.

The phenomenon of the first cut provides very important information about the speed of the maize header unit; the speed ratios and positions of stalk rollers and chopping unit, see on Figure 9. This is an optimal case because the stalk rollers fix the stalk before the first cut, thus, the chance of losing the maize ear is less.

To analyze the working quality of the chopping unit the size distribution and shape of the chopped material are usable. On Figure 10, two types of chopped material are observable: fibrous and cylindrical. In both cases the DEM predicted shape and the real one is very similar. For the fibrous chopped material the stalk was torn up during cutting or the chopping blade hit it several times, while the cylindrical chopped material is the result of two perfect cut on the stalk.

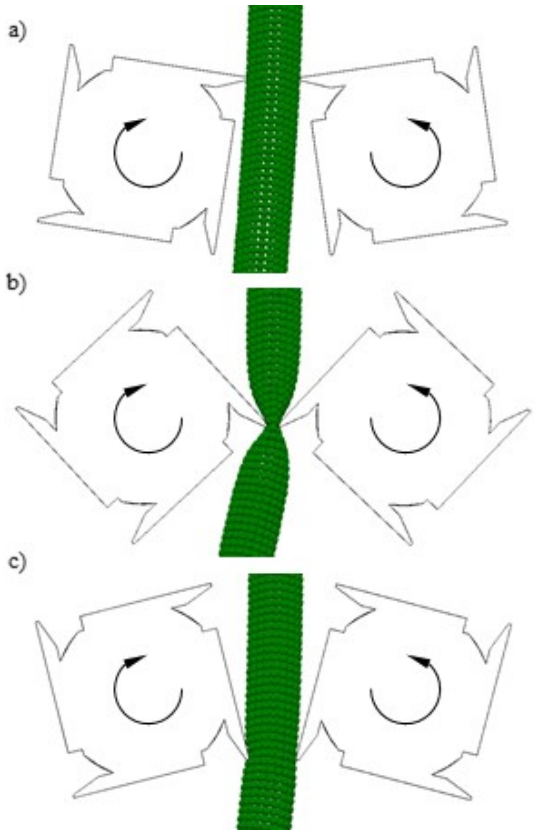


Figure 8: Image-based qualitative analysis on the working efficiency of the stalk rollers during simulation of maize harvesting: a) approaching; b) maximal compression; c) distancing

Figure 9: Image-based qualitative analysis on the phenomena of the first cut during simulation of maize harvesting

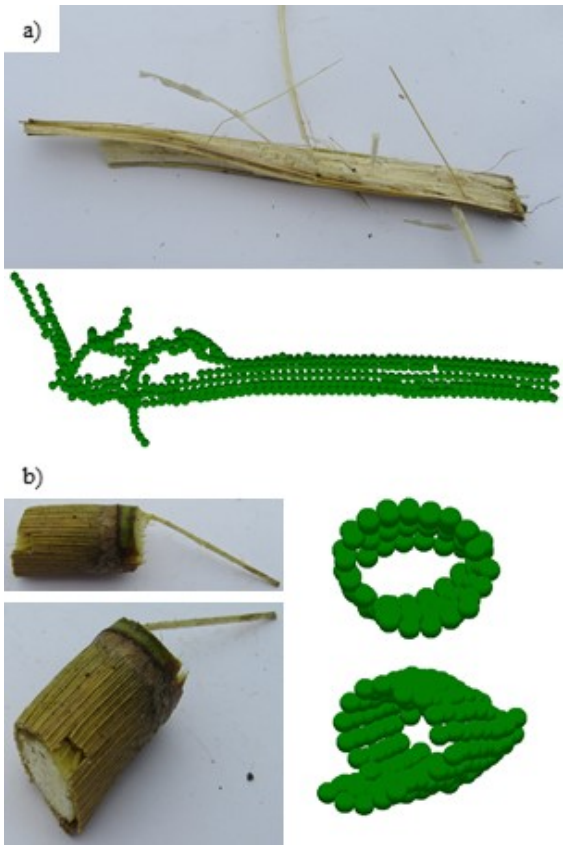


Figure 10: Image-based qualitative analysis on the chopping material during simulation of maize harvesting: a) fibrous chopped material; b) cylindrical chopped material

Quantitative results from FEM analysis

Through the FE analysis the left stalk roller was examined in case of the first cut when the external torque on the rollers reaches its maximum (see Figure 4).
During the simulation the same mesh configuration was used as in the DEM analysis. The FE mesh consisted of 10-nodes tetrahedron elements with 2 mm size. The material of the stalk roller and its blades was considered to be structural steel. The modulus of elasticity was 200 GPa, the Poisson's ratio was 0.3. Because the analysis only focused on one time moment, when the external torque is at its maximum, two fixed constrains were applied both sides of the roller where the hydro drives connecting to. The load imported from the DEM solution as nodal forces using the EDEM® – Ansys® Add-In.
The resulted von Mises equivalent stress can be seen in Figure 11. Comparing the FEM results with the DEM results (e.g. Figure 8 and 9) it can be concluded that the maize branch when it is contacting with the stalk roller causes stresses on the roller's blade. Furthermore, it is also visible in Figure 11 that the branch is pressing not only the edge of the blade but also the top surface of it while the roller's blade is cutting.

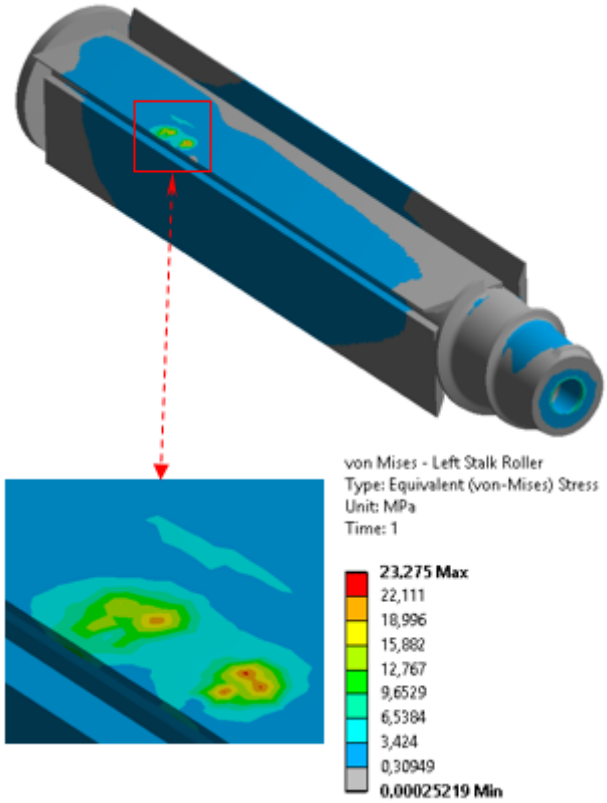


Figure 11: Image-based qualitative analysis of the stresses on the left stalk roller of the first cut during simulation of maize harvesting

CONCLUSIONS

A numerical and experimental study of maize harvesting was undertaken using coupled discrete and finite element method in order to simulate the interactions among the parts of a common maize header unit and a maize stalk.
The course of harvesting was analyzed through quantitative and image-based qualitative evaluation methods: the external torque; the power requirement and the working quality of stalk rollers and chopping blade were analyzed in detail. To provide data for structural design of stalk rollers they were also analyzed by finite element method.
Based on our results the following conclusions could be drawn:

(1) The applied coupled DEM-FEM simulation method is suitable to provide input for development of maize headers.
(2) The geometrical model for maize is usable to analyze the interactions among the maize header unit and the maize stalk during maize harvesting.
(3) The CAD model about a common maize header unit is usable for further simulations on maize harvesting.
(4) The quantitative evaluation method is directly usable to analyze the external loads on different parts of the maize header, thus, to provide input data for design.

(5) The qualitative evaluation method is usable to analyze the working quality of different parts of a maize header, thus, to provide input data for design.
(6) By using coupled DEM-FEM simulations it is possible to examine the stress distribution during the maize harvesting not just on the stalk rollers but in the future also on the cutting blades.

In the future, current results and models can be adapted to more detailed and realistic simulations on maize harvesting process. These techniques can be extended with wear simulations. Together with these simulations the whole harvesting process can be examined not just on the agricultural side but also on the full mechanical engineering side.

ACKNOWLEDGEMENTS

The authors thank Dr. Gábor Szebényi from the Department of Polymer Engineering, Dr. Imre Orbulov from the Department of Material Science and Engineering, Szabolcs Berezvai from the Department of Applied Mechanics and Dr. Tibor Poós from the Department of Building Services and Process Engineering at BUTE for the help they provided in the laboratorial tests.

REFERENCES

Brown N.J.; J-F. Chen; J. Y. Ooi. 2014. "A bond model for DEM simulation of cementitious materials and deformable structures." *Granular Matter*, No. (2014) 16, 299–311.

Celik H.K.; M. Topakci; M. Canakci; A. E. W. Rennie; I. Akinci. 2010. "Modal analysis of agricultural mechineries using finite element method: A case study for a V-belt pulley of a fodder crushing machine." *Journal of Food, Agriculture & Environment*, No. 8, 439-446.

Cundall P.A.; R.D. Hart. 1993. "Numerical Modeling of Discontinua." *Analysis and Design Methods*, No. 1993, 231-243.

da Silva E.P., F.M. de Silva, R.R. Magalhaes. 2014. "Application of Finite elements method for structural analysis in a coffee harvester." *Engineering*, Vol. 6, No. 3, 138-147.

Kemper S.; T. Lang; L. Frerichs. 2014. "The overlaid cut in a disc mower - results from field tests and simulation." *Landtechnik*, No. 69(4), 171-175.

Keppler, I.; L. Kocsis; I. Oldal; A. Csatár. 2011. „Determination of the discrete element model parameters of granular materials." *Hungarian Agricultural Engineering*, No. 23/2011, 30-32.

Kovács Á.; K. Kotrocz; Gy. Kerényi. 2015. "The adaptability of discrete element method (DEM) in agricultural machine design." *Hungarian Agricultural Engineering*, No. 27, 14-19.

Kovács Á.; GY. Kerényi. 2016. "Comparative analysis of different geometrical structures of discrete element method (DEM) for fibrous agricultural materials." In *4th CIGR International Conference of Agricultural Engineering* (Aarhus, Denmark, June 26-29), 1-8.

Kovács Á.; GY. Kerényi. 2017. "Modeling of corn ears by discrete element method (DEM)." In *31st European Conference on Modelling and Simulation* (Budapest, Hungary, May 23-26 2017.

Liang Z.; Y. Li; L. Xu; Z. Zhao. 2016. "Sensor for monitoring rice grain sieve losses in combine harvesters." *Biosystems engineering*, No. 147, 51-66.

Pasha M.; C. Hare; M. Ghadiri; A. Gunadi; P. M. Piccione. 2016. "Effect of particle shape on flow in discrete element method simulation of a rotary batch seed coater." *Powder Technology*, Volume 296, 29–36.

AUTHOR BIOGRAPHIES

ÁDÁM KOVÁCS was born in Debrecen, Hungary and went to Budapest University of Technology and Economics, where he studied agricultural machine design and obtained his MSc. degree in 2016. Currently he is a PhD student in the same institution and his topic is discrete element modeling of maize. He worked for the WIGNER Research Center for Physics at the Department of Plasma Physics for two years, where he designed diagnostic devices for fusion reactors. His e-mail address is: kovacs.adam@gt3.bme.hu and his Web-page can be found at http://gt3.bme.hu/en.

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/zwierczykpetertamas.

INVESTIGATION THE EFFECT OF THE MODEL DIMENSION IN SOIL-CONE PENETROMETER DISCRETE ELEMENT SIMULATIONS

Krisztián Kotrocz
György Kerényi
Department of Machine and Product Design
Budapest University of Technology and Economics
H-1111, Muegyetem rkp. 3-9., Budapest, Hungary
E-mail: kotrocz.krisztian@gt3.bme.hu

KEYWORDS
soil model, cone penetrometer, 3D DEM, discrete element simulation.

ABSTRACT

One of the most common methods to measure soil strength in-situ is by cone penetrometer. This paper is about the development of a 3 dimensional (3D) discrete element model (DEM) for simulation of the soil-cone penetrometer interaction. To simulate cohesive soil the Hertz-Mindlin with bonding contact model was used in EDEM 2.7 discrete element software. The parameters were calibrated to real test result and after that the effect of the model dimension (namely the diameter of the cylinder) on soil penetration resistance was analysed. The results of the calculations were very similar to the real soil penetration resistances therefore the model can be used to simulate the cone penetrometer in-situ test. Another result of the paper that a suggestion for proper model size is given in soil-cone penetrometer discrete element simulations.

INTRODUCTION

To measure soil's penetration resistance during vertical load, cone penetrometers are widely applied (McKyes 1985 and Laib 2002). With this equipment the normal force can be measured while the cone is pressed into the soil vertically. The cone index (C. I.) can be calculated thereafter by dividing the measured vertical force value with the projected area of the cone. Then the cone index can be used to compare the individual soils by their bearing capacity.

Thanks to the evolution of the information technology in the 20th-21st century numerical simulations can be used to investigate real materials virtually. Earlier researches showed that cohesive soil can be modelled properly using the Discrete Element Method (DEM) (Chen et al 2013, Tamás et al 2013) which was invented by Cundall and Strack in 1979 (Cundall and Strack 1979). To our best knowledge limited papers were published in subject of investigation of the soil-cone penetrometer interaction with DEM. One of the exceptions is the work of Tanaka et al (Tanaka et al 2000) who developed a 2D discrete element model to simulate this phenomena for non-cohesive soil.

The main difficulties of the discrete element simulations are the calibration of the contact properties to real, measurable soil mechanical properties and to choose a correct model size where the effect of the model boundary walls on simulation results is negligible.

In this paper the EDEM 2.7 Academic software, available from DEM Solutions Ltd. was used to simulate cohesive soil's behaviour during cone penetration test. The aim of the paper is to investigate the effect of model size (namely the diameter of the soil model) on penetration resistance. From the results the correct model diameter can be choosen for simulation of a soil-cone penetrometer interaction.

MATERIALS AND METHODS

To simulate soil-cone penetrometer interaction with DEM the whole penetration process has to be divided into small timestep of dt. In every timestep the individual displacements of each particle can be calculated using Newton's 2nd law which needs the values of the forces and moments acting on the given elements. The software calculates these values with the help of the contact models which define the behaviour of the contacts between two or more particles. Therefore the contact models play an important role in every discrete element simulations.

In our simulations the Hertz-Mindlin with bonding model were used to simulate cohesive material which is based on the work of Hertz (Hertz 1882), Mindlin (Mindlin 1949), Mindlin and Deresiewicz (Mindlin and Deresiewicz 1953) and Potyondy and Cundall (Potyondy and Cundall 2004). In this contact model the contact normal and shear directions are individual, the forces and moments (as well as the damping forces and moments) are calculated according to the normal and shear directions of the contact coordinate system. To calculate the contact normal force the normal overlap of the elements (δ_n), the equivalent Young's modulus (E^*) and – radius (R^*) should be known. These parameters can be calculated with the following formulas where E, υ and R are the Young's modulus, the Poisson-ratio and the radius of the elements, respectively and the indexes represents the contacting particles as 1 and 2:

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

$$E^* = \left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \right)^{-1}$$

$$R^* = \left(\frac{1}{R_1} + \frac{1}{R_2} \right)^{-1}.$$

If the parameters above are available the contact normal force can be determined with Equation (1):

$$F_n = \frac{4}{3} \cdot E^* \cdot \sqrt{R^*} \cdot \delta_n^{\frac{3}{2}} \qquad (1)$$

In shear (or tangential) direction the contact force can be calculated from the tangential overlap of the elements (δ_t):

$$F_t = -S_t \cdot \delta_t$$

where S_t is the tangential stiffness which can be determined with the equivalent shear modulus (G^*), equivalent radius and the normal overlap of the elements using the following formulas:

$$S_t = 8 \cdot G^* \cdot \sqrt{R^* \cdot \delta_n}$$

$$G^* = \left(\frac{1-\nu_1}{G_1} + \frac{1-\nu_2}{G_2} \right) \qquad (2)$$

In Equivation (2) the G_1 and G_2 are the shear modulus of the contacting elements nr. 1 and nr. 2, respectively. However, the contact tangential force has a limit, namely the value of the force can not be higher than the maximum which can be calculated from the Coulomb friction:

$$F_t \le F_n \cdot \mu_s$$

where μ_s is the static frictional coefficient between the elements.

In the Hertz-Mindlin with bonding contact model the damping effect can be handled as a damping force in contact normal (F_n^d) and tangential direction (F_t^d) as well.

$$F_n^d = -2 \cdot \sqrt{\frac{5}{6}} \cdot \beta \cdot \sqrt{S_n \cdot m^*} \cdot v_n^{rel}$$

$$F_t^d = -2 \cdot \sqrt{\frac{5}{6}} \cdot \beta \cdot \sqrt{S_t \cdot m^*} \cdot v_t^{rel}$$

The equivalent mass (m^*) can be determined from the mass of the two contacting particles (m_1 and m_2):

$$m^* = \left(\frac{1}{m_1} + \frac{1}{m_2} \right)^{-1}$$

v_n^{rel} and v_t^{rel} are the relative normal and tangential velocities of the contacting elements, respectively and the normal stiffness (S_n) and β can be determined as:

$$S_n = 2 \cdot E^* \cdot \sqrt{R^* \cdot \delta_n}$$

$$\beta = \frac{\ln e}{\sqrt{\ln^2 e + \pi^2}}$$

where e is the so-called coefficient of restitution.

Another important contact parameter in the Hertz-Mindlin with bonding model is the contact radius ($R_{contact}$) which determines the boundary of the particle's contacting volume. So if the contact radius is set up greater than the element's radius the particle can be in contact numerically with another one without being in contact with it physically.

To simulate cohesive behaviour of the soil the so-called Parallel Bonds are added to the default Hertz-Mindlin with bonding contact model. So the forces and moments calculated with Equation (3-6) are summed to the corresponding normal and tangential components:

$$\Delta F_n = -S_n^B \cdot A \cdot \Delta\delta_n \qquad (3)$$

$$\Delta F_t = -S_t^B \cdot A \cdot \Delta\delta_t \qquad (4)$$

$$\Delta M_n = -S_n^B \cdot J \cdot \Delta\theta_n \qquad (5)$$

$$\Delta M_t = -S_t^B \cdot \frac{J}{2} \cdot \Delta\theta_t \qquad (6)$$

The area and the polar moments of inertia of the bond's cross section can be determined as:

$$A = \pi \cdot R_B^2$$

$$J = \frac{1}{2} \cdot \pi \cdot R_B^4$$

where R_B is the radius of the bond. In addition the S_n^B and S_t^B are the bond normal- and tangential stiffness, $\Delta\delta_n$ and $\Delta\delta_t$ are the relative normal- and tangential displacements, $\Delta\Theta_n$ and $\Delta\Theta_t$ are the relative normal- and tangential rotations of the contacting elements, respectively. These parameters are calculated at the bond formation time (t_{Bond}) when the Parallel bonds are added to the model. These bonds can break if the value of the normal or tangential bond stresses become higher than their limits (the so-called bond normal and tangential strength):

$$\sigma_{\max} = \frac{-\Delta F_n}{A} + \frac{2 \cdot M_t}{J} \cdot R_B$$

$$\tau_{\max} = \frac{-\Delta F_n}{A} + \frac{M_n}{J} \cdot R_B.$$

The in-situ cone penetrometer tests

Cone penetrometer measurements were conducted near Mohács, Hungary in the October of 2015 using the Eijkelkamp penetrologger can be seen in the following figure.

Figure 1: The used Eijkelkamp cone penetrometer with the additional accessories

The cone bevel angle was 60°and the projected area was 1e-4 m^2. 10 measurements were performed by pressing the penetrometer cone into the soil with average velocity of 1e-2 $m \cdot s^{-1}$. During the test the resistance of the soil (e. g. the normal force) were collected in every depth of 1e-2 m. Finally the average penetration resistance were calculated by averaging the 10 individual force values in each depth. The results of the test can be seen in the Figure (2). According to figure the soil penetration resistance has a high standard deviation which can be experienced because of the relative small projected area of the cone (Laib 2002).

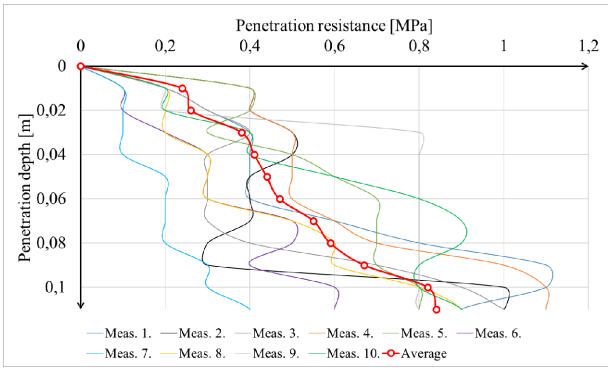


Figure 2: Results of the in-situ penetration test

Development of the 3D discrete element model

A 3D discrete element model was developed to simulate cone penetrometer in-situ tests of cohesive soil. First the initial geometries of soil models were created by allowing the elements to fall down under Earth gravity. To investigate the effect of the model size on penetration resistance five individual geometries (with five individual diameters) of soil models were generated. The parameters of the models were detailed in the Table 1, where the so-called area-ratio can be calculated by dividing the area of the soil model with the area of the penetrometer cone (1e-4m^2):

$$\text{Area ratio}\,(-) = \frac{A_{\text{soil model}}}{A_{\text{cone}}} = \frac{A_{\text{soil model}}}{1e-4}$$

The height of the models was 1.15e-2 m in each simulation.

Table 1: The geometrical properties of the individual discrete element soil models

Diameter of the soil model cross section (m)	*Area ratio (-)*
Ø6.77e-2	36
Ø9.03e-2	64
Ø1.128e-1	100
Ø1.354e-1	144
Ø1.58e-1	196

After the particles' maximum velocity become small enough ($v_{max} \leq$1e-4 $m \cdot s^{-1}$) the Parallel Bonds were added and the same geometry of penetrometer cone to that of used in the in-situ test was imported to the model from an .stl file. The contact properties of the Hertz-Mindlin with bonding model were chosen according to our latest research (Kotrocz and Kerényi 2017) where the direct shear test of cohesive soil were modelled with the discrete element method. Only the value of the bond radius was set up higher to reach closer results to the penetration resistance of the in-situ tests. The contact parameters were detailed in Table 2.

Table 2: The contact properties of the discrete element model

Parameter	Value
Geometrical properties	
Particle radius distribution (m)	2e-03…4.5e-03
Contact radius ($R_{contact}$) (m)	2.67e-03…6e-03
Properties of the Herzt-Mindlin with bonding contact model	
Bulk density (kg m^{-3})	1800
Shear modulus (Pa)	2.88e+06
Poisson ratio (-)	0.3
coefficient of restitution (e) (-)	0.5
Friction coefficient between ball and ball (μ_{ball}) (-)	0.5
Friction coefficient between ball and walls (μ) (-)	0.5
Bond radius (R_B) (m)	3e-03
Bond normal stiffness (S_n^B) (Pa m^{-1})	1.2e+07
Bond shear stiffness (S_t^B) (Pa m^{-1})	1.2e+07
Bond normal strength (Pa)	7.738e+4
Bond shear strength (Pa)	3.869e+4

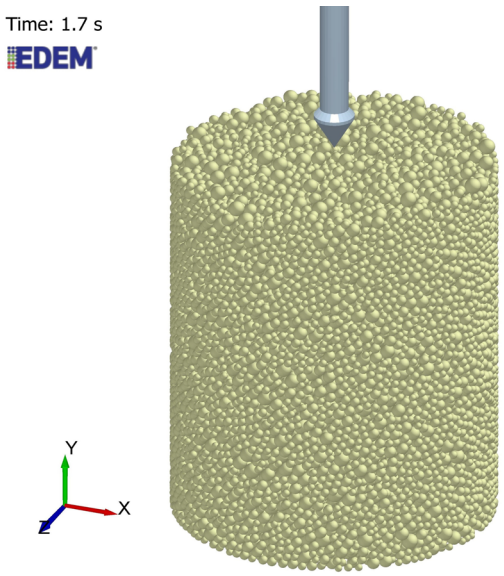


Figure 3: The 3D discrete element model of soil with diameter of Ø9.03e-2 m for simulation of the cone penetration test

Then the same velocity of 1e-2 $m \cdot s^{-1}$ was set up to press down the cone into the soil model in each simulation. The timestep of the calculations were 1e-5 s which is approximately the 11.7 % of the Rayleigh timestep. This value is within the range of 10…20 % according to the EDEM 2.7 software user manual. Finally the results were saved in every timestep of 5e-2 s. The initial geometry of the simulation of soil model with diameter of Ø9.03e-2 m (area ratio of 64) can be seen in Figure (3).

RESULTS AND DISCUSSION

First the results of the soil-cone penetrometer simulations were analysed qualitatively. This means that the velocity of the particles, the contact forces and the place where the bonds breaks were investigated in each calculations. In the left side of Figure 4 (a) the elements were coloured as the value of the compressive force, in the right side (b) the particle's colours were chosen according to their velocities and finally at the bottom of the figure (c) the broken bonds were illustrated as blue lines. According to the work of Tanaka et al (Tanaka et al 2000) the highest compressive force values arise near the penetrometer cone and the highest particle's velocity can be experienced here as well. This can be seen in the figure and can be observed because of the friction between the balls and the penetrometer cone.

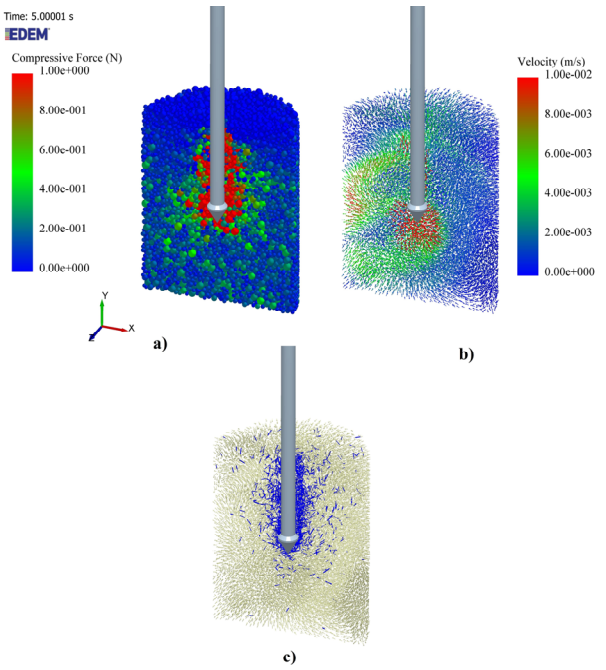


Figure 4: a) The compressive force between the elements and b) The velocity of the elements and (c) The broken bonds (as blue lines) near the penetrometer cone, both in case of simulation of soil model with diameter of Ø9.03e-2 m.

Similar results were experienced in the other discrete element simulations as well. According to these results the calculations seem to be good qualitative approximations of the real soil-cone penetrometer in-situ test.

After that the results were analysed quantitavely as well. The penetration resistance of the soil model was illustrated as a function of penetration depth in Figure (5) in case of soil model with diameter of Ø9.03e-2 m (area ratio of 64). According to the figure it can be seen that the penetration resistance values fluctuated a little bit which can be experienced because of the particle's geometry. In discrete element

calculations this result occurs often however it can be reduced by using smaller elements in the simulations. This will lead to increase the number of the particles and the calculation time as well which is not acceptable.

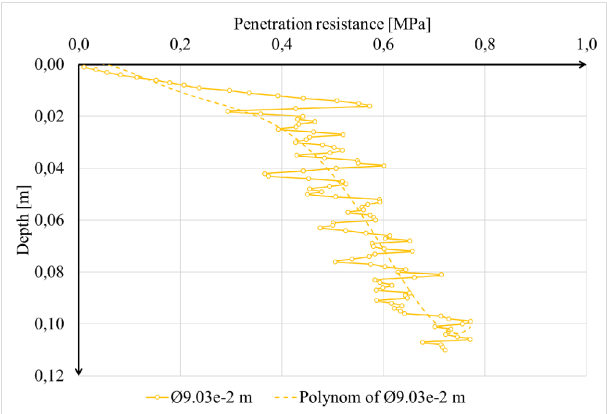


Figure 5: Results of the cone penetrometer discrete element simulation of soil model with diameter of Ø9.03e-2 m

In Figure (5) a polynom trendline can be seen as well which was fitted to the simulation results using the Ordinary Least Squares available in Microsoft Excel 2016 software. The type of the trendline has choosen to reach high R^2 values of the line fitting. Similar was done in each simulations the results can be seen in the following table:

Table 3: The results of the trendline fitting in all discrete element simulations.

Diameter of the soil model cross section (m)	*Area ratio (-)*	*Value of R^2 (-)*
Ø6.77e-2	36	91,4 %
Ø9.03e-2	64	88,0 %
Ø1.128e-1	100	85,4 %
Ø1.354e-1	144	88,9 %
Ø1.58e-1	196	83,3 %

In addition these trendlines can be seen in Figure (6) where the average result of the in-situ cone penetrometer test was illustrated as blue lines as well. This figure also can be used to investigate the effect of the model size on penetration resistance. According to Figure (6) the highest soil resistance value occurs in case of simulation of soil model with diameter of Ø6.77e-2 m (or with area ratio of 36). This is in accordance with our expectations because the cross section of the model become smaller and smaller, the penetration resistance should get higher and higher as well. This is because in case of small diameter (small area ratio) the elements can not move away from the penetrometer cone and become contact with the boundary walls earlier than they do in case of soil model with larger cross section. This will lead to increase the value of contact forces between the elements and therefore the value of the vertical forces acting on the cone as well.

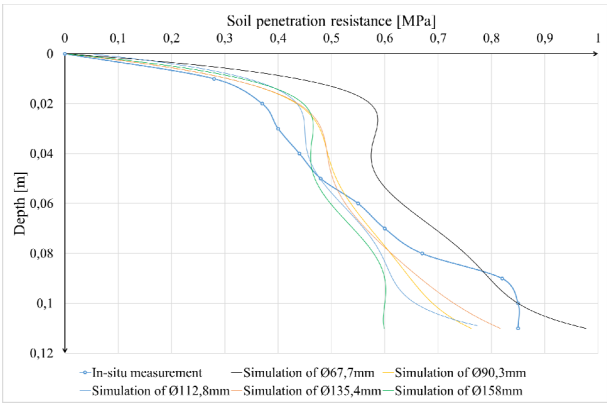


Figure 6: Results of the cone penetrometer discrete element simulations and of the in-situ test

It can be seen as well that the simulations of soil model with higher cross sections gave similar penetration resistance values which means that the boundary of the simulations will no longer has great effect on the results if a soil model with minimum diameter of Ø9.03e-2 m (area ratio of 64) will be used in the discrete element simulations.

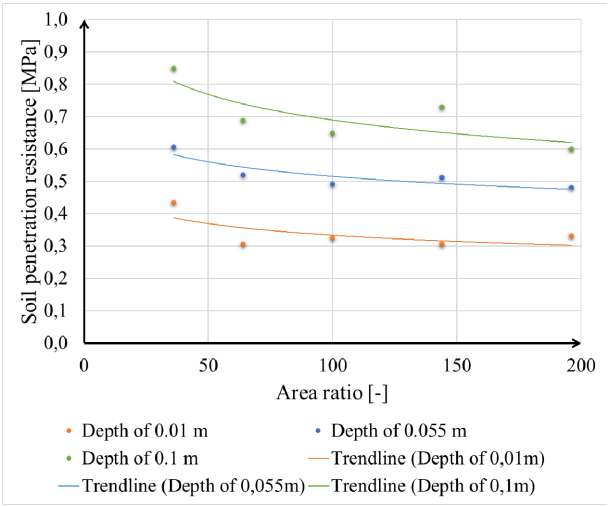


Figure 7: The effect of the model size on soil pemetration resistance

In Figure (7) the penetration resistance values from the simulations at given depth (e. g. the soil model's C. I. index) were illustrated as a function of area ratio. Similar to mentioned above can be said: the highest penetration resitance values can be experienced in case of simulation of soil model with area ratio of 36. In the other simulations the resulst are similar to each other therefore the soil model with area ratio of 64 (diameter of Ø9.03e-2 m) can be chosen for correct model size for simulation of soil-cone penetrometer interaction.

Choosing highest diameter for the soil model is unnecessary because only the calculations time will be increase but the model's penetration resistance will not change significantly.

ACKNOWLEDGEMENT

The authors will gratefully acknowledge the assistance of the staff of NARIC Institute of Agricultural Enginnering of Gödöllő, to provide their Eijkelkamp penetrologger for the in-situ measurements.

REFERENCES

Chen, Y., Munkholm, L. J., Nyord, T., 2013. A discrete element model for soil sweep interaction in three different soils. Soil & Tillage Research, 126, 34-41.

Cundall, P. A. and Strack, O. D. L. 1979. "Discrete numerical model for granular assemblies." *Geotechnique*, 29(1), 47-65.

Hertz, H. 1882. "On the contact of elastic solids." *J. reine und angewandte Mathematik* 92, 156-171.

Kotrocz, K.; Kerényi, Gy. 2017. "Numerical discrete element simulation of soil direct shear test." *European Conference on Modelling and Simulation (ECMS) '17*, 2017.

Laib, L. (Editor), 2002. Terepen mozgó járművek (Moving off-road vehicles). Szaktudás Kiadó Ház, Budapest, Hungary (in Hungarian.)

McKyes, E. 1985. *Soil Cutting and Tillage*. Elsevier, New York, USA.

Mindlin, R. D. 1949. "Compliance of elastic bodies in contact." *Journal of Applied Mechanics* 16, 259-268.

Mindlin, R. D. and Deresiewicz H. 1953. "Elastic spheres in contact under varying oblique forces." *ASME*, September, 327-344.

Potyondy, D. O.; Cundall, P. A. 2004. "A bonded-particle model for rock". *International Journal of Rock Mechanics & Mining Sciences*, 41, 1329-1364.

Tanaka H., Momozu M., Oida A., Yamazaki M., 2000. "Simulation of soil deformation and resistance at bar penetration by the Distinct Element Method." *Journal of Terramechanics*, 37, 41-56.

Tamás, K.; Jóri, J. I. and Mouazen, A. M. 2013. "Modelling soil-sweep interaction with discrete element method." *Soil & Tillage Research*, 134, 223-231.

AUTHOR BIOGRAPHIES

KRISZTIÁN KOTROCZ was born in Salgótarján, Hungary and went to the Budapest University of Technology and Economics, where he studied mechanical engineering and obtained his MSc degree in 2012. After that he started his PhD studies and worked in the Budapest University of Technology and Economics, Department of Machine and Product Design where he is an assistant lecturer curretnly. His research area id soil modelling using discrete element method. His e-mail address is: kotrocz.krisztian@gt3.bme.hu and his Web-page can be found at http://gt3.bme.hu/en.

GYÖRGY KERÉNYI studied agricultural machine design at Szent István University, Gödöllő and after that he went to Budapest University of Technology and Economics, where he obtained his PhD degree in 1997. Currently he is an associate professor and deputy head of Department of Product and Machine Design in the same institution and his research topic is numerical methods in agricultural machine design. His e-mail address is: kerenyi.gyorgy@gt3.bme.hu and his Web-page can be found at http://gt3.bme.hu/en.

AUTOMATIC CALIBRATION OF DISCRETE ELEMENT MODELS

Ferenc Safranyik and Istvan Keppler
Department of Informatics; Department of Mechanics
Eotvos Lorand University; Szent Istvan University
Szombathely, Hungary; Godollo, Hungary
safranyikf@gmail.com; keppler.istvan@gmail.com;

KEYWORDS

Mechanics of granular materials, discrete element method, calibration

ABSTRACT

Calibration means the determination of parameters governing the mechanical interaction of the individual particles and walls making up our discrete element model (DEM). Since the elaboration of DEM [1], calibration is the most difficult part of the DEM modeling process.

INTRODUCTION

Direct measurement of the parameters governing the particle–particle, particle–wall interactions would be the best solution, but in most of the cases it is impossible. It can also happen that even the measured parameters would not be suitable for modeling purposes, as the applied constitutive equations in the numerical calculations are only approximations. In most of the cases the proper measured values of these parameters are not needed, but a combination of parameters ensuring the modeled macro behavior to be the same as the measured one.

Standard shear testing technique

The standard shear technique [2] is one of the most commonly used calibration method [3], [4]. The standard shear testing technique for particulate solids is based on the so called Jenike shear cell (see fig.1). Material sample is placed into the shear apparatus and before shear a vertical force F_N is applied to the upper plate and hence to the particulate solid within the cell to pre-compress the material sample. A horizontal force is applied to the bracket by a mechanically driven measuring stem which is driven forwards at a steady rate of 1-3 $\frac{mm}{min}$. This stem is attached to the drive system through a force transducer which measures the shear force F_T. During the shear operation the shear ring moves from the original offset position to the opposite. During shear a shear zone develops inside the sample, and in this way we can create a shear stress shear strain plot. By knowing the shear stress values corresponding to a given compressive normal stress, the parameters of the failure line can be determined in the form of $T = \varphi N + c$, where T is the shear stress, N is the normal stress, φ is called the internal friction of the assembly and c is the cohesion (see fig. 2). Naturally, the failure curve of the material is not always linear, but the linear approximation of the failure curve is a common method in the practice[5].

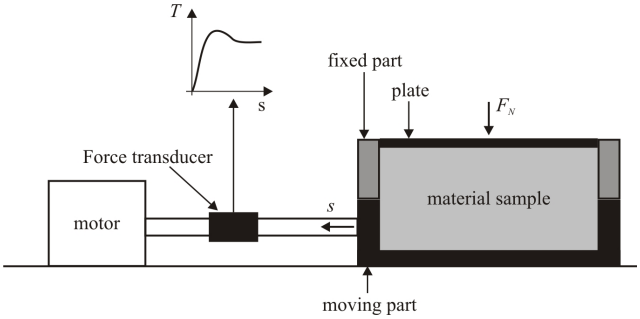


Fig. 1. Shear cell

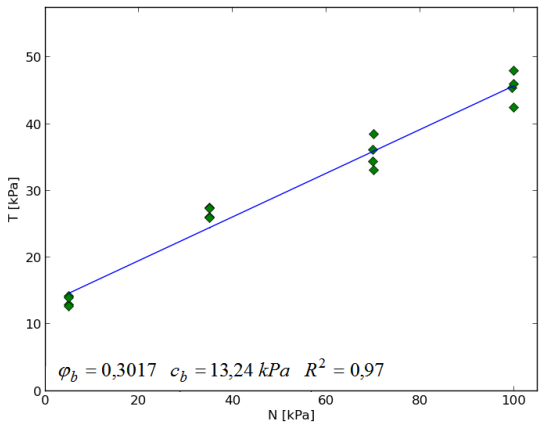


Fig. 2. Typical failure line

Discrete element modeling

DEM models the mechanical behavior of bulk materials by applying and solving the equation of motion on each singular particle of the bulk material assembly. The out of balance forces arising during particleparticle and particlewall interactions are calculated during the simulation circle, which is a cycle with repeated application of the Newtonian laws of motion to obtain the acceleration, velocity and displacement. The displacement is then used to evaluate the contact forces and moments acting due to the interactions between

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

the particles in their new position. In this article, we use Yade discrete element method software for numerical modeling [6]. In YADE the approximate collision detection filters out the impossible collisions, and after this step, a more computationally expensive collision detection algorithm evaluates the possible interactions between these individual particles.

After having exact collision detection, it is possible to model the interaction between the particles in contact. DEM interaction model uses two stiffnesses: K_N normal stiffness and K_T shear stiffness. K_N is related to the Youngs modulus of the particles material, and K_T is defined as a given fraction of K_N. Yade evaluates the K_N stiffness by modeling the two particles in contact as a serial connection of two springs having length equal to the radius of particles in contact:

$$k_N = 2\frac{E_1 r_1 E_2 r_2}{E_1 r_1 + E_2 r_2}.$$

The kinematic variables (displacements) of the contact are called as strains in Yade terminology. To evaluate the normal strain, there is a reference distance d_0 (or equilibrium distance) used to convert the evaluated displacements to dimensionless strain: $d_0 = |\mathbf{C}_2 - \mathbf{C}_1|$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the initial position vectors of the two contacting spheres centers. Because of the possible overlap between the two contacting spheres, there are reduced distances $d_1 = r_1 + \frac{1}{2}(d_0 - r_1 - r2)$ and $d_2 = d_0 - d_1$ used for further strain evaluations. For the constitutive laws Yade uses an equivalent cross-section $A_{eq} = \pi \min(r_1, r_2)^2$ to convert stresses into forces. By knowing the normal and shear displacements u_N , u_T, normal and shear forces are computed in the following way:

$$F_N = K_N u_N,$$

$$F_T = \begin{cases} F_N \tan\varphi & \text{if } |K_T u_T| > F_N \tan\varphi, \\ K_T u_T & \text{otherwise.} \end{cases}$$

If we have cohesive model, cohesive interactions bonds between particles are used to represent the materials cohesive properties. The normal force is $F_N = \min(K_N u_N, a_N)$, where a_N is the normal adhesion. The tangential shear force is $F_T = K_T u_T$, the plasticity condition defines the maximum value of the shear force, by default $F_T^{\max} = F_N \tan\varphi + a_S$, where a_S is the shear adhesion. If the maximum tensile or maximum shear force is reached, the cohesive link is broken.

The forces and torques evaluated above are used to integrate the equations of motion (linear- and angular momentum theorem) of the particles in contact. From the equations of motion by integration the new particle positions can be determined, and based on the new particle positions, the simulation circle can start again from collision detection.

In case of simulating quasi-static phenomena, the kinetic energy of the particle assembly must be dissipated in some way. Since the constitutive law does not include velocity dependent damping, Yade uses artificial numerical damping: the forces, which increase the particle velocities are decreased by $(\Delta F)_d$. Yade uses for the artificial damping the following:

$$\frac{(\Delta F)_{dw}}{F_w} = -\lambda_d \text{sign} F_w \left(\dot{u}_w (t - \Delta t) + \frac{1}{2}\ddot{u}_w (t) \Delta t \right).$$

For numerical stability reasons, there is an upper limit Δt_{crit} on the simulation timestep calculated from the highest eigenfrequancy of the system. The so called unbalanced force is used to keep the modeling process to be close to quasistatic conditions, the unbalanced force ratio will tend to zero as simulation stabilizes.

Discrete element model of the shear test

We used a slightly modified version of Jenike shear cell for the discrete element modeling purposes, as in our case the lid is rectangular (see fig. 3). We evaluated

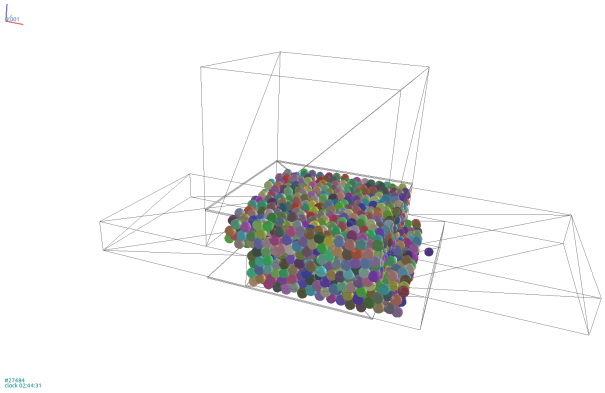


Fig. 3. DEM model of the shear apparatus

the normal-force shear force diagrams in case of four different pre-compressing forces four times (16 simulations for one failure line) to get one failure line similar to the example showed in fig. 2. The block diagram of the shear test can be seen on fig. 4. It can be seen, there, that the compression of the material sample starts only after the granular material poured into the shear box is in a state of rest. The constant compressive force value is maintained by up- and downwards motion of the compressing plate.

AUTONOMOUS CALIBRATION ALGORITHM

The inputs of the calibration algorithm are the desired values of the shear failure line φ and c, and the micromechanical parameters related to a starting point. As a first step, the slope of the failure line is calibrated (see fig. 5), and then the calibration of the cohesion is done (see fig. 6). A sensitivity test is important part of the calibration, because we have to decide, which parameters should be changed to reach the desired parameters of the failure line. The details of the sensitivity test are available in [3].

To calibrate by hand the micromechanical parameters related to the desired cohesion and internal friction values presented in fig. 2, we needed 312 hours of computation time. The automatic calibration algorithm managed to reach the desired micromechanical values in 60 hours.

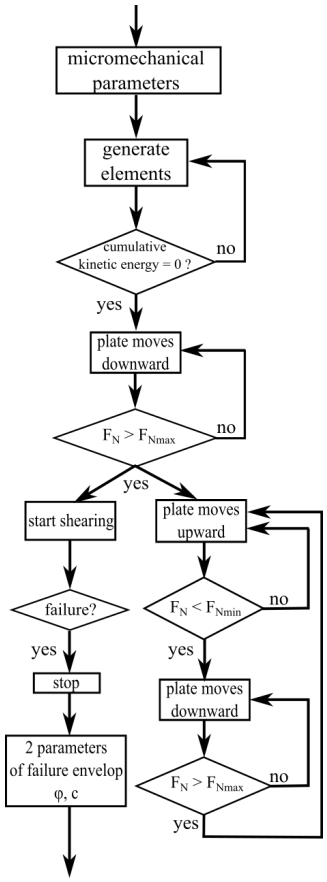


Fig. 4. Block diagram of shear test

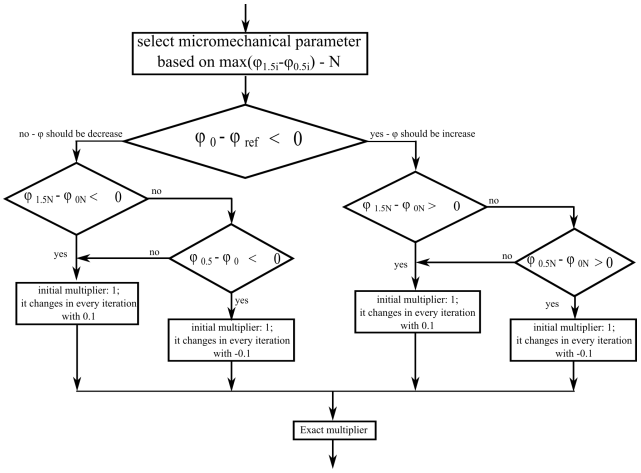


Fig. 5. Automatic calibration for φ

CONCLUSIONS

Calibration of discrete element based models is still a challenging question of the engineering practice. It would be preferable to have a standard calibration method available for use. Our opinion is that the standard shear testing technique is a well-grounded method applicable to ensure properly calibrated discrete element models. The calibration itself is a time consuming, monotonic process which is (in many cases) based on trial and error methods. We suppose that this process can be automatized. For the automatization, the process must be rigorously controlled; the trial and

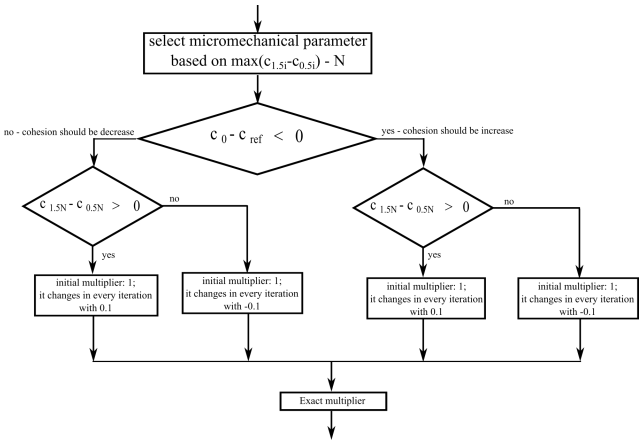


Fig. 6. Automatic calibration for c

error process must be superseded by sensitivity test based gradient methods. In this article, we demonstrated, that this is possible. The discrete element calibration can be automatized. A highly autonomous algorithm can be constructed, which is capable to find desired macromechanical behavior by systematic modification of micromechanical parameters. The change of micromechanical parameters must be based on initial sensitivity analysis to realize relatively short calibration time.

References

[1] **Cundall, P.A. and Strack, O.D.L.**: *A discrete numerical model for granular assemblies*, Geotechnique, Vol. 29 No. 1, pp. 47-65. 1979.

[2] **Jenike, A.W.**: *Storage and flow of solids*, Bulletin No. 123, Utah Engineering Experiment Station, University of Utah, Salt Lake City, UT. 1964.

[3] **Keppler I., Safranyik F., Oldal I.**: *Shear test as calibration experiment for DEM simulations: a sensitivity study*, Engineering Computations, Volume 33, Issue 3, pp. 742 758, 2016.

[4] **Kotrocz K, Kerenyi G**: *Numerical discrete element simulation of soil direct shear test*, ECMS-European Council for Modelling and Simulation, 2017. pp. 510-515.

[5] **The Institution of Chemical Engineers**: *Standard Shear Testing Technique*, ISBN 0 85295 232 5, The Institution of Chemical Engineers, Rugby. 1989.

[6] **V.Smilauer, E. Catalano, B. Chareyre, S. Dorofeenko, J. Duriez, A. Gladky, J. Kozicki, C. Modenese, L. Scholts, L. Sibille, J. Strnsk, and K. Thoeni**: *Yade Documentation*, The Yade Project, 1st ed., 2010.

Dr. Istvan Keppler (1972) Ph.D., associate professor at Szent Istvan University, Hungary. The field of his scientific interest is engineering mechanics, mechanics of granular materials, discrete element modeling.

Dr. Ferenc Safranyik (1988) received his M. S. degree of mechanical engineering in 2013, Ph.D. degree in 2017 from Szent Istvan University (SZIU), Godollo, Hungary. He is now assistant professor at Eotvos Lorand University. The topic of his research is mechanics of granular materials.

Investigation of soil-sweep interaction in laboratory soil bin and modelling with discrete element method

Kornél Tamás, Zsófia Oláh, Lilla Rácz-Szabó

Budapest University of Technology and Economics
Department of Machine and Product Design
H-1111 Budapest, Műegyetem rkp. 3.
MG building 300.
tamas.kornel@gt3.bme.hu

Zoltán Hudoba

NARIC Institute of Agricultural Engineering
Gödöllő, H-2100, Tessedik Sámuel u. 4

KEYWORDS

Soil-sweep interaction, Soil bin, DEM, Clumps

ABSTRACT

Nowadays, it is important to preserve the moisture content of the soil which can be executed with cultivator tools. The objective of this study was the investigation of the energetic and working quality characteristics of sweep tools by soil bin experiments conducted in laboratory conditions and the simultaneous development of a discrete element model (DEM) validated by comparing the measured and the simulated values of draught forces. Three sweeps were analyzed differing in aspects of shape and size; the tool widths were 100, 230 and 300 mm, respectively. Experiments were conducted in the NARIC Institute of Agricultural Engineering soil bin facility filled with sandy soil with volumetric moisture content of 2% and 15%, at three different working speeds (0.5, 1, 2 m/s) at the working depth of 0.15 m. The improved DEM model contain a soil model consisting of discrete particles created within the Yade software and 3D laser scanned model of the sweep tools. Inserting the analyzed tool geometry into the simulation made it possible to investigate the characteristics of the tillage tools and also the effect of the different geometries appearing in the soil cutting forces. The investigation of the effect of the shape of the particles on the energy dissipation and studying the differences between the validated frictional and the cohesive soil models based on the soil bin test were the aim within this study.

INTRODUCTION

Agriculture is an important part of our daily life and supplies food to the world's population. Therefore, appropriate soil conditions must be ensured for the crops grown. In order to increase productivity, there are several possibilities to influence the processes in the soil using different mechanical methods. One of the appropriate tools for this is the sweep tool that does not mix the soil, so the moisture content in the affected soil layer is not exposed to the surface and the soil does not dry out. This is especially important, because the aim of the loosening process with a sweep tool is the seedbed preparation, cutting plants' roots beside the reduction of moisture loss of the cultivated soil.

The development of computer technology put thorough investigation of soil-tool interaction within reach and made it possible to study certain areas of soil mechanics and analyze results which are difficult to achieve in experiments or solely with analytical methods, considering the complexity of the matter. Among the numerical methods, the FEM (finite element method) spread initially: both two- and three-dimensional models were developed. As the computing capacity increased, the calculation time was reduced, enabling the appearance of more complex numerical simulations such as the CFD (computational fluid dynamics), which is a proper technique to investigate the soil movement from a flow point perspective and the SPH (smooth particle hydrodynamics) method, which is suitable to analyze the soil loosening processes (Urbán et al. 2012; Zachár et al. 2016).

Nevertheless, it is difficult to use the FEM or the aforementioned other methods when researchers encounter problems in which the discrete nature of the particles of the material is crucial. In the case of these problematic areas, where the discrete nature of the granulated material is key, the DEM (discrete element method) can be a proper technique to use (Bagi 2007; Tamás 2016).

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

Initially, researchers mainly investigated the behavior of spherical particles, although in reality granular particles are generally non-spherical, therefore the utilization of particles with a more complex shape was needed. The internal structure of the granular material is transformed by the particles rolling down and interlocking, and thus the behavior of the set is greatly influenced by the shape of the particles, which can also be considered as an input parameter to be calibrated. Researchers also use ellipsoids, polyhedrons and clumps as particles. The shape of the material particles is responsible for the extent of the energy dissipation in the material set, so the examination of the effect of the particle shape has also played an important role in this research.

The previous models of soil-tool interaction mostly presumed friction in the soil, but because of its moisture content the soil also shows a cohesive behavior. By increasing the moisture content, the draught force is also increasing, so the introduction of cohesive models was necessary. The focus of this study was the examination of the soil-tool interaction: the main aim was to create a three dimensional DEM model simultaneously with the execution of a laboratory soil bin experiment. An another aim was to carry out the calibration process of the DEM model that was created, with particular emphasis on finding the most appropriate particle shape regarding its effect on the extent of energy dissipation of the interaction. The additional objective was to improve the micromechanical parameter set used in the DEM model of soil-sweep tool interaction by comparing laboratory and simulation results with the use of relative error method.

THE SOIL BIN TEST

In order to create a DEM model, in the first step soil bin study was conducted (Figure 1). The measurements were carried out in the soil bin facility of NARIC Institute of Agricultural Engineering, Gödöllő, Hungary. The laboratory soil bin was filled with sandy soil with the following content: 93% of sand, 4.66% of silt and 2.06% of clay. The dry bulk density was calculated by the results of a previously conducted oedometer test, and was determined as 1424.12 kg/m^3. For the investigation of soil-tool interaction three different sweep geometry were used with the tool width (W) of 100, 230 and 300 mm, respectively, henceforward referring to them as small, medium and large sweep tools (Gürsoy et al. 2017). The experiment focused on the relation between draught force and working speed. Three different working speeds were used, in average 0.5, 1, 2 m/s. The soil bin study was made at the tool's working depth of 0.15 m.


a,
b,

Figure 1: The soil bin of the NARIC Institute of Agricultural Engineering a) the sweep tool in the soil and b) the soil loosening process

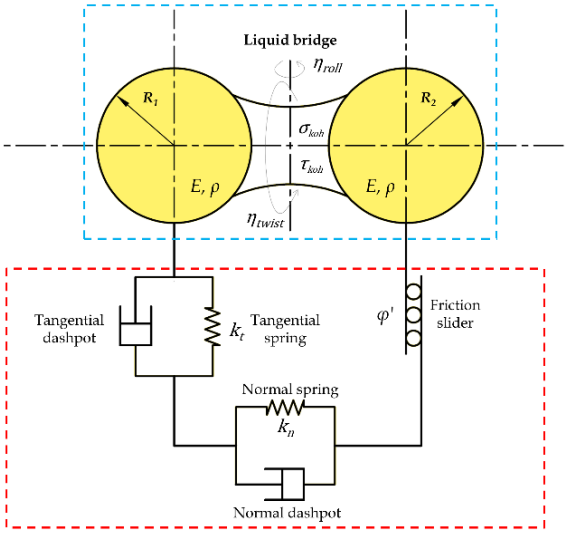


Figure 2: The friction and the cohesion connection in the utilized CohFrictMat contact model (Bourrier et al. 2013, Horváth 2017)

Primarily the relation between draught force and working speed was analyzed. Also another aim of this measurement was to study soil particles' displacements in a vertical direction, and the effect of soil moisture content.

Following the proper preparation of the sandy soil, the penetration resistance and volumetric moisture content measurements were conducted. In

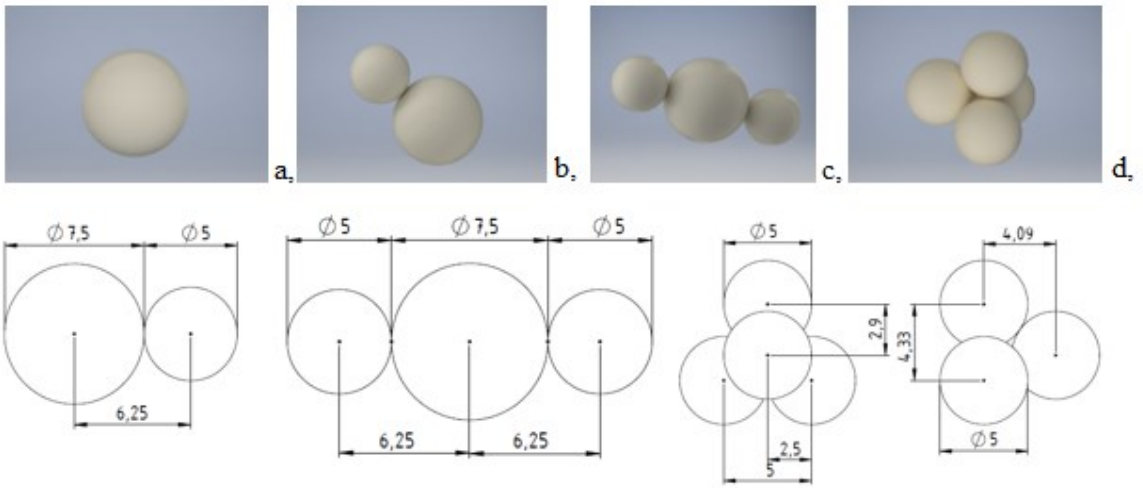


Figure 3: The created and used clumps in the model a) sphere, b) clump made of 2 spheres, c) clump made of 3 spheres, d) clump made of 4 spheres

case of dry soil, the volumetric moisture content of 2% was observed, and in case of wet soil a value of 15% was measured.

In the case of the dry sandy soil, vertical penetration resistance was about 1.12 MPa, and in wet sandy soil it fluctuated around 0.72 MPa. The initial soil profile was then recorded.

Along the measuring section the draught force was monitored, the time required to take the measuring section and the working speed. The sweep, with a rigid tine, was mounted with supports to the tool's draught force with calibrated strain gauges.

The soil profiles formed after the measurements were also recorded using a laser scanner. These were suitable for comparison with the initial soil profile to conclude how the analyzed tool geometries differ in lifting and mixing the soil at a given speed and how far the ground is lifted vertically by them. As the soil particles can only move upwards during the loosening process influenced by sweep tools, therefore a volume increase can be observed in the cultivated layer of the soil.

GENERAL INFORMATION OF DEM MODELING

For developing the DEM model, the Yade open-source software (Šmilauer et al. 2015) was used, in which several different contact models were available regarding the mechanical relation between discrete particles, including FrictMat (frictional) and CohFrictMat (frictional and cohesive) models (Figure 2).

The applied software performs a collision detection to analyze the interactions between the particles involved in the simulation and uses timestep, after which all particles can change their mutual positions. As a first step of the process, the software sets the forces to default, then updates the bindings. After that detects the impact and then calculate the forces based on the displacements. Then speed and position of particles are updated. Running various motors, recording data, and increasing the time step are the following steps.

Geometrical model of soil particles

Within the Yade software, the particles created in this study are so called clumps, composed of spheres (Figure 3). They are connected to one another and form an independent, discrete, perfectly rigid unit.

Also the clumps forming the set are rigid bodies and can't suffer deformations. In the simulations, different set of particles were examined handling them as input parameters (Coetzee 2016). These different particle shapes (clumps) can be seen on Figure 3. There were several preliminary simulations where the appropriate set of particle

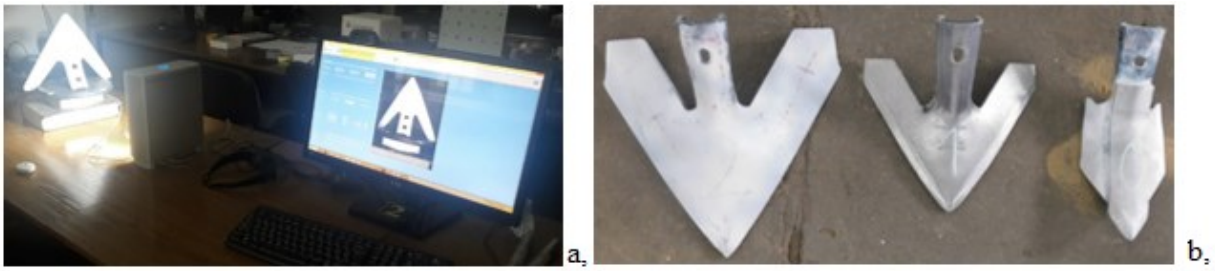


Figure 4: a) The scanning process with the NextEngine 3D Laser Scanner, b) the geometry of the sweeps with different width (W) used in the soil bin experiment (large-W:100mm, medium-W:230mm, small-W:100mm)

assembly for energy dissipation were studied. The particle assembly used in the final soil model was consisting of the following ratio of the given clumps: b: 5%; c: 90%; d is 5% see on Figure 3. Clump particles were designed previously, as a template but the particle assembly was generated in two steps. In the first step the assembly was created with the use of spherical particles in uniform distribution with radius of 20-24 mm. Then in the next step these particles were transformed to clumps with the use of the template data. Therefore, the result of these clump generation process was the clump assembly, where the volume of the clumps were the same as the spheres were generated previously. Finally, the spheres' radius of the prepared clumps were between 10.88-19.26 mm. The particle density was set as 2600 kg/m^3. The density was homogeneous within the clumps. The particle assembly (as the model of soil) was settled with the utilization of the local damping of 0.8 for faster sedimentation. The settling process was finished when the unbalanced force was lower than 0.001. This local damp was set as zero in the soil-sweep simulations later where the CohFrictMat contact model was utilized.

Geometrical models of sweep tools

The sweeps were scanned using a 3D laser scanner in the Virtual Laboratory of Budapest University of Technology and Economics, Department of Machine and Product Design with the utilization of NextEngine 3D Laser Scanner (Figure 4) to create a simulation environment that is as close as possible to reality (Figure 5).

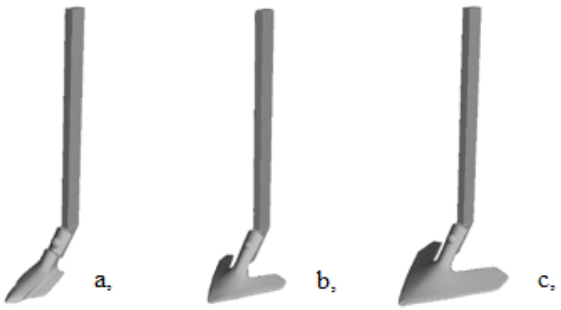


Figure 5: The final geometry of the tools used in the simulations a) small-sweep tool (W:100 mm) b) medium-sweep tool (W:230 mm) c) large-sweep tool (W:300 mm) W=Width

The shape of the cracks in the soil is partly influenced by the tool geometry, so it is important to insert into the simulations the realistic shape of the sweeps used in the measurements. To easily insert the tool geometries into the simulations, the scanned files were previously edited with Autodesk Meshmixer (2017) and the holes in the mesh were filled and the triangle count was reduced than exported as STereoLitography (STL) file format.

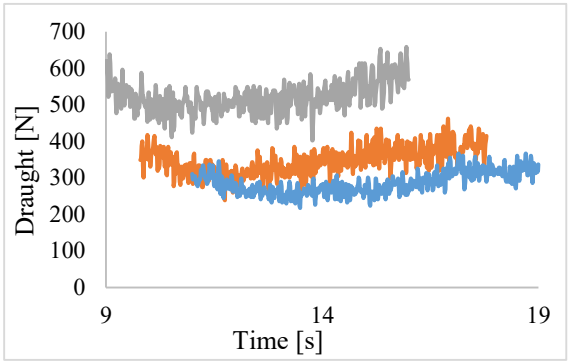


Figure 6: Comparison of the different tool geometries, at moisture content of 2% and speed of 2 m/s grey/upper curves – large sweep tool, orange/middle curves – medium sweep tool, blue/lower curves – small sweep tool

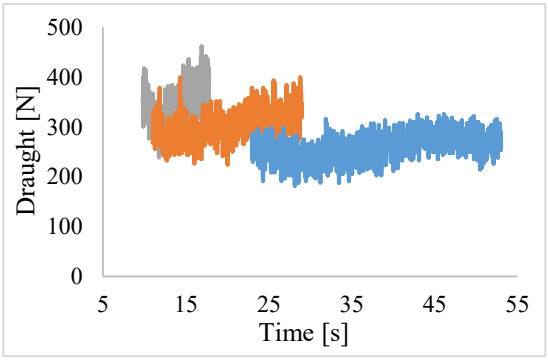


Figure 7: Comparison of different speeds, in moisture content of 2% with the medium sweep tool grey/upper curves – 2 m/s, orange/middle curves – 1 m/s, blue/lower curves – 0.5 m/s

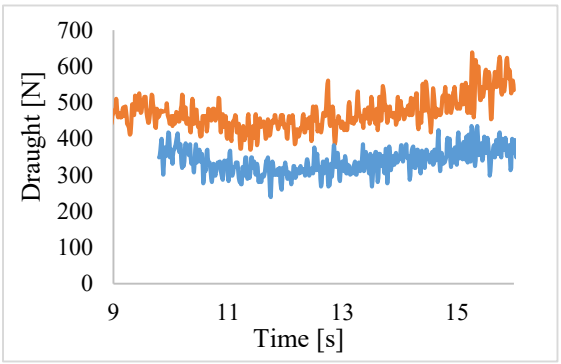


Figure 8: Comparison of different moisture contents at speed of 2 m/s with the medium sweep tool orange/upper curves – moisture content of 15%, blue/lower curves – moisture content of 2%

The scanned files were completed with the shank (Figure 5) was created by Dassault Systems Solidworks (2014) software. The sweep models were applied at 0.15 m working depth under the surface of the particle assembly.

RESULTS OF THE SOIL BIN TEST

The sweep tool creates a compaction, deformation zone in the soil and surface lifting during the soil cutting, and the resistance of the sweep tool is periodically variable at each breaking surface, the cutting resistance decreases dramatically, and after the subsequent compaction work, it starts to grow again. This is also illustrated by the diagrams showing the measurement results the draught forces fluctuate around a mean value (Tamás 2018).

Comparing the sweep tool geometries, at the same moisture content, the large sweep had the highest required draught force, followed by the medium sweep, while in case of the small sweep the smallest draught forces were measured on average (Figure 6). This was expected, given the size of the sweep's contact surfaces with the soil. Increasing the working speed (Figure 7) of the sweep tool increases the amount of energy transferred to the soil and thus the draught force, which determines how much force is needed to move the sweep horizontally in the soil in the given conditions. The draught force was higher at moisture content of 15%, than 2% (Figure 8) independent of the sweeps' widths.

The soil profiles were recorded at speed of 2 m/s. When the 2% moisture content in dry soil condition was measured the draught force and the surface vertical lifting were lower (Figure 9) than in moisture content of 15% (Figure 10), where the cohesion was higher and was effected by the liquid bridges between soil particles. When comparing the effect of the three different sweep tool geometry on profiles, it can be made the conclusion that the width of the deformation zone was the largest in the case of the large sweep and the smallest in the case of the small sweep (Figures 9-10). While examining the resulted soil profiles after the tillage process, it can be concluded that in case of wet soil the traces of the compacting tool is better preserved in the soil surface.

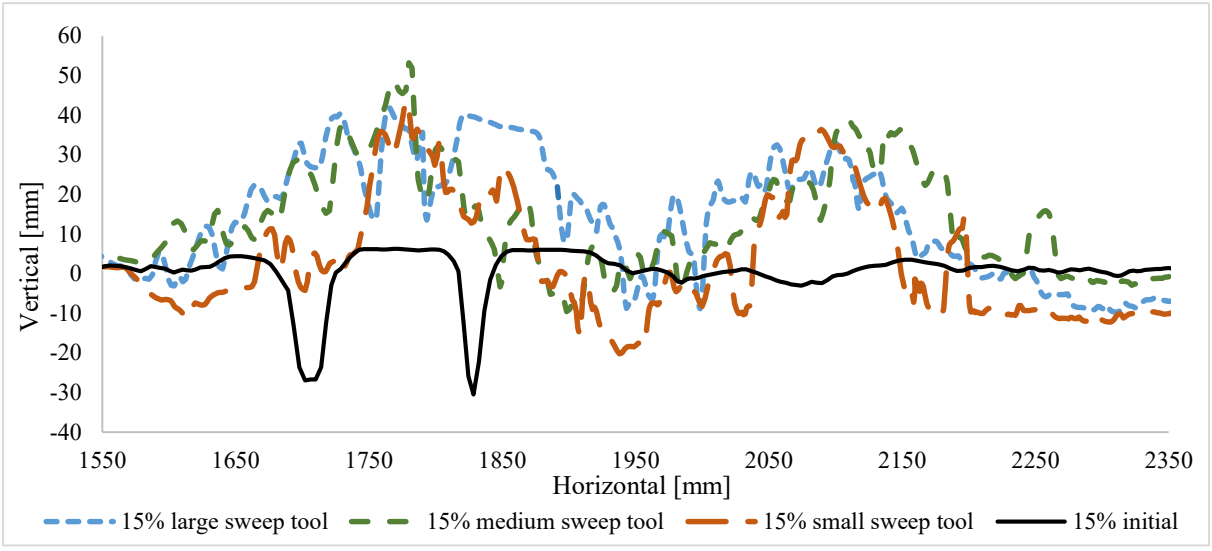


Figure 9: Soil profiles at the moisture content of 2%

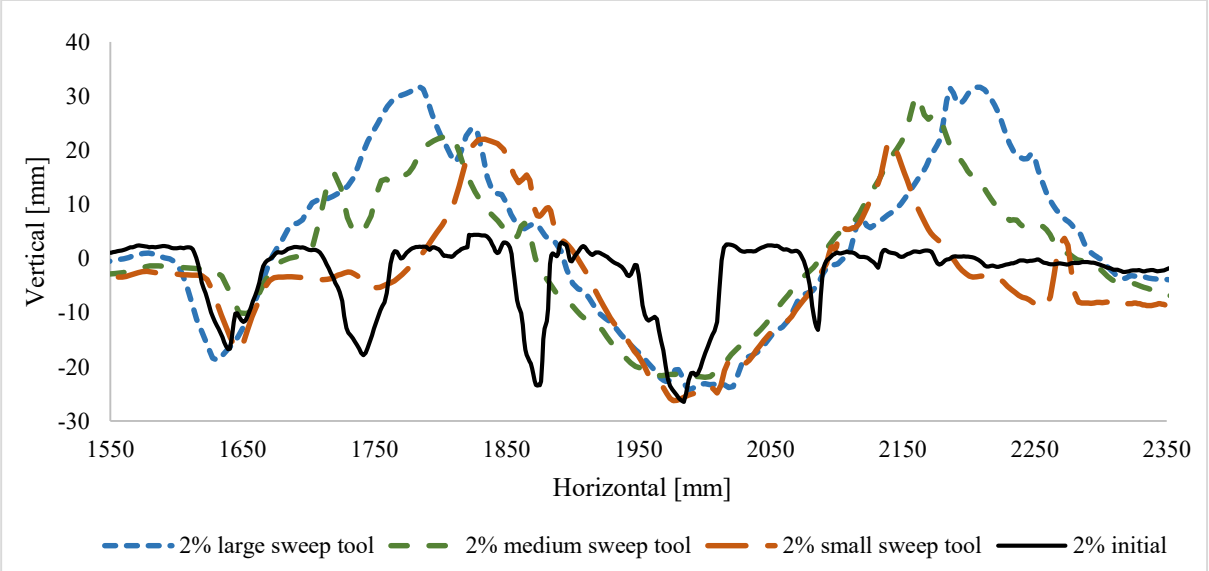


Figure 10: Soil profiles at the moisture content of 15%

RESULTS OF THE DEM SIMULATIONS

In the simulation model, the length of the soil bin was 2 m, its width and height were both 1 m. The validated parameter set, which were based on the experiences of the preliminary simulations and used during the validation process are summarized in Table 1.

Table 1: The utilized parameter set in DEM model

Parameter	Value
Young's modulus [Pa]	$2*10^5$
Rolling friction(etaroll)	0.001
Poisson ratio	0.4
Friction angle [°]	30
Density [Kg/m^3]	2600
Cohesion in normal direction (σ) [Pa]	$2*10^3$
Cohesion in shear direction (τ) [Pa]	$1*10^3$
Number of elements	20 000
Local damping	0.0
Timestep [s]	0.1*PWaveTimeStep()

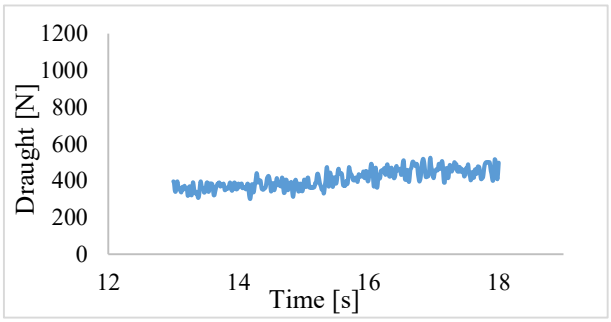


Figure 11: The measured draught forces with the medium tool at working speed of 2 m/s, at moisture content of 2%

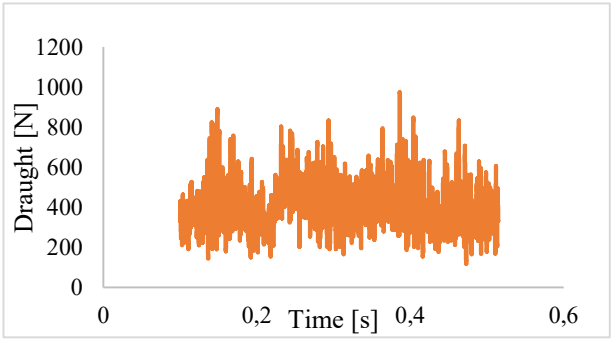


Figure 12: The simulated draught forces with the medium tool at working speed of 2 m/s, in the friction model

The applicability of the DEM model to the estimation of the draught force of the different sweep tools was validated by comparing the results of the soil bin experiments and the simulations (Figures 11-14). The resulting deviation of the draught force was

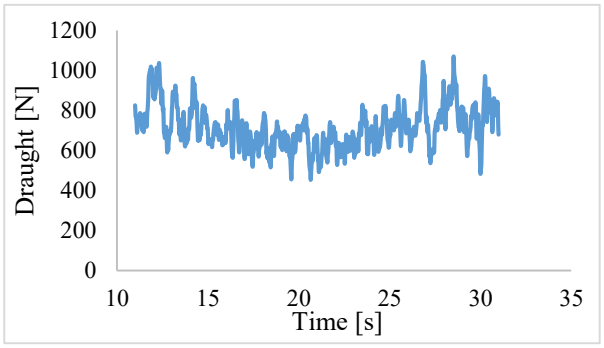


Figure 13: The measured draught forces with the medium tool at working speed of 2 m/s, at moisture content of 15%

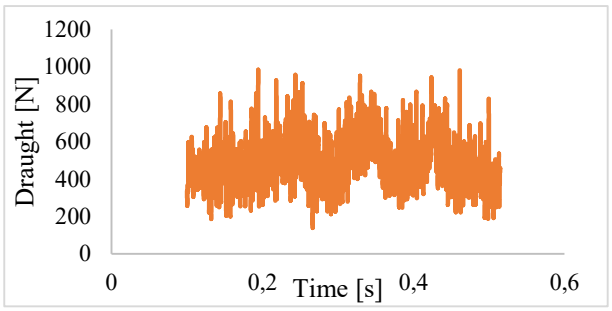


Figure 14: The simulated draught forces with the medium tool at working speed of 2 m/s, in the cohesion model

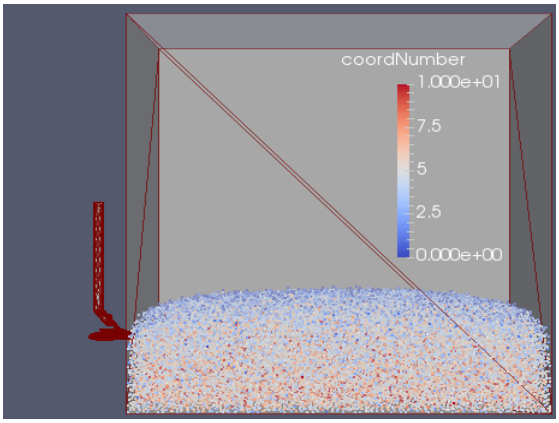


Figure 15: The soil domain in the improved DEM model of soil-sweep interaction. Working depth: 0.15 m. (The colour of particles indicates the coordination number.)

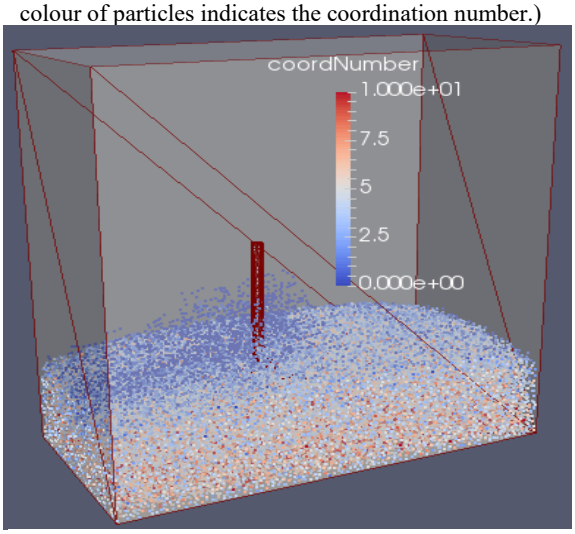


Figure 16: The simulation of the DEM model of soil-sweep interaction. (The colour of particles indicates the coordination number.)

large compared to the measured values, but the trend of the graphs was comparable (Figures 11-14).

Comparison of laboratory measurements and simulation results with the aim of validating the DEM model was carried out with the utilization of the method of the relative error between the measured and the simulated values (Figures 17-18).

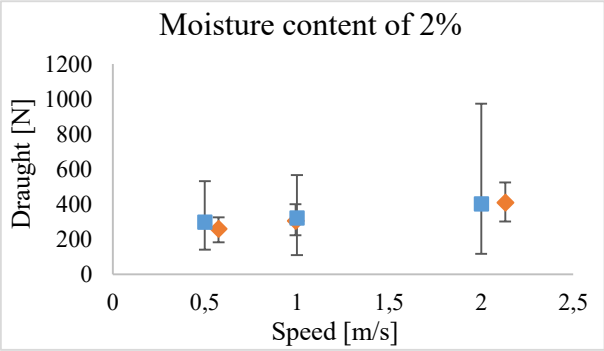


Figure 17: The measured and simulated draught forces and their standard deviation with the medium tool at the moisture content of 2 %

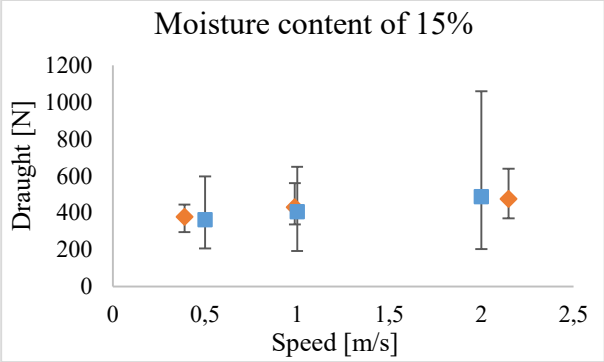


Figure 18: The measured and simulated draught forces and their standard deviation with the medium tool at the moisture content of 15 %

The relative error was determined by the following formula:

$$RE = \frac{100}{n} \sum \left| \frac{S-M}{M} \right| \quad (1)$$

where RE: relative error [%], n: number of data, S: simulated value, M: measured value.

Table 2: Relative errors of sweep tools (negative sign indicates the underestimation)

Speed [m/s]	0.5	1	2
Small sweep tool			
		Friction	
Relative error [%]	*17.52*	*4.23*	*15.92*
		Cohesion	
Relative error [%]	*5.25*	*-2.70*	*7.36*
Medium sweep tool			
		Friction	
Relative error [%]	*14.30*	*5.82*	*-1.73*
		Cohesion	
Relative error [%]	*-3.82*	*-5.82*	*2.47*
Large sweep tool			
		Friction	
Relative error [%]	*-1.75*	*-5.49*	*-2.84*
		Cohesion	
Relative error [%]	*-24.65*	*-28.82*	*-24.59*

Based on the calculated relative errors (Table 2.), it can be concluded that the CohFrictMat model overestimated the draught force results of the measurements conducted in case of wet sandy soil regarding the small and medium sweeps, while underestimation occurred in case of the large sweep. The solutions to reduce the draught force overestimation at the utilization of lower tool's surface area and the underestimation at higher tool's surface area could be the inaccurate parameter set of bonds, or inaccurate shape of the utilized clumps.

Since the calibration process was performed on a medium sweep with a working speed of 2 m/s, the lowest relative error value was also achieved at this setting (RE=1.73%). Concerning the small sweep, the simulations with the cohesion model resulted in minor differences, and in regard to the large sweep, simulations with the friction model were the most effective. In further tasks, parameter sensitivity studies shall be made for reducing the effect of the size of the utilized clumps on the fluctuation of draught forces. Thus, research should be done in qualitative aspect where the bonds' effects on the occurrence of fractures resulted by micro cracks in the assembly have to be investigated, where the micro crack means the bond breakage. Moreover, an accurate further study need, where the displaced and redistributed soil profiles are investigated to verify the appropriate displacement in the particle assembly.

CONCLUSIONS

A DEM model was developed in this research, which is suitable to utilize of the optimization of the geometry of sweep tools used in tillage. In the developed soil-sweep DEM model it could be taken into account the soil's moisture content the particles' shape, the tool's geometry, and

working speed, as well in the given working depth. Adequate input parameters were also selected and verified. However further development of the model is necessary by further examination of the effect of the Young's modulus settings or more appropriate clumps' geometry and size could be a way to improve the model. Therefore, better approximation of the actual draught forces could be realized with the applying of the simulation method. In addition, the aim of future research can also be to modelling the plant residues and roots in the improved soil model beside the particles' angular shape.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the contribution of the NARIC Institute of Agricultural Engineering in soil bin test and the virtual laboratory of Department of Machine and Product Design in the scanning process of sweep tools.

REFERENCES

Urbán, M., Kotrocz, K., & Kerényi, G. 2012. Investigation of the soil-tool interaction by SPH (Smooth Particle Hydrodynamics) based simulation. In Power and Machinery. International Conference of Agricultural Engineering-CIGR-AgEng 2012: agriculture and engineering for a healthier life, Valencia, Spain, 8-12 July 2012. CIGR-EurAgEng.

Zachár, A., Keppler, I., & Oldal, I. 2016. Investigation of the applicability and efficiency of different mathematical modeling and numerical simulation methods for soil-tool interaction. Journal of Computational & Applied Mechanics, 11(1).

Tamás K. 2016. A talaj és a kultivátorszerszám egymásra hatásának modellezése, Budapest, in Hungarian

Bagi K. 2007. A diszkrét elemek módszere, BME Tartószerkezetek Mechanikája Tanszék, Budapest, in Hungarian

Bourrier, F., Kneib, F., Chareyre, B., & Fourcaud, T. 2013. Discrete modeling of granular soils reinforcement by plant roots. Ecological Engineering, 61, 646-657.

Gürsoy, S., Chen, Y., & Li, B. 2017. Measurement and modelling of soil displacement from sweeps with different cutting widths. Biosystems Engineering, 161, 1-13.

Šmilauer, V. et al. 2015. "Yade Documentation 2nd ed." In The Yade Project (http://yade-dem.org/doc/)

Horváth D. 2017. Statikus berendezésekben mozgó szemcsés anyaghalmaz modellezése (TDK dolgozat, BME, inHungarian

Coetzee, C. J. 2016. Calibration of the discrete element method and the effect of particle shape. Powder Technology, 297, 50-70.

Tamás, K. 2018. The role of bond and damping in the discrete element model of soil-sweep interaction. Biosystems Engineering, 169, 57-70.

AUTHOR BIOGRAPHIES

KORNÉL TAMÁS is an assistant professor at Budapest University of Technology and Economics where he received his MSc degree and then completed his PhD degree. His professional field is the modelling of granular materials with the use of discrete element method (DEM). His Web-page can be found at http://gt3.bme.hu/tamaskornel.
His e-mail address is: tamas.kornel@gt3.bme.hu

ZSÓFIA OLÁH was born in Eger, Hungary. She is currently doing her Industrial Design Engineering BSc studies at the Budapest University of Technology and Economics. She is a member of a research group of discrete element modeling at the university.
Her email address is: foltos822@hotmail.com

LILLA RÁCZ-SZABÓ was born in Budapest, Hungary. She is an Industrial Design Engineering Student at the Budapest University of Technology and Economics, where she is currently doing her BSc studies. She is a member of a research group of discrete element modeling at the university.
Her email address is: rszlilla@gmail.com

ZOLTÁN HUDOBA was born on 27th of december in 1961. Obtained seafaring qualification in Budapest, agricultural mechanical engineer degree on Szent István Egyetem. He has an engineer teacher degree too. Second-year PhD student in SZIE doctoral school. He works in HIAE since 2007, currently as an institute engineer.
His email address is: hudoba.zoltan@mgi.naik.hu

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems - Special Session

Concrete vs. Symbolic Simulation to Assess Cyber-Resilience of Control Systems

Giuseppina Mùrino, Armando Tacchella

KEYWORDS

Simulation of Control Systems, Artificial Intelligence, Cyber-Security and Critical Infrastructure Protection

ABSTRACT

State-of-the-art industrial control systems are complex implements featuring different spatial and temporal scales among components, multiple and distinct behavioral modalities, context-dependent and human-in-the-loop interaction patterns. Most control systems offer entry-points for malicious users to disrupt their functionality severely, which is unacceptable when they are part of the national critical infrastructure. Cyber-resilience, i.e., the ability of a system to sustain — possibly malicious — alterations while maintaining an acceptable functionality, is recognized as one of the keys to understand how much damage can be brought to a system and its surrounding environment in case of a successful cyber-attack. In this paper we compare methods to assess resilience considering both concrete simulation and symbolic simulation. Our ultimate goal is to provide maintainers and other stakeholders with a dynamic and quantitative measure of cyber-resilience. Here we present some results on a case study related to waste-water treatment, in order to provide initial evidence that concrete and symbolic simulation can be used in a complementary way to analyze the security of industrial control systems.

INTRODUCTION

When considering industrial control systems, one may observe that current state-of-the-art systems are complex implements intertwining physical processes, hardware, software, and communication networks — the term *cyber-physical system* (CPS) is often used in this context. With respect to "classical" embedded systems, a CPS adds elements of complexity including different spatial and temporal scales among components, multiple and distinct behavioral modalities, context-dependent and human-in-the-loop interaction patterns [Lee08]. Examples of CPSs include heterogeneous systems of systems such as water treatment plants, electric grids (power plants and associated distribution networks), industrial plants, transportation vehicles, and smart homes.

Wireless communication among components and external network access for supervisory control and data acquisition (SCADA) make any control system an ideal target for cyber-attacks. It has been demonstrated that many such systems offer potential entry-points for malicious users to gain control of the controlled system and/or disrupt its functionality severely. This is true also of most CPSs which are part of national critical infrastructure (CI) and, as such, intentional or accidental incidents that alter the regulation of their parameters, feedback lines and/or set points can have dramatic effects on the safety of citizens [WFD10].

Among other security-related issues, resilience is recognized as one of the keys to understand how much damage can be brought to a system and its surrounding environment in case of a successful cyber-attack [DRKS08]. The concept of resilience — literally defined as "*the capacity to recover quickly from difficulties; toughness*" or "*the quality of being able to return quickly to a previous good condition after problems*" — emerges as an additional target, complementary to protection from external threats, but not subordinate to it. This line of thought is pervasive in the Presidential Policy Directive 21 [Oba13] about CI security, 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 incidents*". 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 this paper.

We are interested in a quantitative and succinct measure of resilience, i.e., one that describes as precisely as possible the amount of damage that a system can tolerate before becoming unstable or exhibit undesirable and potentially dangerous behaviors. Mostly outside of cyber-attack scenarios, resilience is a well-studied topic from a control-theoretic standpoint [RKC09], [Far15], [WMW+16], but "classical" approaches are limited to relatively small system components with tractable dynamics. Here we consider simulation techniques to bridge the gap from single components to the system as a whole, and deal with nonlinear and hybrid dynamics. We rely on industry-standard Matlab/Simulink® software to provide modeling and simulation capabilities, and we consider Monte-Carlo style simulations — henceforth referred to as *concrete simulation* — to assess the envelope of system responses with and without attacks and quantify resilience as a ratio of the envelopes. In spite of its expressive power, concrete simulation is often insufficient to foresee all the potential effects of alterations brought by cyber-attacks because the combinatorial explosion of potential states. We propose to complement concrete simulation with algorithmic techniques which, without executing the system, analyze it to find violations of stated requirements — henceforth refereed to as *symbolic simulation*. The approach we consider is Model Check-

Giò Mùrino and Armando Tacchella are with "Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi" (DIBRIS), University of Genoa, Viale Causa 13, 16145 Genoa, Italy. E-mail: giuseppina.murino@edu.unige.it, armando.tacchella@unige.it. The authors wish to thank Leonardo S.p.A. for supporting the research. The corresponding author is Armando Tacchella.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

ing [BKL08], i.e., given the model of a system and a property to check, prove whether the system upholds the property or not. While model checking is well known in hardware and software verification, its application to CPSs is not main-stream yet, so it is fruitful to compare it vs. standard (concrete) simulation techniques.

The main contribution of this paper is twofold. From the methodological point of view, we propose to merge approaches of concrete and symbolic simulation to evaluate the resilience of a system. Here, the goal is to provide maintainers and other stakeholders with a dynamic and quantitative measure of the resilience of a system. From the engineering point of view, we compare both methodologies on a case study related to waste-water treatment. The objective is to provide preliminary evidence that national critical infrastructure security can be enhanced considering our approach both in safety-by-design or safety-by-retrofit frameworks. The rest of the paper is structured as follows. In Section "HYBRID AUTOMATA" we introduce basic terminology and definitions related to the mathematical model we consider for CPSs. In Section "CASE STUDY" we introduce our case study related to a waste-water treatment facility and we describe the model we consider. In Section "CONCRETE VS. SYMBOLIC SIMULATION" we outline the two approaches. In Section "EXPERIMENTAL EVALUATION", we present some results related to the facility herewith described and we conclude the paper in with some final remarks in Section "CONCLUSIONS".

HYBRID AUTOMATA

In order to model CPSs we resort to the formalism of *Hybrid Automata* [ACHH93]. For our purposes, a hybrid automaton can be defined as a tuple $A = (X, V, flow, inv, init, E, jump)$ consisting of the following components:

Variables are a finite ordered set $X = \{x_1, x_2, \dots x_n\}$ of real-valued variables, representing the continuous component of the system's state.

Control modes are a finite set V, representing the discrete component of the system's state.

Flow conditions are expressed with a labeling function $flow$ that assigns a condition to each control mode $v \in V$. The flow condition $flow(v)$ is a predicate over the variables in $X \cup \dot{X}$, where $\dot{X} = \{\dot{x}_1, \dot{x}_2, \dots, \dot{x}_n\}$. The dotted variable $\dot{x}_i$ for $1 \leq i \leq n$ refers to the first derivative of x_i with respect to time, i.e., $\dot{x}_i = \frac{dx}{dt}$.

Invariant conditions determine the constraints of each control mode with the labeling function inv, and *initial conditions* are denoted with the function $init$.

Control switches are a finite multiset $E \in V \times V$. Each control switch (v, v') is a directed edge between a source mode $v \in V$ and a target mode $v' \in V$.

Jump conditions are expressed with a labeling function $jump$ that assigns a jump condition to each control switch $e \in E$. The jump condition $jump(e)$ is a predicate over the variables in $X \cup X'$, where $X' = \{x'_1, x'_2, \dots x'_n\}$. The unprimed symbol x_i, for $1 \leq i \leq n$, refers to the value of the variable x_i before the control switch, and the primed symbol x'_i refers to the value of x_i after the control switch. Thus, a jump condition relates the values of the variables before a control switch to the possible values after the control switch.

Intuitively, the evolution of a hybrid automata can be associated to a sequence of transition between "locations" characterized by specific control modes and values of the variables. In order to frame this concept precisely we define a *state* as a pair $(v, \mathbf{a})$ consisting of a control mode $v \in V$ and a vector $\mathbf{a} = (\mathbf{a_1}, \dots, \mathbf{a_n})$ that represents a value $a_i \in \mathbb{R}$ for each variable $x_i \in X$. The state $(v, \mathbf{a})$ is *admissible* if the predicate $inv(v)$ is true when each variable x_i is replaced by the value a_i. The state $(v, \mathbf{a})$ is *initial* if the predicate $init(v)$ is true when each x_i is replaced by a_i. Consider a pair of admissible states $q = (v, \mathbf{a})$ and $q' = (v', \mathbf{a}')$. The pair (q, q') is a *jump* of A if there is a control switch $e \in E$ with source mode v and target mode v' such that the predicate $jump(e)$ is true when each variable x_i is replaced by the value a_i, and each primed variable x'_i is replaced by the value a'_i. The pair (q, q') is a *flow* of A if $v = v'$ and there is a non-negative real $\delta \in \mathbb{R}_{\geq 0}$ – the duration of the flow – and a differentiable function $\rho : [0, \delta] \to \mathbb{R}^{\kappa}$ – the curve of the flow – such that (i) $\rho(0) = \mathbf{a}$ and $\rho(\delta) = \mathbf{a}'$; (ii) for all time instants $t \in (0, \delta)$ the state $(v, \rho(t))$ is admissible; and (iii) for all time instants $t \in (0, \delta)$, the predicate $flow(v)$ is true when each variable x_i is replaced by the i-th coordinate of the vector $\rho(t)$, and each $\dot{x}_i$ is replaced by the i-th coordinate of $\dot{\rho}(t)$ – where $\dot{\rho} = \frac{d\rho}{dt}$. In words, jumps define the behavior of the automaton when switching from one control mode to another, whereas flows describe the behavior of the automaton inside the control mode. With the concepts above, we can now define executions of the automaton as *trajectories*, i.e., finite sequences $q_0, q_1, \dots, q_k$ of admissible states q_j such that (i) the first state q_0 of the sequence is an initial state of A, and (ii) each pair $(q_j, q_j + 1)$ of consecutive states in the sequence is either a jump of A or a flow of A. A state of A is *reachable* if it is the last state of some trajectory. For the purpose of this paper, the analysis of a hybrid system, amounts to computing the set of reachable states.

Given the expressiveness of the above formalism, it is no surprise that evaluating the reachability of given (sets of) states is, in its most general form, an undecidable problem in hybrid automata [ACHH93]. Even if the control system can be modeled in terms of a *linear hybrid automaton*, i.e., a hybrid automaton where the dynamics of the continuous variables are defined by linear differential inequalities, there is still no guarantee that the exploration of the set of reachable states terminates. The method is still of practical interest, however, because terminations can be enforced by considering the behavior of the system over a bounded interval of time. This technique, known as Bounded Model Checking – see, e.g., [FHT+07] – involves the exploration of increasingly long trajectories until either an unsafe state is reached, or resources (CPU time, memory) are exhausted. Technically, we no longer speak of verification, which is untenable in an infinite state space, but of falsification. Whenever an unsafe state is found, then we have a trajectory witnessing the bug. On the other hand, the unfruitful exploration of increasingly long trajectories is considered an empirical guarantee of safety.

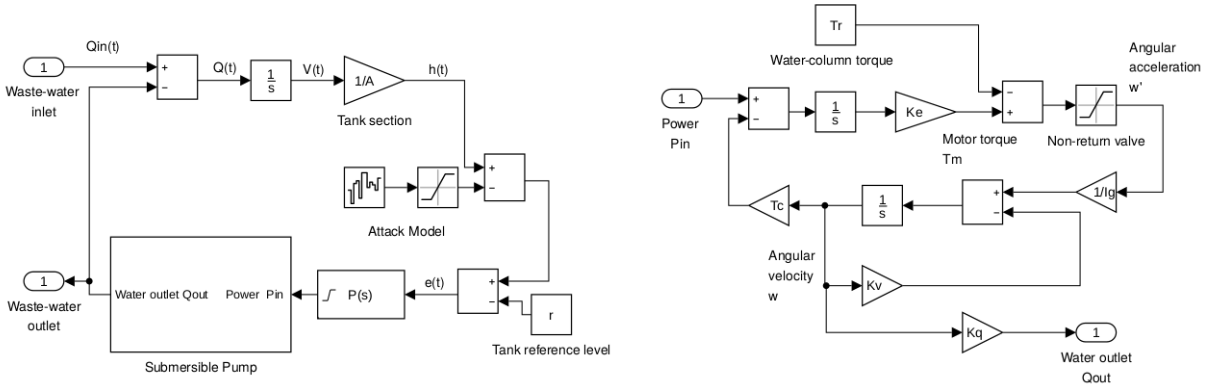


Fig. 1: Model of the nitrification-oxidification tank control system (left) and detail of the submersible pump (right).

CASE STUDY

Our analysis focuses on a waste-water treatment facility whose name and location cannot be disclosed, but whose features and data were made accessible to us to perform this study. In order to achieve a realistic, yet manageable case study, we decided to model only the main waste-water cycle. Components that we did not model include, for instance, sludge treatment and air-quality control. The water cycle is comprised of two main compartments: In the first compartment (pre-treatment), waste-water pumped from the city sewage network is filtered to remove coarse debris, then it is stationed in a tank to remove sand and oil/fat, and finally it is conveyed to a buffer towards further treatment. In the second compartment (biological), waste-water undergoes a denitrification process, then a nitrification-oxidation process, and finally it is conveyed through micro-membranes (MBR) before being released. The subject of our study is the tank wherein nitrification-oxidation (NO) process is carried out. In Figure 1 (left) we show the detailed Matlab/Simulink® model of the NO tank. As we can see from the schematics, we have assumed a simplified linear model for the tank dynamics, whereby the total volume $V(t)$ of fluids contained in the tank is obtained by integrating the net inlet $Q(t)$ which, in turn, is obtained by subtracting the output flow $Q_{out}(t)$ from the tank inlet $Q_{in}(t)$. While the latter is an input to the NO tank, the tank outlet is controlled by electrical pumps driven by a proportional regulator tracking a given reference level r — detail of the motor/pump model is given in Figure 1 (right). The goal of the regulator is to avoid the tank becoming too full, so as to avoid triggering emergency bypasses, or too empty, so as to avoid impairing the chemical process undergoing in the NO tank. Both events are undesirable because bypasses dump to the sea untreated sewage liquor, whereas incomplete chemical processing of waste-water may cause failures in subsequent steps.

We consider the potential modifications that an attacker can bring by gaining access to the system. Since, from a theoretical point of view, the essential mechanism in every CPS is the feedback control loop, an attacker gaining access to the control system can alter the system in three ways only: (a) by changing the set point, (b) by altering the feedback signal, and (c) by changing the regulator parameters. Consider, as an example, the control loop that keeps the level of the tank close to the desired level shown in Figure 1 (left). Here, attack (a) corresponds to changing the desired tank level, attack (b) corresponds to altering the actual tank level, and attack (c) corresponds to changing the proportional gain of the regulator. In practice, an attacker may decide to perform all such actions and in more than one part of the system, as well as other disruptive actions like blocking the functionality of components, e.g., shutting down control devices or flooding them with requests in order to hamper their functionality. As shown in Figure 1 (left), in our simulation we have considered an attack model wherein the feedback signal from the tank is altered — attack (b) — which, in our hypothesis of additive modification, makes it equivalent to alteration of the set point — attack (a). We did not consider attack (c) as well as the possibility of multiple or blocking attacks. The attacker is simulated considering the injection of a faulty signal using a band-limited white noise (BLWN) block, saturated to give output in the range [-1;1]. The idea behind this model is that alterations of the set point or the feedback signal greater than one meter could trigger anomaly alarms — e.g., those guarding against sensor malfunction — but within these bound the attacker can freely modify the feedback value. The impact of the attacker can be regulated by considering the noise power and sample time parameters of the BLWN block. In particular, noise power correlates with the "strength" of the attack, whereas sample time correlates with the "frequency" of the attack. We focus our study of cyber-resilience on the NO tank control system, in the hypothesis that an attacker may gain virtual access to the facility network and act according to the hypotheses above. It is worth noticing that this kind of attack is the same staged by the famous Stuxnet virus [FR11] and therefore, independently from the likelihood of an attacker penetrating the system, it is worth evaluating the impact of such a threat in terms of cyber-resilience.

CONCRETE VS. SYMBOLIC SIMULATION

To be defined as such, an attack must alter the normal behavior of the system in order to cause damage or undesirable effects, which we define as follows:

- Increased frequency of daily power duty cycles to the pump; we consider as a measure of this effect the daily av-

erage of power peaks, defined as the events such that the power delivered to the pump reaches above 80% of the nominal power.

- Increased span of variation in power delivery to the pump; we consider as a measure of this effect the interquartile range[1] of the regulator output signal which in our model corresponds to power delivery to the pump.
- Overflow in the NO tank; technically, such overflows may never happen in practice because of emergency bypasses, but in our model we use the overflow condition to identify when bypasses should be opened.
- Underflow in the NO tank; technically, an underflow in the NO tank may happen in practice due to decreased inlet from the network and excessive outlet of liquor.

All the conditions above may cause damage to the plant or to the surrounding environment to some extent. In particular, the first two conditions may severely reduce the useful life of pump motors and increase the chance of pump failures; the third condition may cause the opening of the bypasses and the dumping of partially-treated liquor at sea; finally, the fourth condition may alter the chemical-biological process, hamper the liquor purification process and cause damage in the NO tank.

As mentioned in the introduction, we consider two different approaches to assess resilience under attack. In particular, starting from the model outlined in the previous section, we wish to compare:

- Empirical analysis of resilience through concrete simulation: since the model in Figure 1 is "executable", this technique is readily available. Its advantages are simplicity, expressiveness and iterative improvability, i.e., running more simulation will yield more precise results. Its main disadvantage is that, due to the combinatorial explosion of inputs, it is impossible to test all potential system configurations to quantify resilience in a precise way.
- Comparing reachable sets through symbolic simulation, i.e., bounded model checking of hybrid automata: given a hybrid automaton modeling the relevant behaviors of the system, it is possible to perform reachability analysis, i.e., (over)estimate the trajectories of the system without actually executing it as required by simulation. Advantages and disadvantages of symbolic simulation are complementary to concrete simulation: since symbolic simulation considers an over-approximation of the reachable sets, no trajectory is possibly left out. However, the computational cost of symbolic simulation is usually much higher than concrete simulation.

As a yardstick for the application of both methodologies, here we consider concrete simulation. In particular, simulation of the plant is performed with and without assuming external attack attempts, and we consider historical data made available from the managing utility to simulate sewage inlet. In order to get meaningful sampling of the potential input space, we consider a daily inlet profile under conditions of maximum utilization - and obtain random variates of such basic profile by adding (band-limited) Gaussian white noise which provides random but realistic deviations from the historical profile. In the following, we call *baseline scenario* the simulation obtained by running the plant without attack, and we define an *attack scenario* where alternations are small, but persistent with the goal of going unnoticed, but still damage the plant in some way. In order to quantify resilience decrease under attack conditions we compare the envelope of the regulator output to the one in the baseline scenario with the following formula

$$R = \frac{\max(P(t))\min(P(t))}{\max(P_a(t))\min(P_a(t))} \tag{1}$$

where $P(t)$ is the regulator power signal in the baseline simulation, and $P_a(t)$ is the same signal in the attack scenario.

To perform resilience evaluation using model checking of hybrid automata we must manually compile, for each control mode, a state equation of the form $\dot{x} = Ax + Bu$, where the vector $x \in \mathbb{R}^n$ represents the state of the system, $u \in \mathbb{R}^m$ is the system input or disturbance, and A, B are matrices of appropriate size. The state equations of specific modes plus (a) a set of potential initial states and (b) boundary conditions for inputs/disturbances has to be supplied to the model checker. While (a) poses no problems and (b) can be shaped in order to take into account the hypothesis of attacks of various intensity brought to the system, the main issue is that the system described in Figure 1 requires at least five control modes to take into account the presence of saturations: one in the pump feedback loop, and another in the motor model. The former saturation is due to the fact that the regulator does not emit "negative" power and does not drive the pump at more than 15 kW — the nominal rating of the pump. The latter saturation is due to the fact that a non-return valve prevents the pump to spin backwards when the motor torque is less than the resistant torque expressed by the water column above the pump. Accounting for these effects requires modeling the system using a hybrid automaton.

When no saturation effect is present, the differential equations modeling the system in Figure 1 are the following:

$$\begin{cases} \dot{h}(t) &= -\frac{K_q}{S}w(t) + \frac{1}{S}Q_{in}(t) \\ \dot{T}_m(t) &= K_e K_{PID} h(t) - K_e T_c w(t) + \\ & \quad -K_e K_{PID}(r(t) + N(t)) \\ \dot{w}(t) &= \frac{1}{I_g}T_m(t) - \frac{1}{K_v}w(t) - \frac{1}{I_g}T_r(t) \end{cases} \tag{2}$$

where:

- $h(t)$ is the height of the liquor in the NO tank: the inlet flow $Q_{in}(t)$ minus the outlet flow, which is a constant K_q multiplied by the rotational speed of the pump $w(t)$, both divided by the area of the tank S, determine the variation of the tank height.
- $T_m(t)$ is the torque of the motor powering the pump: to obtain it, we consider $e(t) = h(t) - r(t) - N(t)$ — called *error signal* in the following; the variation of T_m is influenced by two terms: the product of $e(t)$, the proportional gain of the controller K_{PID}, and the power transfer K_e; the second term is $K_e T_c w(t)$, i.e., a quantity proportional to the speed of the motor which represents the tendency of asynchronous motors to decrease torque when approaching synchronous speed.

[1] The interquartile range of a distribution is the difference between its 75^{th} percentile and 25^{th} percentile. As such, it provides a measure of the dispersion of the distributions values, one which is more resistant than standard deviation to the presence of outliers.

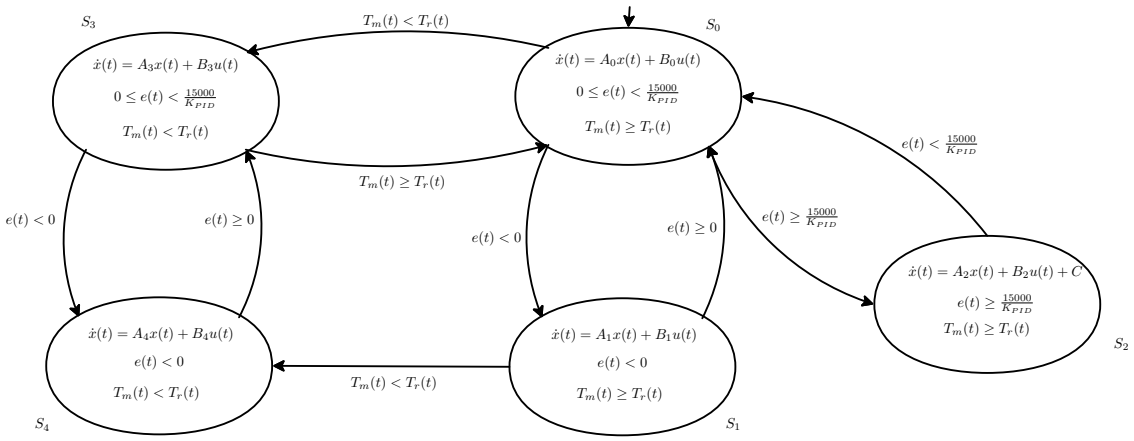


Fig. 2: Hybrid automaton representing the NO tank control system with saturations in power delivery and motor pump. Each oval represents a control mode, i.e., a discrete state S_i with $i \in \{0, 1, 2, 3, 4\}$. Inside the ovals, the flow is described by an ordinary differential equation in matrix form (variables as defined in the text). The invariants are the predicate appearing below the differential equation. Each edge is labeled with the corresponding jump condition, i.e., when the condition becomes true, the control mode is changed. The initial state is denoted with an incoming arrow.

- $w(t)$ is the rotational speed of the pump: the corresponding angular acceleration $\dot{w}$ is determined by the difference between the motor torque T_m on one side, and the resistant torque T_r — due to the water column above the pump — plus the torque due to viscous friction $\frac{w(t)}{K_v}$ on the other side; here, I_g is the rotational inertia of the pump rotor and K_v is the viscous friction coefficient.

If $N(t) = 0$ for all $t \in (0, +\infty)$, then the equations represent the baseline scenario, otherwise $N(t)$ represents the attacker interfering with the set point $r(t)$ in the attack scenario. Let us now consider the *state vector* as the (column) vector $x(t) = [h(t), T_m(t), w(t)]$ and the *input vector* as the (column) vector $u(t) = [Q_{in}(t), r(t) + N(t), T_r(t)]$. Equation (2) can be rewritten as $\dot{x} = A_0x + B_0u$ where A_0 and B_0 are 3×3 matrices:

$$A_0 = \begin{bmatrix} 0 & 0 & -\frac{K_q}{S} \\ K_eK_{PID} & 0 & -K_eT_c \\ 0 & \frac{1}{I_g} & -K_v \end{bmatrix} \tag{3}$$

$$B_0 = \begin{bmatrix} \frac{1}{S} & 0 & 0 \\ 0 & -K_eK_{PID} & 0 \\ 0 & 0 & -\frac{1}{I_g} \end{bmatrix} \tag{4}$$

These equations correspond to the control mode S_0 in Figure 2 where the invariants $0 \le e(t) < \frac{15000}{K_{PID}}$ and $T_m(t) \ge T_r(t)$ are both satisfied, i.e., no saturation effect is present. The other control modes are obtained considering what happens to equation (2) when the invariants above are not satisfied.

In particular, when the error signal $e(t)$ is negative, the power delivered to the motor pump is 0. This dynamic can be described by a system of the form $\dot{x} = A_1x + B_1u$ where the matrices A_1 and B_1 are defined as follows:

$$A_1 = \begin{bmatrix} 0 & 0 & -\frac{K_q}{S} \\ \mathbf{0} & 0 & -K_eT_c \\ 0 & \frac{1}{I_g} & -K_v \end{bmatrix} \tag{5}$$

$$B_1 = \begin{bmatrix} \frac{1}{S} & 0 & 0 \\ 0 & \mathbf{0} & 0 \\ 0 & 0 & -\frac{1}{I_g} \end{bmatrix} \tag{6}$$

As we can see, in both A_1 and B_1 the coefficient due to power transfer from the controller to the pump (K_eK_{PID}) is now 0 (in bold). In Figure 2 this corresponds to control mode S_1. On the other hand, whenever $e(t) \ge \frac{15000}{K_{PID}}$, then the power delivery to the motor of the pump is capped at $K_C = 15$ kW. The corresponding dynamic can now be described by a system of the form $\dot{x} = A_2x + B_2u + C$ where C is the (column) vector $C = [0, K_eK_C, 0]$, $A_2 = A_1$ and $B_2 = B_1$; this configuration corresponds to control model S_2 in Figure 2. Notice that control modes S_0, S_1 and S_2 all satisfy the invariant $T_m(t) \ge T_r(t)$. In principle, in each such state this invariant might not hold, which would result in an automaton with six states. However, in state S_3 since the motor is operating at nominal power, it is not possible to reach a configuration where $T_m(t) < T_r(t)$ unless there is a problem in the pump outlet, e.g., a clog, which is not the subject of our attention.

Control modes S_3 and S_4 in Figure 2 correspond to configurations in which $T_m(t) < T_r(t)$. In case of S_3 this happens when the power signal delivered to the motor is not strong enough to overcome the resistant torque. The matrices corresponding to this mode are:

$$A_3 = \begin{bmatrix} 0 & 0 & -\frac{K_q}{S} \\ K_eK_{PID} & 0 & -K_eT_c \\ 0 & \mathbf{0} & -K_v \end{bmatrix} \tag{7}$$

$$B_3 = \begin{bmatrix} \frac{1}{S} & 0 & 0 \\ 0 & -K_eK_{PID} & 0 \\ 0 & 0 & \mathbf{0} \end{bmatrix} \tag{8}$$

As before, we highlight in boldface the terms which differ with respect to matrices A_0 and B_0 in equations (3) and (4). In this control mode, the pump (if moving) is bound to stop because of the viscous friction. Additionally, is $e(t) < 0$ then the configuration modeled by control mode

NOISE POWER	SAMPLE TIME	NO tank level [m]	MBR tank level [m]	PID input	Power IQR [KW]	Power Peaks (daily average)	MBR input flow [m^3/s]
	60	3.5 - 4.3	0.3 - 1.8	OK	5.118	3431.9	OK
100	600	3.7 - 5.25	0 - 3.4	1.6	3.658	5849.2	0.32
	6000	4.3 - 5	0 - 2.7	OK	0.984	89.24	0.305
	60	3.5 - 4	0.4 - 1.6	OK	5.369	109.23	OK
1000	600	3.5 - 5.3	0 - 3.8	2	3.381	6983.9	0.32
	6000	3.5 - 5.62	0 - 3.8	1.7	1.193	1941.2	0.32
	60	3.7 -4.10	0.3 - 1.8	OK	5.369	109.23	OK
10000	600	3.5 - 5.3	0 - 4.10	1.8	3.610	4690.3	0.32
	6000	3.5 - 5.62	0 - 4.51	2.12	2.841	4331.2	0.32

Fig. 3: Results of concrete simulation in the attack scenario.

S_4 is reached. Here, no power is delivered to the motor and no acceleration is given to the pump. Matrices A_4 and B_4 can be obtained by A_1 and B_1 by zeroing the same entries highlighted in (7) and (8). The only dynamics left in S_4 are those related to the tank filling by effect of the inlet Q_{in}. Notice that, in this configuration it is impossible for $T_m(t)$ to become greater than $T_r(t)$. Therefore, the only possibility is for $e(t)$ to become larger than 0 and then large enough to make $T_m(t) > T_r(t)$ again, i.e., move from control model S_4, back to S_3 and then S_0.

EXPERIMENTAL EVALUATION

In Figure 3 we present a table with the results obtained by simulating the undercover attack scenario in 9 different configurations of the BLWN block simulating the attacker. Simulations run across 100 days starting from regime conditions — corresponding to control mode S_0 in Figure 2 — and statistics are reported on a daily basis. The table is organized as follows. Columns "NOISE POWER" and "SAMPLE TIME" report the corresponding parameters of the BLWN block; columns "NO tank level" and "MBR tank level" report the minimum and maximum levels reached during simulation by the NO and MBR tanks, respectively — the MBR tanks are next to the NO tank, and they are critical for the whole process;"PID input" reports about the error signal $e(t)$ entering the regulator: OK means that the signal is contained within the range $[-4.5; +1]$, a value means that the range was exceeded up to that value; "Power IQR" and "Power Peaks" report about regulator output: "IQR" is the interquartile range of the power signal, and "Peaks" are the number of times in which 80% of the nominal pump power is exceeded on a daily basis; finally, "MBR input flow" refers to the NO tank outlet which is also the MBR tank inlet: OK means that the inlet is less than 0.3 cubic meters per second, a value means that the threshold was exceeded up to that value. From the table we draw the conclusion that all the simulations presented in the scenario qualify as attacks because they meet one or more of the conditions defined in the methodological section. The simulations with exceeding thresholds (red values) are clearly definable as such (and become overt attacks), whereas the simulations where the attack frequency is low (60 seconds of sample time) do not have a detectable impact, but more subtle effects instead. In particular, a baseline simulation without hacker features a power IQR of 0.874 KW, whereas in the simulations herewith considered the power IQR never falls below 5 KW when the sample time is 60 seconds; also the baseline scenario does not feature power peaks — power is consistently regulated to follow the tank inlet — whereas in the attack scenarios the average number of peaks can reach the order of thousands. Considering the simulations above, the value of the ratio in equation (1) is 0.13 in all the cases displayed in Figure 3. This means that, under attack conditions, resilience of the system is reduced to almost one tenth of the system working in normal conditions according to the simulation leg of our methodology. Similar results are obtained with other "sensible" signals, including motor torque and angular velocity.

Considering the techniques described in [ALGK11] and embodied in the tool CORA (COntinuous Reachability Analyzer)[2] we consider the model presented in Figure 2 and try to reach conclusions which are independent from the specific simulation. Modeling hybrid dynamics in CORA requires an effort which is beyond the scope of the paper. In order to get an idea of the extra capabilities granted by CORA — and model checking techniques in general — in Figure 4 we show the plot of reachable sets in the plane where the x-axis is the NO tank height and the y-axis is the motor torque considering control mode S_0 of the automaton depicted in Figure 2. As we can see, in the case of no attack brought to the system (plot on the left), the controller is able to maintain a satisfactory height of the tank, even if initial conditions where slightly off-equilibrium, and this is the case for all potential system trajectories from the set of initial states considered. On the other hand, if an attack is brought to the initial state (plot on the right), CORA shows that the system stabilizes again, but this is an artifact of the model since the motor torque, by construction, cannot exceed 30 Nm at maximum rated power. Albeit in a qualitative fashion, using CORA allows us to evaluate the impact of such attack and potentially quantify resilience once the correct dynamics can be taken into account. More precise

[2] http://www.i6.in.tum.de/Main/SoftwareCORA

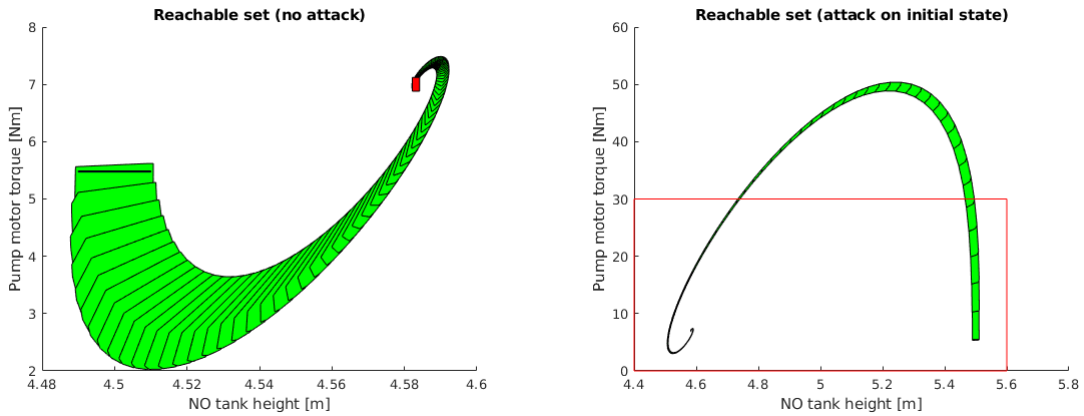


Fig. 4: Analysis of the system with CORA: baseline scenario (left) and attack scenario (right).

quantification can be obtained using a hybrid model, which will be the subject of further study.

CONCLUSIONS

We have shown that, considering a simplified but realistic case study, simulation is able to quantify the decrease of resilience in a system under attack, whereas model checking techniques have the potential to complement it providing further insights. As a future work, we plan to consolidate our methodology by further integrating its elements of the approach and by formalizing the theoretical connections between them. On the engineering side, we wish to extend our analysis to cover the whole waste-water treatment in the facility considered by providing more accurate modeling of various security-critical components. The result of this effort will enable better protection of the infrastructure, and will provide also the first seed of a tool to simulate attacks and responses in order to train facility personnel or test protection devices. We also plan to further validate our methodology by extending it to evaluate resilience of other critical-infrastructure facilities, with a focus on energy production plants and distribution networks.

REFERENCES

[ACHH93] R. Alur, C. Courcoubetis, T.A. Henzinger, and P.H. Ho. Hybrid automata: An algorithmic approach to the specification and verification of hybrid systems. *Lecture notes in computer science*, pages 209–229, 1993.

[ALGK11] Matthias Althoff, Colas Le Guernic, and Bruce H Krogh. Reachable set computation for uncertain time-varying linear systems. In *Proceedings of the 14th international conference on Hybrid systems: computation and control*, pages 93–102. ACM, 2011.

[BHSZ15] Fredrik Björck, Martin Henkel, Janis Stirna, and Jelena Zdravkovic. Cyber resilience-fundamentals for a definition. In *WorldCIST (1)*, pages 311–316, 2015.

[BKL08] Christel Baier, Joost-Pieter Katoen, and Kim Guldstrand Larsen. *Principles of model checking*. MIT press, 2008.

[DRKS08] Salvatore DAntonio, Luigi Romano, Abdelmajid Khelil, and Neeraj Suri. Increasing security and protection through infrastructure resilience: the inspire project. In *International Workshop on Critical Information Infrastructures Security*, pages 109–118. Springer, 2008.

[Far15] Amro M Farid. Static resilience of large flexible engineering systems: Axiomatic design model and measures. *IEEE Systems Journal*, 2015.

[FHT+07] M. Franzle, C. Herde, T. Teige, S. Ratschan, and T. Schubert. Efficient solving of large non-linear arithmetic constraint systems with complex boolean structure. *Journal on Satisfiability, Boolean Modeling and Computation*, 1:209–236, 2007.

[FR11] James P Farwell and Rafal Rohozinski. Stuxnet and the future of cyber war. *Survival*, 53(1):23–40, 2011.

[Lee08] Edward A. Lee. Cyber physical systems: Design challenges. In *11th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC 2008), 5-7 May 2008, Orlando, Florida, USA*, pages 363–369, 2008.

[Oba13] Barack Obama. Presidential policy directive 21: Critical infrastructure security and resilience. *Washington, DC*, 2013.

[RKC09] Dorothy A Reed, Kailash C Kapur, and Richard D Christie. Methodology for assessing the resilience of networked infrastructure. *IEEE Systems Journal*, 3(2):174–180, 2009.

[WFD10] Chunlei Wang, Lan Fang, and Yiqi Dai. A simulation environment for scada security analysis and assessment. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on*, volume 1, pages 342–347. IEEE, 2010.

[WMW+16] Junwei Wang, Raja R Muddada, Hongfeng Wang, Jinliang Ding, Yingzi Lin, Changli Liu, and Wenjun Zhang. Toward a resilient holistic supply chain network system: Concept, review and future direction. *IEEE Systems Journal*, 10(2):410–421, 2016.

GIUSEPPINA MURINO graduated from the University of Genoa, Italy with a "Laurea Magistrale" (MSc equivalent) in Management Engineering in December 2013. She is now a Research Engineer at the Department of Informatics, Bioengineering, Robotics and System Engineering, also pursuing PhD studies in the field of modeling and simulation. She is currently the technical leader of the project aimed at studying the resilience of cyber-physical systems controlling elements of national critical infrastructure.

ARMANDO TACCHELLA graduated from the University of Genoa, Italy with a "Laurea" (MSc equivalent) in Computer Engineering and a PhD in Computer Science and Engineering. Dr. Tacchella was a research associate at Rice University in Houston (US) and then a research fellow at University of Genoa, where he became associate professor of information processing systems in 2005. His research interests are in the field of artificial intelligence and formal methods applied to system engineering.

Performance optimization of edge computing homeland security support applications

Marco Gribaudo
Dipartimento di Elettronica, Informatica e Bioingegneria
Politecnico di Milano
via Ponzio 51
20133, Milano, Italy

Mauro Iacono
Dipartimento di Matematica e Fisica
Università degli Studi della Campania "L. Vanvitelli"
viale Lincoln 5
81100 Caserta, Italy

Agnieszka Jakóbik
Institute of Computer Science
Cracow University of Technology
Warszawska st 24,
31-155 Cracow, Poland

Joanna Kołodziej
Research and Academic Computer Network (NASK)
Kolska st 12, 01-045
Warsaw, Poland

KEYWORDS

Edge computing; Cloud computing; Performance evaluation; Emergency response; Queuing networks; Genetic algorithms.

ABSTRACT

Critical distributed applications have strict requirements over performance parameters, that may affect life of users. This is a limitation that may prevent the exploitation of cost effective solutions such as Cloud Computing (CC) based architectures: in fact, the quality of the connection with the CC facility and the lack of control on cloud resources may limit the overall performances of an application and may cause outages. A way to overcome the problem, and disclose the advantages of CC to critical applications, is provided by Edge Computing (EC). EC adds local support to CC, allowing a better distribution of application tasks according to their timeliness requirements. In this paper we present an innovative Special Weapons And Tactics (SWAT) support application, designed to empower effective operations in wide scenarios, that leverages EC to join CC elasticity and local immediateness, and we exploit Queuing Networks (QN) and Genetic Algorithms (GA) to design and optimize the system parameters for an effective workload distribution.

I. INTRODUCTION

Critical systems are systems characterized by strict requirements (in general, on dependability, performances, or also security attributes), to avoid consequences that may be dramatic in terms of economic loss or threats to human life. In order to satisfy such requirements, the design process of the system should carefully be planned, and the use of models to support the design of performance parameters during the life cycle of the system may provide a significant help in design choices.

In this paper, the focus is on performance modeling of a reference architecture for a novel application, oriented to support complex SWAT operations in high risk scenarios involving hostages and criminals or terrorists. The system, inspired to a firefighting support system presented for the first time in [4], assists a number of SWAT teams on the field by providing sensors, augmented reality tools, tactical information and life protection support to enhance the effectiveness of the team while raising the level of protection for agents, by a real time analysis of the conditions of the field and the possibility of computationally complex coordination and data management applications to lower the impact of possible menaces (e.g., terrorists hiding between hostages or their organized reaction to the SWAT action).

In order to provide a flexible, yet cost effective solution, that allows to combine highly available and timely tasks with heavy and variable workloads, to support both field oriented, tactical and strategic operations, the application is based on EC. This poses the problem of having an optimal dimensioning of the various components of the architecture according to the needs of each mission, or to the evolution of the scenario. This requires the model to be optimized to find the right balance between the different tasks on the system, and the evaluation and optimization have to be feasible with a low computational effort and in a short time, in order to keep the system well tuned along all the duration of the mission.

The EC architecture has been chosen because of its high flexibility [12]. In the EC paradigm CC facilities are complemented with additional computing facilities located where the application has to be used. These additional facilities are "at the edge of the cloud" (hence the name), so that they are close to the user and do not suffer from the problems that affect the communications between the user and the CC facility. The edge facilities may be common computers, mobile devices, low power/low cost devices (e.g. Raspberry Pi), or other devices.

As performances are a critical issue since the first phases of the design cycle and along the life cycle of the system, and the system is subject to be deployed in different alternative configurations in relation to scenarios, we complemented a performance evaluation technique, based on QN, with a parameter optimization technique, based on GA. The choice of QN is due to

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

the fact that QN are a consistent, yet simple way to consider all relevant aspect of the system, easily understandable for non specialists and, in general, widely supported by available tools and easily included into more complex modeling frameworks [2][1].

The choice of GA is due to their generality and the fast and reliable evaluation of the possible parameter configurations. They enable to find the near optimal solution of the NP-hard optimization problems faster then traditional 'hard' computing methods, [16], [17].

In this paper, we present a performance evaluation approach for a complex, scalable, critical EC based SWAT operation support and management system, designed to assist missions involving one or more SWAT teams in high-risk scenarios, that leverages on wearable and deployable sensors, augmented reality devices and CC support for strategical mission supervision and additional support, to allow an effective coordination of SWAT teams, providing life parameters monitoring, extended environmental modeling and real time information enhancement.

The paper is organized as follows: after this introduction, Section II introduces the system, III describes the modeling approach, IV presents performance evaluation. Conclusions follow.

II. A CRITICAL EDGE COMPUTING BASED SYSTEM

The proposed system is inspired to the one presented for the first time in [4] and described in details in [**?**]. In these papers a different QN performance model has been presented, together with some performance related considerations, but no optimization support.

The architecture of the system leverages on the commercial availability of several key technologies, namely *Wireless Sensors Networks* (WSN) (including optical and thermal cameras) and related *Internet of Things* (IoT) and *Mobile Sensors Networks* (MSN) [9][15], *Augmented Reality* (AR), *Wearable Computing* (WC), and the already introduced EC and CC. These technologies have been considered, for their maturity level and the existence of cost effective devices on the COTS market, to satisfy the 4 main functionalities that allow to accomplish the mission of the system: *field control*, *agent support*, *tactical support* and *strategical support*. For a matter of clarity, let us consider an example scenario: a terrorist attack into a skyscraper.

A. System functionalities

Field control consists in the capability of providing real time information about what is happening on the field. In our example scenario, the SWAT teams may deploy wireless and/or mobile sensors and cameras to take control and tactical superiority in the building, and may wear sensors to monitor the environment; additionally, a (partial) supplemental view on the environment may be obtained by means of drones, either moving inside or inside the building, so to obtain a partition of the overall scenario into an observable and a non observable part, in which the observable part is directly monitored by the SWAT team leader. Additionally, the building may have an existing sensor network or cameras, eventually partially damaged by the attackers, that may be somehow remotely accessible by the mission leader, outside the action scope of the SWAT team leader.

Agent support is provided by the equipment that is given to each agent. For the purposes of this paper, this equipment consists of wearable sensors to monitor the environment and the health conditions of each agent, and AR support to provide real time additional data superposed to the physical perception of the environment, eventually integrated with a WC device to preprocess data. Data are exchanged in real time with the SWAT team leader, that reports to the mission leader, by means of the tactical support.

Tactical support consists in allowing each SWAT team leader to coordinate his team, control the observable part of the scenario and to take decisions harvesting the information about the current tactical situation on the field. The SWAT team leader should be able to communicate with his team and to control mobile sensors, including drones and cameras or other complex devices. Tactical support is provided by a real time information system running on a dedicated server, locally hosted on the field (e.g. into a van) and capable of performing real time analysis on gathered sensor data and providing updated AR information to the agents belonging to a SWAT team. With reference to the example scenario, one or more SWAT teams are on the field, each one supported by a van in which the SWAT team leader operates, located in close proximity with respect to the skyscraper so to avoid communication problems with the SWAT team.

Strategical support consists in providing the mission leader, which operates remotely, with updated information about the SWAT teams on the field and the field itself, by merging sensed data and external data sources, and additional advanced data analysis tools (e.g. face recognition applications) to obtain strategical superiority and to evaluate the possible alternative scenarios resulting from possible course of events and the consequences of the various options that the SWAT team leader may decide to examine before taking his decisions. In the example, the commander may use a mission supervision environment leveraging information from the various SWAT teams and from external sources such as the internal sensing system of the skyscraper, accessed by the World Wide Web, and Homeland Security department (or equivalent governmental or non-governmental institutions) databases to simulate the course of events or wider scenarios, and the effects of different actions of the SWAT teams. Processing is not real time, but may require elastic computing resources and interfacing with other local or national authorities, such as police, air force, secret services and the aforementioned Homeland Security authorities.

B. System architecture

In order to implement the 4 functionalities of the system, the architecture is organized into 4 main subsystems, namely *Cloud Backend* (CB), *Edge Frontend(s)* (EF), *Personal Support* (PS), *Sensing Network* (SN): they are meant to provide the implementation platform respectively for the strategical support, the tactical support, the cyberfireman support and the field control. A representation of a possible configuration is in Fig. 1.

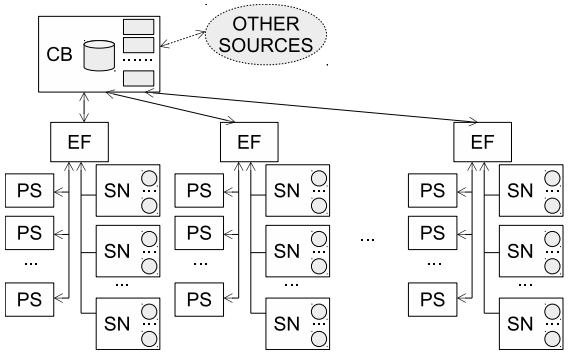


Fig. 1. System architecture layout

CB provides the needed elasticity for the execution of the needed strategical superiority and mission supervision applications, that may vary from a simple decision support system to complementary applications such as signal processing, OLAP, image recognition, simulation, integrating field sensed data and external sources accessed on line, obtained from other authorities in real time and/or other databases already available in the system or special purpose applications. In our example, the CB may run advanced signal processing applications to find signals into the data complex provided by the wearable sensors that agents cannot perceive, such as the voices of nearby hostages or of terrorists, not directly audible by the agents, or can perform data fusion between data from SN and PS and from the building sensor network to provide references about the actual viability of the skyscraper or the position of people by thermal sensing, and integrate AR data or lower the complexity of real time generation of AR data for the EF. Consequently, workload may significantly vary according to the phases of the mission and different needs, that may in principle be not defined before they actually manifest. An important parameter is the volume of service requests coming from the EFs and the amount of supervisory requests generated on the CB by the mission leader (and, possibly, other authorities).

EF provides computing resources for the real time execution of the SWAT team coordination tasks. It is basically provided by a high performance server (e.g. as in [5]), that is connected to the CB by a high speed mobile connection (e.g. 4G or 5G) and is installed on a van. In the following, we will not consider the connection problems with PS and SN that may arise because of the difficult conditions, as it is out of the scope of this paper: we hypothesize, without loss of generality, that the quality of the connection is kept with the use of proper repeaters (e.g. by setting up an ad-hoc network). The EF workload is supposed to be regular and dependent on the number of agents in a SWAT team and on the number and type of sensors in the controlled SN.

PS provides an agent (similarly to what presented in [8]) what is needed to implement wearable sensing and AR by means of a WC system. A minor computing effort is performed locally, basically devoted to service tasks and devices management, and communications with the related EF. Data from sensors (e.g. position, health monitoring, environmental sensors and additional deployable ones), with partial preprocessing, are sent to EF, and commands and AR information are received from the EF. Workload is mainly related to the configuration of the equipment, and sensors contribute to the workload of the EF.

SN consists in a traditional sensor network, in which sensors may be fixed, mobile, disposable, actuable, multimedia or complex sensors. All sensors are controlled by the EF, and affect its workload. The kind and number of sensors in the network may vary, even during a same mission, because of damages, of new needs, of energy exhaustion, of replacement (e.g. see [9][15]).

According to the scenario, the actual configuration may vary, and proper criteria for a correct parameter setting is crucial in order to ensure that performances are sufficient to match the requirements. A configuration will be composed of one CB, one EF per each SWAT team operating in the field, one PS per each agent and one or more SN for each EF.

III. PERFORMANCE MODEL

We model system with the multi-class open queueing network shown in Figure 2. Queuing stations are used to model EF nodes (`Edge Frontend`), network communication (`Edge Network`) and CC nodes (`Cloud Backend`). Arrivals to the `Edge Frontend` nodes represent data acquired from the field. Since different data, of different complexity, can be collected at different speeds, we use costumer classes to model traffic heterogeneity. Without loss of generality, in this work we focus on two different classes of data: the one generated by SN devices, and the one originating from PS nodes. We call λ_{SN} and λ_{PS} respectively the data rate of the two sources. Since we consider scenarios where more EF are present, we model each one with a different queuing station. We call N_{EF} the number of frontend nodes, and we suppose that each of them is characterized by the number of devices it receives from / sends to: for this reason we might have different arrivals for each EF station. For sake of simplicity, we consider EF nodes to be homogeneous, all characterized by the same arrival rate from same classes, and each node providing the same performance and thus serving its requests in the same amount of time. However, the service time for different traffic classes might differ: in particular we call $S_{EF,SN}$ and $S_{EF,PS}$ the average service time required by the two considered data

acquisition classes. To simplify the evaluation of the system, we suppose that EF nodes are equipped by an operating system that, when more requests needs to be served, processes them concurrently. In this way EF nodes can be modeled adopting the processor sharing service discipline.

All nodes send data to the cloud using the same network connection, hence communication is modeled with a single queuing station to which all the EF nodes are connected. Service times might be different for the considered sensor classes, and in particular we set them to $S_{N,SN}$ and $S_{N,PS}$. The main feature of edge computing is that data collected by the sensors is first analyzed by the EF nodes, and only a fraction of them is required to be sent through the network. We model this by allowing jobs of the two classes to immediately leave the EF stations respectively with probabilities $1-p_{SN}$ and $1-p_{PS}$. In this way only a fraction p_{SN} and p_{PS} of data produced by the sensors reaches the cloud.

The cloud part of the application is considered to distributed between N_C equivalent virtual machines that share the arriving load in an uniform way. Besides data received from the EF nodes, the cloud backend is queried also by other users that monitor the event from different institutions. This is modeled by adding an extra class, characterized by a global arrival rate λ_C. Traffic generated by this particular class immediately leaves the system after being served. Data incoming from the sensors (SN and PS classes) go instead back to the EF nodes to allow feedback from the central infrastructure to the server deployed on the field. We call respectively $S_{C,SN}$, $S_{C,PS}$, $S_{C,C}$ the average service time required by the three considered classes. Table III summarizes model parameters and shows the base-line value considered in the following part of the paper. Values have been validated as realistic from experts of the field.

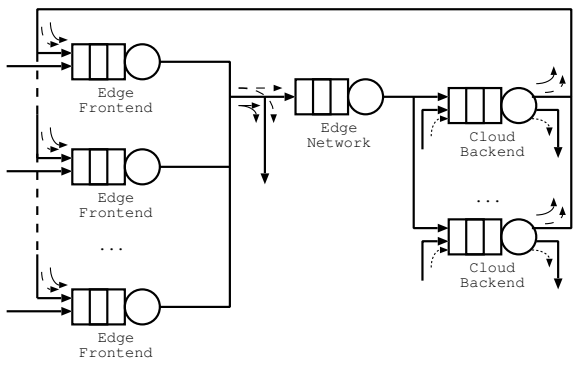


Fig. 2. Queuing network model for the system

IV. SCENARIOS AND ANALYSIS

The model presented in Figure 2 can be used to study several deployment scenarios. We will focus on a couple of them: maximizing the system computational power, while minimizing the cost for acquiring resources; and considering the impact of the coordination activity on the cloud on the system response time.

Param.	Description	Value
N_{EF}	Num. of EF nodes	2
λ_{SN}	Arrival rates of	1.25 data/s.
λ_{PS}	SN and PS classes	1.25 data/s.
$S_{EF,SN}$	Service times	0.2 s.
$S_{EF,PS}$	at the EF nodes	0.35 s.
p_{SN}	Probability of	0.18
p_{PS}	going to the cloud	0.12
$S_{N,SN}$	Network	10 ms.
$S_{N,PS}$	transfer times	15 ms.
N_C	Number of VMs	2
λ_C	Coordination traffic	0.1 req./s.
$S_{C,SN}$	Service times	1.5 s.
$S_{C,PS}$	at the cloud	1 s.
$S_{C,C}$	backend	2 s.

TABLE I: The model parameters

A. MAXIMIZING SYSTEM POWER

As defined by Kleinrock in [19], *system power* corresponds to $\Pi = \frac{X}{R}$: the ratio of the throughput and the response time of a system, or of one of its components. The main rationale about this performance index is that it increases when either the throughput increases, or when the response time decreases. In other words, the configuration in which a component works at its maximum system power corresponds to the point where users receives the best tradeoff between speed and latency. In this first study we want to determine the point that maximizes system power, considering a fixed total data generation rate. In particular, we imagine that data is collected at a total rate of Λ readings per second, which is a structural parameter that depends on the specific sensing technology and communication protocol between the sensors and the EF that is employed. The choice is then to properly balance reading from the two class of sensors. In particular, we call $0 \leq \beta \leq 1$ the *traffic mix* parameter, and we define $\lambda_{SN} = \beta\Lambda$ and $\lambda_{PS} = (1-\beta)\Lambda$. Another parameter we want to take into consideration is the number of EF nodes that is being used. Indeed, the higher the number of EF, the higher will be the system power of the application. However, considering a larger number of EF nodes implies a higher cost: in particular we consider a discount function $g_\mu(N_{EF})$, that describes the impact of having a higher cost due to a larger number of EF nodes, on the average system power we aim to maximize. In this example we set $g_\mu(N_{EF}) = \frac{\mu}{N_{EF}}$, with $\mu > 0$. The rationale is that a smaller number of nodes has an higher impact on the system power we aim to maximize. We thus create an objective function $f_1(\beta, N_{EF})$ defined as:

$$f_1(\beta, N_{EF}) = \alpha \frac{\lambda_{SN}}{R_{SN}(\beta, N_{EF})} + (1-\alpha)\frac{\lambda_{PS}}{R_{PS}(\beta, N_{EF})} + \frac{\mu}{N_{EF}} \quad (1)$$

where $R_{SN}(\beta, N_{EF})$ and $R_{PS}(\beta, N_{EF})$ represent the response time of the two sensor data classes. Constant α is used instead to balance the interest between the sensors ($\alpha = 1$) and the personal devices ($\lambda = 0$).
Figure 3 shows the system power of the two data classes and the value of the objective function 1 for different traffic mixes β and for a different number of nodes N_{EF}. As it can be seen, the system power Π_{SN} of the sensor network has a monotonic behaviour, while the one of the personal support class Π_{PS} shows a maximum when the number of EF nodes $N_{EF} = 1$. The figure also shows the values of the objective function $f_1(\beta, N_{EF})$ for $\alpha = 0.5$, and $\mu = 0.5$ or $\mu = 1$, to consider two different cost scenario where data classes are equally important. As it can be seen, in some cases behaviour is monotonic, while in some other it experiences some internal convexity. This motivates the use of optimization algorithms to study the values of β and N_{EF} where this maximum is reached.

Figure 4 shows the different values of the parameter

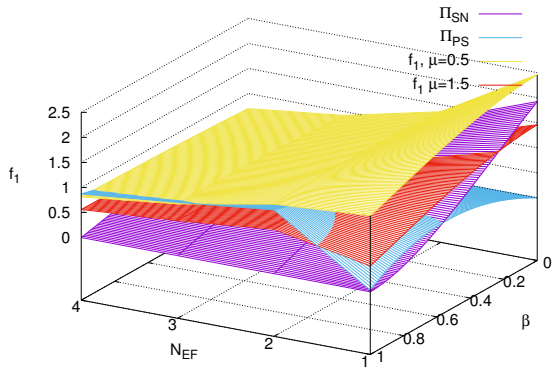


Fig. 3. Ibjective function $f_1(\beta, N_{EF})$ for different values of resource cost parameter μ and $\alpha = 0.5$.

that maximizes the objective function for different cost parameters μ, and the balance parameters α. When the cost is very high, it is better to keep as few EF nodes as possible: however, depending on the balance parameter α there could be optimal non-trivial traffic mixes that optimizes the system. When the cost is very low, it is not always true that it is better to use a larger number of EF nodes: the balance parameter α can make the system work at its optimum configuration even with $N_{EF} = 1$, and $N_{EF} > 2$ are never needed.

B. THE EFFECT OF THE COORDINATORS

Different players in the management of the event monitor the evolution of the scenario accessing the cloud back-end of the application. Depending on their update rate λ_C, they might have a clearer and more up-to-date picture of the scenario, increasing their ability to help and their effectiveness in solving the situation. However, a larger upload rate might overload the system, leading to an unstable behaviour. In this study, we consider the effect on the response time of the traffic generated by the coordinators, varying $0.1 \leq \lambda_C \leq 0.5$

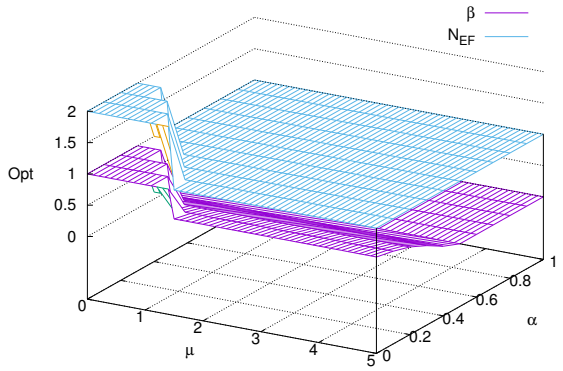


Fig. 4. Optimum values of β and N_{EF} as function of the cost parameter μ, and the balance parameter α.

req./s., and we combine the analysis with the traffic mix introduced in the previous section. We measure the non-effectiveness of the coordination actions with a function $l_\theta(\lambda_C) > 0$: larger values denotes a less effective coordination of forces, while values that tends to 0 denotes optima interactions. In this work we use $l_\theta(\lambda_C) = \frac{\theta}{\lambda_C}$, where $\theta > 0$ is a scaling parameter to make the measure compatible with the expected response times. Again we want to account the cost on the infrastructure in our performance assessment. In this case, however, since response time must be minimized, it must increase with the number of resources. Moreover, since we are studying the coordinators, which mainly affect the cloud, we must account not only for the number of EF nodes N_{EF}, but also for the number of provisioned VMs N_C. In particular, we use a cost function $h_\nu(N_{EF}, N_C)$, and we set it to $h_\nu(N_{EF}, N_C) = \nu \cdot (N_{EF} + N_C)$ (with $\nu > 0$ as a metric conversion parameter), and we set the objective function $f_2(\beta, N_{EF}, N_C, \lambda_C)$ to:

$$
\begin{aligned}
f_2(\beta, N_{EF}, N_C, \lambda_C) \quad = \quad & \frac{\lambda_{SN}}{\Lambda^*} R_{SN}(\beta, N_{EF}, N_C) + \\
& \frac{\lambda_{PS}}{\Lambda^*} R_{PS}(\beta, N_{EF}, N_C) + \\
& \frac{\lambda_C}{\Lambda^*} R_C(\beta, N_{EF}, N_C) + \\
& \frac{\theta}{\lambda_C} + \nu \cdot (N_{EF} + N_C) \qquad (2)
\end{aligned}
$$

where $R_C(\beta, N_{EF}, N_C)$ is the response time of the coordinator class, and $\Lambda^* = \Lambda + \lambda_C$ is the total system throughput.
Figure 5 shows the value of the objective function $f_2(\beta, N_{EF}, N_C, \lambda_C)$ for different traffic mix β and coordinator traffic λ_c in different resource configuration. As it can be seen, there are many cases in which the system is not stable (represented in the figure with $f_2(\beta, N_{EF}, N_C, \lambda_C) = 0$), and the objective function is concave in some cases. Figure 6 shows for which values of β and λ_C the minimum of the objective function is reached. It is interesting to see that, when the cost

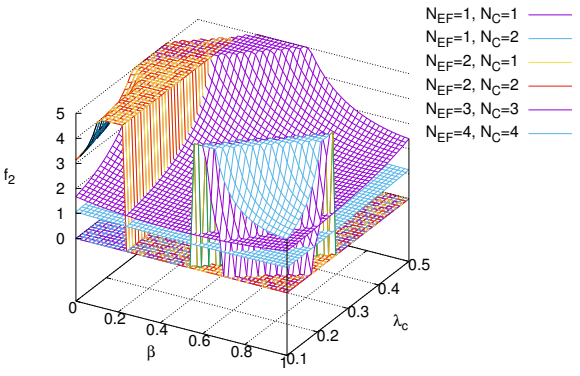


Fig. 5. Objective function $f_2(\beta, N_{EF}, N_C, \lambda_C)$ for different combinations of EF nodes N_{EF} and cloud VMs N_C.

is very high; the minimum is obtained for $\beta = 1$ and $\lambda_C = 0.18$ req./s. When the cost is lower, the minimum is obtained in different combinations of β and λ_C. The optimization algorithm also determined the optimum number of nodes N_{EF} and N_C: even if not explicitly shown in the figure to make it more readable, the minimum of the objective function is reached when the minimum number of resource is used when $\beta < 1$. In the flat area with $\beta = 1$ instead, the optimum is reached when the maximum number of resources is used.

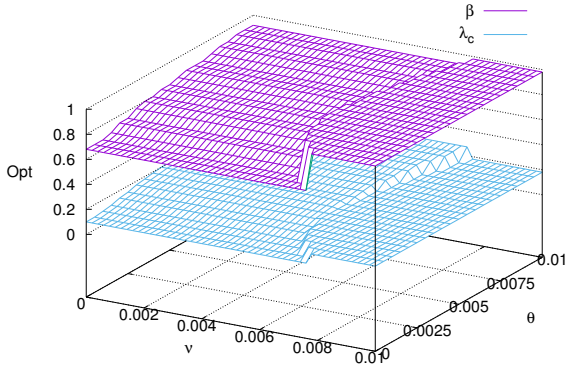


Fig. 6. Optimum values of β and λ_C as function of the cost parameter ν, and coordination effectiveness θ.

V. RELATED WORKS

The integration of Cloud computing, mobile computing and Internet of Things (IoT) allows to exploit a greater flexibility in performance engineering and tuning with respect to cloud architectures, and extends their great available power and elasticity by allowing an efficient collection and transmission of massive and ubiquitous data sources, such as IoT devices and mobile devices, to overcome the limitations of the weak point of cloud architectures, the network connection towards the edge of the cloud. The readers may refer to [23] for a first approach to edge architectures problems and features, while the general context of application of these solution is presented in [12] and [21], that report about security related and law related topics. Of course, a correct balance, and related tuning and adaptation, of system parameters is crucial to achieve desired overall performance figures, as discussed in [22], [24], [26], [20] and [6], that may help readers in understanding the general framework in which responsiveness, scalability, privacy and fault tolerance in edge architectures may be achieved and modeled or measured. Due to the interest about edge-based solutions, there is an intense research activity that spans over performances and system/software management, communication protocols and architectural solutions. In [10] the fog computing perspective and some related applications are presented. In [18] the role of containers in edge-based software architectures is discussed and quantitatively documented, and in [14] the general software architectural aspects are examined. In [7] the focus is on the relation of edge architectures with IoT, with special attention to communication problems. In [11] an autonomic approach to edge architecture management is discussed. In [25] the edge approach is extended towards osmotic computing. For the aspects focusing specifically on cloud related problems and performance modeling, readers may refer to other previous papers of the authors, stating their positions that influenced this work ([16], [17], [3], [13]. In the first two papers authors successfully applied a specifically designed GA for solving NP-hard task scheduling problems for Computational Clouds. Their novel approach assumed generating every consecutive population from 'freezing' the worst parents instead of removing them from the population. New population was of the equal size as the previous population. This enabled to generate more broad sub-optimal solutions space and hence find the solution faster.

VI. CONCLUSIONS AND FUTURE WORK

Results show the effectiveness of the presented modeling approach for the performance evaluation of critical edge based systems, that showed a non trivial behavior. Future work include a better exploitation of GA in the technique and a more detailed characterization of the system.

VII. ACKNOWLEDGEMENTS

This article is partially based upon work from COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet), supported by COST (European Cooperation in Science and Technology)

REFERENCES

[1] S. Balsamo and A. Marin. Queueing networks. In M. Bernardo and J. Hillston, editors, *Formal Methods for Performance Evaluation, 7th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2007, Bertinoro, Italy, May 28-June 2, 2007, Advanced Lectures*, volume 4486 of *Lecture Notes in Computer Science*, pages 34–82. Springer, 2007.

[2] E. Barbierato, M. Gribaudo, and M. Iacono. Exploiting multiformalism models for testing and performance evaluation in SIMTHESys. 2011.

[3] E. Barbierato, M. Gribaudo, and M. Iacono. Modeling and evaluating the effects of Big Data storage resource allocation in global scale cloud architectures. *International Journal of Data Warehousing and Mining*, 12(2):1–20, 2016.
[4] E. Cavalieri d'Oro, S. Colombo, M. Gribaudo, M. Iacono, D. Manca, and P. Piazzolla. Modeling and evaluating performances of complex edge computing based systems: a firefighting support system case study. In *Valuetools 2017,*, 2018.
[5] D. Cerotti, M. Gribaudo, M. Iacono, and P. Piazzolla. Modeling and analysis of performances for concurrent multithread applications on multicore and graphics processing unit systems. *Concurrency and Computation: Practice and Experience*, 28(2):438–452, 2016. cpe.3504.
[6] H. Chang, A. Hari, S. Mukherjee, and T. V. Lakshman. Bringing the cloud to the edge. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 346–351, April 2014.
[7] M. Chiang and T. Zhang. Fog and iot: An overview of research opportunities. *IEEE Internet of Things Journal*, 3(6):854–864, Dec 2016.
[8] S. Colombo, M. Gribaudo, M. Iacono, D. Manca, and P. Piazzolla. A low-cost distributed IoT-based augmented reality interactive simulator for team training. In *Proceedings of the 31th European Conference of Modelling and Simulation, May 23-26 2017, Budapest*, page to appear, 2017.
[9] M. D'Arienzo, M. Iacono, S. Marrone, and R. Nardone. Estimation of the energy consumption of mobile sensors in wsn environmental monitoring applications. In *Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on*, pages 1588–1593, March 2013.
[10] A. V. Dastjerdi and R. Buyya. Fog computing: Helping the internet of things realize its potential. *Computer*, 49(8):112–116, Aug 2016.
[11] M. Desertot, C. Escoffier, and D. Donsez. Towards an autonomic approach for edge computing: Research articles. *Concurr. Comput. : Pract. Exper.*, 19(14):1901–1916, Sept. 2007.
[12] 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.
[13] M. Gribaudo, M. Iacono, and D. Manini. Improving reliability and performances in large scale distributed applications with erasure codes and replication. *Future Generation Computer Systems*, 56:773 – 782, 2016.
[14] 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.
[15] M. Iacono, E. Romano, and S. Marrone. Adaptive monitoring of marine disasters with intelligent mobile sensor networks. In *Environmental Energy and Structural Monitoring Systems (EESMS), 2010 IEEE Workshop on*, pages 38–45, 2010.
[16] A. Jakbik, D. Grzonka, and F. Palmieri. Non-deterministic security driven meta scheduler for distributed cloud organizations. *Simulation Modelling Practice and Theory*, 76:67 – 81, 2017. High-Performance Modelling and Simulation for Big Data Applications.
[17] A. Jakóbik, D. Grzonka, J. Kołodziej, and H. González-Vélez. Towards secure non-deterministic meta-scheduling for clouds. In *30th European Conference on Modelling and Simulation, ECMS 2016, Regensburg, Germany, May 31 - June 3, 2016, Proceedings.*, pages 596–602, 2016.
[18] 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.
[19] L. Kleinrock. Power and deterministic rules of thumb for probabilistic problems in computer communications. In *ICC '79; International Conference on Communications, Volume 3*, volume 3, pages 43.1.1–43.1.10, 1979.
[20] 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.
[21] 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.
[22] M. Satyanarayanan. The emergence of edge computing. *Computer*, 50(1):30–39, Jan 2017.
[23] 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.
[24] W. Shi and S. Dustdar. The promise of edge computing. *Computer*, 49(5):78–81, May 2016.
[25] 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.
[26] 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.

AUTHOR BIOGRAPHIES

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 modeling, queueing networks, mean-field analysis and spatial models. The main applications to which the previous methodologies are applied comes from cloud computing, multi-core architectures and wireless sensor networks.

MAURO IACONO is a tenured assistant Professor and Senior Researcher in Computing Systems at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy. 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. More information is available at http://www.mauroiacono.com.

AGNIESZKA JAKÓBIK (KROK) received her M.Sc. in the field of stochastic processes at the Jagiellonian University, Cracow, Poland and Ph.D. degree in the field of neural networks at Tadeusz Kosciuszko Cracow University of Technology, Poland, in 2003 and 2007, respectively. From 2009 she is an Assistant Professor. Her e-mail address is: agneskrok@gmail.com

JOANNA KOŁODZIEJ is an associate professor in Research and Academic Computer Network (NASK) Institute and Department of Computer Science of Cracow University of Technology. She is a Head of the Department for Sciences and Development. She serves also as the President of the Polish Chapter of IEEE Computational Intelligence Society. She is also a Honorary Chair of the HiPMoS track of ECMS. Her e-mail address is: joanna.kolodziej68@gmail.com. The detailed information is available at www.joannakolodziej.org

Anchor placement in indoor object tracking systems for virtual reality simulations

Marco Gribaudo
Pietro Piazzolla
Dipartimento di Elettronica, Informatica e Bioingegneria
Politecnico di Milano
via Ponzio 51
20133, Milano, Italy

Mauro Iacono
Dipartimento di Matematica e Fisica
Università degli Studi della Campania
"L. Vanvitelli"
viale Lincoln 5
81100 Caserta, Italy

KEYWORDS

Indoor Object Tracking Systems; Performance evaluation; Virtual reality; Augmented reality; Emergency response; Simulation; Fire fighting; Training; Human simulation.

ABSTRACT

Indoor Object Tracking Systems (IOTS) allow sensing moving objects inside a closed space, where GPS is not available. Besides the most popular use, indoor navigation, IOTS may also contribute to extend the operational range and the possible applications of Virtual Reality (VR) and Augmented Reality (AR) based technologies, such as complex training scenarios or entertainment oriented simulations: in fact, providing devices with a reliable IOTS support adds realism and allows a higher degree of safety and interactivity, that allow a high number of people to take part and collaborate in a simulated scene with a very high degree of physical interaction. In this paper we introduce a novel approach for the optimization and the evaluation of movement tracking in a IOTS based system, oriented to VR/AR applications, with special focus on the training of teams. Our proposal is applied to a case study, an AR application designed to assist business buildings workers in fire extinguisher use training. Performances of our proposal are evaluated by means of a simulation, and results are validated in a test scenario based on Ultra Wide Frequency positioning by means of a simulation scenario fed with real data from anchors.

I. INTRODUCTION

Sensing position and movement of people and things is an important problem in different application fields. While applications in open range are common and well known also in the daily experience of a wide public (GPS based navigation or logistic tracking are considered commodities), there is not yet an obvious solution for closed spaces. Although it may be considered as the small scale correspondent problem, there are some specific aspects that make unapplicable the same solutions: typical applications are designed to work inside buildings, that are generally full of obstacles (besides the roof, that shields GPS signals, walls and objects also shield local radio emitters and there is electromagnetic noise because of the activities), involve human beings moving in constricted spaces (again walls, but also furnitures and objects on the floor) and require more precision (as the relevant distances are smaller). Interaction with the environment is thus crucial to provide quality services, specially when IOTS are used to track relative movements of physical objects in the space (e.g. rotation of a stick handled by a subject moving into the environment). The precision of fine movement tracking is specially important for VR/AR applications, as they involve a physical interaction of human subjects with virtual or mixed (virtual and real) objects in a real physical space, that may be not evidently perceptible, with possible harmful consequences if the environment has to be a real context. An optimal coverage of the environment by anchors, the elements that allow position and movement sensing in IOTS, is crucial for the delivery of quality services and to enhance both the realism and the safety of VR/AR applications designed to be executed in real world environments. In this paper we study an optimization and performance evaluation method for anchor positioning in a IOTS. The method is demonstrated by using real anchors that feed a specific simulation tool, in order to avoid experimentation with human subjects in the loop. The test scenario uses an Ultra Wide Band (UWB) radio based IOTS solution, but results can be generalized. The chosen application is devoted to the training of teams of non specialist personnel in their workplace, by means of an AR application, for a safe reaction in case of fire emergencies.

The paper is organized as follows: in next Section we present a short literature survey focusing on indoor positioning and tracking literature. The studied system is presented in subsequent Section III, along with the description of simulation dataset. In Section IV we present the proposed methodology, the simulation approach and the results, with a discussion of the outcomes. Conclusions follow.

II. RELATED WORKS

Indoor positioning [7] and indoor tracking [3] are two related and interconnected (actually, partially super-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

posed) application areas that specialize positioning and tracking problems to closed spaces. Indoor positioning deals with the problem of locating with precision an item in a closed space, typically a building like an office, a school, a warehouse; indoor tracking deals with the problem of contextualizing moving objects, or parts of an object, in an environment that is typically bounded by perimeter walls and in presence of obstacles, like inner walls, other fixed or moving objects or steps. Both problems refer to position or movements in a given coordinate system, that may be based on 2 or 3 dimensions.

While for open range applications GPS based solutions are ubiquous and manage to capture the most of the needs of civil applications, there is a number of different proposals that deal with the peculiarities of closed spaces. In the open range, there is generally no shielding of GPS signals; in indoor positioning and tracking, GPS signals are typically shielded, thus not available, and the system has to rely on local references, namely *anchors*. The presence of obstacles may influence anchor visibility with respect to the position of the target object. Suitable technologies should be able to operate efficiently with no harm to humans that may interact with the environment and should be able to provide high precision, to fit the scale of the environment.

Uses of indoor positioning include detection of goods in a wharehouse, location of people in a building, location of intervention points in plant maintenance, implementation of reference points along paths, presence detection of objects or animals in given areas, location dependent interactions or information delivery. Uses of indoor tracking include monitoring of parts in assembly lines or plants, security and safety monitoring, indoor navigation, indoor autonomous veichles management, warehouse automation, operations in critical or dangerous environments.

Considering radio frequency (RF) based technologies, some of the most popular for indoor positioning are active RFID [9], UWB [11], IEEE 802.11 (also known as WLAN) [5], Bluetooth [4]. Active RFID is very popular for its low cost and for the widespread use of its passive version, that allows information exchange by exploiting the electromagnetic signals exchanged between a powerless tag and a reader. In the active version, this technology is able to transmit an identification code when solicited. The range is 1-2 meters in the first case, tens of meters in the second, the cost is low. UWB

Commercial solutions exist that couple different technologies (e.g. with ultrasounds, GPS, cellular, infrared), or that use different, proprietary solutions.

For sake of space, for a comprehensive review of techniques, metrics and technologies used in indoor positioning and tracking we again suggest to refer to [7] and [3], that also provide a significant list of references.

An evaluation of algorithms for UWB indoor tracking is presented in [2]. In this paper interested readers can find a realistic evaluation of a number of different location and tracking algorithms that include trilateration, least square-multidimensional scaling, extended Kalman filter and particle filter, that are compared by using a realistic ranging model specific for UWB based devices.

III. A CASE STUDY

A. Context

Typical indoor tracking systems rely on the interactions of several devices, divided into two groups. Those belonging to the first group, usually called *anchors*, provide reference points in the physical world space and are usually placed in fixed positions. Anchors are able to communicate with one another by mean of different, manufacturers' specific, communication protocols. By exchanging messages they compute reciprocal distances by leveraging on their response times. Devices belonging to the second group, referred as *tags*, move into the portion of space covered by the communication range of the anchors. By exchanging messages with the anchors, their are able to determine their position in the defined space.

We focus our research on devices able to communicate using radio signals to transmit information, such as UWB based devices. Inference from such radio signals based measurements can be quite challenging due to a number of factors, e.g. their dependence on the relative distances/angles among the devices, environmental obstacles, signal power [3].

B. Application setup

The proposed case study concerns the implementation of IOTS features into the equipment that supports a training application. The individual has to be trained to cooperate in a team during an alarm using a fire extinguisher and consequently enact proper operations to limit the consequences, in a wider scenario such the one described in [1]. The trainee plays his role in a simulation, that is implemented by a real time immersive distributed simulator consisting of a personal support node for each team member. Personal support nodes form a peer to peer system, and provide each other the parameters about the behavior of the team member wearing them. Each node consists of a smartphone, that provides computing power, a smart sensing subsystem, that provides information about the team member and his actions, an AR/VR subsystem, providing immersive visual feedback, and an Arduino based coordinator, that controls the sensing subsystem. The smartphone runs the local software component of the distributed simulator, and produces the information for the AR/VR subsystem, that in turn visualizes the other team members and the additional simulated effects superposed to the real image observed by the perspective of the team member. The smart sensing subsystem include specific sensors placed on the dummy tools (e.g. the extinguishers) that the team members will use, to ensure that they are used properly and to provide precise information to compute their effects on the simulated fire. Further details on this simulation architec-

ture are available in [10]. The architecture is depicted in Figure 1 (from [10]).

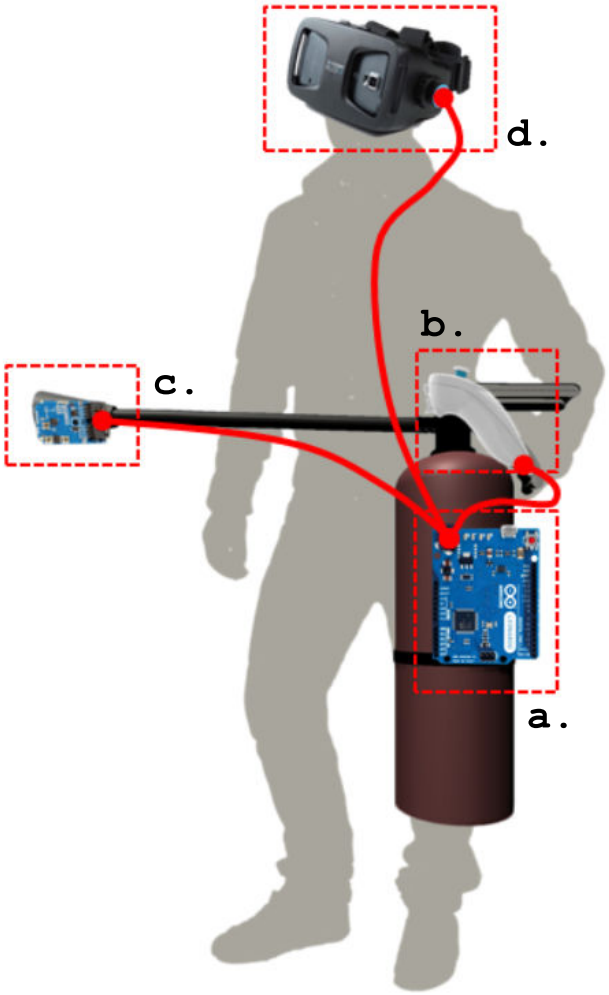


Fig. 1. Peripherals of a personal node, with a coordinator (a.), two sensors (b. and c.) and the AR device (d.)

The smart sensing subsystem also includes a receiver for IOTS, that is capable of exploiting the information coming from the anchors located in the environment. Information is provided to the smartphone, that computes the exact location by means of the proposed metric and exchanges its position with the other nodes to update the simulation and the AR/VR related information.

C. Simulation setup

The dataset for human movements simulation has been chosen among publicly available, sound repositories, collecting real or realistic profiles. The chosen dataset has been selected so to fit the data that are in practice generated by the reference IOTS, to obtain a significant simulation testbed.

This dataset, available at CMU Graphics Lab [6], is a collection of motion capture data recorded from different subjects performing different kind of activities, from sports to everyday tasks and stored in **.bvh* files[8], or Biovision Hierarchy file format. All such activities, because of the capturing tools involved, were recorded indoor thus matching our model assumptions.

Fig. 2. Tag assumed position for our model validation

From the collection, a subset of animation files were selected, in particular we focus on those animations files where the subject was involved with locomotion, especially if there was some sort of interaction the surrounding environment, and on those concerning physical activities such as sports.

The typical indoor tracking system we are investigating in this paper, as seen in Section III, leverages on the use of a tag whose position is triangulated when moving inside a delimited area. This tag is, as usual, applied to the object or person to be tracked. We consider the tag applied to the person at waist level, in close vicinity to her body as it can be seen in Figure 2. The hierarchical structure used to alter the visual appearance of a 3D character to simulate its movement is called *skeleton*. This structure is also used to store data about position and rotation of each body part, as the movement proceeds in time. The elements of this skeleton system are called *bones*, each dedicated to influence a specific portion of the character. We focused data about the hip from the dataset, the bone closest to the waist, where we assumed the tag was positioned. In Figure 3 the bone structure used in our dataset is presented.

In Figure 4 we present a sample positioning system layout. It consists primarly of a given number of UWB emitters (Figure 4-a.) acting as reference points for the tracked tag (Figure 4-b.).

During the setup of the positioning system, each anchor is required to be somehow informed about its distance with respect to the other anchors as well as from the tag (dotted gray lines in Figure 4). The anchors and tag distance from the ground is also required. To this end, a calibration sequence is performed, either man-

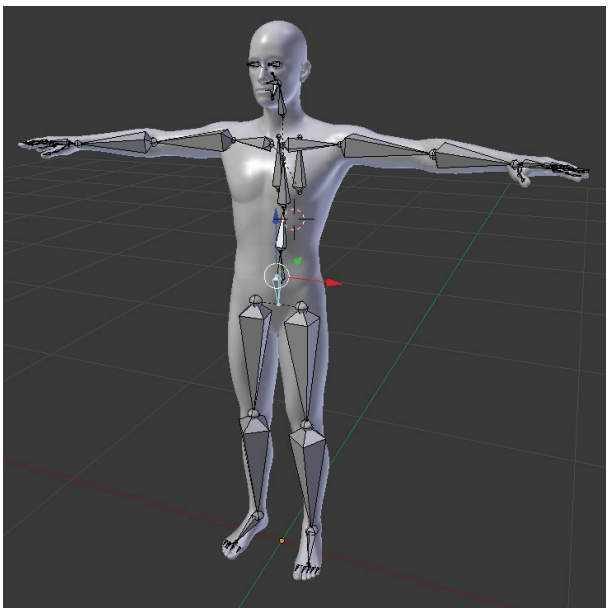


Fig. 3. The skeleton of a 3D character. Each element is called bone. Hips are highlighted in the picture

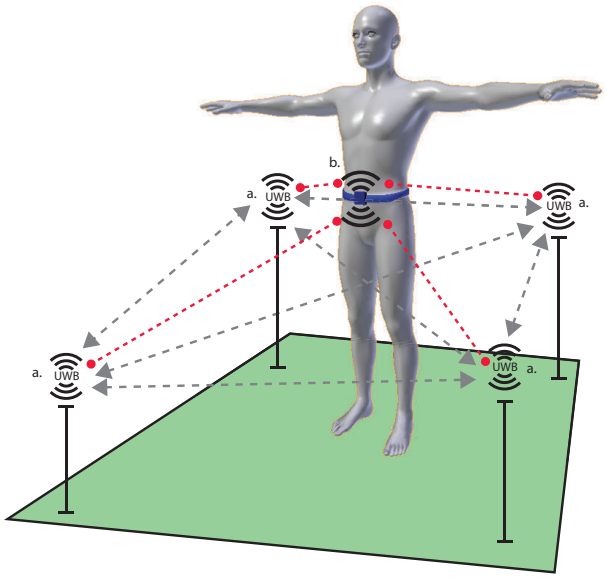


Fig. 4. A sample positioning layout

ually or automatically. In the first case, the required values are provided to the system by human intervention, in the second case an automatic process is started during which the positioning devices compute and exchange information about their relative distances.

The dotted red line represents the tag computing its position relative to the emitters once the system begins working after the calibration procedure.

IV. SIMULATION AND RESULTS

In order to simulate the tracking process, we first present a model of the distance measurement error. The rationale about the model is that it must have an area where measurement occurs at its best, and that the error increases both when the tag is too close or too far away from the anchor. In this work we first start

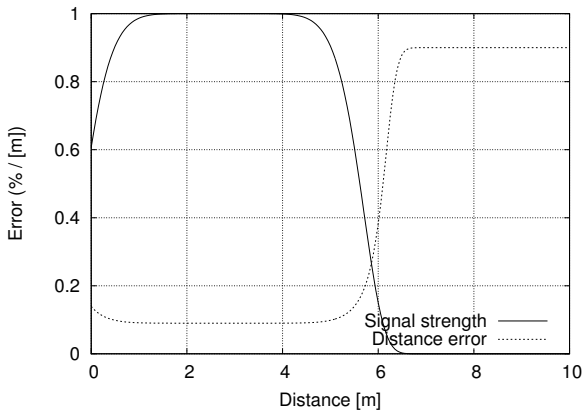


Fig. 5. Distance measurement error

evaluating this measurement quality using the following function $\eta(d)$ of the distance d between the tag and the anchor:

$$\eta(d) = e^{-\left(\frac{d-\mu}{\sigma}\right)^{\gamma}} \tag{1}$$

Parameter μ represents the best distance: the one where the measurement is more precise. Constant σ determines the width of this area, and γ the way in which it fades as the tag moves away from the anchor. In the rest of the paper we have used in Equation 1 $\mu = 2.75$, $\sigma = 2$, $\gamma = 8$. We then suppose that measurements are affected by a gaussian error of a given standard deviation $\nu(d)$ that depends on the distance d. In particular, we suppose that, when the measurement has the best quality, a standard deviation d_0 is achieved. As the quality decreases, the error becomes larger and larger, and in the worst cases it reaches $\frac{d_0}{\alpha}$ with $0 < \alpha \leq 1$. In particular we have defined $\nu(d)$ as follows:

$$\nu(d) = \frac{d_0}{\alpha + (1-\alpha)\eta(d)} \tag{2}$$

In this work we have set $d_0 = 0.09$ and $\alpha = 0.1$ in Equation 2.

Figure 5 shows the evolution of the signal strength $\eta(d)$ and of the distance error $\nu(d)$. Note that both the definitions of $\eta(d)$ and $\nu(d)$ are arbitrary and came from direct experience by working with the technology. More realistic values of ηd and νd can be obtained with an extensive measurement campaign, which was outside the scope of this work that focuses on the methodology of improving the placement of the anchor, but is planned as future work.

The tracking process is corrupted by measurement errors. Following the definition of $\nu(d)$, the error affected distances $\tilde{l}(d)$ are computed as:

$$\tilde{l}(d) = d + N(0,1) \cdot \nu(d) \tag{3}$$

Let us suppose that we have a tag in position $\mathbf{v} = (x, y, z)$. We call N_A the number of anchors: as it has been defined in literature, for having an accurate tracking, $N_A \geq 4$. Let us call $\tilde{l}_i$ the estimated distance of point $\mathbf{v}$ from anchor i, placed at coordinates

$\mathbf{v}_i = (x_i, y_i, z_i)$. We suppose that tracking is performed by a non-linear minimization process. In particular, the tracked position $\hat{\mathbf{v}} = (x, y, z)$ is selected by minimizing the following objective function $f_T(\mathbf{v})$:

$$f_T(\mathbf{v}) = \sum_{i=1}^{N_A} \left(|\mathbf{v} - \mathbf{v}_i| - \tilde{l}_i \right)^2 \quad (4)$$

The minimization step is done using *Successive Quadratic Programming* (SQP), a well known and easy to use optimization technique. In particular, for what concerns function $f_T(\mathbf{v})$, it can be particularly effective since both the gradient vector $(\frac{\partial f_T(\mathbf{v})}{\partial x}, \ldots)$ and the Hessian matrix $(\frac{\partial^2 f_T(\mathbf{v})}{\partial x\, \partial y}, \ldots)$ can be explicitly computed.

The main contribution of this work is studying a procedure for optimizing the positioning of anchors in a room, in order to obtain the best tracking results. The first issue is to find a suitable objective function that, starting from the position of the anchors, allows a non-linear optimization algorithm to find the optimal anchor placement.

This is done by simulating a tracking process, and using the obtained average tracking error as the value of the corresponding objective function. First we select a set of N_S test points $\mathbf{t}_j$ $(1 \leq j \leq N_S)$: in particular we select these points from one of the traces defined in Section III-C . In order to study the average tracking error, we repeat the simulation on N_E experiments, each one involving tracking of the same N_S points, affected however by a different measuring error. This is achieved by considering an uniform set of standard normally distributed random errors $\epsilon_{ijk} \sim N(0, 1)$ $(1 \leq N_A,\ 1 \leq j \leq N_S,\ 1 \leq k \leq N_E)$ for each combination of anchor i, point j, and experiment k. Fixing the normally distributed error allows the minimization process that studies the anchor placement consistency and avoids the risk of highly variable solutions. We thus define the best position of the anchors by minimizing the following objective function $f_A(\mathbf{v}_1, \ldots, \mathbf{v}_{N_A})$:

$$f_A(\mathbf{v}_1, \ldots, \mathbf{v}_{N_A}) = \frac{1}{N_S} \frac{1}{N_E} \sum_{j=1}^{N_S} \sum_{j=1}^{N_E} \left(\mathbf{t}_j - \hat{\mathbf{t}}_{jk} \right)^2 \quad (5)$$

where $\hat{\mathbf{t}}_{jk}$ represents the result of the tracking process for test point $\mathbf{t}_j$ in simulation k, when the distance from each anchor $\mathbf{v}_i$ is affected by error ϵ_{ijk}. In particular:

$$\hat{\mathbf{t}}_{jk} = \arg\min_{\mathbf{t}} \sum_{i=1}^{N_A} \left(|\mathbf{t} - \mathbf{v}_i| - \tilde{l}_{ijk} \right)^2 \quad (6)$$

where:

$$\tilde{l}_{ijk} = |\mathbf{t}_j - \mathbf{v}_i| + \epsilon_{ijk} \cdot \nu(|\mathbf{t}_j - \mathbf{v}_i|) \quad (7)$$

Anchors are constrained to be fixed on the four walls of a rectangular room. To reduce the number of symmetries, anchors are sorted on the perimeter in a counterclockwise order: the optimization process is constrained to maintain this order, avoiding each anchor to "overtake" the following one. Again, optimization is performed using the SQP algorithm; to avoid local minima, a random-restart procedure is used, by repeating the process several time from a different initial anchor placement.

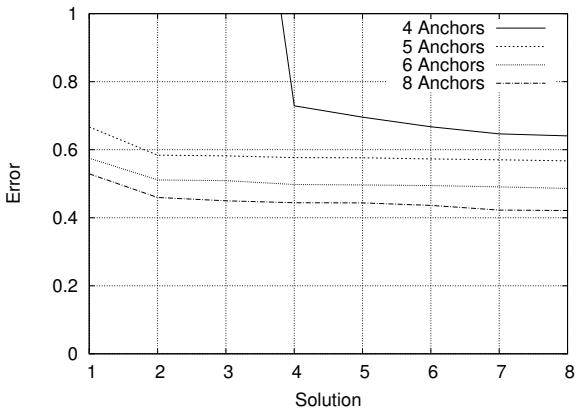


Fig. 6. Optimization of anchor placement

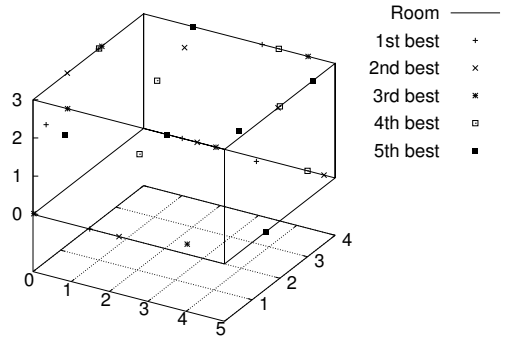


Fig. 7. Placement of the anchors in the room for the best solutions in the 6 anchor cases

Figure 6 shows the value of objective functions found at the end of each iteration of the optimization process, for a different number of anchors, and Figure 7 shows where they are placed into a 4m.×5m.×3m. large room. It is interesting to note that when the number of anchors is high, different placements can lead to equally good tracking results. However, with the minimum number of anchors ($N_A = 4$), an optimal placements of anchors is mandatory to obtain good results.

The effectiveness of the tracking process is shown in Figure 8 for different room sizes. Figures 8.a-d considers the results of the optimization procedure into 4 rooms of different sizes. As it can be seen from the spread of the tracked points from the one of the original trace, the error increases with the size of the room. In particular, Figure 8.e shows a zoomed window of the process for the smallest (S) and the largest rooms, from which it can be clearly appreciated the larger spread of the tracked points in the larger room case.

The optimization step is generally just the beginning of the tracking process: estimated position data is further

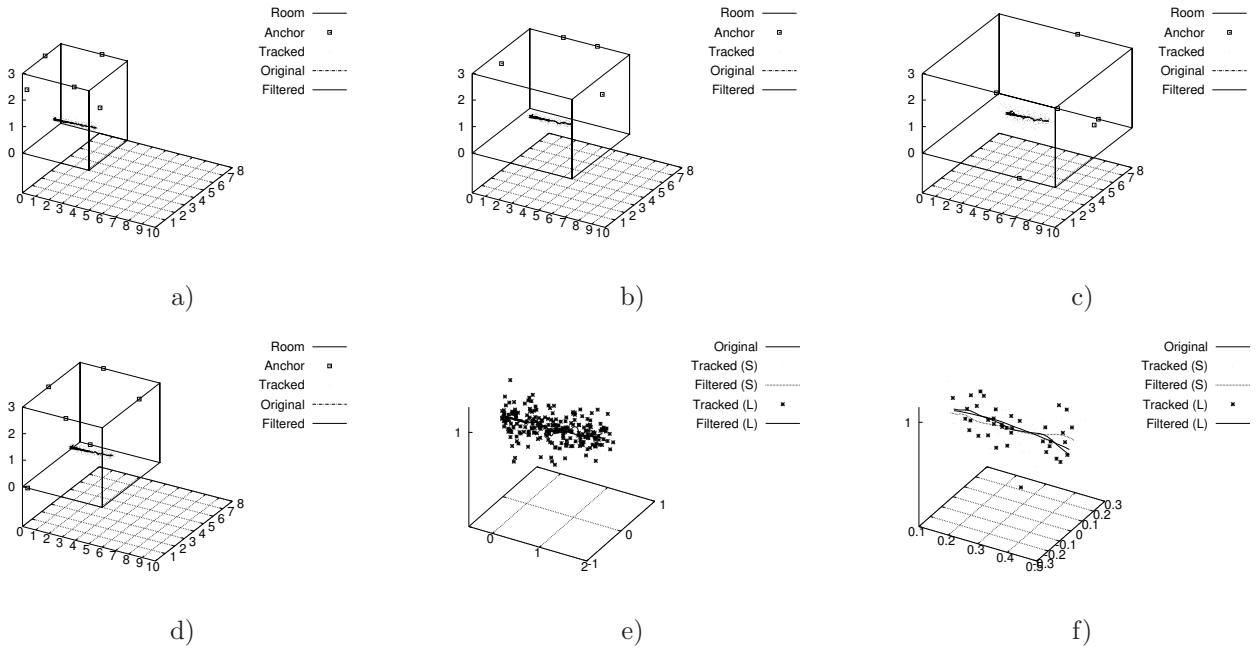


Fig. 8. Tracking process in different rooms: a) 4×5, b) 6×7.5, c) 8×10, d) 6×6, e) Zoom, f) Larger zoom

filtered to obtain more consistent and less noisy data. In many cases tags are also equipped with an IMU (Inertial Measurement Units) that includes an accelerometer, a magnetometer and a gyroscope. Elite tracking algorithms apply complex sensor fusion algorithms and Kalman filters to improve the position estimation. To simulate this process, in this work we applied a simple two pole low-pass Butterworth digital filter, centred at 0.05 of the sampling window. As it can be seen in Figure 8.f, that shows a more enlarged zoom of the tracking process, even this basic digital filter can produce results that are quite close to the original curve, independently from the room size.

To further study the effectiveness of the anchor placement process, Figure 9 reports about the application of the tracking algorithm to different data sets with the anchors placed according to the solution of the optimization process. The small room is considered in Figure 9.a for a longer linear walk, and in Figure 9.b for a dance routine. Results remains in the same error range as the one obtained during the dataset used for training. Figure 9.c considers the dance routing in the larger room: in this case it is visible that larger errors occurs when the dancer is in the outer part of the room, while results are more accurate when he is in the middle.

Figure 10 aims at studying the required number of anchors to achieve a desired level of accuracy. Without filtering, the anchor placement with the optimization produces equally good results for the three considered datasets. As expected, a larger number of anchor produces better tracking results with a lower average estimation error. However, when a good filtering algorithm is used, the number of anchors seems no longer to play an important role, and the difference between using 4 or 8 anchors become minimal. However, these results also show that the effectiveness of the filtering algorithms greatly depends on the considered trace - that is, on the type of movements performed by the tracked anchor.

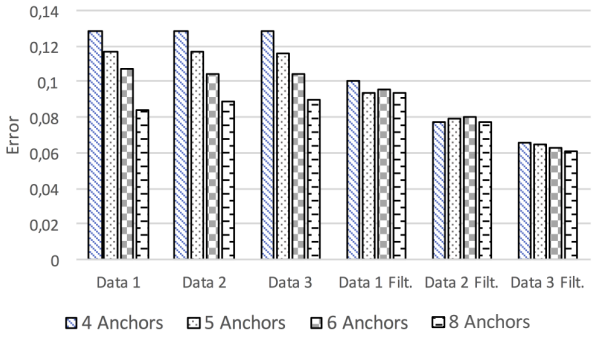
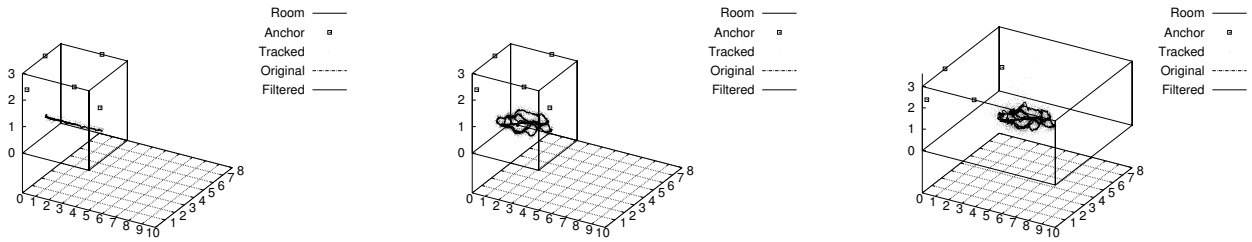


Fig. 10. Tracking error for different datasets as function of the number of anchors

V. CONCLUSIONS AND FUTURE WORK

This work presented preliminary simulations that support the exploration of the effects of different anchor setups in a training application. Simulations show some interesting results, that provide a general framework in which the experimental phase of the development of the application will be set up. Results also povide a guideline for the experiments that will take place to validate simulations. Future work includes the implementation of the studied scenarios, both to validate the approach and include it into a standard application setup procedure, and to tune the system for human-in-the loop tests. More, less trivial scenarios are planned, heading to consolidate the methodology to be used in real working environments including obstacles.



a) b) c)

Fig. 9. Tracking process for different datasets: a) longer walk in a 4×5 room, b) dance in a 4×5 room, c) dance in a 8×10 room

References

[1] E. Barbierato, M. Gribaudo, M. Iacono, and A. H. Levis. *Modeling Crowd Behavior in a Theater*. Springer International Publishing, 2018. In press.

[2] J. Chliz, A. Hernndez-Solana, and A. Valdovinos. Evaluation of algorithms for uwb indoor tracking. In *2011 8th Workshop on Positioning, Navigation and Communication*, pages 143–148, April 2011.

[3] D. Dardari, P. Closas, and P. M. Djuri. Indoor tracking: Theory, methods, and technologies. *IEEE Transactions on Vehicular Technology*, 64(4):1263–1278, April 2015.

[4] C. Frost, C. S. Jensen, K. S. Luckow, B. Thomsen, and R. Hansen. Bluetooth indoor positioning system using fingerprinting. In J. Del Ser, E. A. Jorswieck, J. Miguez, M. Matinmikko, D. P. Palomar, S. Salcedo-Sanz, and S. Gil-Lopez, editors, *Mobile Lightweight Wireless Systems*, pages 136–150, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[5] G. Kul, T. Ozyer, and B. Tavli. IEEE 802.11 WLAN based Real Time Indoor Positioning: Literature Survey and Experimental Investigations. *Procedia Computer Science*, 34:157 – 164, 2014. The 9th International Conference on Future Networks and Communications (FNC'14)/The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC'14)/Affiliated Workshops.

[6] C. G. Lab. CMU motion capture database. `http://mocap.cs.cmu.edu/`. Accessed: 2018-02-12.

[7] H. Liu, H. Darabi, P. Banerjee, and J. Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080, Nov 2007.

[8] M. Meredith and S. Maddock. Motion capture file formats explained. `http://www.dcs.shef.ac.uk/intranet/research/public/resmes/CS0111.pdf`. Accessed: 2018-02-12.

[9] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil. Landmarc: Indoor location sensing using active rfid. *Wireless Networks*, 10(6):701–710, Nov 2004.

[10] P. Piazzolla, M. Gribaudo, S. Colombo, D. Manca, and M. Iacono. A low-cost distributed iot-based augmented reality interactive simulator for team training. In Z. Z. Paprika, P. Horák, K. Váradi, P. T. Zwierczyk, Á. Vidovics-Dancs, and J. P. Rádics, editors, *European Conference on Modelling and Simulation, ECMS 2017, Budapest, Hungary, May 23-26, 2017, Proceedings.*, pages 591–597. European Council for Modeling and Simulation, 2017.

[11] G. Shi and Y. Ming. Survey of indoor positioning systems based on ultra-wideband (uwb) technology. In Q.-A. Zeng, editor, *Wireless Communications, Networking and Applications*, pages 1269–1278, New Delhi, 2016. Springer India.

AUTHOR BIOGRAPHIES

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 modeling, queueing networks, mean-field analysis and spatial models. The main applications to which the previous methodologies are applied comes from cloud computing, multi-core architectures and wireless sensor networks. His email is marco.gribaudo@polimi.it

MAURO IACONO is a tenured assistant Professor and Senior Researcher in Computing Systems at Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy. 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 is available at http://www.mauroiacono.com.

PIETRO PIAZZOLLA received his M.Sc. in Multimedia and Performing Arts at Universit degli Studi di Torino, Torino, Italy and Ph.D. degree in Business & Management, Economics Faculty at Universit degli Studi di Torino, Torino, Italy, in 2007 and 2012, respectively. From 2011 he is a Research Associate at Politecnico di Milano, Milano, Italy. His email is pietro.piazzolla@polimi.it

New fuzzy numbers comparison operators in energy effectiveness simulation and modeling systems

Wojciech T. Dobrosielski
Casimir the Great University in Bydgoszcz
Department of Computer Science
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
E-mail: wdobrosielski@ukw.edu.pl

Jacek M. Czerniak
Casimir the Great University in Bydgoszcz
Department of Computer Science
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
E-mail: jczerniak@ukw.edu.pl

Hubert Zarzycki
University of Information Technology and Management Copernicus
ul. Inowroclawska 56, 53-648 Wroclaw, Poland
E-mail: hzarzycki@wsiz.wroc.pl

Janusz Szczepański
Institute of Fundamental Technological Research
Polish Academy of Sciences
ul.Pawinskiego 5B, 02-106 Warsaw, Poland
E-mail: jszczepa@ippt.pan.pl

KEYWORDS
Fuzzy logic, comparison, OFN, GR, ML, TR

ABSTRACT

Energy efficiency is often a key optimization problem. Many control systems use fuzzy logic and as a result applying compare operators to fuzzy numbers. The article deals with the issue of comparing fuzzy numbers. The similarity relation is most probably the most frequently used and the most difficult to precisely determine the convergence measure. Analysis of the similarity of two objects is a basic assessment tool and constitutes the basis for reasoning by analogy. It also directly affects the energy effectiveness of the universe that it controls. This article presents the methods for determining the similarity used in fuzzy logic. Many of these methods were dedicated only to fuzzy triangular or trapezoidal numbers (Dobrosielski et al. 2017, Chi-Tsuen Yeh 2017, Abbasbandy and Hajjar 2009). This was a computational inconvenience and posed a question about the axiological basis of this type of comparison. The authors proposed two new approaches for comparing fuzzy numbers using one of the known extensions of fuzzy numbers (Kacprzyk and Wilbik 2009, 2005). This allowed to simplify the operation and eliminate the duality (Zadrożny, 2004).

INTRUDUCTION

In all fields of science for a long time it was necessary to compare certain objects. While some branches of science sought to answer the question about the nature of the similarities, others need precise, formal definition. Comparison of two objects or occurrences can be seen as an attempt to determine the relation between them (Piegat, 2015, Piegat et al., 2015). The most important and most frequently used relations between objects are similarity, difference and inclusion. In the literature, most attention is dedicated to the issue of the similarity of objects (Stachowiak and Dyczkowski 2013, Wenyi et al 2016). In recent decades, the theory of fuzzy sets has been used in many areas of science and everyday life (Czerniak et al. 2017a, Zarzycki et al. 2017a, Dobrosielski et al. 2017, Ewald 2018a, Apiecionek et al. 2018, Marszałek 2014). The need to compare fuzzy sets emerged naturally from the very beginning of the theory. There are plenty of methods, often based on those used for conventional sets. Intensive development of fuzzy logic and its applications often need to identify new ways of comparing objects (Lebiediewa et al. 2016, Zarzycki et al. 2017b) or issues as Linguistic Summaries (Kacprzyk et al 2005, 2006, 2009, Zadrożny 2004). This issue is particularly important in the computer aided decision support, classification and processing of natural language. Although the issue of the comparison is crucial for many applications of fuzzy set theory, still we failed to clearly formalize the basic concepts such as similarity or inclusion (Adabitabar et al. 2012, Khorshidi and Nikfalayar 2017, Dabashree 2010). While some researchers concerned with fuzzy logic seeks to define precisely the concepts, others questioned this approach, saying that imposing rigid framework limits the practical applications (Czerniak et al., 2017).
Through years of development of fuzzy logic, many researchers has been developing methods of comparing sets and fuzzy numbers. Among them, it is worth to recall Several fuzzy number comparison methods and indices have been researched since 1977 Zadeh, Yager and Kaufman, Chang, and Amado (Abbasbandy, 2009). Bortolan and De-gani and Dadgostar reviewed some of the methods for ranking fuzzy sets (Tran et al., 2002a, 2002b), including Yager's first, second and third indexes, Chang's algorithm, Adamo's method, Baas and Kwakernaak's method, Baldwin and Guild's method, Kerre's method (Wenyi et al, 2016), Jain's method and Dubois and Prade's four grades of dominance (PD,PSD, ND, NSD). Dadgostar and Kerr (Dabashree 2010) proposed a consistent method, called Partial Comparison Method (PCM) (Adabitabar et al. 2012). Wang and Kerre proposed several axioms as reasonable properties to determine the rationality of a fuzzy

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

ordering or ranking method and systematically compared a wide array of fuzzy ranking methods (Guixiang 2017).

THREE NEW DEFUZZYFICATORS

In this paragraph, the mathematical foundations of the three methods of defuzzyficating will be presented. They are all sensitive to directing. This means that the methods applied to numbers in the OFN notation generally provide different defuzzyficating results for numbers with opposite directing on the same co-ordinates (Ewald et al., 2018).

Golden Ratio (GR) defuzzyfication operator

Fibonacci series is based on the assumption that it starts with two ones, and each consecutive number is the sum of the previous two. The proposal for the Golden Ratio method of defuzzification is based on the proportion of the golden ratio (Dobrosielski et al., 2017). As a result of dividing each of the numbers by its predecessor, we always obtain quotients oscillating around the value of 1.618 the golden ratio number. The exact value of the limit is the golden number itself:

$$GR = \frac{min(supp(A)) + |supp(A)|}{\Phi}$$
$$where\ \Phi = 1,618033998875\ldots \quad (1)$$

where: GR is the defuzzification operator, supp(A) is support for fuzzy set A in
universe X.

Mandala Factor (ML) defuzzyfication operator

Buddhist monks can create amazing pictures of colored sand grains. Those pictures are called mandala. The Mandala Factor defuzzyfication operator is inspired by mandala. Calculation of the R value using the Mandala Factor for the rising edge, falling edge and core set function integral. Then the obtained value should be scaled from the center of the coordinate system by adding it to the start of the support value of the fuzzy number (Czerniak et al., 2018). When defuzzyfication is performed in the OFN arithmetic, then in the case of a positive order, one should proceed as described below, while in the case of a negative order, one should deduct the calculated value from the first coordinate of the OFN number corresponding to the outermost right side of the OFN support.

$$MF(A) = \begin{cases} c + r\ , if\ order\ (A)\ is\ positive \\ c - r\ , if\ order\ (A)\ is\ negative \end{cases} \quad (2)$$

where

$$r = \frac{1}{d-c}\int_c^d x\,dx - \frac{c}{d-c}\int_c^d dx + \frac{f}{f-e}\int_f^e dx - \frac{1}{f-e}\int_e^f x\,dx + \int_d^e dx \quad (3)$$

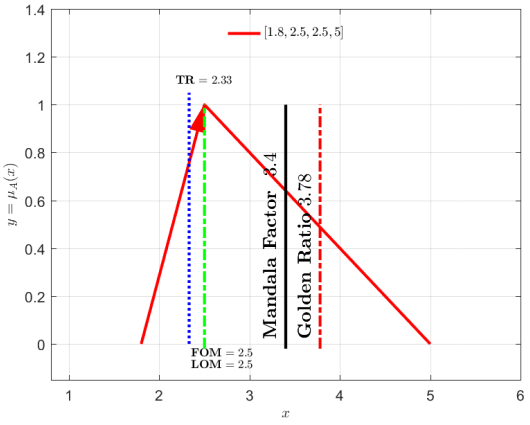
Triangular Expanding (TR) defuzzyfication operator

The above considerations include formal description of the proposed method in the OFN number defuzzyfication process that comes down to determining the equations for the intersection of two circles and the intersection of two linear functions (Dobrosielski et al., 2017).

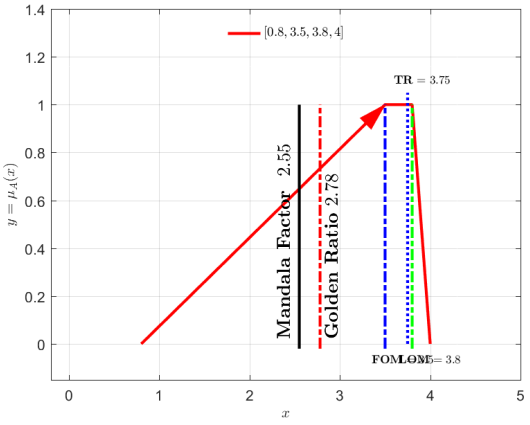
$$x_w = \frac{f(0)(f(1)s_1 - g(0)s_1 + g(0)) - g(0)x_1}{f(0) + f(1)s_1 - g(0)s_1 - x_1} \quad (4)$$

where: x1, s1 are the determined coordinates of the point crisp value

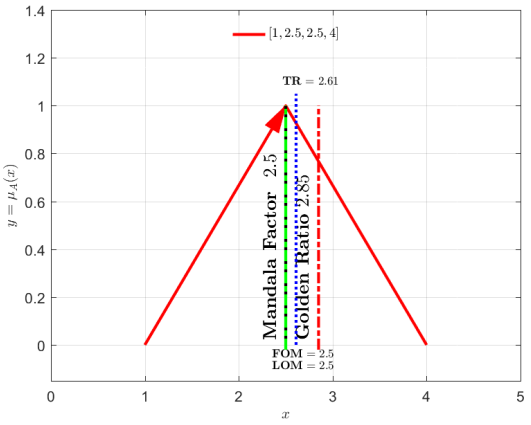
The following set of drawings presents the visualization of the seven fuzzy numbers. We are going to compare these numbers, which is not a trivial task in the case of fuzzy logic.



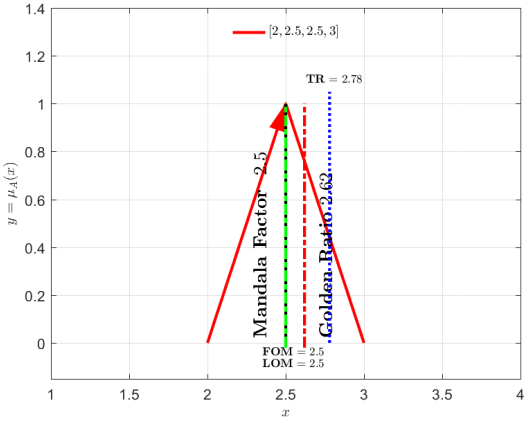
Figures 1: Crisp value of bA1 by GR,ML and TR



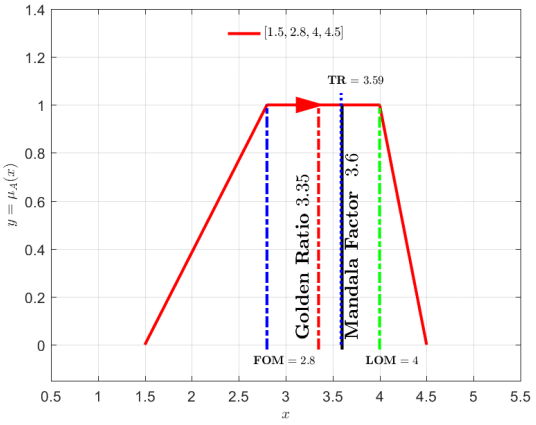
Figures 2: Crisp value of bA2 by GR,ML and TR



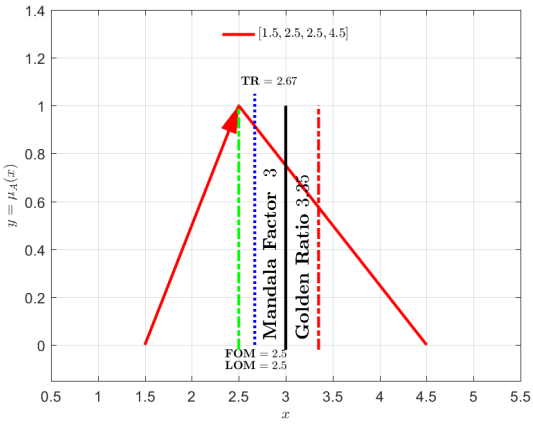
Figures 3: Crisp value of dA1 by GR,ML and TR



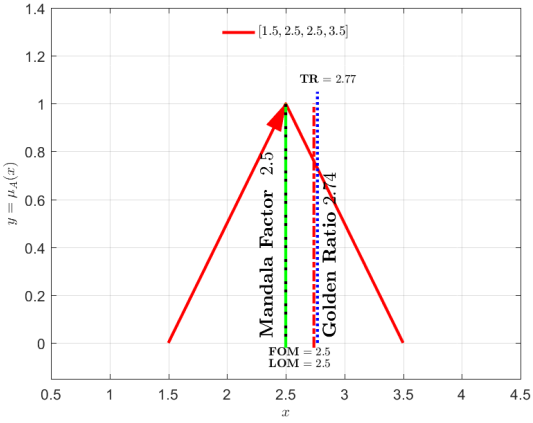
Figures 4: Crisp value of dA2 by GR,ML and TR



Figures 5: Crisp value of eA1 by GR,ML and TR



Figures 6: Crisp value of eA2 by GR,ML and TR

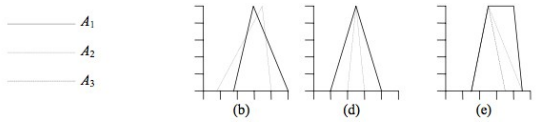


Figures 7: Crisp value of eA3 by GR,ML and TR

EXAMPLES OF FUZZY NUMBER COMPARISON

This paragraph presents the results obtained using the ML, GR and TR methods for the following set of compared numbers.

Table 1: Fuzzy number comparison methods



Methods		A_1	A_2	A_1	A_2	A_1	A_2	A_3
Yager	F1	**0.61**	**0.53**	**0.50**	**0.50**	**0.62**	**0.56**	**0.50**
	F2	**0.66**	**0.69**	**0.61**	**0.54**	**0.81**	**0.64**	**0.58**
	F3	**0.58**	**0.56**	**0.50**	**0.50**	**0.62**	**0.54**	**0.50**
Chang		**0.40**	**0.34**	0.29	0.10	**0.56**	**0.33**	**0.20**
Adamo	0.9M	0.55	0.66	**0.53**	**0.51**	**0.81**	**0.54**	**0.52**
	0.9m	0.55	0.66	**0.53**	**0.51**	**0.81**	**0.54**	**0.52**
	0.5	0.75	0.72	**0.65**	**0.55**	**0.85**	**0.70**	**0.60**
Baas-Kwakernaak		**0.84**	**1**	**1**	**1**	1	1	1
Baldwin-Guild	lap.	0.42	0.33	**0.27**	**0.28**	**0.45**	**0.37**	**0.27**
	g.	0.44	0.37	**0.30**	**0.24**	**0.53**	**0.40**	**0.28**
	r.a.	0.34	0.24	**0.20**	**0.23**	**0.31**	**0.28**	**0.21**
Kerre		**0.96**	**0.89**	**0.91**	**0.91**	**1**	**0.85**	**0.75**
Own results	GR	3.78	2.78	2.85	2.62	3.35	3.35	2.74
	ML	3.4	2.55	2.5	2.5	3.6	3	2.5
	TR	2.33	3.75	2.61	2.78	3.59	2.67	2.77
Dubois-Prade	PD	0.84	1	**1**	**1**	1	1	1
	PSD	0.54	0.46	0.73	0.24	0.80	0.20	0
	ND	0.54	0.46	0.27	0.76	0.50	0.50	0.50
	NSD	0	0.16	0	0	0	0	0
Lee-Li	U.m.	0.61	0.53	0.50	0.50	**0.62**	**0.56**	**0.50**
	U.g.	-	-	0.12	0.04	-	-	-
	P.m.	0.53	0.58	0.50	0.50	**0.63**	**0.55**	**0.50**
	P.g.	-	-	0.09	0.03	-	-	-
Dadgostar-Kerr	PCM	**0.57**	**0.43**	**0.50**	**0.50**	**0.63**	**0.37**	**0.25**
Dorohonceanu-Marin	B2	**0.59**	**0.41**	**0.50**	**0.50**	**0.65**	**0.60**	**0.40**
	B2 x	**0.59**	**0.41**	**0.50**	**0.50**	**0.68**	**0.61**	**0.39**

As can be seen from the table above, the use of defuzzyfication operators GR, ML and TR allowed to reduce the uncertainty of fuzzy numbers and allowed to compare them. Thus, in the first group (b) where the A1 and A2 numbers are compared, individual defuzzyficators allow to detect the following relations:
GR: A1>A2, ML: A1>A2, TR: A1<A2.
In group (d) defuzzyfication allowed detection of the following relations:
GR: A1>A2, ML: A1=A2, TR: A1<A2.
The third group (e) of fuzzy numbers, the elements of which have been defuzzyfied, will present the following set of relations:
GR: A1=A2>A3, ML: A1>A2>A3, TR: A1>A3>A2

CONCLUSION

The article presents three defuzzyficating methods that can be applied both to classic fuzzy numbers and to numbers in OFN notation. This time, the discussed defuzzyficators were used, ie GR, ML and TR as operators preparing data for comparing fuzzy numbers.
The novelty that the article brings is, in addition to the development of three proprietary defuzzyfication methods, applying them to the fuzzy numbers mentioned in the introduction, commonly used as peculiar benachmas of fuzzy logic. As a result, a new application of a novel defuzzyfication operators is shown.
The GR and ML defuzzyfication operators showed in group (b) the same results as most compared operators.
In group (d), only the ML operator indicated the same type of relationship as the majority of known operators, and GR and TR were among the minority compared operators. However, in the third group (e) the ML operator indicated the relation that the majority indicated and the GR and TR operators signaled another type of relation not mentioned in the group.

REFERENCES

Adabitabar Firozja M.; G.H. Fath-Tabar; Z. Eslampia. 2012. "The similarity measure of generalized fuzzy numbers based on interval distance", Applied Mathematics Letters, Volume 25, Issue 10, pp.1528-1534, Elsevier.

Apiecionek Ł.; H. Zarzycki; J.M. Czerniak; W.T. Dobrosielski; D. Ewald. 2018. "The Cellular Automata Theory with Fuzzy Numbers in Simulation of Real Fires in Buildings". In: Atanassov K. et al. (eds) Uncertainty and Imprecision in Decision Making and Decision Support: Cross-Fertilization, New Models and Applications. IWIFSGN 2016. Advances in Intelligent Systems and Computing, vol 559. Springer, Cham.

Chi-Tsuen Yeh. 2017. "Existence of interval, triangular, and trapezoidal approximations of fuzzy numbers under a general condition". Fuzzy Sets and Systems, Volume 310, pp. 1-13, Elsevier

Czerniak J.M.; Zarzycki H. 2017. "Artificial Acari Optimization as a new strategy for global optimization of multimodal functions". Journal of Computational Science, Volume 22f, pp. 209-227, Elsevier.

Czerniak J.M.; Zarzycki H.; Apiecionek Ł.; Palczewski W.; Kardasz P. 2018. "A Cellular Automata-Based Simulation Tool for Real Fire Accident Prevention" Mathematical Problems in Engineering 2018, Article ID 3058241, 12 pages, Hindawi.

Dobrosielski W.; J. Szczepański.; H. Zarzycki. 2016. A Proposal for a Method of Defuzzification based on the Golden Ratio - GR, In: Atanassov K. et al. (eds) Novel Developments in Uncertainty Representation and Processing. Advances in Intelligent Systems and Computing, Volume 401, pp 75-84, Springer, Cham.

Dobrosielski W.; J. M. Czerniak; J. Szczepański; and H. Zarzycki. 2017. "Triangular expanding, a new defuzzification method on ordered fuzzy numbers," Advances in Intelligent Systems and Computing, Volume 642, pp. 605–619, Springer.

Debashree Guha; Debjani Chakraborty. 2010. "A new approach to fuzzy distance measure and similarity measure between two generalized fuzzy numbers", Applied Soft Computing, Volume 10, Issue 1, pp. 90-99, Elsevier.

Dyczkowski K. 2007. "A Less Cumulative Algorithm of Mining Linguistic Browsing Patterns in the World Wide Web". EUSFLAT Conf., Volume 2, pp.129-135.

Ewald D.; J.M. Czerniak; H. Zarzycki. 2018. "OFNBee Method Used for Solving a Set of Benchmarks". In: Kacprzyk J., Szmidt E., Zadrożny S., Atanassov K., Krawczak M. (eds) Advances in Fuzzy Logic and Technology 2017. IWIFSGN 2017, EUSFLAT 2017. Advances in Intelligent Systems and Computing, vol 642. Springer, Cham.

Guixiang Wang; Jing Li. 2017. "Approximations of fuzzy numbers by step type fuzzy numbers", Fuzzy Sets and Systems, Volume 310, pp. 47-59, Elsevier.

Abbasbandy S.; T. Hajjari. 2009. "A new approach for ranking of trapezoidal fuzzy numbers".

Jiang Wen; Xin Fan; Dejie Duanmu; Deng Yong. 2011. "A modified similarity measure of generalized fuzzy numbers", Procedia Engineering, Volume 15, pp.2773-2777, Elsevier.

Kacprzyk J.; Wilbik A. 2009 "Using Fuzzy Linguistic Summaries for the comparison of time series: an

application to the analysis of investment fund quotations". IFSA/EUSFLAT Conf., pp.1321-1326, 2009.
Kacprzyk J., Wilbik A., Zadrożny S., On some types of linguistic summaries of time seties, in: Proceedings of 3rd International IEEE Conference Intelligent Systems, London, UK, Sept. 4-6, 2006, IEEE Press, pp.373-378
Kacprzyk J.; Zadrożny S. 2005. "Fuzzy Linguistic Summaries in Text Categorization for Human-Consistent Document-Driven Decision Support Systems". In: Reusch B. (eds) Computational Intelligence, Theory and Applications. Advances in Soft Computing, volume 33, pp. 1373-1380, Springer.
Khorshidi Hadi Akbarzade; Sanaz Nikfalazar. 2017. "An improved similarity measure for generalized fuzzy numbers and its application to fuzzy risk analysis", Applied Soft Computing, Volume 52, pp. 478-486, Elsevier.
Lebiediewa S.; H. Zarzycki; W.T. Dobrosielski. 2016. "A New Approach to the Equivalence of Relational and Object-Oriented Databases". In: Atanassov K. et al. (eds) Novel Developments in Uncertainty Representation and Processing. Advances in Intelligent Systems and Computing, vol 401. Springer, Cham.
Macko M.; J. Flizikowski. 2010. "The method of the selection of comminution design for non-brittle materials". AIChE Annual Meeting, Conference Proceedings 2010.
Marszałek A.; T. Burczyński. 2014. "Modeling and forecasting financial time series with ordered fuzzy candlesticks". Information Science, Volume 273, pp.144-155, Elsevier.
Mikołajewska E.; D. Mikołajewski. 2011. "Exoskeletons in Neurological Diseases - Current and Potential Future Applications". Advances in Clinical and Experimental Medicine, Volume 20, pp.227-233, 2011.
Piegat A. 2005. "A new definition of the fuzzy set", Appl. Math. Comput. Sci. Volume 15, No. 1, pp. 125-140.
Piegat A.; M. Pluciński. 2015. "Computing with words with the use of inverse RDM models of membership functions". International Journal of Applied Mathematics and Computer Science, 25(3), pp. 675-688, De Gruyter
Rojek I. 2016. "Technological process planning by the use of neural networks", Artificial Intelligence for Engineering Design, Analysis and Manufacturing, Volume 31, No. 1, pp. 1-15, Cambridge University Press.
Stachowiak A.; Dyczkowski K. 2013. "A similarity measure with uncertainty for incompletely known fuzzy sets". 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), pp. 390-394.
Szmidt E.; Janusz Kacprzyk. 2000. "Distances between intuitionistic fuzzy sets, Fuzzy Sets and Systems", Volume 114, Issue 3, pp.505-518, Elsevier.
Śmigielski G.; W. Toczek; R. Dygdała. 2016. "Metrological Analysis of Precision of the System of Delivering a Water Capsule for Explosive Production of Water Aerosol". Metrology and Measurement Systems, Volume 23, No. 1, pp. 47-58, De Gruyter.
Śmigielski G., R. Dygdała, H. Zarzycki, and D. Lewandowski, "Real-time system of delivering water-capsule for firefighting," Advances in Intelligent Systems and Computing, vol. 534, pp. 102–111, Springer.
Tofigh Allahviranloo, M. Adabitabar Firozja, Ranking of fuzzy numbers by a new metric, Soft Computing, 2010, Volume 14, Number 7, Page 773-782
Tran L., L. Duckstein, Multiobjective fuzzy regression with central tendency and possibilistic properties, Fuzzy Sets and Systems 130 (2002) pp.21–31.
Tran L., L. Duckstein, Comparison of fuzzy numbers using a fuzzy distance measure, Fuzzy Sets and Systems 130 (2002) pp.331–341
Wenyi Zeng; Deqing Li; Qian Yin. 2016. "Distance and similarity measures between hesitant fuzzy sets and their application in pattern recognition", Pattern Recognition Letters, pp.267-271, Volume 84, Elsevier.
Zarzycki H.; Czerniak J.M.; and Dobrosielski W.T. 2017. "Detecting Nasdaq Composite Index Trends with OFNs". In: Prokopowicz P., Czerniak J., Mikołajewski D., Apiecionek Ł., Slezak D. (eds) Theory and Applications of Ordered Fuzzy Numbers. Studies in Fuzziness and Soft Computing, vol 356. Springer, Cham.
Zarzycki H.; J.M. Czerniak; D. Lakomski, P. Kardasz. 2017. Performance Comparison of CRM Systems Dedicated to Reporting Failures to IT Department. In: Madeyski L., Śmiałek M., Hnatkowska B., Huzar Z. (eds) Software Engineering: Challenges and Solutions. Advances in Intelligent Systems and Computing, vol 504. Springer, Cham.
Zadrożny S., Kacprzyk J., On the use of linguistic summaries for text categorization, in: Proceedings of IPMU'2004 – International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, 2004, vol. 2, pp.1373-1380

AUTHOR BIOGRAPHIES

JACEK M. CZERNIAK received Ph.D. degrees in Computer Science in 2005 from Technical University of Szczecin in Poland. In 2000, Dr. Czerniak received a M.Eng. degree in Computer Science from the Technical University of Szczecin. Dr. Czerniak is currently employed as an Assistant Professor in the Computer Science Department at Casimir the Great University in Bydgoszcz. Dr. Czerniak is a founding director the AIRlab Artificial Intelligence and Robotics Laboratory. He is also founding Editor-in-Chief journal of Studies and Materials in Applied Computer Science. Dr. Czerniak has been awarded twice by the President of Casimir the Great University in Bydgoszcz. First in october 2009, awarded Outstanding Teacher Award, and the second time in november 2015 and third in 2017 awarded Outstanding Researcher Award.
His e-mail address is : JCzerniak@ukw.edu.pl and his Web-page can be found at
http://www. JCzerniak.ukw.edu.pl

HUBERT ZARZYCKI servers as an assistan proffesor at University of Information Technology and Management Copernicus in Wroclaw. He received his

MSc and PhD (2006) degree in the discipline of Computer Science from Technical University of Szczecin. He also holds a master's degree in the field of Finance and Banking. His research interests developed while working in different companies and research centers in Poland, the United States, Germany and Italy. He collaborates with renowned institutions, domestic and foreign. He has many years of experience in the information technology industry. His research interests relate to artificial intelligence, fuzzy numbers, computer system architecture, databases and programming. His e-mail address is : hzarzycki@wsiz.wroc.pl.

.

WOJCIECH T. DOBROSIELSKI graduated from the Faculty of Computer Science at Technical University of Szczecin in specialty programming techniques and information systems. He worked as a research assistant at the Casimir the Great University in Bydgoszcz and Systems Research Institute of Polish Academy of Sciences. His research interests are related to Artificial Intelligence, with the main focus is associated with fuzzy logic. More specifically his scientific work often discusses applications of ordered fuzzy numbers. Related research interest are: control systems, semantic networks, swarm intelligence - ACO, robotics, parallel programming, cryptography, software engineering and electronics. His e-mail address is : wdobrosielski@ukw.edu.pl and his Web-page can be found at http://www.it.ukw.edu.pl

JANUSZ SZCZEPAŃSKI received the M.Sc. degree in mathematics (1979) at the Warsaw University, the Ph.D. degree in applied mathematics (1985) and habilitation degree in computer sciences (2007) in Polish Academy of Sciences. Currently, he is a professor

at the Institute of Fundamental Technological Research, PAS and deputy-chairmen of the Scientific Council. In 2001–2004 he was a Consultant on Cryptography with the Polish Certification Authority ‘‘Centrast’’ Co. For Public Key Infrastructure. In 2000 and 2003 he was a Visiting Scientist at the Miguel Hernandez University, Elche, Spain. In 2004 he was a Visiting Scholar at the University of California, San Diego. His research interests include information theory, neuroscience, application of dynamical systems and stochastic processes to biological systems and cryptography. Prof. Szczepanski received the PAS Award in 1989.

His e-mail address is : jszczepa@ippt.pan.pl and his Web-page can be found at http:// http://www.ippt.pan.pl/staff/jszczepa

Stackelberg Game-based Models in Energy-aware Cloud Scheduling

Damián Fernández-Cerero,
Alejandro Fernández-Montes,
Department of Computer Languages and Systems
University of Seville
Av. Reina Mercedes s/n, 41012 Seville, Spain

Agnieszka Jakóbik,
Cracow University of Technology
Warszawska st 24, 31-155 Cracow, Poland
Joanna Kołodziej
Research and Academic Computer Network
Kolska st 12, 01-045 Warsaw, Poland

KEYWORDS

Computational clouds; Cloud computing; Tasks scheduling; Energy saving; Cloud services modelling; Stackelberg Game

ABSTRACT

Energy-awareness remians the important problem in today's cloud computing (CC). Optimization of the energy consumed in cloud data centers and computing servers is usually related to the scheduling problems. It is very difficult to define an optimal scheduling policy without negoative influence into the system performance and task completion time. In this work, we define a general cloud scheduling model based on a Stackelberg game with the workload scheduler and energy-efficiency agent as the main players. In this game, the aim of the scheduler is the minimization of the makespan of the workload, which is achieved by the employ of a genetic scheduling algorithm that maps the workload tasks into the computational nodes. The energy-efficiency agent selects the energy-optimization techniques based on the idea of switchin-off of the idle machines, in response to the scheduler decisions. The efficiency of the proposed model has been tested using a SCORE cloud simmulator. Obtained results show that the proposed model performs better than static energy-optimization strategies, achieving a fair balance between low energy consumption and short queue times and makespan.

I. Introduction

New paradigms, such as cloud computing, and the ever-growing web applications and services, have imposed new challenges to traditional high-performance computing (HPC) systems. In the same time, HPC infrastructures that provide the core foundation for the parallel computing solutions have grown drastically in recent years to satisfy the ever-evolving user requirements. Modern large-scale HPC systems are composed of thousands of computational distributed servers. The energy consumed by such HPC systems may be compared to the energy utilized by the small towns and large factories. Data centers account for more than 1.5% of global energy consumption [16].

Several hardware and infrastructure models have been recently developed for the successful reduction of the energy consumption in real-life large-scale data centers. The most popular models and technologies include: (i) cooling and temperature management [22] [4], (ii) memory and CPU power proportionality [19] [5], (ii) construction of the efficient new-generation green hard disks [1]; and new techniques in energy transportation [6]. Also the resource management and scheduling models in clouds are defined with the energy optimization modules. Energy utilization policies may be based on various power related physical models, however the most popular scenario is to switch off idle servers. Although such power-off strategy is commonly used in small-area grids and clusters [20], in realistic CC systems, the existing power-off models need to be improved especially in the case of dynamical changes in the task workloads and cloud resource infrastructure [8].

In this work, the balance between two opposed needs of the data-center environment is modelled by means of a Stackelberg Game (SG). On one hand, the performance side, represented by the *Scheduling Manager* (end users experience), wants tasks to be processed as fast as possible, while the efficiency side (CC provider), represented by the *Energy-Efficiency Manager*, wants the minimization of the energy consumption of the data center.

In our SG model, the *Scheduling Manager*, that is the leader of the game, processes firstly every task to make its decision (move). Once the particular *Task* is processed by the leader, then the follower, that is the *Energy-efficiency Manager*, handles it to make its move. This competition process is implemented in a trustworthy simulation tool focused on simulating realistic large-scale data-center scenarios.

The paper is organized as follows. In Section II we briefly describe the most usual and relevant strategies for achieving energy-efficiency in CC systems. In Section III, the shut-down decisions under consideration in this paper are presented. In Section IV, we formally define the proposed Stackelberg Game model for the balance between energy consumption and performance in CC systems, which is the theoretical core of this work. The used simulation tool and the experimentation performed, including the experimental environment and

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

workload, as well as the results obtained are presented in Section V. Finally, the conclusions and future work are discussed in Section VI

II. Related work

One of the most popular method of saving the energy in the distributed computational environments is deactivation of the idle servers, since data centers in cloud environments usually operate at 20-30% of their capacity [2], [21] and the workload pressure and requirements suffer from fluctuations, such as those derived of day/night or weekdays/weekend patterns. Such servers are switched into the 'sleeping' mode in the idle periods or simply switched off. This strategy was used as the main energy conservation method in data center clouds and was fundamental for the other methodologies and models used in cloud scheduling and data and tasks processing. We present below a very simple state-of-the-art analysis of the main trends and achievements in the domain of energy-awareness in the data center clouds:

Several energy-aware scheduling policies have been developed to raise server utilization [13], [17]. These policies are useful to minimize the number of machines that process workload rather than spreading the tasks among the maximum number of available servers. This enables the application of several energy-efficiency approaches, such as DVFS and hibernating nodes in an idle state. However, these strategies are static and can not easily address drastic workload pattern changes without having a negative impact on the performance of the CC system.

Consolidation and migration of virtual machines (VMs) in the cloud clusters are other well explored models of energy conservation in the cloud computational environments [23], [3]. In this work we adopt a different strategy in order to achieve efficiency: the modeling of the data-center environment as a Stackelberg Game. However, these models may be incorporated as a part of the Stackelberg Game in a future work.

Other authors [14], [18] propose the application of various energy-efficient techniques in cloud computing subsystems, such as the distributed file system, and other paradigms such as Grid Computing [9], in order to improve cluster power proportionality. However, these approaches focus usually on only one side of the whole CC system, which makes them sub-optimal when a complex cloud-computing operation process is under consideration.

The major contribution of this work is a model for the dynamic application of energy-efficiency policies based on the Stackelberg Game model. We model the opposed requirements of any energy-efficient data center, that are performance and energy efficiency, as the sides of this game. The application of the proposed model results on the balance between fast and reliable task execution and low energy consumption.

III. "Switch-off" decision policies

We assume in our model that the energy conservation policies do not have a notable negative impact on the performance of the whole cloud network. Therefore we define in our model a *Central Energy-efficiency Manager* that decides the power-off strategy to be applied, which deactivates the servers in an idle mode. It should be noted that *Always* strategy cannot be kept active when a machine compute tasks and send/receive data. In the case of huge workloads, where tasks and data may leave and arrive dynamically from and to the cloud servers, the active servers may be overloaded and the whole task execution process can be significantly delayed. Therefore, there is a need of the development of the decision model which allows us to activate the *Always* power-off strategy in the optimal periods. The following shut-down decision policies have been implemented in our model:

- **Margin** – this decision strategy activates the *Always* power-off strategy only if, at least, a specified amount of resources (servers) is ready to accept the incoming tasks.
- **Random**– in this case, the *Always* power-off strategy is activated randomly. This strategy is usually defined together with the *Never* shut-down policy, where all servers are kept in the active mode (it happens usually in realistic cloud data centers) and the *Always* shut-down scenario, where all idle machines are switched-off.
- **Gamma** – in this case, the *Always* shut-down strategy is activated depending on the probability of incoming tasks of oversubscribing the available resources. This probability is computed by the means of the Gamma distribution.

The utilization of the *Energy-efficiency Manager* in our model does not guarantee the fair reduction of the energy consumed by the cloud system. Therefore, we define another component of the model, that is the *Scheduling Manager*. This component allows the optimal schedule of tasks onto the cloud servers based on the energy-conservation criterion. In this work, we focus on the problem of the independent tasks scheduling. We use the genetic cloud scheduler developed in [11] and ETC Matrix scheduling model described in [10]. The makespan constitutes the most representative parameter of the performance, and hence it becomes the scheduling goal.

IV. Stackelberg Game model

In the model presented in this work we used the Stackelberg Game framework for the optimization of the balance between two main and opposed components of the model: *Scheduling Manager* – Leader and Energy-efficiency Manager –Follower.

Let us define first a 2-players non-zero symmetric game Γ_n as follows:

$$\Gamma_n = ((N, \{S_i\}_{i \in N}, \{Q_i\}_{i \in N}) \quad (1)$$

where:

- $N = \{1, \ldots, 2\}$ is the set of players,

- $\{S_1, \ldots, S_2\}$ $(cardS_i \geq 2; i = 1, \ldots, 2$) is the set of strategies for them
- $\{H_1, \ldots, H_n\}; H_i : S_1 \times \cdots \times S_2 \to \mathbb{R}; \forall_{\beth=\nVdash,\ldots,\ltimes}$ is the set of payoff functions for each player players.

Each player in this game may make its own decisions. A single decision is one from the set of possible actions. In this game, the strategy is defined by the set of actions that the player considers beneficial for him. Both pure strategies and mixed strategies are considered in our model, see [24]. A pure strategy specifies the most beneficial actions for a given situation, thus, pure strategies are deterministic. Mixed strategies extend pure strategies by the assignation of a probability to each pure strategy. The usage of a mixed strategy enables a player to randomly select a single pure strategy from the set of available strategies. Let us denote by s_i the **Pure strategy** of the player i and the set of all pure strategies specified for player i is denoted by S_i. The **mixed strategy of the player** i is denoted by $\sigma_i \in S_i \subset \Delta S_i$ and may be defined as follows:

$$\sigma_i = \{\sigma_i(s_{i_1}), \sigma_i(s_{i_2}), ..., \sigma_i(s_{i_m})\}, \tag{2}$$

where $\sigma_i(s_i)$ is the probability that the player i choses the pure strategy s_i.

Randomization in the game is provided by the probability distribution $\sigma_i(s_i)$.

The result of following a given strategy is the **expected payoff** of the player i in the 2-players game. Let this pay-off function be defined as:

$$H_i(s_i, \sigma_{-i}) := \sum_{s_{-i} \in S-i} \sigma_{-i}(s_{-i}) H_i(s_i, s_{-i}) \tag{3}$$

It is assumed in that game, that player i plays the pure strategy $s_i \in S_i$ and his opponents plays the mixed strategy $\sigma_{-i} \in \Delta S_{-i}$.

The **expected payoff** of the player i when playing the mixed strategy $\sigma_i \in \Delta S_i$ and when his enemy plays the mixed strategy $\sigma_{-i} \in \Delta S_{-i}$ is defined as:

$$H_i(\sigma_i, \sigma_{-i}) = \sum_{s_i \in S_i} \sigma_i(s_i) H_i(s_i, \sigma_{-i}) \tag{4}$$

$$= \sum_{s_i \in S_i} \Big(\sum_{s_{-i} \in S_{-i}} \sigma_i(s_i) \sigma_{-i}(s_{-i}) H_i(s_i, s_{-i}) \Big) \tag{5}$$

In Stakelberg Games (SG), one player (the leader) may play first, and the rest of the players (the followers) are obliged to follow the leader and make their decisions after him [24]. The main objective of the game for each player is to maximize his expected payoff by finding and playing the optimal strategy.

During the game proposed in this paper the leader and the follower evolve their strategies alternately. Thus, each player reacts to the decisions made by the opponent. We assumed only two players in the game, therefore $i = 1$ or 2 and $-i = 1$ or 2.

In our model, we define independently the utility functions of both players. It is modeled as a non-zero sum game and allow us to generate a separate game model for each player.

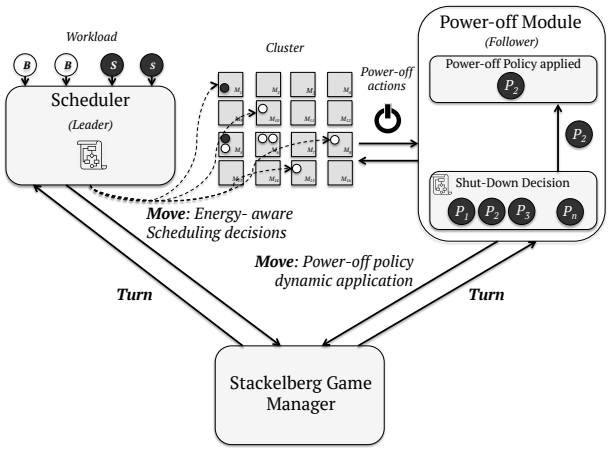


Fig. 1: Stackelberg Game workflow, scheduling workflow, B - Batch type task, S - Service type task, M - Virtual Machine

TABLE I: Players in the proposed game, their roles in the cloud and in the game model

Player 1	**Player 2**
Scheduling Unit	Energy-Efficiency Manager
Choses one of several schedules	Choses one of several energy policies
Leader	**Follower**

A. Leader Payoff and decisions

The leader in the proposed model is the *Scheduler* component, which perform the scheduling logic and dispatches tasks among the *Computing Nodes*. These *Computing Nodes* are grouped into *Computational Units*, denoted as $CU_1, CU_2, ...CU_P$. The *Scheduler* is responsible for processing incoming *Jobs*, which are composed of a set of *Tasks*, and for deploying these *Tasks* on the available *Computing Nodes* of a given *Computational Unit*.

Therefore there are P possible decisions. The strategy vector $\sigma_i(s_i)$ represents the probability for a *Job* to be assigned to the CU_p, for $p = 1, 2, ..., P$. The s_i may be taken from the set $1, 2, ..., P$.

The expected payoff of the *Scheduler* in the Stackelberg Game depends on the completion time of all the *Tasks* in the scheduled *Job*, thus, the makespan of that *Job*. In order to minimize the makespan, we employed a Monolithic Scheduler [15] which makes scheduling decisions based on the Expected Time to Compute (ETC) matrix, defined as follows:

$$ETC = [ETC[j][i]]_{j=1,\ldots,n}^{i=1,\ldots,m^p} \tag{6}$$

where

$$ETC[j][i] = wl_j / cc_i^p \tag{7}$$

In this equation, cc_i^p denotes the computational capacity of the i-th Computing Node (CN) in the pth Computing Unit (CU) in Giga Flops per Second (GFLOPS) and wl_j represents the workload of j-th

task in Flops (FLO); n and m^p denote the number of tasks and number of Computing Nodes in the pth Computing Unit respectively, see [12]. The main goal of the scheduling strategy is the minimization of the *Job* makespan:

$$C_{\max}(wl_1, ..., wl_n, cc_1^p, ..., cc_{m^p}^p, m^p, n, p) = \quad (8)$$

$$= \min_{S \in Schedules} \left\{ \max_{j \in Tasks} C_j \right\}, \quad (9)$$

where C_j is the completion time of the j-th task. $Tasks$ represents the set of tasks in the *Job*, and $Schedules$ is the set of all possible schedules that can be generated for the *Tasks* of that *Job*. The shortest makespan is achieved by the means of the Expected Time to Compute (ETC) matrix. In this matrix, the cell corresponding to the ith row and the jth column shows the completion time of the jth task if deployed on the ith CN. The lower values are to be considered, since the higher the value, the longer the makespan of the *Job*. Once the optimal schedule is computed, the obtained makespan value for that *Job* is taken as the utility function value for the game leader:

$$H_1(\sigma_1, \sigma_2) = \sum_{p=1,...,P} \sum_{l=1,..,L} \sigma_1^p \sigma_2^l C_{\max}(wl_1, ..., wl_n, cc_1^p, ..., cc_{m^p}^p, m^p, n, p) \quad (10)$$

$$= C_{\max} = \min_{S \in Schedules} \left\{ \max_{j \in Tasks} C_j \right\}, \quad (11)$$

where L indicates the number of decisions that may be taken by the game follower. The value of the makespan depends on the computational power of the CNs. These parameters may be modified by the second player. Therefore, we may rewrite the eq. (5) in the form of:

$$H_1(\sigma_1, \sigma_2) = \sum_{p=1,...,P} \sum_{l=1,..,L} \sigma_1^p \sigma_2^l C_{\max}(wl_1, ..., wl_n, cc_1^p(l), ..., cc_{m^p}^p(l), m^p(l), n, p(l)) \quad (12)$$

Thus, each player decisions infers the decision of the other player.

B. Follower Payoff and decisions

The follower in the game is the *Central Energy-efficiency Manager*, which applies energy policies to all the CNs in the data center. Those policies may change the availability of the CNs, which directly impact on their computational capacity. Therefore, the follower may decide about the $cc_i^p(l)$ and the $m^p(l)$. Those decisions will influence not only the follower's utility value, but also the behavior of the leader. The payoff for the follower player is the energy consumed by the CC system for the execution of the scheduled computed by the *Scheduler* and can be defined by the following equation:

$$H_2(\sigma_1, \sigma_2) = \sum_{i=1,...,m} \sum_{j=1,..,n} \sigma_1^j \sigma_2^i E(wl_1, ..., wl_n, cc_1^p, ..., cc_{m^p}^p, m^p, n, p, schedule) \quad (13)$$

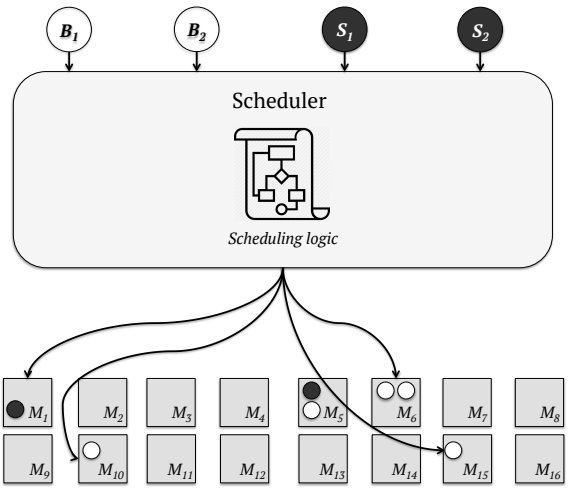


Fig. 2: Leader payoff computing, B - Batch type task, S - Service type task, M - Computing Node

The above equation shows that the payoff of the follower player depends on both the workload of the *Job* under consideration, and the schedule resulting of the leader player move. This means that every decision made by the leader player influences the next decision of the follower player. This scenario is similar to a chess play. Thus, each move of a given player changes the game environment and enforces the other player to make a decision.

In order to calculate the energy consumed by the CC during computing the given tasks according the the chosen schedule, the following values were introduced:

- E_{total} - the total energy spent by particular *Job*
- t_{idle}^i - c;
- t_{busy}^i - the time that the i-th CN spends on computing tasks;
- P_{idle}^i - the power a CN requires to operate in a idle state;
- P_{busy}^i - the power a CN requires to compute tasks

Therefore, we may express the time that the i-th CN spends on computing tasks by calculating:

$$t_{busy}^i = max_{j \in Tasks\ scheduled\ \ for\ \ CN_i} C_j \quad (14)$$

and the time that the i-th CN spends on computing tasks as follows:

$$t_{idle}^i = C_{max} - t_{busy}^i \quad (15)$$

The total energy consumption may be expressed in

the following way:

$$E_{total} = \sum_{i=1}^{m} \int_{0}^{C_{max}} Pow_{CN_i}(t)dt = \sum_{i=1}^{m} (P^i_{idle} * t^i_{idle} + P^i_{busy} * t^i_{busy} \quad (16)$$

These utility functions model the competition between two strategies that are usually contradictory: that of the *Scheduler*, which tries to compute tasks as fast as possible and that of the *Central Energy-efficiency Manager*, which tries to apply the more optimal power states to the CNs in order to maximize the energy efficiency.

V. Experimental analysis of the Stackelberg Game model

A. Simulation tool

The analysis of the described energy-efficiency strategies in real-life large-scale data centers is not feasible in such an immature stage. To overcome this limitation, in this work we chose a simulation tool designed to trustfully simulate energy-aware large-scale data centers called SCORE [7], which provides us with the means to reproduce realistic heterogeneous workloads and to easily implement various energy-efficiency policies.

In this paper, we extended the SCORE simulation tool in order to implement the Stackelberg Game process, which dynamically switches between energy-efficiency policies. The resulting architecture is shown in Figure 3.

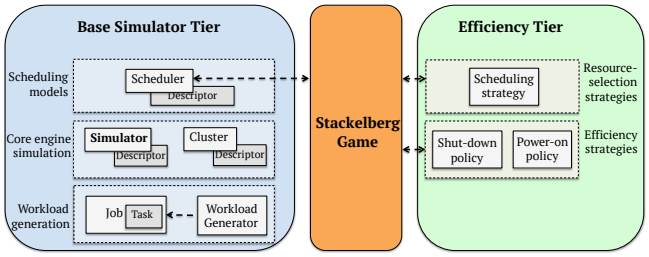


Fig. 3: Simulation tool architecture

B. Simple example for Always and Never power-off policies in SCORE simulator

In this experiment, we aim to empirically show a simple strategy where a dynamic change of *Power-off* policy could represent a significant improvement of energy-efficiency.

We used the SCORE simulator [7] to perform a simple experiment that simulates seven days of operation time of a data center composed of 1,000 heterogeneous machines of 4 CPU cores and 8GB RAM and one central monolithic scheduler. In this experiment, we chose an heterogeneous day-night patterned mixed workload. This workload uses 30% of the data center computational resources on average, with peak loads that achieve 60% of utilization. This heterogeneous workload is composed of the following kind of jobs:

- **Batch jobs** perform a given amount of computational work and then are completed. Thus, this kind of job has a given start and end. In this experiment, *Batch* jobs are composed of 50 homogeneous tasks which consume 0.3 CPU cores and 0.5 GB of memory, and last for 90 seconds on average.
- **Service jobs** represent long-running jobs which serve end users. Due to this, this kind of job has an undetermined finish time. In this experiment, *Service* jobs are composed of 9 homogeneous tasks which consume 0.5 CPU cores and 1.2 GB of memory, and last for 2,000 seconds on average.

The Stackelberg process described previously is applied for every scheduling decision in the system. In this experiment, the *Shut-down decision policy* used to switch the *Power-off* policy is made based on cluster available resources. Every time that the idle resources exceed a given threshold, the *Always* power-off policy is applied. On the other hand, when the amount of available resources is lower than that threshold, the *Never* power-off policy is applied. The results of the application of the Stackelberg game against the static energy policies are presented in Tables II and III.

This experimentation shows that the Stackelberg approach applies almost no negative impact in terms of queue times compared to the *Always* strategy at the cost of approximately 9% of more energy consumption. On the other hand, the *Always* strategy achieves an approximately 10% lower final average makespan time while achieving approximately 20% higher queue times.

C. Extended example in SCORE simulator

In this section, we extended the simple experimentation presented in Section V-B. In order to keep results comparable, we reused all the configuration parameters taken for the large-scale CC system shown in Section the V-B. However, in this experiment the *Central Energy-efficiency Manager* switches dynamically between the *Never* and the *Always* power-off policies by applying every *Shut-down decision policy* described in Section III. The results obtained are shown in Table IV and V.

In general, the Stackelberg process may apply a negative impact in terms of makespan due to that the *Power-off* policy may suddenly change. This change can impact on two consecutive scheduling processes of a single job, which could apply a performance penalty if there are no sufficient resources to immediately execute the job tasks. This negative impact can be mitigated by the scheduler when only one static *Power-off* policy is applied. It should be borne in mind that only *Batch* jobs would suffer from this negative impact since *Service* jobs have no determined finish.

This experimentation shows that the results of the Stackelberg approach depends directly on the *Shut-*

TABLE II: Energy-efficiency results for the simple Stackelberg experiment

Strategy	MWh consumed	MWh saved	Savings (%)	# shut downs	kWh saved per shut-down	Idle resources (%)
Never	58.53	0	0	0	N/A	69.94
Always	30.95	27.82	47.34	16,071	1.7314	3.52
Stackelberg	**36.07**	**23.00**	**38.94**	**742**	**31.00**	**13.96**

TABLE III: Performance results for the simple Stackelberg experiment

Strategy	Workload	Queue time until all tasks scheduled (ms)	Queue time until first task scheduled (ms)	Scheduler busy time (h)	Final makespan avg. (s)	Epoch 0 makespan avg. (s)
Never	Batch	74.11	74.11	9.28	136.16	175.46
Never	Service	73.72	73.72	0.66	N/A	N/A
Always	Batch	125.27	88.26	10.52	143.65	184.33
Always	Service	126.12	88.99	0.70	N/A	N/A
Stackelberg	**Batch**	**75.14**	**74.32**	**9.30**	**162.08**	**178.09**
Stackelberg	Service	78.31	74.40	0.66	N/A	N/A

down decision policy. More conservative probabilistic approaches, such as Gamma, achieve 26% faster queue times than a *Random* approach by consuming approximately 7% more energy. On the other hand, strategies that rely on leaving a security margin of free resources, such as *Margin*, could achieve approximately 22% faster queue times than a *Random* approach, and it would only consume less than 5% more energy. It can be noticed that conservative strategies such as Gamma apply almost no stress to the hardware, performing less than 1,000 shut-downs in a week of operation time, which represents a 10% of those performed by the *Random* decision policy.

VI. SUMMARY

In this paper, we presented a method that focus on the balance between two opposite needs of every energy-efficient CC system: high performance throughput and low energy consumption.

The proposed model is based on a non-zero sum Stackelberg Game with the leader player, the *Scheduling Manager*, which tries to minimize the makespan with its scheduling decisions while the follower player, the *Energy-efficiency Manager*, responds to the leader player move with the application of energy-efficiency policies that may shut-down the idle machines. These strategies are represented by the independent utility functions for each player. Our model enables the dynamic application of energy-efficiency strategies depending on the current and predictable workload.

The results of our simple experimental evaluation show that the proposed model perform better than the application of only one energy-efficiency policy, both in terms of energy-efficiency and performance. This means that the Stackelberg Game model can balance better between opposed needs (performance and energy efficiency) and can adapt better to heterogeneous workloads.

It could be also observed in the experimental analysis, that probabilistic decision strategies that try to predict the short-term future workload can balance better between energy consumption and performance impact.

The presented model is just our first step towards the development of the new scheduling and resource allocation policies in order to optimize the energy utilization in the whole cloud distributing system . The model improvement plans include: a) exploration of more advanced energy policies; b) introduction of multiple players in order to play several games simultaneously without any central energy manager; c) examination of more scheduling models, such as two-level or shared-state models; and d) testing more complex and dynamic scheduling strategies.

ACKNOWLEDGEMENT

This article is based upon work from COST Action IC1406 "High-Performance Modelling and Simulation for Big Data Applications" (cHiPSet), supported by COST (European Cooperation in Science and Technology) and by the VPPI - University of Sevilla.

References

[1] D. G. Andersen and S. Swanson. Rethinking flash in the data center. *IEEE micro*, 30(4):52–54, 2010.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[3] A. Beloglazov and R. Buyya. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*, 24(13):1397–1420, 2012.

[4] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder. Temperature management in data centers: why some (might) like it hot. *ACM SIGMETRICS Performance Evaluation Review*, 40(1):163–174, 2012.

[5] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *ACM SIGARCH Computer Architecture News*, volume 35, pages 13–23. ACM, 2007.

[6] M. E. Femal and V. W. Freeh. Boosting data center performance through non-uniform power allocation. In *Second International Conference on Autonomic Computing (ICAC'05)*, pages 250–261. IEEE, 2005.

TABLE IV: Energy-efficiency results for the extended Stackelberg experiment

Strategy	Switch Policy	MWh consumed	MWh saved	Savings (%)	# shut downs	kWh saved per shut-down	Idle resources (%)
Never	N/A	57.18	0.00	0	0	N/A	69.90
Always	N/A	30.54	26.74	46.68	13,680	1.95	3.87
Stackelberg	Margin	33.59	23.80	41.47	1,331	17.89	10.90
Stackelberg	Random	30.80	26.47	46.21	9,763	2.71	4.57
Stackelberg	Gamma	35.18	22.36	38.85	959	23.31	14.31

TABLE V: Performance results for the extended Stackelberg experiment

Strategy	Workload	Switch Decision Policy	Queue time until all tasks scheduled (ms)	Queue time until first task scheduled (ms)	Scheduler busy time (h)	Final makespan avg. (s)	Epoch 0 makespan avg. (s)
Never	Batch	N/A	71.07	71.07	9.16	139.66	177.25
Never	Service	N/A	73.89	73.89	0.66	N/A	N/A
Always	Batch	N/A	121.02	84.61	10.24	141.98	184.97
Always	Service	N/A	111.06	85.44	0.70	N/A	N/A
Stackelberg	Batch	Margin	78.33	72.61	9.24	159.98	179.74
Stackelberg	Service	Margin	77.92	75.88	0.66	N/A	N/A
Stackelberg	Batch	Random	100.79	79.58	9.88	141.95	183.85
Stackelberg	Service	Random	95.40	82.05	0.69	N/A	N/A
Stackelberg	Batch	Gamma	74.20	71.76	9.20	163.81	179.36
Stackelberg	Service	Gamma	78.47	75.69	0.66	N/A	N/A

[7] D. Fernández-Cerero, A. Fernández-Montes, A. Jakóbik, J. Kołodziej, and M. Toro. Score: Simulator for cloud optimization of resources and energy consumption. *Simulation Modelling Practice and Theory*, 82:160–173, 2018.

[8] A. Fernández-Montes, D. Fernández-Cerero, L. González-Abril, J. A. Álvarez-García, and J. A. Ortega. Energy wasting at internet data centers due to fear. *Pattern Recognition Letters*, 67:59–65, 2015.

[9] A. Fernández-Montes, L. Gonzalez-Abril, J. A. Ortega, and L. Lefèvre. Smart scheduling for saving energy in grid computing. *Expert Systems with Applications*, 39(10):9443–9450, 2012.

[10] A. Jakóbik, D. Grzonka, and J. Kołodziej. Security supportive energy aware scheduling and scaling for cloud environments. 2017.

[11] A. Jakóbik, D. Grzonka, J. Kolodziej, A. E. Chis, and H. González-Vélez. Energy efficient scheduling methods for computational grids and clouds. *Journal of Telecommunications and Information Technology*, (1):56, 2017.

[12] A. Jakobik, D. Grzonka, and F. Palmieri. Non-deterministic security driven meta scheduler for distributed cloud organizations. *Simulation Modelling Practice and Theory*, in press. (available online 4 November 2016).

[13] F. Juarez, J. Ejarque, and R. M. Badia. Dynamic energy-aware scheduling for parallel task-based application in cloud computing. *Future Generation Computer Systems*, 2016.

[14] R. T. Kaushik and M. Bhandarkar. Greenhdfs: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster. In *Proceedings of the USENIX annual technical conference*, page 109, 2010.

[15] J. Kołodziej. *Evolutionary Hierarchical Multi-Criteria Metaheuristics for Scheduling in Large-Scale Grid Systems*. Springer Publishing Company, Incorporated, 2012.

[16] J. Koomey. Growth in data center electricity use 2005 to 2010. *A report by Analytical Press, completed at the request of The New York Times*, 9, 2011.

[17] Y. C. Lee and A. Y. Zomaya. Energy efficient utilization of resources in cloud computing systems. *The Journal of Supercomputing*, 60(2):268–280, 2012.

[18] X. Luo, Y. Wang, Z. Zhang, and H. Wang. Superset: a non-uniform replica placement strategy towards high-performance and cost-effective distributed storage service. In *Advanced Cloud and Big Data (CBD), 2013 International Conference on*, pages 139–146. IEEE, 2013.

[19] A. Miyoshi, C. Lefurgy, E. Van Hensbergen, R. Rajamony, and R. Rajkumar. Critical power slope: understanding the runtime effects of frequency scaling. In *Proceedings of the 16th international conference on Supercomputing*, pages 35–44. ACM, 2002.

[20] E. Niewiadomska-Szynkiewicz, A. Sikora, P. Arabas, and J. Kołodziej. Control system for reducing energy consumption in backbone computer network. *Concurrency and Computation: Practice and Experience*, 25(12):1738–1754, 2013.

[21] S. Ruth. Reducing ICT-related carbon emissions: an exemplar for global energy policy? *IETE technical review*, 28(3):207–211, 2011.

[22] R. K. Sharma, C. E. Bash, C. D. Patel, R. J. Friedrich, and J. S. Chase. Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Computing*, 9(1):42–49, 2005.

[23] S. Sohrabi, A. Tang, I. Moser, and A. Aleti. Adaptive virtual machine migration mechanism for energy efficiency. In *Proceedings of the 5th International Workshop on Green and Sustainable Software*, pages 8–14. ACM, 2016.

[24] A. Wilczyński and A. Jakóbik. Using Polymatrix Extensive Stackelberg Games in Security–Aware Resource Allocation and Task Scheduling in Computational Clouds. *Journal of Telecommunications and Information Technology*, 1, 2017.

AUTHOR BIOGRAPHIES

AGNIESZKA JAKÓBIK (KROK) received her M.Sc. in the field of stochastic processes at the Jagiellonian University, Cracow, Poland and Ph.D. degree in the field of neural networks at Tadeusz Kosciuszko Cracow University of Technology, Poland, in 2003 and 2007. She is an Assistant Professor. Her e-mail address is: agneskrok@gmail.com

JOANNA KOŁODZIEJ is an associate professor in Research and Academic Computer Network (NASK) Institute and Department of Computer Science of Cracow University of Technology. She serves also as the President

of the Polish Chapter of IEEE Computational Intelligence Society. Her e-mail address is: joanna.kolodziej68@gmail.com

DAMIÁN FERNÁNDEZ CERERO received the B.E. degree and the M.Tech. degrees in Software Engineering from the University of Sevilla. In 2014, he joined the Department of Computer Languages and Systems, University of Seville, as a PhD. Student. Currently he both teaches and conducts research at University of Sevilla. He has worked on several research projects supported by the Spanish government and the European Union. His research interests include energy efficiency and resource scheduling in data centers. His e-mail address is: damiancerero@us.es

ALEJANDRO FERNÁNDEZ-MONTES GONZÁLEZ received the B.E. degree, M. Tech. and International Ph.D. degrees in Software Engineering from the University of Sevilla, Spain. In 2006, he joined the Department of Computer Languages and Systems, University of Sevilla, and in 2013 became Assistant Professor. His research interests include energy efficiency in distributed computing, applying prediction models to balance load and applying on-off policies to Data Centers. His e-mail address is: afdez@us.es

ANN-based secure task scheduling in computational clouds

Jacek Tchórzewski
AGH University of Science and Technology
30-059 Cracow, Poland
Cracow University of Technology
31-155 Cracow, Poland

Ana Respício
CMAFCIO
Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal

Joanna Kołodziej
Research and Academic Computer Network (NASK)
Kolska st 12, 01-045 Warsaw, Poland

KEYWORDS

Computational Clouds; Cloud Security; Tasks Scheduling; Artificial Neural Networks; Intelligent Security System.

ABSTRACT

Assuring the security of services in Computational Clouds (CC) is one of the most critical factors in cloud computing. However, it can complicate an already complex environment due to the complexity of the system architecture, the huge number of services, and the required storage management. In real systems, some security parameters of CC are manually set, which can be very time-consuming and requires security expertise.

This paper proposes an intelligent system to support decisions regarding security and tasks scheduling in cloud services, which aims at automating these processes. This system comprises two different kinds of Artificial Neural Networks (ANN) and an evolutionary algorithm, and has as main goal sorting tasks incoming into CC according to their security demands. Trust levels of virtual machines (VMs) in the environment are automatically set to meet the tasks security demands. Tasks are then scheduled on VMs optimizing the makespan and ensuring that their security requirements are fulfilled.

The paper also describes tests assessing the best configurations for the system components, using randomly generated batches of tasks. Results are presented and discussed. The proposed system may be used by CC service providers and CC consumers using Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) Cloud Computing models.

I. INTRODUCTION

Validation of security demands (SD) of tasks processed in Computational Clouds (CC) is a crucial part of the CC workload management process [5], [9]. Personalization of cloud services enables users to change the features of Virtual Machines (VMs) that provide the computational power for executing pools of tasks. Especially the security Trust Levels (TLs) offered by VMs may be changed and adapted to fit the SD of tasks. Another important aspect is to ensure the proper assignment of tasks to suitable VM offering the proper TL to fulfill the SD required by tasks.

This paper presents an intelligent system for management of tasks submitted into the cloud. Fig. 1 displays a representation of the proposed system, which considers three computational cloud units:

- the edge of the cloud,
- the cloud computing center, and
- the cloud storage center.

Additional units could be considered, however these three provide a sufficient level of detail for testing our system.

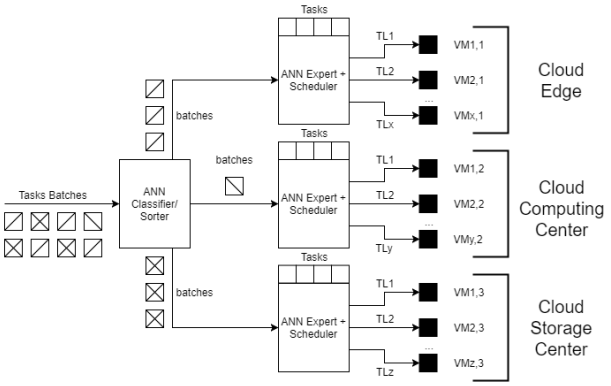


Fig. 1: Model of the proposed system

Tasks are submitted to the cloud in batches, which are classified into security classes according to their tasks SDs. According to this classification, batches are then delivered to be processed at the proper cloud units, thus allowing to determine the workload required in each cloud unit and further to decide about the configuration of the available VMs. These VMs are then customized to serve the TLs that fit the tasks SDs. Finally, tasks in each of the considered cloud units are scheduled on VMs offering proper security levels while minimizing the total processing time.

The presented system is composed of two Artificial Neural Networks (ANNs) and an Evolutionary Scheduler. One of the ANNs is a classifier/sorter ANN, which classifies and sorts batches incoming into the cloud;

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

the other is an Expert ANN that predicts the VMs configurations. The Evolutionary Scheduler optimizes the scheduling of tasks relying on an evolutionary algorithm.

The paper is organized as follows. Section II presents the concepts underlying task processing in CC systems considering security aspects. Section III is devoted to describing the proposed intelligent system and its components. In Section IV, we describe in detail the system testing and discuss its results. The paper ends with Section V, which contains a summary, conclusions based on the experiments results, and ideas for future work.

II. SECURITY OF TASK PROCESSING IN CC SYSTEMS

Processing tasks with security requirements in CC involves many steps [10]. End users start by specifying security demands for their tasks before uploading them into the cloud. Then one has to decide in which component of the cloud infrastructure should tasks be processed. For instance, some tasks may be computed at the cloud edge, while others have to be processed inside the cloud. The collected tasks are then scheduled into available VMs which offer proper security trust levels. Finally, the results of tasks processing are returned to the end users or delivered to another service from CC.

Assuring the security of the users data and cloud infrastructure is a complex process [18], [9]. In the IaaS and PaaS models, often the biggest responsibility for ensuring security in the cloud is on the cloud providers side [3]. Moreover, as there still exist many vulnerabilities in cloud systems, international organization have produced standards providing guidance regarding cloud security [16], [8].

A. *Mapping security requirements into VMs*

Considering a given cloud unit for which there are n tasks to be processed, and that there m available VMs in that CC unit, we define an SD vector describing the security requirements of these tasks [5]:

$$SD = [sd_1, ...sd_n], \quad (1)$$

where sd_j is the security demand value of the jth task to be processed in the cloud unit. Accordingly, we define the TL vector representing the security levels offered by the VMs in that CC unit [5]:

$$TL = [tl_1, ..., tl_m], \quad (2)$$

where tl_i is the trust level offered by the ith VM in the CC unit. A task can only be scheduled on a particular VM that offers a TL equal or higher than the SD of the task. According to the NIST guidelines [2] we propose four levels of security demand/trust level, that is: $sd_j, tl_i \in \{0, 1, 2, 3\}$.

To differentiate security requirements as far as cryptography is concerned, we introduce three classes:

- Class 1, containing only small tasks, processed online using strong but fast enough cryptography algorithms. The RSA asymmetric cipher with 1024 bits key should be used for ciphering and deciphering tasks [15].
- Class 2, requiring operational cryptography [19]. This class of security demands is designed for most tasks, which may have varied workload and are processed mostly off-line, using advanced cryptography protocols without security compromises. If it is possible, these protocols should be designed to be as fast as possible. The RSA asymmetric cipher with 2048 bits key should be used for ciphering and deciphering tasks.
- Class 3, corresponding to data at rest cryptography [4]. This class contains tasks with very heavy workload, mainly ciphering data to be stored, and using the strongest cryptography methods designed for processing data at rest. The RSA asymmetric cipher with 2048 bits keys is used for ciphering and deciphering tasks. Furthermore, tasks are signed with Elliptic Curve Digital Signature Algorithm (ECDSA) based on curve defined over 521 bits field [1].

B. *Security computational overhead*

Each security operation adds a computational overhead that has to be considered during tasks processing and comprises two parts as follows.

1. Part of this overhead influences the task scheduling and includes pre and post-processing security operations, such as verification of the task integrity and ciphering results of task processing. The bias required to deliver tl_i to task j, requiring a security demand of sd_j, is assumed as an estimated time (sec.) denoted by [11]:

$$b_{i,j} = b(sd_j, wl_j, tl_i, cc_i, inputSize_j, outputSize_j) \quad (3)$$

where wl_j is the workload of task j, $inputSize_j$ and $outputSize_j$ are the sizes (in bytes) of the files characterizing the task and storing the corresponding result, respectively, and cc_i is the computational capacity of VM i, $i = 1, 2, .., m, j = 1, 2..., n$. This value can be approximated by the sum of the number of instructions to compute the cryptographic requirements and the workload.

2. The remaining part of the computational overhead is associated with complying the security protocols and performing operations that do not influence the scheduling process, such as, ciphering data stored in the data center (before sending the task back to the end user) and verifying the digital signature of the end user who wants to recover some results from the Cloud Computing system.

C. *Scheduling tasks into VMs by matching SD and TL*

Tasks scheduling is the process of assigning tasks to the available VMs assuring that each task is processed by a VM offering a TL value of at least the required SD and optimising the utilisation of VMs by minimizing the makespan. Our system uses the scheduler previously developed in [10], [11], which has as objective

the minimization of the makespan - the time of conclusion of the last task. As an additional criterion, here we introduce the condition of compliance with the required SD.

III. THE INTELLIGENT SYSTEM TO SUPPORT SECURITY DECISIONS

This section describes the proposed system and its components: the sorter ANN, the expert ANN, and the scheduler. The input of the system is a stream of batches of tasks. Each batch is considered separately. The system works according to the schema illustrated in Fig. 1.:

- The first stage consists of classifying batches and sorting them to the appropriate destination in the cloud system considering their security requirements.
- The second stage concerns setting the VMs parameters according to the tasks workloads and security requirements so that all tasks can be computed with adequate security levels and without spare computing time or energy losses.
- The last stage is assigning the tasks from each batch into the created VMs ensuring the proper security level.

The output from the system is a vector describing the TL of all VMs and the complete schedule.

A. *The sorter ANN*

The system starts by automatically classifying the incoming traffic into security classes, which define the cryptography level required by tasks in a particular batch, without any additional knowledge about them.

The classified batches are then sorted into pools corresponding to their security classes and defining where the tasks should be processed in the CC infrastructure:

- Class 1: tasks that may be processed in the cloud edge,
- Class 2: tasks to be executed by fast VMs inside the cloud, and
- Class 3: tasks that have to be stored in the cloud longer and then processed inside cloud storage centers (when enough computational power is available).

To define the training set for the pattern recognition problem, a set of batches was previously classified. The numbers of tasks requiring each SD value l were counted separately in each batch:

$$N(t)_l = card\{j : sd_j = l\}, l = 0, 1, 2, 3 \quad (4)$$

where t indexes batches entering the CC system, $t = 1, 2, ..., T$ and $card$ represents the set cardinality.

The SD statistics for the t-th batch are stored in vector $input^{recog}(t)$, which is considered as the input for the ANN classifier/sorter.

$$input^{recog}(t) = [N(t)_l]. \quad (5)$$

The security class of each batch t is declared by specifying its target value

$$target^{recog}(t) = k. \quad (6)$$

The training and testing sets were created from data obtained from different batches coming during system operations:

$$TSET^{recog} = \{(input^{recog}(t), target^{recog}(t))^{t=1,2...,T}\} \quad (7)$$

The original set was split into three parts: the training $TSET^{recog}_{train}$, containing information from 70% of all batches; the validation set $TSET^{recog}_{valid}$, and testing set $TSET^{recog}_{test}$, each containing 15% of randomly chosen batches (not used for training).

A shallow feed-forward Neural Network was then trained to classify inputs according to the target classes defined before. We used a two-layer feed-forward network with sigmoid hidden and softmax output neurons [12], see Fig. 2.

Additionally, the sorter may help to detect anomalies in batches, for instance, by allowing the identification of differences in security demands from past patterns, therefore supporting detection of hostile tasks or users abnormal behavior.

B. *Expert system for setting Trust Levels*

The aim of the expert system is to decide the proper TL values for the VMs based on the SD levels required by the tasks. For each cloud unit, an expert ANN was designed to get knowledge about the computational capacities cc of all VMs, as represented in Fig.1.

We examined several strategies of assigning TL values to VMs, for example, by using arbitrary human decisions or Stackelberg Game solutions [12]. Then we trained the ANN to mimic these decisions. To formulate the input for this expert system, we analyzed the individual tasks workload; wl_j denotes the workload of task j. The total workload of tasks requiring the SD level l in batch t is represented by $W(t)_l$

$$W(t)_l = \sum_j \{wl_j : sd_j = l\}, l = 0, ..., 3, t = 1, 2, ..., T \quad (8)$$

.

Vector

$$input^{expert}(t) = [W(t)_l] \quad (9)$$

represents the workload for each SD value l in batch t, and is given as the input to the ANN expert system.

Target values were defined to indicate the trust levels tl for all the virtual machines in the system, in ascending computer capacity order.

Vector

$$target^{expert}(t) = [tl_1, tl_2, ..., tl_m] \quad (10)$$

represents the decisions of an expert for a particular batch t. This vector assigns the trust level tl_i to the i-th VM according to it's computing capacity cc_i and the amount of work that has to be done using trust level tl_i, $i = 1, ..., m$.

The training, validation and testing sets were also formulated considering data obtained from different

batches entering the system, through the definition of the set

$$TSET^{expert} = \{(input^{expert}(t), target^{expert}(t))^{t=1,2...,T}\} \quad (11)$$

which was split into training set $TSET^{expert}_{train}$, validation set $TSET^{expert}_{valid}$ and $TSET^{expert}_{test}$, analogously to the process described in Section III-A for the sorter ANN.

A backpropagation feed-forward NN was then trained with the defined inputs and used for targets prediction. We used a two-layer ANN, with sigmoid hidden and linear output neurons [7], as represented in Fig.4.

The quality of prediction was assessed through the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (12)$$

. where:
- y is the given set of data,
- $\hat{y}$ represents the calculated values of y, and
- $\bar{y}$ represents mean value of y.

A properly trained expert system may be an automatic alternative for human security decisions made by CC administrators. Once the ANN is trained, it can also deal with unknown situations, without requiring additional customization.

C. *Evolutionary scheduler*

Our system uses the Independent Batch Scheduler [13] as the main method of mapping tasks into VMs. The Security Biased Expected Time to Compute (SBETC) matrix with security biases is computed using Eq. (13):

$$SBETC = [ETC[j][i] + SB(SD,TL)(i,j)]^{i=1,...,m}_{j=1,...,n} \quad (13)$$

where

$$ETC[j][i] = wl_j/cc_i \quad (14)$$

in which cc_i is the computational capacity of the i-th VM in Giga Flops per Second (GFLOPS) and wl_j is the workload of j-th task in Flops (FLO); n and m are the number of tasks and the number of VMs, respectively.

The Security Bias Matrix (SB) is obtained by aggregation of security biases in a matrix form:

$$SB(SD,TL)(i,j) = [b(sd_j, wl_j, tl_i, cc_i, inputSize_j, outputSize_j)]^{j=1,2...,n}_{i=1,2...,m}. \quad (15)$$

The full description of this model can be found in [11]. The main scheduling objective is the minimization of the makespan, which can be defined as follows:

$$C_{\max} = \min_{S \in Schedules} \left\{ \max_{j \in Tasks} C_j \right\}, \quad (16)$$

where C_j is the conclusion time of the j-th task, $Tasks$ is the set of tasks in the batch, and $Schedules$ is the set of all possible schedules which can be generated for the tasks from that batch. The scheduler is based on the evolutionary algorithm solution proposed in [10] and [11], using only a particular subset of its features.

IV. NUMERICAL EVALUATION OF THE SYSTEM PERFORMANCE

Cloud Sim test bed (www.cloudbus.org) was used as a testing tool. All security algorithms were implemented in Java. Processing of different dimensions pictures was considered as a tasks set. The examined workload was based on proposed day and night pattern (to represent the load of a CC system). All designed ANN were implemented in MATLAB 2017 (www.mathworks.com). We examined different ANN sizes together with different learning algorithms, to assess the quality of the solutions.

A. *Trust Levels*

The following trust levels of VMs were considered:
- Trust Level 0 (tl=0) - bare tasks are processed without any cryptographic computational overhead (no security required).
- Trust Level 1 (tl=1) - corresponds to the TL required by tasks of Class 1 as defined in Section III.
- Trust Level 2 (tl=2) - corresponds to the TL required by tasks of Class 2 as defined in Section III.
- Trust Level 3 (tl=3) - corresponds to the TL required by tasks of Class 3 as defined in Section III.

B. *Tasks and their security demands*

The security demand of each task was defined according to the trust level values defined in Section III. Tab. 1 presents the sizes of ten pictures that were used for the tests. For each picture, the size is presented in pixels and with a qualitative classification (Small, Medium and Big).

TABLE 1: Pixel size of blurred JPG pictures

Picture Number [PN]	Picture size [pixels]	Size
PN1	200x200	Small
PN2	400x400	Small
PN3	600x600	Medium
PN4	800x800	Medium
PN5	1000x1000	Medium
PN6	1200x1200	Medium
PN7	1400x1400	Medium
PN8	1600x1600	Medium
PN9	1800x1800	Medium
PN10	2000x2000	Big

As bare task we considered a Gaussian Blur operation with 5x5 mask made on each of these pictures. A task resulted from combining a picture size [PN1-PN10] with an SD value. Tab. 2 presents the characteristics of these tasks. For each task, column (1) is the task ID, columns (2) and (4) present the pair (PN, sd), column (3) is the workload in terms of the number of instructions (without security), column (5) is the number of

instructions required to process the SD (bias), and column (6) is the total number of instructions to process the task (size).

TABLE 2: All possible tasks and their characteristics in terms of workload without bias, SD value, bias workload and workload with bias

task ID (1)	PN (2)	Instr. bare task wl (3)	sd (4)	Instr. bias b (5)	Instr. TOTAL wl $wl+b$ (6)
1	PN1	180281484	0	0	180281484
2			1	58248804234	58429085718
3			2	189148318241	189328599725
4			3	193351681261	193531962745
5	PN2	366694660	0	0	366694660
6			1	233237558475	233604253135
7			2	762857236471	763223931131
8			3	767753680777	768120375437
9	PN3	582192590	0	0	582192590
10			1	549063149835	549645342425
11			2	1723705647330	1724287839920
12			3	1729619478414	1730201671004
13	PN4	935976091	0	0	935976091
14			1	984022696944	984958673035
15			2	3071839716809	3072775692900
16			3	3079432624626	3080368600717
17	PN5	1381754383	0	0	1381754383
18			1	1467324300905	1468706055288
19			2	4807412731944	4808794486327
20			3	4817016916348	4818398670731
21	PN6	1907429016	0	0	1907429016
22			1	2218939482253	2220846911269
23			2	6930461839807	6932369268823
24			3	6942582060238	6944489489254
25	PN7	2541112486	0	0	2541112486
26			1	2877893318993	2880434431479
27			2	9441396183684	9443937296170
28			3	9456608456473	9459149568959
29	PN8	2943581287	0	0	2943581287
30			1	3761249543741	3764193125028
31			2	12339533813479	12342477394766
32			3	12357819370145	12360762951432
33	PN9	3672489994	0	0	3672489994
34			1	4821170159062	4824842649056
35			2	15973646688412	15977319178406
36			3	15995539895451	15999212385445
37	PN10	5027151620	0	0	5027151620
38			1	5668347564768	5673374716388
39			2	19726744097608	19731771249228
40			3	19752079020993	19757106172613

C. *Classifier/sorter ANN tests*

The batches were classified as follows:

- Batch Type from Class 1 [BT1] - containing only small tasks which demand on-line and fast cryptography, which can be calculated e.g. on the edge of the cloud.
- Batch Type from Class 2 [BT2] - containing mixed big, medium and small tasks. Each task has to be considered separately.
- Batch Type from Class 3 [BT3] - comprising only big tasks that have to be sent to the cloud and stored until there is enough computational capacity available for the cryptographic bias.

The classifier/sorter ANN was designed to classify each batch into the proper type (BT1, BT2 or BT3). The batches workload in all three classes was generated according to following the day-night pattern function:

$$f(x) = \begin{cases} 25\sin(\frac{x\pi}{12} + 75) & \text{when } \sin(\frac{x\pi}{12}) > 0 \\ 75\sin(\frac{x\pi}{12} + 75) & \text{when } \sin(\frac{x\pi}{12}) \leq 0 \end{cases} \quad (17)$$

The classifier/sorter ANN uses the Scaled Conjugate Gradient (SCG) backpropagation learning method, which is appropriate for classification [6]. Neural Networks were tested with different numbers of neurons: 5, 10, 15, 20, 25, 30, 50, 100. Our focus is on assessing the True Positive Rate (TPR) for BT3 because this is the most critical classification. In our system, the worst situation occurs when a BT3 batch is classified as BT2 or BT1, meaning that a batch requiring high computational power could be delivered to a VM not offering enough computational power.

For each NN configuration, we made 10 measurements and computed the TPR of BT3 mean and standard deviation values. The corresponding results are shown in Tab. 3.

TABLE 3: Classifier/sorter ANN: TPR of BT3 average, standard deviation for each tested number of neurons in the hidden layer

Avr % ± St. dev.%	No. neurons
86,8 % ± 4,7%	5
85,8% ± 9,0%	10
86,9% ± 6,5%	15
90,9% ± 4,2%	20
86,5% ± 8,9%	25
88,7 % ± 6,1%	30
86,9 % ± 9,0%	50
86,4 % ± 8,9%	100

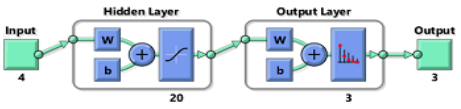


Fig. 2: Sorter ANN, a two-layer network with 20 sigmoid hidden neurons and 3 softmax output neurons.

The highest TPR for BT3 appeared when the SCG algorithm was applied for the ANN with 20 neurons (90,9% ± 4,2%). So this ANN was chosen as the most accurate one (Fig. 2).

The NNs results were also validated using confusion matrices for the training and testing process. Fig. 3 displays an example of the confusion matrix for the sorter ANN with 20 neurons. Green cells show the numbers of correct classifications (for each class separately) and the corresponding percentage of all data. Red cells represent these figures for incorrect predictions (for each class separately). Gray cells in the last column indicate the percentage of correct predictions for each class, while gray cells in the last row present the percentage of correctly classified cases. The blue cell shows the percentage of overall correct and incorrect predictions. For validation, the gray cell in the red frame was used, as it indicates the percentage of properly classified class 3 tasks. In this example the TPR for BT3 is 93.8%.

D. *Expert ANN tests*

The expert ANN has to allocate proper security levels to virtual machines (a machine with a given security level can compute only tasks with the same or lower security level). This network was created and tested with three learning methods appropriate for prediction [6]:

- SCG backpropagation algorithm,
- Bayesian Regularization (BR), and
- Levenberg-Marquardt (LM).

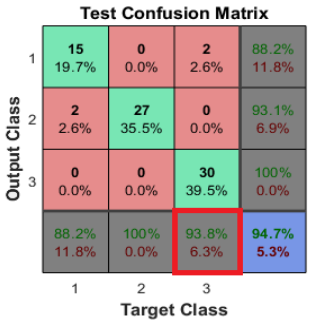


Fig. 3: Confusion Matrix for the Sorter ANN with 20 Neurons

For each of these methods, an ANN with different hidden layer size (5, 10, 15, 20, 25, 30, 50, 100) was ran ten times. To evaluate the Expert ANN performance we used the coefficient of determination (Eq. (12)). We computed it on the testing set (15% of all data) to determine which configuration of the method and the number of neurons present the highest R^2 value. Tab. 4 displays the average and standard deviation R^2 values in the ten runs of each configuration.

TABLE 4: Expert ANN: Average and standard deviation values of R^2 for learning methods LM, BR and SCG with different number of neurons

LM Avr % ± St. dev.	BR Avr % ± St. dev.	SCG Avr % ± St. dev.	Neurons No. of neurons
95% ± 2%	96,3% ± 2%	91,8% ± 2%	5
92% ± 3%	87,5% ± 10%	90% ± 4%	10
92% ± 3%	87,6% ± 14%	87,9% ± 4%	15
91% ± 3%	91,7% ± 3%	87,5% ± 5%	20
90% ± 2%	88,1% ± 8%	88,1% ± 5%	25
89% ± 4%	87% ± 3%	88,2% ± 3%	30
88% ± 5%	85,2% ± 5%	87,6% ± 4%	50
81% ± 3%	85,1% ± 5%	73,9% ± 7%	100

The best value of the coefficient of determination was obtained for the BR algorithm and an the ANN with 5 neurons (96,3% ± 2%). This ANN was chosen as the final solution, as illustrated by Fig. 4.

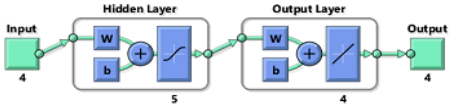


Fig. 4: Expert ANN, a two-layer network with 5 sigmoid hidden neurons and 4 linear output neurons.

E. Evolutionary scheduler tests

The aim of these tests was to examine how much time we may gain by scheduling tasks for generated VMs instead of submit them randomly.
For this purpose we have simulated a multi-cloud environment consisting of three types of cloud architectures, namely: cloud edge, cloud computing center and cloud storage center (see: Fig. 1). This environment is based on a real cloud characteristics from public (cloud storage center) and private, academic (cloud edge and cloud computing center) infrastructure. Tab. 5) presents the characteristics of the simulated environment: the VMs simulated in each CC unit, their types, their computational capacity (in GFLOPS) and their range of TLs.
Batches of tasks were generated and classified by the Classifier/sorter ANN and allocated to one of the simulated environment parts. For each of these components, the Expert ANN was determining the proper TLs for the corresponding VMs. Finally, the evolutionary scheduler was assigning tasks into particular VMs of each of the three parts of the tested system.

TABLE 5: Characteristics of the simulated Cloud, see Fig.1, and of the VMs used for simulation

Cloud Edge	cc	tl$\in 0,1,2,3$
VM number, type	[GFLOPS]	min:max
1, m1.tiny	0.293799	0
2, m1.tiny	0.293799	0
3, m1.tiny	0.293799	0
4, m1.tiny	0.293799	0
Cloud Computing Center	cc	tl$\in 0,1,2,3$
VM number, type	[GFLOPS]	min:max
5, m1.tiny	0.293799	0:1
6, m1.small	1.227791	0:2
7, m1.medium	1.856781	0:2
8, m1.large	3.444585	0:3
Cloud Storage Center	cc	tl$\in 0,1,2,3$
VM number, type	[GFLOPS]	min:max
9, m3.medium	97,6	0:1
10, m3.large	396,8	0:2
11, m3.xlarge	787,2	0:2
12, m3.2xlarge	1523,2	0:3

Tab. 6 presents the value of gain as far as makespan Eq. (16) is concerned. The presented results were obtained for 20 runs of the implemented evolutionary algorithm after 300 epochs.The results reveal that the maximum gain achieved by the application of the evolutionary scheduler was reached for the cloud storage center, being greater than twice the gain for the cloud edge.

TABLE 6: Results of applying the evolutionary scheduler: mean values of makespan gain in percentage for the Random task sender (Rand) and the Evolutionary scheduler (Evol)

CC part	Gain Rand	Gain Evol
Edge	0%	15%
Computing Center	0%	35%
Storage Center	0%	42%

V. SUMMARY

In this paper we presented an intelligent system for supporting security services in Computational Clouds (CC), which can improve the quality of security cloud services. The system is composed of two different kinds of Artificial Neural Networks (ANN) and an evolutionary algorithm. The first stage of system operation concerns sorting incoming batches of tasks according to their security demands. This allows one to divide the traffic into streams that can be processed by different parts of the CC environment. Our system uses the classifier/sorter ANN to perform this stage. The second stage, which is performed by the expert ANN, consists of fitting the security services offered by the VMs into the security demands of tasks in each of the CC components. Finally, the scheduler, based on the evolutionary algorithm, maps tasks into VMs, minimizing the makespan of tasks processed in the corresponding CC component. This scheduler considers the tasks characteristics in terms of size and security requirements, as well as the particular VM security services. Additional possible feature of our system is the possibility of detection of deviations in traffic incoming into CC. Therefore, in the future, it can be used to support the detection of security threats, like tasks injection or malicious workload.

The experimental results presented in this paper demonstrate and confirm the effectiveness of the system. The system is designed for CC service providers and CC consumers using Infrastructure as a Service or Platform as a Service Cloud Computing models.

In the future, we would like to introduce genetic algorithms for supporting the automatic detection of security threats.

ACKNOWLEDGEMENT

This article is based upon work from COST Action IC1406 "High-Performance Modelling and Simulation for Big Data Applications" (cHiPSet), supported by COST (European Cooperation in Science and Technology).

References

[1] W. A. Al-Hamdani. Elliptic curve for data protection. In *Proceedings of the 2011 Information Security Curriculum Development Conference*, InfoSecCD '11, pages 1–14, New York, NY, USA, 2011. ACM.

[2] P. Bowen, J. Hash, and M. Wilson. NIST Special Publication 800-100, Information Security Handbook: A Guide for Managers, 2006.

[3] CSA. *Security Guidance for Critical Areas of Focus in Cloud Computing V3.0*. Cloud Security Alliance, 2011.

[4] Google. Encryption at rest in google cloud platform, 2016.

[5] D. Grzonka, A. Jakóbik, J. Kołodziej, and S. Pllana. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security. *Future Generation Computer Systems*, 2017.

[6] M. T. Hagan, H. B. Demuth, and M. Beale. *Neural Network Design*. PWS Publishing Co., Boston, MA, USA, 1996.

[7] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[8] ISO/IEC. ISO/IEC 27002:2005 – Information Technology - Security Techniques - code of practice for information security management, 2005.

[9] A. Jakóbik. Big data security. In F. Pop, J. Kołodziej, and B. Di Martino, editors, *Resource Management for Big Data Platforms: Algorithms, Modelling, and High-Performance Computing Techniques*, pages 241–261, Cham, 2016. Springer International Publishing.

[10] A. Jakóbik, D. Grzonka, J. Kołodziej, and H. González-Vélez. Towards secure non-deterministic meta-scheduling for clouds. In *30th European Conference on Modelling and Simulation, ECMS 2016, Regensburg, Germany, May 31 - June 3, 2016, Proceedings.*, pages 596–602, 2016.

[11] A. Jakóbik, D. Grzonka, and F. Palmieri. Non-deterministic security driven meta scheduler for distributed cloud organizations. *Simulation Modelling Practice and Theory*, 76:67 – 81, 2017.

[12] A. Jakóbik and A. Wilczynski. Using polymatrix extensive stackelberg games in security-aware resource allocation and task scheduling in computational clouds. *Journal of Telecommunications and Information Technology (JTIT)*, pages 1–71, 2017.

[13] J. Kołodziej. *Evolutionary Hierarchical Multi-Criteria Metaheuristics for Scheduling in Large-Scale Grid Systems*. Springer Publishing Company, Incorporated, 2012.

[14] S. Matsuda and S. Moriai. Lightweight cryptography for the cloud: Exploit the power of bitslice implementation. In E. Prouff and P. Schaumont, editors, *Cryptographic Hardware and Embedded Systems – CHES 2012*, pages 408–425, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[15] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot. *Handbook of Applied Cryptography*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1996.

[16] T. E. Network and I. S. Agency. *Cloud Computing Benefits, risks and recommendations for information security*. ENISA, 2009.

[17] NIST. Publication 500-291, Version 2, NIST Cloud Computing Standards Roadmap. Technical report, 2013.

[18] P. K. S. Gupta. *Taxonomy of cloud security*, volume 3. 2013.

[19] B. Schneier. *Applied Cryptography (2Nd Ed.): Protocols, Algorithms, and Source Code in C*. John Wiley & Sons, Inc., New York, NY, USA, 1995.

AUTHOR BIOGRAPHIES

JACEK TCHÓRZEWSKI received his B.Sc. and M.Sc. degrees with distinctions in Computer Science at Cracow University of Technology, Poland, in 2016 and 2017, respectively. Currently he is Research and Teaching Assistant at Cracow University of Technology and Ph.D. student at AGH Cracow Univeristy of Science and Technology. His e-mail address is: jacek.tchorzewski@onet.pl

ANA RESPÍCIO is Assistant Professor at the Informatics Department, Faculty of Science, University of Lisboa, and Senior Researcher of CMAFCIO. She is vice-chair of the IFIP WG 8.3: Decision Support. Her e-mail address is: alrespicio@fc.ul.pt

JOANNA KOŁODZIEJ is an associate professor in Research and Academic Computer Network (NASK) Institute and Department of Computer Science of Cracow University of Technology. She serves also as the President of the Polish Chapter of IEEE Computational Intelligence Society. Her e-mail address is: joanna.kolodziej68@gmail.com

EFFICIENCY ANALYSIS OF RESOURCE REQUEST PATTERNS IN CLASSIFICATION OF WEB ROBOTS AND HUMANS

Grażyna Suchacka
Institute of Mathematics and Informatics
Opole University
ul. Oleska 48
45-052 Opole, Poland
E-mail: gsuchacka@uni.opole.pl

Igor Motyka
Institute of Mathematics and Informatics
Opole University
ul. Oleska 48
45-052 Opole, Poland
E-mail: igor_motyka@mail.com

KEYWORDS
Internet Robot, Web Bot, Web Crawler, Web Server, Web Traffic, HTTP Traffic, Classification

ABSTRACT

The paper deals with the problem of classification of Web traffic generated by robots and humans on e-commerce websites. Due to the still growing proliferation and specialization of bots, a large body of research into characterization and recognition of their traffic has been conducted so far. In particular, some approaches to classify bot and human sessions on websites have been proposed in the literature. In this paper we verify and discuss the efficiency of such recently proposed approach, which uses differences in resource request patterns of bots and humans. We reconstructed Web sessions from actual HTTP log data for three different e-commerce sites, varying in the traffic intensity and proportions of bot sessions in the overall traffic. Two heuristic procedures for labeling sessions as driven by a bot or a human were proposed and implemented. Resource request patterns for both session classes, using both session labeling procedures, were analyzed and their potential to differentiate between bot and human sessions was investigated. Results show that the broader session labeling procedure allows one to capture more bot sessions and that resource requests patterns are a good discriminant of bots and humans on e-commerce sites.

INTRODUCTION

Nowadays the traffic on many Web servers is dominated by Web robots. A Web robot (bot, crawler) is an autonomous software tool that can traverse the Web using the structure of hyperlinks and carry out specific tasks on visited sites. According to the Bot Traffic Report 2016 (Zeifman 2017), bots comprise the majority of online traffic, with an average of almost 52%. About 23% of the overall traffic may be attributed to good bots (feed fetchers, search engine bots, commercial crawlers, monitoring bots) and 29% – to bad ones (impersonators, hacking tools, content scrapers, spammers). The most active bots are impersonators, i.e., bots which assume false identities to circumvent security solutions and perform attacks – typically DDoS (distributed denial of service) attacks. It is common for such bots to hide behind user agents in HTTP/s headers to present themselves as legitimate Web clients (i.e., Web browsers driven by human users).

To cope with the undesirable presence of bots on Web servers, much research into the analysis and characterization of their traffic has been conducted so far (Almeida et al. 2001; Doran et al. 2013; Suchacka 2014). The goal has been to distinguish specific bots' features and to develop efficient bot recognition strategies. In (Doran and Gokhale 2011) four types of bot recognition approaches were distinguished taking into consideration the information used and techniques applied: syntactic log analysis, traffic pattern analysis, analytical learning techniques, and Turing test systems. In general, one can distinguish bot detection methods operating offline (Doran and Gokhale 2016; Lee et al. 2009; Saputra et al. 2013; Stassopoulou and Dikaiakos 2009; Stevanovic et al. 2011; Suchacka and Sobków 2015) and online (Balla et al. 2011; Doran and Gokhale 2016).

In this paper we focus on offline bot detection at a Web server, when a decision on session classification (bot or human) is made given a description of the whole session (as a sequence of HTTP requests). The key motivation for the offline bot detection is gaining the possibility of assessing an impact of bots on server performance and security, as well as gaining an insight into properties and behavioral patterns of bot traffic. This is useful to develop online bot detection methods.

In this study we implemented and tested the offline bot classification method proposed in (Doran and Gokhale 2016). This approach classifies bot and human sessions on a Web server based on the information on resource request patterns in session. In contrast to most of other classification methods, this approach is conceptually simple and relatively easy to implement (it does not involve a time-consuming learning phase to infer traffic patterns). Doran and Gokhale (2016) argued that their approach is effective because resource request patterns of robots and humans constitute an intrinsic distinction between these two types of sessions which is not expected to change over time. They evaluated the efficiency of their approach using log data from three various Web servers and compared it against some other analytical learning classifiers, achieving very good classification results of their approach in terms of recall, precision, and F1.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

The motivation for our experimental analysis of the approach proposed by Doran and Gokhale (2016) were some shortcomings of their experimental study. First, they used relatively old log data, which dated from 2008 to 2011, depending on a dataset. Second, only one out of three datasets was for an e-commerce website. Third, their procedure for session labeling resulted in a significant percentage of unlabeled sessions (20-36%, depending on a dataset), which were therefore excluded from the analysis.

To address the aforementioned shortcomings, we implemented the offline bot recognition method based on resource request patterns and carried out an experimental analysis of its performance for multiple e-commerce sites. We used actual log data for three online stores (data had been recorded in December 2015, January 2016, and May 2016) from various domains of the Web. The analyzed stores offered various types of products/services online and applied various online marketing techniques to attract potential customers. They were also differentiated in the software used to implement the store, the website structure, and consequently, the type of server resources accessed. Furthermore, they were differentiated in terms of the website popularity, the Web traffic intensity, and the share of bot requests in the overall traffic. This allows us to generalize results of our experimental analysis to some extent.

The main contribution of the paper is:

- the experimental analysis of resource request patterns on various e-commerce sites,
- a proposal and verification of a broad heuristics for bot session labeling.

The rest of the paper is organized as follows. The next Section discusses the experimental methodology, including the implemented classification approach, our procedures for session labeling, and measures applied to evaluate classification results. Then we describe our datasets, basic statistics on the reconstructed sessions and distributions of resource requests. Afterwards we discuss classification results obtained using both labeling procedures. The last Section concludes the paper.

METHODOLOGY

Basic Concepts

We consider Web traffic incoming and being processed on a typical Web server according to the HTTP protocol. Typically, HTTP requests are issued by Web browsers used by human users to access consecutive pages of a website. For each page demand a browser generates a sequence of requests, in which a page request is followed by hits for objects embedded in the page (images, text documents, compressed files, etc.). Some part of the incoming traffic, however, is generated not by human users' browsers but by Web robots. A sequence of bot's requests in session does not reflect the logical structure of the site but it depends solely on the algorithm underlying the bot implementation.

A *session* is defined as a sequence of HTTP requests (containing more than one request) received from a Web client (identified with an IP address and a user agent field combined), assuming that the time between any two consecutive requests does not exceed 30 minutes. Each request corresponds to a specific server resource, uniquely identified by its URI (Unified Resource Identifier). Each resource may be assigned to some resource type depending on the file extension in URI. We followed the partition of resource types proposed by Doran and Gokhale (2016), except the type corresponding to requests for directory contents, because such requests are not typical for e-commerce sites. Eight resource types were distinguished:

- *text* – text-formatted files (e.g., txt, xml, sty),
- *web* – page files and scripts (e.g., html, php, cgi),
- *img* – graphic files (e.g., jpg, png, tiff),
- *doc* – rich-text documents (e.g., doc, pdf, dvi),
- *av* – multimedia files (e.g., avi, mp3, mpg),
- *prog* – program files (e.g., exe, dll, dat),
- *compressed* (*zip*) – compressed files (e.g., zip, gzip, rar, 7z),
- *malformed* – malformed requests or unknown file extensions.

Thus, each session being a sequence of requests may be represented as a sequence of resource types, i.e., a *resource request pattern*.

Session Classification Based on Resource Request Patterns

We apply the offline classification approach as it was proposed in (Doran and Gokhale 2016). In this approach resource request patterns in sessions are represented as a model based on a first-order discrete time Markov chain (DTMC). A DTMC model is a tuple (**s**, **P**). **s** is a vector of probabilities of starting a session at individual resource types so that the i^{th} element of vector **s** is the probability that a session starts at resource type i. **P** is a matrix of transition probabilities between the resource types in session so that $\boldsymbol{p}_{i,j}$ at the position (i, j) is the probability that after the request for resource type i there will occur a request for resource type j. A DTCM model is trained based on resource request patterns of sessions allocated to a training set.

Let $\boldsymbol{S} = (x^1, x^2, \ldots, x^a)$ be the resource request pattern of a new session containing a requests. The log-probability that a DTMC will generate $\boldsymbol{S}$ is given by the formula:

$$\log \boldsymbol{Pr}(\boldsymbol{S}|\mathbf{s}, \mathbf{P}) = \log \mathbf{s}_{x^1} + \sum_{i=2}^{a} \log p_{x^{i-1},x^i} \qquad (1)$$

For each session class (robot or human) a separate DTMC model is trained. Let $\mathbb{R} = (\mathbf{s}_r, \mathbf{P}_r)$ be the DTMC trained with robot sessions and $\mathbb{H} = (\mathbf{s}_h, \mathbf{P}_h)$ the DTMC trained with human sessions. A class of a new session $\boldsymbol{S}$ is determined using (1) according to the following rule: if $\log \boldsymbol{Pr}(\boldsymbol{S}|\mathbb{R}) > \log \boldsymbol{Pr}(\boldsymbol{S}|\mathbb{H})$, $\boldsymbol{S}$ is classified as a robot and as a human otherwise.

Session Labeling

In order to evaluate the efficiency of a classification approach, sessions have to be labeled as driven by a bot or a human. We labeled the sessions using the following heuristic approach.

A base for session labeling was an online database of user agent fields and IP addresses known to correspond to various kinds of robots and Web browsers. Since a database used to session labeling by Doran and Gokhale (2016) has not been available anymore, we used the *Udger* database (Udger 2017). It contained 2,832 known user agent fields and 996,657 known IP addresses of Web robots, as well as 843 user agent fields of known browsers. Besides, we used an older, freely available database of *User agents* (User-agents 2014). In our basic labeling procedure (*labeling procedure 1*) a session of a given Web client was marked as a robot when:

- the client's user agent was found in the *Udger* database and the corresponding class was "crawler", "validator", "e-mail client", "library", "multimedia player", or "offline browser" or
- the client's user agent was found in the *User-agents* database and it was labeled as robot or
- the client's user agent contained a keyword suggesting a robot or
- the client's IP address was found in the *Udger* database and the corresponding class was "crawler", "known attack source - mail", "fake crawler", "known attack source - http", or "known attack source - ssh" or
- the file "robots.txt" was requested in session.

Otherwise, the session was marked as a human when the client's user agent was found in the *Udger* database and the corresponding class was "browser" or "mobile browser". Sessions that were not marked as a robot or a human, were labeled as unknown and were excluded from the experimental analysis.

Since in reality many robots are impersonators retaining legitimate user agent fields (Zeifman 2017), we decided to broaden our heuristics and to additionally label some sessions as robots based on some session features that are untypical for humans. In the broader labeling procedure (*labeling procedure 2*) a session was additionally marked as a robot ("probable bot") when:

- the image to page ratio in session was zero or
- all page requests in session had empty referrer fields or
- all responses in sessions were 4xx or
- all requests were of type HEAD.

Applying the two labeling procedures resulted in two scenarios in our experimental analysis, referred to as scenario 1 and scenario 2 for the labeling procedure 1 and 2, respectively.

Experimental Setup

Log file parsing, data preprocessing, and session reconstruction was done for each of the three online stores. For each store session labeling was performed using the labeling procedures 1 and 2, leading to two session datasets, used in the corresponding two experimental scenarios. The experimental analysis was conducted separately for each store and for each scenario. To evaluate the efficiency of resource request patterns-based classification for each scenario, a five-fold cross-validation was applied. A given session dataset was partitioned into five subsets so that each subset contained the same number of bot sessions and the same number of the human ones. In each round of cross-validation another subset was a testing set whereas the remaining four subsets aggregated made a training set. Finally, classification results (i.e., performance measures) were averaged over the five rounds.

Performance Measures

Three standard measures were adopted to evaluate the classification efficiency: recall, precision, and F1. *Recall* is the number of correctly classified robot sessions divided by the number of all robot sessions – thus, it reflects the percentage of correctly classified bot sessions. *Precision* is the number of correctly classified robot sessions divided by the number of all sessions classified (correctly or incorrectly) as robots – thus, this measure reflects the scale of wrong classification of human sessions as robots. Finally, F1 summarizes precision and recall into a single value as their harmonic mean – thus, it reflects the overall quality of the classifier. Moreover, we visualized matrices of transition probabilities between the resource types in sessions of both classes (without the cross-validation) to illustrate and assess the strength of differences in robot and human resource request patterns.

RESULTS AND DISCUSSION

Dataset Description

In the experimental study we used access log data for three various e-commerce websites from different domains over the Internet:

1) *Auto* – the online store offering car parts and accessories,
2) *Books* – the site of a publishing house with its own online bookstore, offering books, films, and multimedia,
3) *Elderly* – the online store offering products and services for elderly people.

The data covered different time spans in 2015 and 2016. The intensity of session arrivals was very differentiated across the websites: 148, 921, and 2,154 sessions per day had arrived on average at *Auto*, *Books*, and *Elderly* websites, respectively (Tab. 1). Although these three websites may be not representative of all e-commerce sites, they differ in many aspects – especially in resources request patterns.

Tab. 2 and Tab. 3 summarize the number of robot, human, and unknown sessions determined by using the labeling procedure 1 (without "probable bots") and 2 (with "probable bots"), respectively. One can notice that

regardless of the labeling procedure used, the proportion of robot traffic is very differentiated across the websites.

Table 1: Basic Information on the E-Commerce Datasets Used in the Experiments

	Auto	***Books***	***Elderly***
Time of data collection	Oct 1-Dec 27, 2015	May 1-31, 2016	Jan 1-17, 2016
No. of sessions	13,012	28,565	36,618
Avg. no. of sessions per day	147.9	921.4	2,154.0

When we consider only known bots, identified with a user agent field or an IP address (Tab. 2), bot sessions constitute only 7% of all sessions on *Elderly* site compared to 22% on *Books* site and as much as 71% on *Auto* site. However, when bots are additionally identified based on some specific session features (Tab. 3), the share of their sessions grows a bit for *Auto* (by 9%) and for *Books* (by 10%) and increases really drastically for *Elderly* (from 7 to as much as 86%, i.e. by 79%). This increase is due only to a small extent to sessions in which all requests are of type HEAD or all responses are 4xx – the main reason is a huge number of sessions in which all page requests have empty referrers and sessions with zero image to page ratio.

Table 2: Number of Human, Robot, and Unknown Sessions Identified by the Labeling Procedure 1 (Scenario 1 – without "Probable Bots")

	Auto	***Books***	***Elderly***
Robot	9,256 (**71.1%**)	6,432 (**22.5%**)	2,671 (**7.3%**)
Human	3,644 (28.0%)	21,339 (74.7%)	33,643 (91.9%)
Unknown	112	3	5

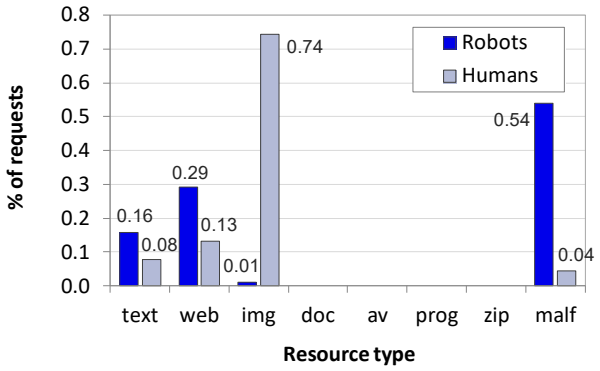
Table 3: Number of Human, Robot, and Unknown Sessions Identified by the Labeling Procedure 2 (Scenario 2 – with "Probable Bots")

	Auto	***Books***	***Elderly***
Robot	10,399 (**79.9%**)	9,086 (**31.8%**)	31,487 (**86.0%**)
Human	2,613 (20.1%)	19,476 (68.2%)	5,126 (14.0%)
Unknown	0	3	5

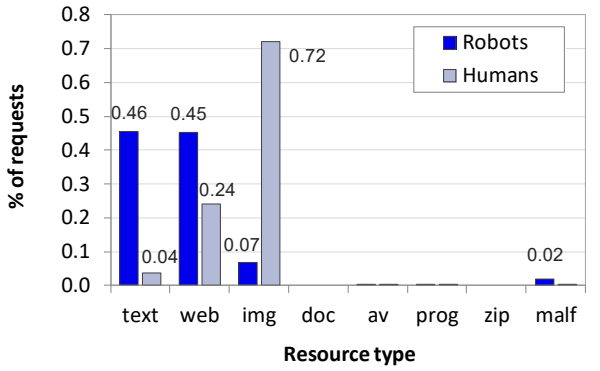
Resource Request Distribution

Fig. 1 presents distributions of resource request types on the analyzed websites for robots and humans. Only distributions for scenario 2 are shown due to the space limit. The corresponding results for scenario 1 lead to the same conclusions. The most clear differences were in bot sessions (1) for *Books*: shares of text and web requests 0.68 and 0.19 instead of 0.46 and 0.45, respectively, and (2) for *Elderly*: shares of web and malformed requests 0.86 and 0.08 instead of 0.79 and 0.15, respectively.

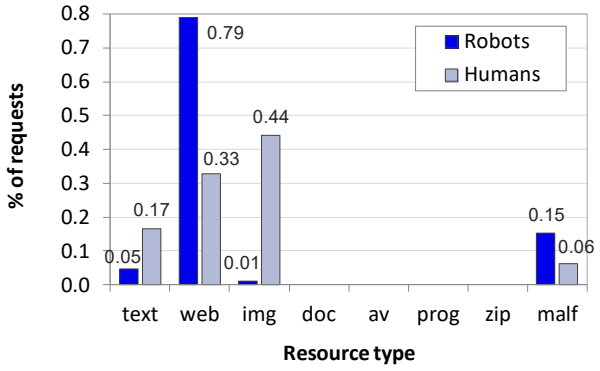
The main observation from Fig. 1 is that in human sessions similar resource types are requested across the websites (there are similar shares of requests of type img, web, text, and malformed) whereas for robot sessions access patterns differ across the websites (bot traffic on *Auto* site is dominated by malformed request, on *Elderly* site – by web request, and on *Books* site – by both text and web requests). A common feature of bot sessions is a very small share of image requests. Another observation is that on the analyzed e-commerce sites requests regarding resource type doc, av, prog, and compressed are extremely rare. This observation is consistent with results obtained in (Doran and Gokhale 2016) for the e-commerce site (this case of their experiments is referred to in our paper as "Ref_EC").



(a) *Auto*



(b) *Books*

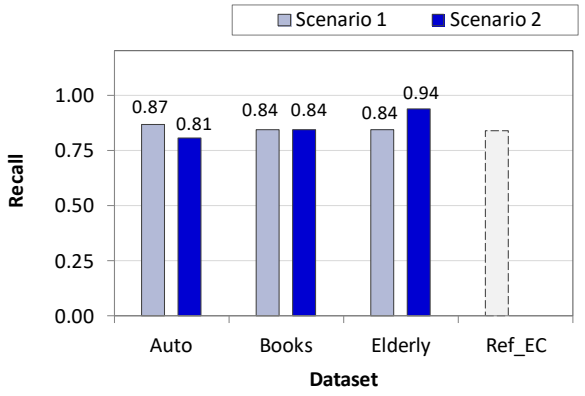


(c) *Elderly*

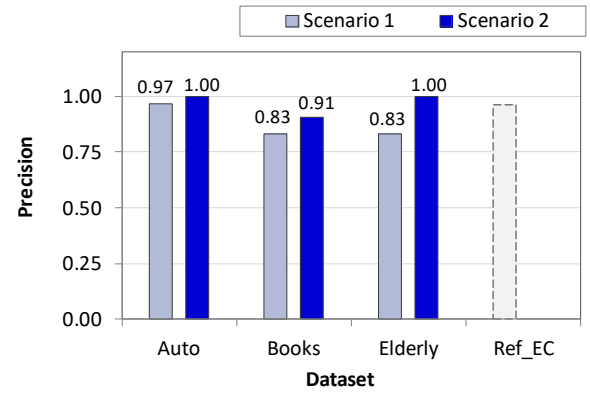
Figure 1: Distribution of Resource Requests for the Three E-Commerce Websites Broken Down by Session Class (Scenario 2).

Classification Results

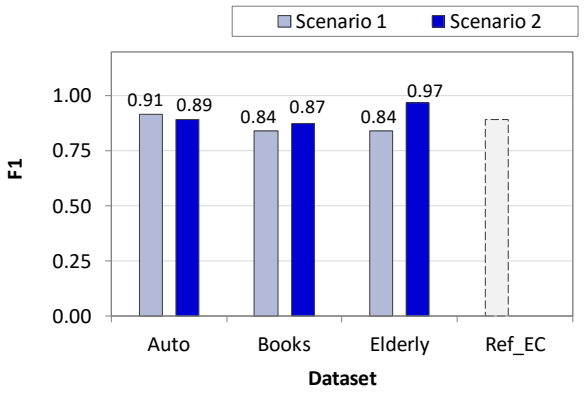
The experiments resulted in very high values of all three performance measures for both scenarios across all the datasets (Fig. 2). Recall is in the range of 0.84 to 0.87 for scenario 1 and 0.81 to 0.94 for scenario 2. This result is comparable with recall for "Ref_EC" (Fig. 2a).



(a) Recall



(b) Precision



(c) F1 measure

Figure 2: Classification Performance Compared to the Performance for the E-Commerce Dataset ("Ref_EC") Reported in (Doran and Gokhale 2016)

The tested approach is especially effective in terms of precision for scenario 2 (precision is in the range of 0.83 to 0.97 for scenario 1 and 0.91 to 1 for scenario 2). The high precision corresponds to the low false positive rate – this means that very few humans are wrongly classified as bots. This finding is also consistent with the results of the original experimental study reported for the e-commerce site ("Ref_EC").

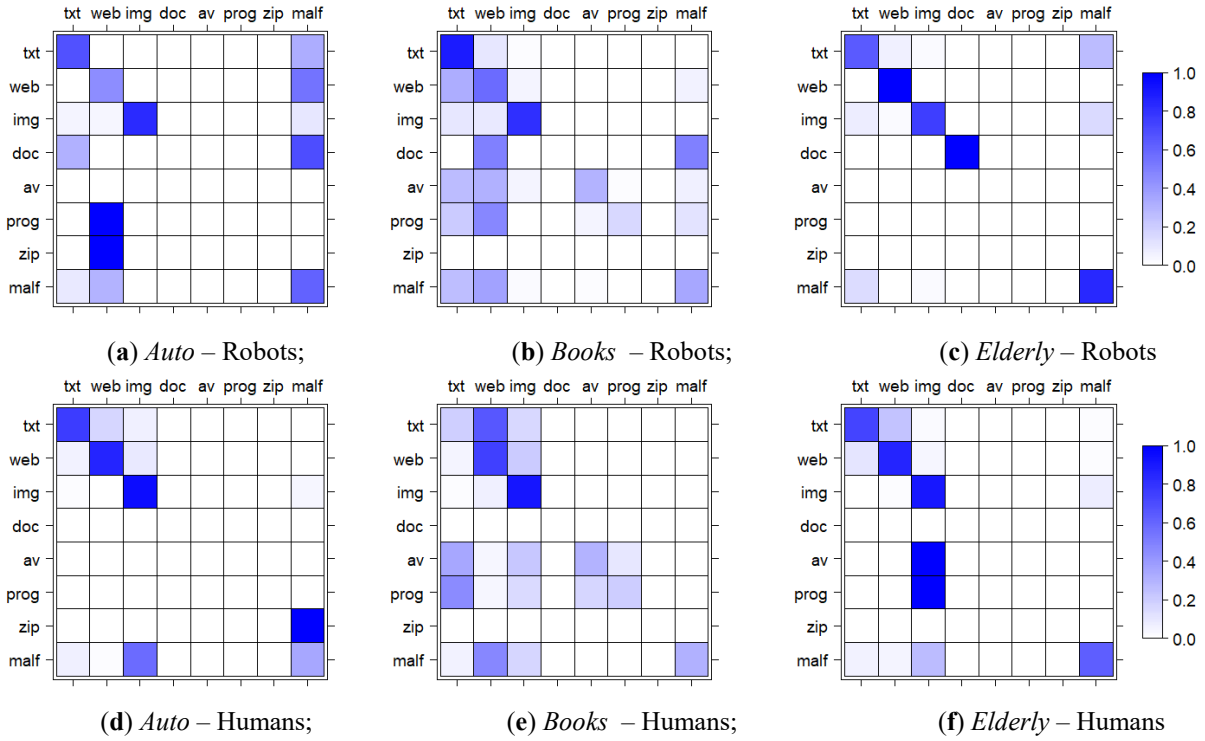
The resulting recall values do not allow us to clearly state that one of our labeling procedures is better than the other regarding classification capabilities of the tested approach. However, the resulting precision values for all three datasets demonstrate an evident superiority of labeling procedure 2. The overall quality of the classifier (F1 measure) also tends to favor the broader session labeling procedure.

This conclusion is confirmed by the visual inspection of transition probabilities between the resource types in sessions of both classes for scenario 1 (Fig. 3) and 2 (Fig 4). Transition matrices, computed for all sessions of a given class, are presented as level plots. The higher transition probability between two request types in the matrix **P** is, the darker color the corresponding cell has on the level plot. The most significant changes between scenario 1 and 2 are marked in Fig. 4 with black ovals.

A general look at the figures and the comparison of the plots in a vertical dimension (*Auto* vs. *Books* vs. *Elderly*) allows one to observe that each website has its own specific resource request patterns. Similarly, the comparison of the plots in a horizontal dimension (robots vs. humans) indicate clear distinctions between resource request patterns for both session classes at the same website. *Books* website is characterized by much larger diversity in the types of resource requested than other two websites – it should be emphasized, however, that the total numbers of requests for types doc, av, prog, and compressed are very low (cf. Fig. 1b).

There are some common features of robot and human patterns across the three websites. First, robots tend to issue requests of malformed type after requests of other types much more often than humans (there are more shaded cells in the rightmost columns of their matrices). This distinction is even clearer for scenario 2 (Fig. 4) than scenario 1 (Fig. 3). Second, both robot and human sessions reveal a high transition probability from type web to web and from type img to img. Third, for robot sessions transitions from type web to img almost do not happen, in contrast to human sessions. Also this feature is more evident for scenario 2.

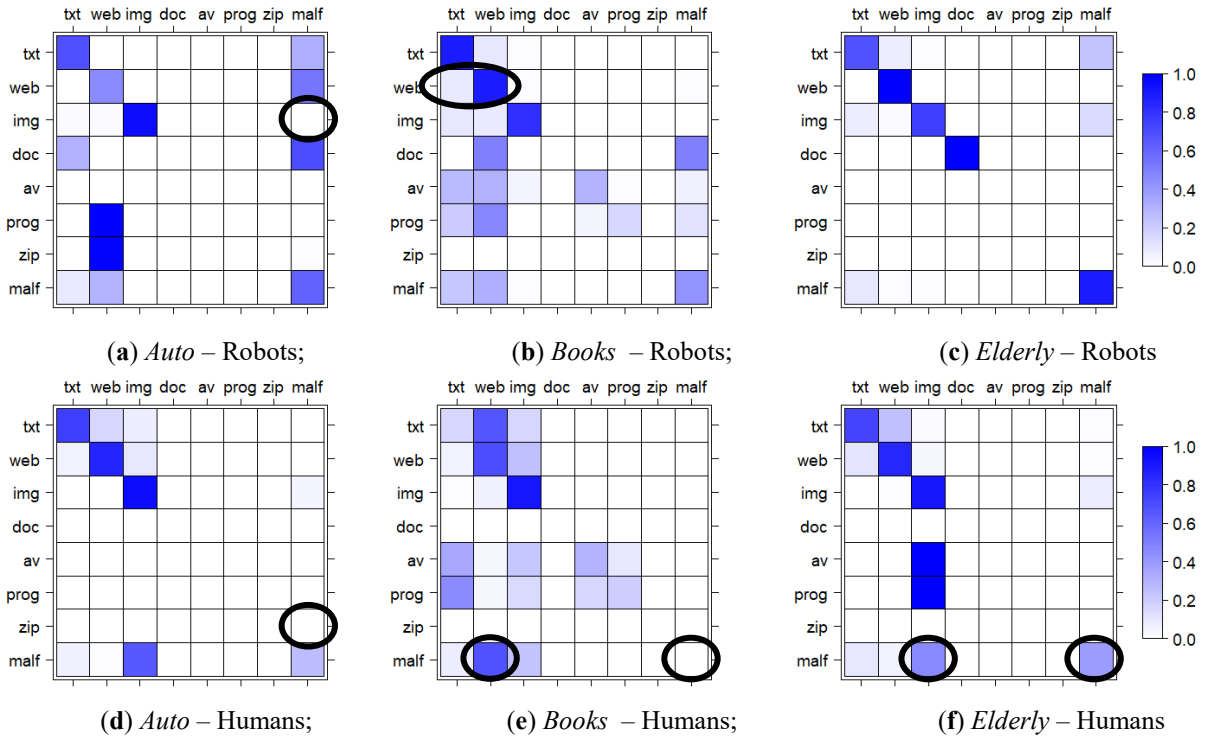
Generally, the level plots show that distinctions between resource request patterns of bots and humans are more significant when labeling procedure 2 is applied to identify bots than for procedure 1. In particular, broadening a subset of bot sessions led to the clarification of transition matrices for humans. For example, for human sessions it turned out that there are not transitions from compressed to malformed types on *Auto* website (Fig. 4d), there are not transitions from malformed to malformed types on *Books* website (Fig. 4e), and there are significantly less transitions from malformed to malformed on *Elderly* website (Fig. 4f).



(**a**) *Auto* – Robots; (**b**) *Books* – Robots; (**c**) *Elderly* – Robots

(**d**) *Auto* – Humans; (**e**) *Books* – Humans; (**f**) *Elderly* – Humans

Figure 3: Transition Probabilities Between Request Types in Session for the Three E-Commerce Websites for the Scenario 1: For Robots (Top) and Humans (Bottom).



(**a**) *Auto* – Robots; (**b**) *Books* – Robots; (**c**) *Elderly* – Robots

(**d**) *Auto* – Humans; (**e**) *Books* – Humans; (**f**) *Elderly* – Humans

Figure 4: Transition Probabilities Between Request Types in Session for the Three E-Commerce Websites for the Scenario 2: For Robots (Top) and Humans (Bottom).

CONCLUSIONS

In the study discussed in this paper we experimentally evaluated the efficiency of the session classification approach known from the literature, which uses differences in resource request patterns of bots and humans. We used real actual log data from three e-commerce sites from different domains of the Web. We proposed and implemented two heuristic procedures of labeling bots' and humans' sessions: (1) the labeling procedure based solely on known user agent fields and IP addresses and (2) the procedure which additionally uses intuition regarding bot session features untypical for human sessions. The broader session labeling procedure allowed us to label almost all the sessions in the analyzed e-commerce datasets, in contrast to other labeling approaches reported in the literature. Thus, we were able to verify the efficiency of the classification approach for almost all the sessions on the analyzed websites.

Moreover, differences between resource request patterns of bots and humans are more clear when the labeling procedure 2 is used. The general conclusion is that for the purpose of labeling sessions as bots or humans it is worth including session features untypical for humans, like the image to page ratio equal to zero, all page requests with empty referrers, all responses erroneous, and all requests of type HEAD.

Our analysis showed that bot and human sessions reconstructed from HTTP log data reveal significant differences in resource request patterns. Our findings confirmed high performance of the tested classification approach across multiple e-commerce websites from multiple domains, especially in terms of precision and F1 measure. Furthermore, our results indirectly confirmed that in reality many bots are camouflaged and use legitimate names of known Web browsers in their user agent string fields.

As part of future work we will address the problem of online Web bot detection and investigate the efficiency of resource request patterns in recognizing bot sessions in real time. In terms of the practical application it would be worth striving for elimination of false positives and increase in recall.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Centre (NCN) in Poland under Grant No. DEC-2017/01/X/ST6/01070.

REFERENCES

Almeida, V.; D. Menascé; R. Riedi; F. Peligrinelli; R. Fonseca; and W. Meira Jr. 2001. "Analyzing Robot Behavior in E-Business Sites". In *Proc. of ACM SIGMETRICS* (Cambridge, Massachusetts, USA, Jun.16-20). ACM, New York, NY, USA, 338-339.

Balla, A; A. Stassopoulou; and M.D. Dikaiakos. 2011. Real-Time Web Crawler Detection. In *Proc. of the 18th* ICT'11 (Ayia Napa, Cyprus, May 8-11). IEEE, Piscataway, N.J.

Doran, D. and S.S. Gokhale. 2011. "Web Robot Detection Techniques: Overview and Limitations." *Data Mining and Knowledge Discovery* 22, No. 1-2, 183–210.

Doran, D. and S.S. Gokhale. 2016. "An Integrated Method for Real Time and Offline Web Robot Detection." *Expert Systems* 33, No. 6, 592-606.

Doran, D.; K. Morillo; and S.S. Gokhale. 2013. "A Comparison of Web Robot and Human Requests". In *Proc. of the IEEE/ACM* ASONAM'13 (Niagara, ON, Canada, Aug.25-29). IEEE, Piscataway, N.J., 1374-1380.

Lee, J.; S. Cha; D. Lee and H. Lee. 2009. "Classification of Web Robots: An Empirical Study Based on Over One Billion Requests." *Computers & Security* 28, No.8, 795-802.

Saputra, C.H.; E. Adi; and S. Revina. 2013. "Comparison of Classification Algorithms to Tell Bots and Humans Apart." *Journal of Next Generation Information Technology* 4, No.7, 23-32.

Stassopoulou, A. and M.D. Dikaiakos. 2009. "Web Robot Detection: A Probabilistic Reasoning Approach." *Computer Networks* 53, No.3, 265-278.

Stevanovic, D.; A. An.; and N. Vlajic. 2011. "Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers." *Foundations of Intelligent Systems*. ISMIS 2011, LNCS 6804.

Suchacka, G. 2014. "Analysis of Aggregated Bot and Human Traffic on E-Commerce Site." In *Proc. of* FedCSIS'14 (Warsaw, Poland, Sep.7-10), ACSIS, Vol. 2. IEEE, Piscataway, N.J., 1123-1130.

Suchacka, G. and M. Sobków. 2015. "Detection of Internet Robots using a Bayesian Approach". In *Proc. of* CYBCONF'15 (Gdynia, Poland, Jun.24-26). IEEE, Piscataway, N.J., 365-370.

Udger. 2017. https://udger.com (access: September 4, 2017).

User-agents. 2014. http://www.user-agents.org (access: September 4, 2017).

Zeifman, I. 2017. "Bot Traffic Report 2016". Imperva Incapsula, https://www.incapsula.com/blog/bot-traffic-report-2016.html.

AUTHOR BIOGRAPHIES

GRAŻYNA SUCHACKA received the M.Sc. degrees in Computer Science and in Management, as well as the Ph.D. degree in Computer Science from Wroclaw University of Technology, Poland. Now she is an assistant professor in the Institute of Mathematics and Informatics at Opole University, Poland. Her research interests include analysis and modeling of Web traffic, Web mining, and Quality of Service with special regard to electronic commerce support and Web bot detection. Her e-mail address is: `gsuchacka@uni.opole.pl`.

IGOR MOTYKA is a student of Computer Science at the Faculty of Mathematics, Physics and Computer Science at Opole University, Poland. He is passionate about object-oriented programming, especially in C# and Java. His current area of interest is primarily development of WPF applications and Android software. His e-mail address is: `igor_motyka@mail.com`.

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

-

Special Session

SIMULATION OF LARGE-SCALE QUEUEING SYSTEMS

Sergey Vasilyev, Galina Tsareva
Department of of Applied Probability and Informatics
RUDN University
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
Email: vasilyev_sa@rudn_university

KEYWORDS

Countable Markov chains; Large-scale queueing systems; Dobrushin approach; Singular perturbed systems of differential equations; Differential equations of infinite order; Small parameter;

ABSTRACT

In this paper we consider the dynamics of large-scale queueing systems with an infinite number of servers. We assume that a Poisson input flow of requests with intensity $N\lambda$. We suppose that each incoming request selects two any servers randomly and a next step of an algorithm includes sending this request to the server with the shorter queue instantly. A share $u_k(t)$ of the servers that have the queues lengths with not less than k can be described using an system of ordinary differential equations of infinite order. We investigate this system of ordinary differential equations of infinite order with a small real parameter. A small real parameter allows us to describe the processes of rapid changes in large-scale queueing systems. We use the simulation methods for this large-scale queueing systems analysis.

INTRODUCTION

The recent research of large-scale queueing systems with complex routing discipline in [16], [25], [26], [27], transport networks [1],[7], [8] and the asymptotic behavior of Jackson networks [21] faced with the problem of proving the global convergence of the solutions of certain infinite queueing systems of ordinary differential equations to a time-independent solution. Scattered results of these studies, however, allow a common approach to their justification. In work [17] the countable systems of differential equations with bounded Jacobi operators were studied and the sufficient conditions of global stability and global asymptotic stability were obtained. In [15] it was considered finite closed Jackson networks with N first come, first serve nodes and M customers. In the limit $M \to \infty$, $N \to \infty$, $M/N \to \lambda > 0$, it was got conditions when mean queue lengths are uniformly bounded and when there exists a node where the mean queue length tends to ∞ under the above limit (condensation phenomena, traffic jams), in terms of the limit distribution of the relative utilizations of the nodes. It was deriven asymptotics of the partition function and of correlation functions. In papers [5], [11], [20] the authors built various models of large-scale queueing systems and considered their dynamics.

Cauchy problems for the systems of ordinary differential equations of infinite order was investigated A.N.Tihonov [22], K.P.Persidsky [18], O.A.Zhautykov [28], [29], Ju.Korobeinik [10], A.M.Samoilenko, Yu.V.Teplinskii [19] other researchers. For example, Markus Kreer, Aye Kzlers and Anthony W. Thomas [13] investigated fractional Poisson processes, a rapidly growing area of non-Markovian stochastic processes, that are useful in statistics to describe data from counting processes when waiting times are not exponentially distributed. They showed that the fractional KolmogorovFeller equations for the probabilities at time t could be represented by an infinite linear system of ordinary differential equations of first order in a transformed time variable. These new equations resemble a linear version of the discrete coagulationfragmentation equations, well-known from the non-equilibrium theory of gelation, cluster-dynamics and phase transitions in physics and chemistry. The singular perturbed systems of ordinary differential equations with a small parameter was studied by A.N. Tihonov [23], A.B.Vasil'eva [24], S.A. Lomov [14] other researchers.

In paper [2] we investigated the singular perturbed systems of ordinary differential equations of infinite order of Tikhonov-type $\epsilon\dot{x} = F(x(t, g_x), y(t, g_y), t)$, $\dot{y} = f(x(t, g_x), y(t, g_y), t)$ with the initial conditions $x(t_0) = g_x$, $y(t_0) = g_y$, where x, $g_x \in X$, $X \subset l_1$ and y, $g_y \in Y$, $Y \in \mathbf{R^n}$, $t \in [t_0, t_1]$ $(t_0 < t_1)$, $t_0, t_1 \in T$, $T \in \mathbf{R}$, g_x and g_y are given vectors, $\epsilon > 0$ is a small real parameter.

In this paper we apply Dobrushin approaches from [26]. We consider the dynamics of large-scale queueing systems that consists of infinite number of servers with a Poisson input flow of requests of intensity $N\lambda$. We can use an algorithm that selects two any servers for each incoming request and sent it to the one of a server with the shorter queue instantly. We suppose that service time has mean $1/\mu$ with exponential distribution. In this case a share $u_k(t)$ of the servers that have the queues lengths with not less than k can be described using an system of ordinary differential equations of infinite order. We investigate this system of ordinary differential equations of infinite order with a small real parameter. A small real parameter allows us to describe the processes of rapid changes in large-scale queueing

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

systems. Tikhonov type Cauchy problem for this system with small parameter ϵ and initial conditions is investigated.

We investigate the truncation system of this ordinary differential equations of infinite order with a small real parameter order N. Tikhonov type Cauchy problem for this truncation system with small parameter ϵ and initial conditions is used for the simulation of behavior solutions and for analysis of large-scale queueing systems with taking into account parameters λ, μ, ϵ.

LARGE-SCALE QUEUEING SYSTEMS MODEL

The basic model considered there is a queueing system S_N, with N identical infinite-buffer FCFS (First-Come, First-Served) single-servers, with a Poisson arrival flow of rate $N\lambda$ and with i.i.d. exponential service times of mean $1/\mu$, where $0 < \lambda < \mu$. Upon its arrival each task chooses m servers at random (i.e., independently of the pre-history of the queueing system (QS) and with probability $1/(N^m)$) and then selects, among the chosen ones, the server with the lowest queue-size, i.e., the lowest number of tasks in the buffer (including the task in service). If there happen to be more than one server with lowest queue-size, the task selects one of them randomly.

One is interested in the 'typical' behavior of a server in S_N, as $N \to \infty$. Formally, it means that $\forall\, t \geq 0$ and $k = 0, 1, \ldots$, we consider the fraction $q_k(t) = M_k(t)/N$ where $M_k(t)$ is the (random) number of servers with the queue-size k at time t. Clearly, $0 \leq q_k(t) \leq 1$, $\sum_k q_k(t) = 1$; and $Q(t) = (q_k(t))$, $t \geq 0$, forms a Markov process (MP). Technically, it is more convenient to pass to the tail probabilities $r_k(t) = \sum_{j \geq k} Q_k(t)$; the state space of the corresponding MP $U_N(t) = (f_k(t)), t \geq 0$, is the set $\mathbf{U}_N$ of non-increasing non-negative sequences $\mathbf{u} = (u_k, k = 0, 1, \ldots)$ with $u_0 = 1$, $\sum_{k>1} u_k < \infty$ and with the u_k's multiple of $1/N$, which implies that $u_k = 0$ for all k large enough. It is convenient to prolong the sequences $\mathbf{u} \in \mathbf{U}_N$ to the negative k's by the value 1.

The generator of $\{U_N(t)\}$ is an operator $\mathbf{A}$ acting on functions $f : \mathbf{U}_N \to C^1$ and given by

$$\mathbf{A}_N f(\mathbf{u}) = N \sum_{k>0} (u_k - u_{k+1}) \left[f\left(\mathbf{u} - \frac{\mathbf{e}_k}{N} - f(\mathbf{u})\right)\right] + \\ +\lambda N \sum_{k>0} \left((u_{k-1})^2 - (u_k)^2\right) \left[f\left(\mathbf{u} + \frac{\mathbf{e}_k}{N} - f(\mathbf{u})\right)\right]. \quad (1)$$

Here, $\mathbf{e}_k$ stands for the sequence with the k-th entry 1 and all others 0, the addition of the sequences is componentwise. Process $\{U_N(t)\}$ is positive-recurrent and thus possess a unique invariant distribution, π_N; given any initial distribution ϖ, the distribution of $U_N(t)$ approaches π_N as $t \to \infty$. The main result of [25] is that, as $N \to \infty$, the expected value $\mathbf{E}_{\pi_N} r_k(t)$ converges to the value $\{a_k\}$, where

$$a_k = \left(\frac{\lambda}{\mu}\right)^{(m^k-1)/(m-1)}, \; k \geq 0. \quad (2)$$

Pictorially speaking, it means that, as $N \to \infty$, an 'average' server in the QS will have k or more tasks in the buffer with probability a_k.

It is interesting to compare $\mathbf{S}_N$ with another queueing system $\mathbf{L}$, where the arriving task chooses the server completely randomly (i.e., independently of the pre-history and with probability $1/N$). Clearly, $\mathbf{L}$ is equivalent to an isolated $M/M/\infty$ queue with the arrival and service rates λ and μ, respectively, which justifies omitting subscript N in this notation. More precisely, the average server in $\mathbf{L}$ will have k or more tasks in the buffer with the geometrical probability

$$a_k^0 = \left(\frac{\lambda}{\mu}\right)^k \; k \geq 1, \quad (3)$$

(independently of N), which is much larger than a_k.

In fact, as was shown in [25], the whole process $\{U_N(t)\}$ is asymptotically deterministic as $N \to \infty$. More precisely, let $\mathbf{U}$ denote the set of the non-increasing non-negative sequences $\mathbf{u} = (u_k, k \in \mathbf{Z})$ with $u_k = 1$ for $k \leq 0$ and $\sum_{k \leq 0} u_k < \infty$.Then, if the distribution ϖ of initial state $U_N(0)$ approaches a Dirac delta-measure concentrated at a point $\mathbf{g} = \{g_k\} \in \mathbf{U}$, the distribution of $\{U_N(t)\}$ is concentrated in the limit at the 'trajectory' $\mathbf{u}(t) = u_k(t)$, $t \geq 0$, giving the solution to the following system of differential equations

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \\ +\lambda\left((u_{k-1}(t))^2 - (u_k(t))^2\right), \quad (4)$$

$$u_0(t) = 0, u_k(0) = g_k \geq 0, \; k = 1, 2, \ldots, \; t \geq 0,$$

where $g = \{g_k\}_{k=1}^{\infty}$ is a numerical sequence ($1 = g_1$, $g_k \geq g_{k+1}$, $k = 1, 2, \ldots$) [25]. Point $\mathbf{a} = (a_k)$ (see (2)) is a (unique) fixed point for system (4) in $\mathbf{U}$.

These results illustrate the essence of the mean-field approximation for QS S_N. Equations (4) describe a 'self-compatible' evolution of vector $\mathbf{u}(t)$, or, equivalently, of the probability distribution $\mathbf{q}(t) = \{q_k(t)\}$ defined by $q_k(t) = u_k(t) - u_{k+1}(t)$, $t \geq 0$, $k = 0, 1, \ldots$ As before, $\mathbf{u}(t)$ is simply the sequence of the tail probabilities for $\mathbf{q}(t)$.

We can compare system (4) with the linear system

$$\dot{y}_k(t) = \mu\left(y_{k+1}(t) - y_k(t)\right) + \lambda\left(y_{k-1}(t) - y_k(t)\right), \quad (5)$$

(where $k \geq 1$) describing the evolution of the probability distribution $\mathbf{q}^0(t) = (q_k^0(t), q_k^0(t) = y_k(t) - y_{k+1}(t))$ in a standard $M/M/1/\infty$ queue with the arrival and service rates λ and μ, respectively. The μ-terms in (4) and (5) are the same; they correspond with the departure of the tasks and 'push' the probability mass in $\mathbf{q}(t)$ and $\mathbf{q}^{(0)}(t)$ towards $k = 0$. On the other hand, the λ-terms (different in both SQ) correspond with the arrival of the tasks; these terms shift the probability mass to larger k's. The λ-term in (4) is smaller than the one in (5) when $u_k(t)$ is small; pictorially speaking, system (4) provides (for the same values of λ and μ) more 'protection', for large k, against the shift to the right, which may lead to an 'explosion', when the relation $\sum_{k>1} u_k(t) < \infty$ or $\sum_{k>1} y_k(t) < \infty$ may fail as

$t \to \infty$. Because of this, the entries a_k of sequence $\mathbf{a}$ (see (2)) giving the fixed point of (4) decrease 'super-exponentially', in contrast with the exponential decay of the tail probabilities in the fixed point $\mathbf{a}^0 = (a_k^0)$ of (5).

LARGE-SCALE QUEUEING SYSTEMS MODEL WITH A SMALL PARAMETER

Let's consider a system that consists of N servers with a Poisson input flow of requests of intensity $N\lambda$. Each request arriving to the system randomly selects two servers and is instantly sent to the one with the shorter queue. The service time is distributed exponentially with mean $\bar{t} = 1/\mu$. Let $u_k(t)$ be a share servers that have the queues lengths with not less than k. It is possible to investigate the asymptotic distribution of the queue lengths as $N \to \infty$ and $\lambda < 1$ [25]. The considered system of the servers is described by ergodic Markov chain. There is a stationary probability distribution for the states of the system and if $N \to \infty$ the evolution of the values $u_k(t)$ becomes deterministic and the Markov chain asymptotically converges to a dynamic system the evolution of which is described by system of ordinary differential equations of infinite order

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \\ +\lambda\left((u_{k-1}(t))^2 - (u_k(t))^2\right). \quad (6)$$

For this system of ordinary differential equations of infinite order we can formulate Cauchy problem in the form

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \\ +\lambda\left((u_{k-1}(t))^2 - (u_k(t))^2\right), \quad (7)$$
$$u_0(t) = 0,\ u_k(0) = g_k \geq 0,\ k = 1, 2, \ldots,\ t \geq 0,$$

where $g = \{g_k\}_{k=1}^{\infty}$ is a numerical sequence ($1 = g_1$, $g_k \geq g_{k+1}$, $k = 1, 2, \ldots$) [25].

We can investigate Cauchy problem for system of ordinary differential equations of infinite order with small parameter such form

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \lambda\left((u_{k-1}(t))^2 -,\right.$$
$$\left. -(u_k(t))^2,\ k = 0, 1, \ldots, n-1, \right. \quad (8)$$
$$\dot{u}_n(t) = \mu\left(U_{n+1}(t) - u_n(t)\right) + \lambda\left((u_{n-1}(t))^2 - (u_n(t))^2\right),$$
$$\epsilon\dot{U}_k(t) = \mu\left(U_{k+1}(t) - U_k(t)\right) + \lambda\left((U_{k-1}(t))^2 -\right.$$
$$\left. -(U_k(t))^2,\ k = n+1, n+1, \ldots,\right.$$
$$u_k(0) = g_k \geq 0,\ k = 0, 1, 2, \ldots, n,$$
$$U_k(0) = g_k \geq 0,\ k = n+1, \ldots,$$

where $\epsilon > 0$ is a small parameter that bring a singular perturbation to the system (7), which allows us to describe the processes of rapid change of the systems.

Using (8) we can write Tikhonov problems for systems of ordinary differential equations of infinite order with a small parameter ϵ and initial conditions

$$\dot{u} = f(u(t, \mu, \lambda, g_u), U(t, \mu, \lambda, g_U), t),$$
$$\epsilon\dot{U} = F(U(t, \mu, \lambda, g_U), t); \quad (9)$$
$$u(0, \mu, \lambda, g_u) = g_u,\ U(0, \mu, \lambda, g_U) = g_U,$$

where $u, f \in X$, $X \in \mathbf{R^n}$ are n-dimensional functions; $U, F \in Y$, $Y \subset l_1$ are infinite-dimensional functions and $t \in [0, T_0]$ ($0 < T_0 \leq \infty$), $t \in T$, $T \in \mathbf{R}$; $g_u \in X$ and $g_U \in Y$ are given vectors ($g_u = \{g_k\}_{k=0}^{n}, g_U = \{g_k\}_{k=n+1}^{\infty}$), $\epsilon > 0$ is a small real parameter; $u(t, g_u) = g_u = \{u_k\}_{k=0}^{n}$ and $U(t, g_U) = \{u_k\}_{k=n+1}^{\infty}$ are solutions of (9). Given functions $f(u(t, \mu, \lambda, g_u), U(t, \mu, \lambda, g_U), t)$ and $F(U(t, \mu, \lambda, g_U), t)$ are continuous functions for all variables

$$f_k(u(t, \mu, \lambda, g_u), t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \\ +\lambda\left((u_{k-1}(t))^2 - (u_k(t))^2\right), k = 0, 1, \ldots, n-1,$$
$$f_n(u(t, \mu, \lambda, g_u), U(t, \mu, \lambda, g_U), t) = \\ \epsilon\left(U_{n+1}(t) - u_n(t)\right) + \lambda\left((u_{n-1}(t))^2 - (u_n(t))^2\right), \quad (10)$$
$$F_k(U(t, \mu, \lambda, g_U), t) = \epsilon\left(U_{k+1}(t) - U_k(t)\right) + \\ +\lambda\left((U_{k-1}(t))^2 - (U_k(t))^2\right), k = n+1, n+2, \ldots$$

Let S is an integral manifold of the system (9) in $X \times Y \times T$. If any point $t^* \in [0, T_0]$ $(u(t^*), U(t^*), t^*) \in S$ of trajectory of this system has at least one common point on S this trajectory $(u(t, G), U(t, g), t) \in S$ belongs the integral manifold S totally.

If we assume in (9) that $\epsilon = 0$ than we have a degenerate system of the ordinary differential equations and a problem of singular perturbations

$$\dot{u} = f(u(t, \mu, \lambda, g_u), U(t), t),$$
$$0 = F(u(t, \mu, \lambda, g_u), U(t, \mu, \lambda), t); \quad (11)$$
$$u(0, \mu, \lambda, g_u) = g_u,$$

where the dimension of this system is less than the dimension of the system (9), since the relations $F(u(t, \mu, \lambda), U(t, \mu, \lambda), \lambda, t) = 0$ in the system (11) are the algebraic equations (not differential equations). Thus for the system (10) we can use limited number of the initial conditions then for system (9). Most natural for this case we can use the initial conditions $u(0, \mu, \lambda, g_u) = g_u$ for the system (11) and the initial conditions $U(0, \mu, \lambda, U_y) = g_U$ disregard otherwise we get the overdefined system. We can solve the system (11) if the equation $F(u(t, \mu, \lambda), U(t, \mu, \lambda), \mu, \lambda, t) = 0$ has roots. If it is possible to solve we can find a finite set or countable set of the roots $U_q(t, \mu, \lambda, g_u) = u_q(u(t, \mu, \lambda, g_u), t)$ where $q \in \mathbf{N}$. If the implicit function $F(u(t, \mu, \lambda), U(t, \mu, \lambda), \mu, \lambda, t) = 0$ has not simple structure we must investigate the question about the choice of roots. Hence we can use the roots $U_q(t, \mu, \lambda, g_u) = u_q(u(t, \mu, \lambda, g_u), t)$ ($q \in \mathbf{N}$) in (11) and solve the degenerate system

$$\dot{u_d} = f(u_d(t, \mu, \lambda, g_u), u_q(u_d(t, \mu, \lambda, g_u), t), \lambda, t); \quad (12)$$
$$U_d(0, \mu, \lambda, g_u) = g_u.$$

Since it is not assumed that the roots $U_q(t, \mu, \lambda, g_u) = u_q(u(t, \mu, \lambda, g_u), \mu, \lambda, t)$ satisfy the initial conditions of

the Cauchy problem (9) ($U_q(0) \neq g_u$, $q \in \mathbf{N}$), the solutions $U(t, \mu, \lambda, g_U)$ (9) and $U_q(t, \mu, \lambda, g_u)$ do not close to each other at the initial moments of time $t > 0$. Also there is a very interesting question about behaviors of the solutions $u(t, \mu, \lambda, g_u)$ of the singular perturbed problem (9) and the solutions $u_d(t, \mu, \lambda, g_u)$ of the degenerate problem (11). When $t = 0$ we have $u(0, \mu, \lambda, g_u) = u_d(0, \mu, \lambda, g_u)$. Do these solutions close to each other when $t \in (0, T_0]$? The answer to this question depends on using roots $U_q(t, \mu, \lambda, g_u) = u_q(u(t, \mu, \lambda, g_u), t)$ and the initial conditions, which we apply for the systems (9) and (12).

TRUNCATION LARGE-SCALE QUEUEING MODEL AND NUMERICAL ANALYSIS

Using (8) we can rewrite system of differential equations order N in the form

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \lambda\left((u_{k-1}(t))^2 - \right.,$$
$$\left. -(u_k(t))^2,\ k = 0, 1, \ldots, n-1, \right. \qquad (13)$$
$$\dot{u}_n(t) = \mu\left(U_{n+1}(t) - u_n(t)\right) + \lambda\left((u_{n-1}(t))^2 - (u_n(t))^2\right),$$
$$\epsilon\dot{U}_k(t) = \mu\left(U_{k+1}(t) - U_k(t)\right) + \lambda\left((U_{k-1}(t))^2 - \right.$$
$$\left. -(U_k(t))^2,\ k = n+1, n+1, \ldots, N. \right.$$

For this truncation system of ordinary differential equations of order N we can formulate Cauchy problem in the form

$$\dot{u}_k(t) = \mu\left(u_{k+1}(t) - u_k(t)\right) + \lambda\left((u_{k-1}(t))^2 - \right.,$$
$$\left. -(u_k(t))^2,\ k = 0, 1, \ldots, n-1, \right. \qquad (14)$$
$$\dot{u}_n(t) = \mu\left(U_{n+1}(t) - u_n(t)\right) + \lambda\left((u_{n-1}(t))^2 - (u_n(t))^2\right),$$
$$\epsilon\dot{U}_k(t) = \mu\left(U_{k+1}(t) - U_k(t)\right) + \lambda\left((U_{k-1}(t))^2 - \right.$$
$$\left. -(U_k(t))^2,\ k = n+1, n+1, \ldots, N, \right.$$
$$u_k(0) = g_k \geq 0,\ k = 0, 1, 2, \ldots, n,$$
$$U_k(0) = g_k \geq 0,\ k = n+1, \ldots, N.$$

The numerical analysis was carried out using the adaptive step Runge-Kutta integration method, which is one of the most commonly used methods for the numerical solution of the singularly perturbed system of differential equations.

The numerical example is presented in the figure (see Fig. 1, 2) where $n = 7$, $N = 10$, $\lambda = 0.5$, $\mu = 1.1$, $g_0 = 1$ $g_k = 1 - 0.1k$, $k = \overline{0,9}$ and a small parameter $\epsilon = 0.1$ (Fig. 1) , $\epsilon = 0.01$ (Fig. 2), $\epsilon = 0.001$ (Fig. 3). In these numerical examples we can see the existence of steady state conditions for evolutions $u_i(t)$, $i = \overline{0,5}$ and quasi-periodic conditions with boundary layers for evolutions $u_i(t)$, $i = \overline{6,10}$.

The numerical simulation show that the solution of the singularly perturbed systems of differential equations have an area of rapid change of the function, which is usually located in the initial point of the problem. This area of rapid function change is called the area of the mathematical boundary layer. The thickness of the boundary layer depends on the value of a small parameter, and when the small parameter decreases, the thickness of the boundary layer decreases. The integration area is divided into external (outside the boundary layer) and internal (inside the boundary layer). The solution of the singularly perturbed equation is sought in the form of a solution suitable for the outer domain, which is then refined in the vicinity of the boundary point where the boundary layer is located. The numerical examples are shown the existence of steady state conditions for evolutions $u_i(t)$ and quasi-periodic conditions with boundary layers for evolutions $u_i(t)$.

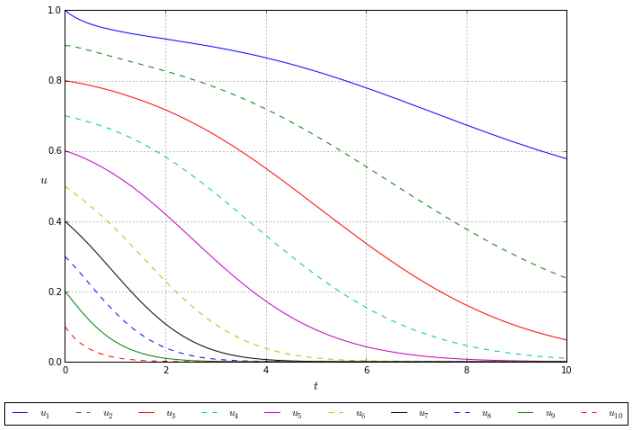


Fig. 1. Evolution analysis of u_k ($\epsilon = 0.1$).

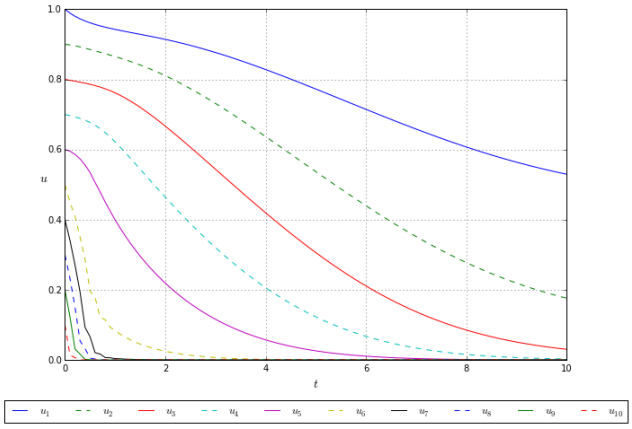


Fig. 2. Evolution analysis of u_k ($\epsilon = 0.01$).

CONCLUSIONS

We investigate the dynamics of large-scale queueing systems that consists of infinite number of servers with a Poisson input flow of requests of intensity $N\lambda$. Each request arriving to the system randomly selects two servers and this request is instantly sent to the one with the shorter queue. We suppose that service time has mean $1/\mu$ with exponential distribution. In this case a share $u_k(t)$ of the servers that have the queues lengths with not less than k can be described using an system of differential equations of infinite order. Tikhonov type Cauchy problem for this system with small parameter ϵ. Tikhonov type Cauchy problem for this system with small parameter ϵ and initial conditions is inves-

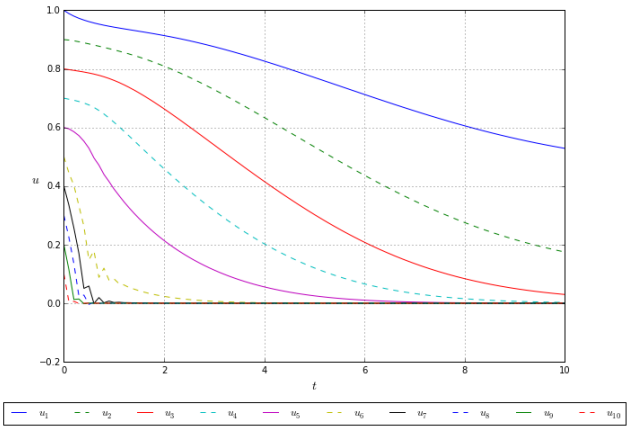


Fig. 3. Evolution analysis of u_k ($\epsilon = 0.001$).

tigated. We use the simulation methods for behavior solutions analysis with taking into account parameters λ, μ, ϵ. The numerical examples are shown the existence of steady state conditions for evolutions $u_i(t)$ and quasi-periodic conditions with boundary layers for evolutions $u_i(t)$.

REFERENCES

Acknowledgments

The publication has been prepared with the support of the RUDN University Program 5-100 and partially funded by RFBF grants No 15-07-08795, No 16-07-00556.

References

[1] Afanassieva, L.G., Fayolle, G., Popov, S. Yu. 1997. "Models for Transportation Networks". *J. Math. Science*, Vol.84, Issue 3, pp. 1092–1103.

[2] Bolotova, G.O., Vasilyev, S.A., Udin, D.N. 2016. "Systems of Differential Equations of Infinite Order with Small Parameter and Countable Markov Chains". *Distributed Computer and Communication Networks – 19th International Conference, DCCN 2016, Moscow, Russia, November 21–25, 2016. Communications in Computer and Information Science*, Vol. 678. Publisher: Springer Verlag, pp. 565–576.

[3] McDonald, D.R., Reynier, J. 2002. "A mean-field model for multiple TCP connections through a buffer implementing RED". *Performance Evaluation*, Vol. 49, Issues 14, pp. 77–97.

[4] Daletsky, Y.L., Krein, M.G. 1970. "Stability of solutions of differential equations in Banach space". oscow, Science Pub.

[5] Gaidamaka, Y., Sopin, E., Talanova, M. 2016. "Approach to the analysis of probability measures of cloud computing systems with dynamic scaling". *Communications in Computer and Information Science*, Vol. 601, pp. 121—131.

[6] Henry, D. 1981. "Geometric theory of semilinear parabolic equations. Lecture Notes in Mathematics". *Springer-Verlag*, Berlin.

[7] Khmelev, D. V., Oseledets, V.I. 1999. "Mean-field approximation for stochastic transportation network and stability of dynamical system". *Preprint No. 434 of University of Bremen.*

[8] Khmelev, D. V. 2001. "Limit theorems for nonsymmetric transportation networks". *Fundamentalnaya i Prikladnaya Matematika*, Vol. 7, no. 4, 1259-1266.

[9] Kirstein, B. M., Franken, D. E., Stoian, D. 1977. "Comparability and monotonicity of Markov processes". *Theory of probability and its applications*, Vol. 22, Issue 1, pp.43–54.

[10] Korobeinik, Ju. 1970. "Differential equations of infinite order and infinite systems of differential equations". *Izv. Akad. Nauk SSSR Ser. Mat.*, Vol. 34, pp. 881-922.

[11] Korolkov,a A.V., Eferina, E.G., Laneev, E.B., Gudkova, I.A., Sevastianov, L.A., Kulyabov, D.S. 2016. "Stochastization of one-step processes in the occupations number representation". *Proceedings - 30th European Conference on Modelling and Simulation*, ECMS 2016, pp. 698 — 704.

[12] Krasnoselsky, M.A., Zabreyko, P.P. 1984. "Geometrical methods of nonlinear analysis". *Springer-Verlag*, Berlin.

[13] Kreer, M., Kzlersu Ayseand, Anthony, W. 2014. "Thomas Fractional Poisson processes and their representation by infinite systems of ordinary differential equations". *Statistics and Probability Letters*, Vol. 84, pp. 2732.

[14] Lomov, S. A. 1968. "The construction of asymptotic solutions of certain problems with parameters". *Izv. Akad. Nauk SSSR Ser. Mat.*, Vol. 32, pp. 884913.

[15] Malyshev, V., Yakovlev, A. 1996. "Condensation in large closed Jackson networks". *Ann. Appl. Probab.*, Vol. 6, no. 1, pp. 92–115.

[16] Mitzenmacher, M. 1996. "The Power of Two Choices in Randomized Load Balancing". *PhD thesis, University of California at Berkley.*

[17] Oseledets, V. I., Khmelev, D. V. 2000. "Global stability of infinite systems of nonlinear differential equations, and nonhomogeneous countable Markov chains". *Problemy Peredachi Informatsii (Russian)*, Vol. 36 , Issue 1, pp. 60–76.

[18] Persidsky, K.P. 1948. *Izv. AN KazSSR, Ser. Mat. Mach.*, Issue 2, pp. 3–34.

[19] Samoilenko, A. M., Teplinskii, Yu. V. 2003. "Countable Systems of Dierential Equations". *Brill*, Leiden.

[20] Samouylov, K., Naumov, V., Sopin, E., Gudkova, I., Shorgin, S. 2016. "Sojourn time analysis for processor sharing loss system with unreliable server". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9845, pp. 284—297.

[21] Scherbakov, V.V. 1997. "Time scales hierarchy in large closed Jackson networks". *Preprint No. 4. Moscow: French-Russian A.M. Liapunov Institute of Moscow State University.*

[22] Tihonov, A. N. 1934. "Uber unendliche Systeme von Differentialgleichungen". *Rec. Math.*, Vol. 41, Issue 4, pp. 551–555.

[23] Tihonov, A. N. 1952. "Systems of differential equations containing small parameters in the derivatives". *Mat. Sbornik N. S.*, Vol. 31, Issue 73, pp. 575586.

[24] Vasil'eva, A. B. 1963. "Asymptotic behaviour of solutions of certain problems for ordinary non-linear differential equations with a small parameter multiplying the highest derivatives". *Uspehi Mat. Nauk.*, Vol. 18, Issie 111, no. 3 , 15-86.

[25] Vvedenskaya, N.D., Dobrushin, R.L., Kharpelevich, F.I. 1996. "Queueing system with a choice of the lesser of two queues the asymptotic approach". *Probl. inform.*, Vol. 32, Issue 1, pp.15–27.

[26] Vvedenskaya, N.D., Suhov, Yu.M. 1997. "Dobrushin's Mean-Field Approximation for a Queue with Dynamic Routing". *Markov Processes and Related Fields*, Issue 3, pp. 493–526.

[27] Vvedenskaya, N.D. 1998. "A large queueing system with message transmission along several routes". *Problemy Peredachi Informatsii*, Vol. 34, no. 2, pp. 98–108.

[28] Zhautykov, O. A. 1959. "On a countable system of differential equations with variable parameters". *Mat. Sb. (N.S.)*, Vol. 49, Issue 91, pp. 317330.

[29] Zhautykov, O. A. 1959. "Extension of the Hamilton-Jacobi theorems to an infinite canonical system of equations". *Mat. Sb. (N.S.)*, Vol. 53, Issue 95, pp. 313328.

SERGEY VASILYEV received the B.S./M.S. degrees in physics from the Moscow State University, Moscow, Russia, in 1993 and 1995, respectively and received the Ph.D. degree in mathematical modeling, numerical methods, and software systems from the Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, in 2004. His research interests include differential equations of infinite order, systems of differential equations of infinite order with a small parameter, Markov chains with continuous time, numerical methods for solving problems of mathematical physics and mathematical methods in economics. Sergey Vasilyev is an Associated Professor at the Department of Applied Probability and Informatics of RUDN University since 2013.

GALINA TSAREVA received the B.S./M.S. degrees in mathematics from the Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, in 2012 and 2014, respectively. Her research interests include simulation of telecommunication systems, Markov chains and QoS problems in telecommunication systems. Galina Tsareva is a PhD Student at the Department of Applied Probability and Informatics of RUDN University since 2014.

GLOBAL AND LOCAL SYNCHRONIZATION IN PARALLEL SPACE-AWARE APPLICATIONS

Franco Cicirelli, Agostino Forestiero
Andrea Giordano, Carlo Mastroianni
ICAR-CNR, Rende (CS), Italy
Email: {cicirelli,forestiero,giordano,mastroianni}
@icar.cnr.it

Rostislav Razumchik
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia,
Peoples' Friendship University of Russia
(RUDN University), Moscow, Russia
Email: rrazumchik@ipiran.ru,
razumchik_rv@rudn.university

KEYWORDS

Synchronization algorithms, parallel computing, performance evaluation

ABSTRACT

Space-aware applications are characterized by an explicit representation of a spatial environment in which some entities live and operate by interacting with each other and with the hosting territory. A relevant space-aware application domain is the so-called urban computing, embracing issues like the simulation and implementation of public transportation systems, traffic management, urban monitoring and control. The execution of such applications is often distributed on parallel computing nodes, which need to cooperate and exchange data among each other, thus raising synchronization issues. In this paper we analyze time-related characteristics of the computational process in a space-aware application in the case when each node does not need *global synchronization* (i.e. synchronization with all other nodes) but requires only *local synchronization* (i.e. synchronization with a subset of neighbor nodes). Performance is evaluated both analytically and numerically. We provide the analytical support to an important conclusion: the mean computation time per step remains finite irrespective of the number of nodes under local synchronization, while under global synchronization it grows unboundedly as the number of nodes increases. In practical scenarios this corresponds to significantly better scalability properties of local synchronization.

INTRODUCTION

The need for parallelizing a computation can be caused by the necessity to increase the efficiency of a very complex application or can be inherent and related to the scenario in which the application is defined. The latter case, on which we focus here, occurs in a large variety of "space-aware applications" (SAAs), for which data and computation are inherently distributed and rely on the explicit representation of a territory, that is, a spatial environment on which data and objects are defined (see Shook et al. (2013)). For example, in an urban environment, data is generated by the users that move in a city, and there can be the need for aggregating and processing the data both at a local level – e.g., at each city neighborhood – and at the a global level, e.g., to derive general knowledge concerning the whole environment.

In such contexts, it is natural to parallelize the computation on distributed nodes and to perform the partitioning by utilizing the topological properties of the territory itself. Specifically, different regions of the territory can be assigned to different computing nodes which can process local data in parallel. Geology, biology, hydrology, social sciences, logistics and transportation, smart electrical grids, are significant examples of application fields strongly related to SAAs (see Cicirelli et al. (2016); Gong et al. (2013); Tang et al. (2011)). Another one is the urban-computing field mentioned above, for which data contains information regarding the mobility of people or vehicles, air quality, safety issues, water/electricity consumptions, etc., and can be profitably used to improve urban services and environments (see Zheng et al. (2014); Blecic et al. (2014)). Two more application fields that are emerging recently are the "Internet of Things" (IoT) (see Atzori et al. (2010); Lee and Lee (2015)) and some new distributed forms of Cloud Computing, sometimes referred to as Fog Computing or Edge Computing (see Krishnan et al. (2015); Hu et al. (2017)), where the computation is brought closer to the user's end and/or where the data is generated.

In general, space-aware applications are not "embarrassingly parallel" (Ekanayake and Fox, 2010), i.e., computation at the single nodes cannot be performed in isolation because parallel tasks need to exchange data during computation. This means that the computation performed by different nodes needs to be synchronized (Fujimoto, 2000), i.e., at certain time instants, one node must wait for the data coming from other nodes before proceeding to the next piece of computation. In this paper, for the sake of simplicity, we focus on the very common case of step-based computation (i.e., the computation is organized in work units which we call "steps"), and synchronization occurs at the end of each step.

Given the growing interest in space-aware applica-

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

tions, there is a strong need for methodological approaches that help assessing the performance of their parallel execution. In particular, it is a well-established empirical fact that the methodology adopted for synchronization has a notable impact on the performance. With the widely adopted all–to–all synchronization approach, henceforth also referred to as *global*, a computation step can be executed only as soon as all the nodes have completed the previous step. However, in many contexts this requirement can be relaxed, and a node can proceed to the next step after synchronizing with a limited number of nodes, e.g., those which are assigned to adjacent portions of the territory in a urban computing scenario. This type of synchronization is henceforth referred to as *local synchronization*.

One example of an application that can profitably adopt local synchronization is the computation of frequent mobility patterns (Yuan et al., 2013), i.e., the most common routes that are followed by vehicles, as these patterns can be extracted by concatenating the local patterns discovered in different city districts (Harri et al., 2009). The opportunity emerges of synchronizing the computation only among a limited number of parallel nodes, without the need for a central coordinator node. The computation at one node can proceed after being notified about the patterns discovered in neighbor nodes, and can then concatenate these patterns, thus allowing mobility patterns to be available much more rapidly than in the scenario where the computation is synchronized globally and data is delivered to a central node. The patterns regarding the whole territory are achieved by progressively extending the area covered by local patterns.

In Mastroianni et al. (2017) it was shown that local synchronization performs better than global synchronization in terms of computation efficiency. However, there is the need for building a theoretical foundation that is able to accurately predict the performance, and also to analyze the scalability properties. The scalability issue arises when the number of parallel nodes increases, either because the analyzed territory is extended (for example, the city area for the mobility pattern application mentioned before) or because the same area is divided into a larger number of regions in order to improve the accuracy of the computation. An analytical study is essential to tackle important engineering issues, such us: what is the number of parallel nodes needed to execute the computation in a given time interval? what is the impact of synchronization degree, i.e., the number of nodes with which a single node must synchronize?

In this paper, we provide an analytical representation of the model and its key ingredients: the overall execution time and the time to perform a single step when the synchronization overhead is taken into account. In particular, we prove that in the general case, the average time to perform a computation step on each node converges under local synchronization, i.e., it is bounded when the number of nodes increases, while under global synchronization it grows unboundedly. Therefore, the adoption of local synchronization is of utmost importance when the computational load or the involved scenario requires the use of a large number of nodes.

The paper is organized as follows. The next section describes the local synchronization model and provides an analytical formulation for the execution time. Then we assess the performance of local and global synchronization: at first analytically by using the extreme value theory and the max-plus algebra, then numerically by simulation. In conclusion, we summarize the work and indicate some interesting research avenues for future work.

LOCAL SYNCHRONIZATION MODEL

A natural way to optimize the execution of algorithms working on spatial data is to partition the territory and use this partitioning to decompose and parallelize the computation. The idea is to individuate a number of *regions* of the territory, and assign each region, along with the contained entities, to a *computing node* that will be in charge of performing the computation pertaining to that portion of the territory. Partitioning favors system scalability in that as the size of the territory increases, more computing nodes can be used to speed up the execution. A territory can be partitioned through either a *one-dimensional* or a *bidimensional* schema, as shown in Figure 1.

Let us denote by N the number of nodes, by $l_i(k)$ the time needed by node i, $1 \leq i \leq N$, to execute the local computation at the step k, and by $T_i(k)$ the time elapsed from the beginning of the computation at node i (i.e., start of the step 1) until the end of the step k.

It is important to define how the computing nodes synchronize with each other. In many parallel/distributed systems, as reported in the current literature, synchronization is global or, in other words, *all–to–all*, i.e., it is performed by constraining each node to start the execution for a given step only when all the nodes have finished their execution of the step before. The left part of Figure 2 shows an example of the dynamics of a system composed of seven nodes, for two consecutive steps. It is seen in the figure that, at each step, all the nodes must wait for the slowest one before advancing to the next step. In the figure, node 5 is the slowest node at step 1 while node 3 is the slowest at step 2.

For many SAAs scenarios, global synchronizations is not required. Indeed, as mentioned in the introduction section, it can be that a node needs to communicate and synchronize only with a set of neighbor nodes. Figure 3 shows the loop executed by each computing node, at each step, when adopting such local synchronization. The loop includes three phases. First, the node executes the local computation, i.e., the computation related to the specific region for the current step. Afterwards, the node sends data to its neighbor nodes. Finally, the node waits for the analogous data coming from its neighbors, i.e., the nodes managing the neighbor regions.

Resuming the execution advancement shown in the left part of Figure 2, corresponding to the case of global syn-

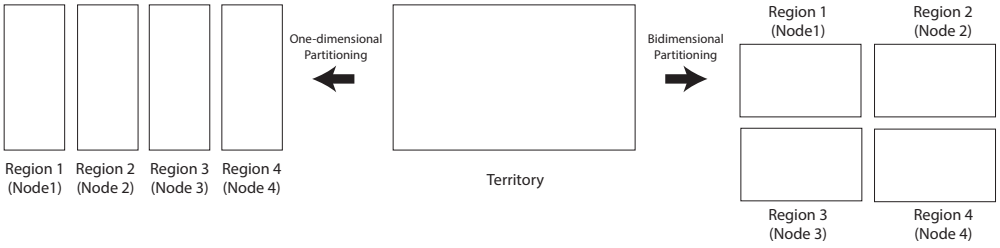


Figure 1: A territory partitioned into regions which are associated with parallel computing nodes. Two alternative types of partitioning are shown, one-dimensional and bidimensional.

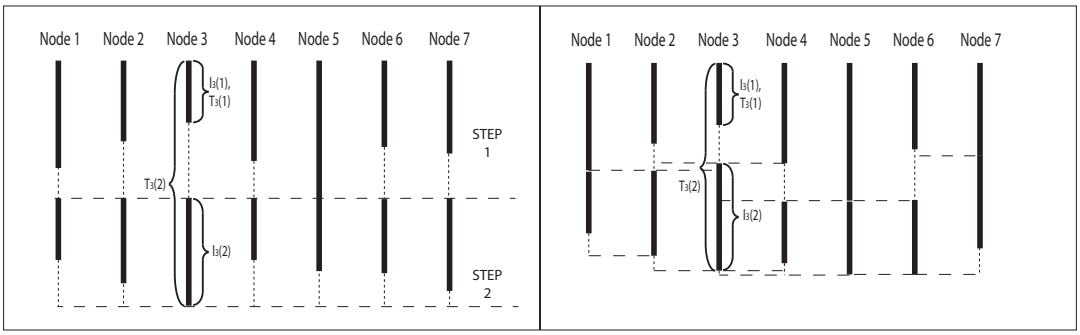


Figure 2: Dynamics of seven nodes for two steps using global synchronization (left) and local synchronization (right). The solid vertical lines represent the execution times, the dashed vertical lines are the waiting times and the horizontal dashed lines represent the synchronization points.

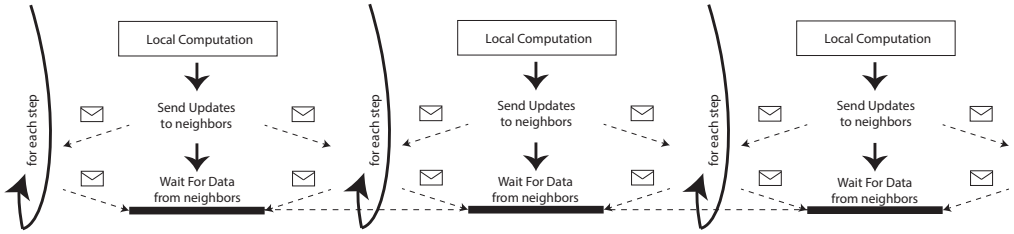


Figure 3: Execution loop under local synchronization.

chronization, we take the same local computation times, $l_i(k)$, and in the right part of Figure 2 we show the corresponding execution advancement when adopting the local synchronization. It can be seen that the times $T_i(k)$ tend to be shorter when compared to the case of global synchronization.

Under local synchronization, $T_i(k)$ are determined from the following recursive formula:

$$T_i(k+1) = \max\left(T_i(k), T_{i-1}(k), T_{i+1}(k)\right) + \\ + \; l_i(k+1), \; 1 \le i \le N, \quad (1)$$

where $T_0(k) = T_{N+1}(k) = 0$.

In (1) we have implicitly assumed that the time for transmitting the data between the neighbor nodes is negligible. Let $c_{i,j}(k)$ be the communication time needed for transmitting the data from the node i to the node j at the end of step k. When the communication time is not negligible, the recursive formula (1) is transformed to:

$$T_i(k+1) = \max\left(T_i(k), T_{i-1}(k) + c_{i-1,i}(k), \right. \\ \left. T_{i+1}(k) + c_{i+1,i}(k)\right) + \; l_i(k+1), \; 1 \le i \le N. \quad (2)$$

In the case of bidimensional partitioning, using a grid with R rows and C column, and $N = R \cdot C$, let us call $i_{r,c}$ the node located in row r, $1 \le r \le R$, and column c, $1 \le c \le C$. For example, in the right part of Figure 1, $i_{1,1}$ is Node 1 and $i_{1,2}$ is Node 2. Accordingly, $l_{r,c}(k)$ and $T_{r,c}(k)$ are, respectively, the local computation time and the time elapsed from the beginning of the computation to the end of the step k at the node $i_{r,c}$, while $c_{r,c,r',c'}(k)$ is the time for transmitting data from the node $i_{r,c}$ to the node $i_{r',c'}$ at the end of step k. In the bidimensional case, $T_{r,c}(k)$ are computed from the recursion:

$$T_{r,c}(k+1) = \max(T_{r,c}(k), T_{r-1,c-1}(k) + c_{r-1,c-1,r,c}(k), \\ T_{r-1,c}(k) + c_{r-1,c,r,c}(k), T_{r-1,c+1}(k) + c_{r-1,c+1,r,c}(k), \\ T_{r,c-1}(k) + c_{r,c-1,r,c}(k), T_{r,c+1}(k) + c_{r,c+1,r,c}(k), \\ T_{r+1,c-1}(k) + c_{r+1,c-1,r,c}(k), T_{r+1,c}(k) + c_{r+1,c,r,c}(k), \\ T_{r+1,c+1}(k) + c_{r+1,c+1,r,c}(k)) + \; l_{r,c}(k+1) \\ 1 \le r \le R, \; 1 \le c \le C, \quad (3)$$

where $T_{0,c}(k) = T_{R+1,c}(k) = T_{r,0}(k) = T_{r,C+1}(k) = 0$.

Despite the relative simplicity of the systems of equations (2) and (3), which govern the behavior of the synchronization model, it turns out to be very hard to come up with the analytic analysis of its performance characteristics. In the next section we dwell on only one aspect of this problem: analysis of the mean computation time per step.

PERFORMANCE OF GLOBAL AND LOCAL SYNCHRONIZATION

In what follows we give the analytic support to the following conclusion: global synchronization leads to unbounded mean computation time per step as the number of nodes increases, whereas the local synchronization guarantees that the mean computation time per step remains finite irrespective of the number of nodes.

In what follows, for the sake of ease of exposition, we dwell on the simple case of the model: one-dimensional partitioning and negligible communication times. In addition, we assume that the computation times $l_i(k)$ depend only on the nodes but do not depend on the step number and that $l_i(k) = l_i$, $1 \le i \le N$, are i.i.d. random variables. In the case of global synchronization we have

$$T_i(k+1) = (k+1)\max(l_1, \ldots, l_N),\ k \ge 0, \qquad (4)$$

and in the case of local synchronization we have (1), which is readily reduced to

$$T_i(k+1) = \max(T_i(k), T_{i-1}(k), T_{i+1}(k)) + \\ + l_i,\ k \ge 0,\ 1 \le i \le N. \qquad (5)$$

Note that the random sequence $\{\vec{T}(k)/k,\ k \ge 1\}$, where $\{\vec{T}(k) = (T_1(k), \ldots T_N(k))\}$ and $T_i(k)$ are defined by (1) (and, of course, (4)), falls into the framework of stochastic equations as described in Borovkov (1979). Using the results from Borovkov (1979) it can be shown that when N is finite and $\{\vec{l}(k) = (l_1(k), \ldots, l_N(k)),\ k \ge 1\}$ are independent, the sequence $\{\vec{T}(k)/k,\ k \ge 1\}$ is ergodic and stable.

Analysis of global synchronization

Let us start with the analysis of (4). The conclusion about the behavior of the global synchronization case follows from the well-known results for the order statistics and extreme value theory (see, for example, David and Nagaraja (2003); Ang and Tang (1984); Madala and Sinclair (1991)). If the positive random variable l_i has any continuous distribution with the support[1] on a semi-infinite interval, then as the number of nodes N grows, the mean computation time per step $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$ grows as well and in the limit as $N \to \infty$ we have that

$$\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k)) = \mathbf{E}(\max(l_1, \ldots, l_N)) \to \infty.$$

For example, if l_i are i.i.d random variables distributed exponentially with mean μ, then for sufficiently large N we have[2] $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k)) \approx \mu\,(\ln N + 0.5772)$.

Such nice expressions for extreme values are not available for all distributions. Yet for quite a large class of distributions (those having pure-exponential and non-pure exponential tails like gamma distribution) the general expressions for $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$ can be found in Whitt et al. (2007).

So in the case of global synchronization, if the computation times are random with unbounded support, then as the number of nodes N increases, on average, we wait longer and longer in order to make the next computation step. The computation of the cumulative distribution function (c.d.f.) of $\frac{1}{k}\mathbf{E}(T_i(k))$ is straightforward when l_i are i.i.d. It should also be noticed that deeper insight into the global synchronization can be gained by looking at the relation between the global synchronization model and the two types of queueing models: split-merge queues[3] (see, for example, Altiok and Perros (1986)) and faucet queues as described in Lebedev (2003, 2004).

Analysis of local synchronization

Once we give up the global synchronization and allow the node to proceed to the next computation step once a finite number of neighbors finish their computations, the conclusion is changed: the mean computation time per step $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$ becomes finite irrespective of the number of nodes N. In the following we consider the case of synchronization with two neighbors[4], which is described by (5).

The model (5) falls into the general framework of discrete event dynamic systems and it is convenient to describe its evolution in terms of the max-plus algebra (Heidergott et al., 2006). As soon as it is done, one can use the well-known results of the max-plus theory to study the values of $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$.

Define the following "(max,+)" notations:

$$\forall x, y, \in \mathbb{R} \cup \{-\infty\},\ x \oplus y = \max(x, y),\ x \otimes y = x + y.$$

Define also the $N \times N$ matrix $\mathbf{T}(k) = [\mathbf{T}(k)]_{ij}$, $1 \le i, j \le N$. With this notation, remembering that $\vec{T}(k) = (T_1(k), \ldots T_N(k))$, equation (5) can be rewritten as

$$\vec{T}^{\,T}(k+1) = \mathbf{T}(k) \otimes \vec{T}^{\,T}(k),\ k \ge 0, \qquad (6)$$

where $\cdot^T$ stands for transpose, the matrix-vector product is defined by $[\mathbf{T}(k) \otimes \vec{T}^{\,T}(k)]_i = \max_{1\le j\le N}([\mathbf{T}(k)]_{ij} + T_j(k))$

[1] In the case that l_i are independent (not necessarily identically distributed) random variables with a continuous distribution having support on a bounded interval, the mean computation time $\lim_{k\to\infty} \mathbf{E}(T_i(k))/k$ is always a constant, irrespective of the number of nodes N.

[2] It is well-known that the right part is the approximation of the expected maximum of N i.i.d. random variables with exponential distribution equal to $\mu \sum_{i=1}^{N} i^{-1}$, with $\sum_{i=1}^{N} i^{-1}$ being the N^{th} harmonic number.

[3] The sojourn time of the k^{th} customer in a split-merge queue in the light-traffic regime is equal to $\frac{1}{k}\mathbf{E}(T_i(k))$.

[4] But the analysis and conclusions remain valid also in the case of more than two neighbours and in the case of (at least simple) bidimensional partitioning.

and the matrix $\mathbf{T}(k)$ is defined by

$$\mathbf{T}(k) = \begin{pmatrix} l_1 & l_1 & -\infty & -\infty & \dots & -\infty & -\infty \\ l_2 & l_2 & l_2 & -\infty & \dots & -\infty & -\infty \\ -\infty & l_3 & l_3 & l_3 & \dots & -\infty & -\infty \\ -\infty & -\infty & l_4 & l_4 & \dots & -\infty & -\infty \\ \vdots & \vdots & \vdots & \vdots & \ddots & \dots & \ddots \\ -\infty & -\infty & -\infty & -\infty & \dots & l_{N-1} & l_{N-1} \\ -\infty & -\infty & -\infty & -\infty & \dots & l_N & l_N \end{pmatrix}.$$

Here the initial condition $\vec{T}^T(0)$ is simply the column-vector of zeros. Now we can make use of the well-known asymptotic results from the max-plus theory (see, for example, Baccelli and Konstantopoulos (1992)). The matrix $\mathbf{T}(k)$ has at least one finite entry on each row, which is the necessary and sufficient condition for $T_j(k)$ to be finite. From (Baccelli and Konstantopoulos, 1992, Lemma 6.1) we find[5] that there exists $\gamma > 0$ such that $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k)) = \gamma$, $1 \le i \le N$. In the case when l_i are stochastically bounded by a single variable (say L), having moment generating function (say $L(s)$), the upper bound for the value of γ is (see (Baccelli and Konstantopoulos, 1992, Proposition 6.2)):

$$\gamma \le \inf\{x > \mathbf{E}(L) \text{ such that } M(x) > \ln 3\}, \qquad (7)$$

where $M(x) = \sup_{\theta\in\mathbb{R}}(\theta x - \ln L(\theta))$ and $\mathbf{E}(L) = L'(0)$.

What this result tells us is that in the case of local synchronization the mean computation time remains finite for any number of nodes N. We could also proceed with the establishing of the upper bound in (7) right from (5) without resorting to the results of the max-plus algebra. Indeed, assuming, as above, that all l_i are stochastically bounded by L with the moment generating function $L(s)$, we can apply stochastic ordering techniques to find the following upper bound for $\lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$: $(\ln 3 + \ln L(s))/s$, where s is the solution to the equation $L(s) = 3$.

The exact computation of γ in the considered model is difficult and we are unaware of any well-established procedure (either in max-plus world or conventional probability theory) to perform such computation. On the contrast to the global synchronization model, where the exact value of γ can be easily written out, the main problem in the local synchronization is the inter-dependence between the values of $T_i(k)$ and $T_j(k)$, $i \neq j$. If we drop this dependence by assuming that $\{T_i(k),\ 1 \le i \le N\}$ are independent for each k, the computations of the c.d.f. of $T_i(k)$ are possible but they may lead to highly underestimated or overestimated values of γ. If we keep the dependence structure, then at each step k we have to perform the computations of the N-dimensional c.d.f. of the vector $\vec{T}(k)$, which already for small values of k and N become infeasible. At last, we note that the structure of (5) suggests that for the construction of the N-dimensional c.d.f. it is appealing to apply the dependence trees technique (see Chow and Liu (1968)), which takes into account order dependence relationships between the random variables.

[5] Baccelli and Konstantopoulos (1992) gives even stronger results of convergence of $\lim_{k\to\infty} \frac{1}{k}(T_i(k))$ to the same γ with probability 1.

Numerical results

In the following we report some numerical results both for global and local synchronizations under: one-dimensional partitioning, negligible communication times and i.i.d. computation times l_i. We have used Matlab to simulate the computational behavior modeled by (4) and (5) with different number of nodes N. We have also considered the extended local synchronization scenario in which a node synchronizes with more than two neighbors (the number of neighbors with which a node synchronizes on each of the two sides (left and right) is referred to as N_b). For l_i we have considered[6] exponential and hyper-gamma distributions. The performance is assessed by computing[7] the mean computation time per step $T_{step} = \lim_{k\to\infty} \frac{1}{k}\mathbf{E}(T_i(k))$.

First, we consider the case in which l_i has an exponential distribution with the mean equal to 6.185, for a fair comparison with the hyper-gamma distribution discussed later. Figure 4 shows the values of T_{step} versus the number of nodes N in the case of global and local synchronization with different values of N_b. The figure also reports the theoretical value for the case of global synchronization, which corresponds to the N^{th} harmonic number. The experimental values are consistent with the theoretical bounds discussed in the previous section, and it can be seen that the use of local synchronization allows T_{step} to be notably reduced with respect to global synchronization, even when the number of neighbors N_b increases. In Figure 5 we focus on the "exponential" scenario with N_b=1 and report the bound obtained with the

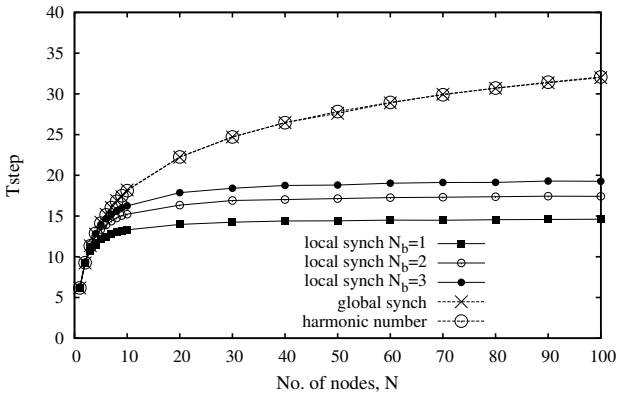


Figure 4: Values of T_{step} as function of N under global and local synchronizations with different values of N_b. The local computation times have exponential distribution with mean 6.185.

[6] Our choice is motivated by (Lublin and Feitelson, 2003), where hyper-gamma distribution is shown to be well-suited workload model in parallel computing systems.

[7] In order to compute each time the steady state value of T_{step} we used a single simulation run with $k = 10000$ and the batch-means method.

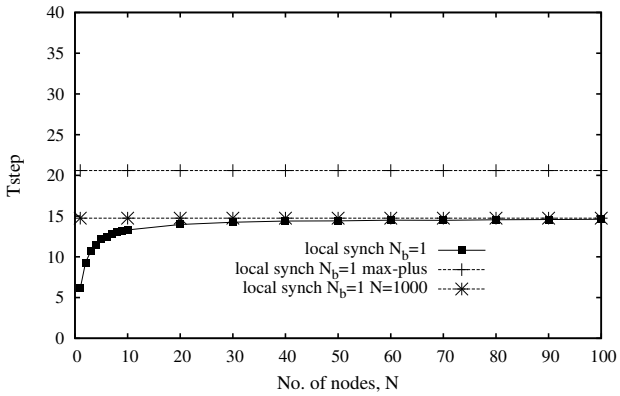


Figure 5: Values of T_{step} as function of N under local synchronization and N_b=1. The local computation times have exponential distribution with mean 6.185. The plot shows the bound obtained with max-plus algebra and the numerical value with N=1000.

max-plus algebra. We also show the value for $N = 1000$ nodes. From the figure it can be seen that the exact numerical bound is already reached when the number of nodes is 100.

Figure 6 shows the results that are analogous to those of Figure 4, this time assuming that the local computation times have hyper-gamma distribution with the pdf

$$p(x) = p\frac{1}{\Gamma(\alpha_1)}\beta_1^{\alpha_1} x^{\alpha_1-1} x^{-\beta_1 x} + (1-p)\frac{1}{\Gamma(\alpha_2)}\beta_2^{\alpha_2} x^{\alpha_2-1} x^{-\beta_2 x},$$

and the parameters p, α_i, β_i taken from (Lublin and Feitelson, 2003, Table 2):

$$p = 0.55, \alpha_1 = 6, \beta_1 = 1.51, \alpha_2 = 68.5, \beta_2 = 7.692.$$

It is seen that the advantage of local synchronization is larger with the exponential distribution due to its larger variance. This is a general result that we have also found in other experiments not shown here.

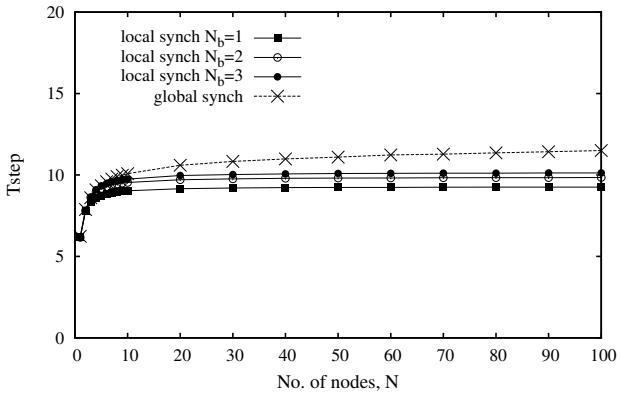


Figure 6: Values of T_{step} as function of N under global and local synchronizations with different values of N_b. The local computation times have hyper-gamma distribution, with mean equal to 6.185 and variance equal to 8.011.

CONCLUSION

In this paper two different synchronization strategies are being compared, namely local and global, for the execution of distributed space-aware applications. We have provided an analytical support, based on the max-plus theory, to conclusions about the finiteness of the computation time per step under local synchronization in general, and its unboundedness (for infinite number of nodes) under global synchronization. In practical scenarios, this corresponds to a much better scalability behavior of local synchronization, as confirmed by numerical experiments. We deem that this result can be profitably considered when designing a real computational infrastructure, for example for the support of urban computing applications. The problem of exact computation of moments of computation time per step in the asymptotic case is still open and requires further study. Also, the benefits deriving from local synchronization need to be better assessed in cases when communication times are not negligible and the computation load is not equally partitioned among the nodes and/or varies with time.

REFERENCES

Altiok, T. and Perros, H. (1986). Open networks of queues with blocking: Split and merge configurations. *IIE Transactions*, 18(3):251–261.

Ang, A.-S. and Tang, W. (1984). *robability Concepts in Engineering Planning and Design Vol. II*. Rainbow Bridge.

Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. *Computer networks*, 54(15):2787–2805.

Baccelli, F. and Konstantopoulos, P. (1992). Estimates of cycle times in stochastic petri nets. In *Applied Stochastic Analysis*, pages 1–20. Springer Berlin Heidelberg.

Blecic, I., Cecchini, A., Trunfio, G. A., and Verigos, E. (2014). Urban cellular automata with irregular space of proximities. *Journal of Cellular Automata*, 9(2-3):241–256.

Borovkov, A. A. (1979). Ergodicity and stability theorems for a class of stochastic equations and their applications. *Theory of Probability & Its Applications*, 23(2):227–247.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Cicirelli, F., Forestiero, A., Giordano, A., and Mastroianni, C. (2016). Transparent and efficient parallelization of swarm algorithms. *ACM Trans. Auton. Adapt. Syst.*, 11(2):14:1–14:26.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, Third Edition*. John Wiley.

Ekanayake, J. and Fox, G. (2010). High performance parallel computing with clouds and cloud technologies. In *Cloud Computing*, pages 20–38. Springer.

Fujimoto, R. (2000). *Parallel and distributed simulation systems*. John Wiley.

Gong, Z., Tang, W., Bennett, D. A., and Thill, J.-C. (2013). Parallel agent-based simulation of individual-level spatial interactions within a multicore computing environment. *International Journal of Geographical Information Science*, 27(6):1152–1170.

Harri, J., Filali, F., and Bonnet, C. (2009). Mobility models for vehicular ad hoc networks: a survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 11(4):19–41.

Heidergott, B., Olsder, G. J., and van der Woude, J. (2006). *Max Plus at Work: Modeling and Analysis of Synchronized Systems: A Course on Max-Plus Algebra and Its Applications*. Princeton University Press.

Hu, P., Dhelim, S., Ning, H., and Qiu, T. (2017). Survey on fog computing: architecture, key technologies, applications and open issues. *Journal of Network and Computer Applications*, 98:27 – 42.

Krishnan, Y. N., Bhagwat, C. N., and Utpat, A. P. (2015). Fog computing- network based cloud computing. In *2nd IEEE International Conference on Electronics and Communication Systems (ICECS)*, pages 250–251.

Lebedev, A. V. (2003). The gated infinite-server queue with unbounded service times and heavy traffic. *Problems of Information Transmission*, 39(3):309–316.

Lebedev, A. V. (2004). Gated infinite-server queue with heavy traffic and power tail. *Problems of Information Transmission*, 40(3):237–242.

Lee, I. and Lee, K. (2015). The internet of things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons*, 58(4):431 – 440.

Lublin, U. and Feitelson, D. G. (2003). The workload on parallel supercomputers: modeling the characteristics of rigid jobs. *Journal of Parallel and Distributed Computing*, 63(11):1105 – 1122.

Madala, S. and Sinclair, J. B. (1991). Performance of synchronous parallel algorithms with regular structures. *IEEE Trans. on Parallel and Distributed Systems*, 2(1):105–116.

Mastroianni, C., Cesario, E., and Giordano, A. (2017). Efficient and scalable execution of smart city parallel applications. *Concurrency and Computation: Practice and Experience*. Early view, http://dx.doi.org/10.1002/cpe.4258.

Shook, E., Wang, S., and Tang, W. (2013). A communication-aware framework for parallel spatially explicit agent-based models. *International Journal of Geographical Information Science*, 27(11):2160–2181.

Tang, W., Bennett, D. A., and Wang, S. (2011). A parallel agent-based model of land use opinions. *Journal of Land Use Science*, 6(2–3):121–135.

Whitt, W., Crow, C., and Goldberg, D. (2007). Two-moment approximations for maxima. *Operations Research*, 55(3):532–548.

Yuan, J., Zheng, Y., Xie, X., and Sun, G. (2013). T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):220–232.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):1–55.

AUTHOR BIOGRAPHIES

Franco Cicirelli Ph.D, is a researcher at the ICAR-CNR Institute, Italy, since 2015. He earned a Ph.D. in System Engineering and Computer Science at the University of Calabria (Italy). He was a researcher fellow at the University of Calabria (Italy) from 2006 to 2015. Research topics include agent-based systems, distributed simulation, parallel and distributed systems, real-time systems, workflow management systems, Internet of Things and cyber-physical systems.

Agostino Forestiero is a researcher at ICAR-CNR, Italy, since 2010. He received his Laurea degree and his Ph.D. degree in Computer Engineering from the University of Calabria, Italy, in 2002 and 2007. He co-authored over 50 papers published in international journals, among which IEEE/ACM TON, IEEE TEVC and ACM TAAS, and conference proceedings. His areas of interest are distributed systems, Cloud Computing, P2P, swarm intelligence, multi-agent systems and cyber-physical systems.

Andrea Giordano is a researcher at ICAR-CNR, Italy, since 2011. He earned a Masters degree in Computer Engineering and a Ph.D in System Engineering and Computer Science at the University of Calabria, Italy. His research work focuses on agent-based systems, parallel and distributed systems, swarm intelligence, distributed simulation and cyber-physical systems.

Carlo Mastroianni is a senior researcher at ICAR-CNR, Italy. He received his Laurea degree and his Ph.D. degree in Computer Engineering from the University of Calabria, Italy, in 1995 and 1999, respectively. He authored over 100 papers published in international journals, among which IEEE/ACM TON, IEEE TCC, IEEE TEVC and ACM TAAS, and conference proceedings. His research focuses on Cloud Computing, P2P, bio-inspired algorithms, multi-agent systems.

Rostislav Razumchik received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as a leading research fellow at Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS). Currently he also holds the associate professor position at Peoples' Friendship University of Russia (RUDN University). His current research activities are focused on queueing theory and its applications for the evaluation of stochastic systems.

SOFTWARE PACKAGE FOR THE ACTIVE QUEUE MANAGEMENT MODULE MODEL VERIFICATION

Tatyana R. Velieva, Anna V. Korolkova, Migran N. Gevorkyan, Sergey A. Vasilyev
Department of Applied Probability and Informatics,
Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
Email: velieva_tr@rudn.university, korolkova_av@rudn.university,
gevorkyan_mn@rudn.university, vasilyev_sa@rudn.university

Ivan S. Zaryadov
Department of Applied Probability and Informatics,
Peoples' Friendship University of Russia
(RUDN University),
6 Miklukho-Maklaya St, Moscow,
117198, Russian Federation
and Institute of Informatics Problems,
FRC CSC RAS, IPI FRC CSC RAS,
44-2 Vavilova str., Moscow, 119333, Russia
Email: zaryadov_is@rudn.university

Dmitry S. Kulyabov
Department of Applied Probability and Informatics,
Peoples' Friendship University of Russia
(RUDN University),
6 Miklukho-Maklaya St, Moscow,
117198, Russian Federation
and Laboratory of Information Technologies
Joint Institute for Nuclear Research
Joliot-Curie 6, Dubna, Moscow region, 141980, Russia
Email: kulyabov_ds@rudn.university

KEYWORDS

Active queue management, simulation, NS2, Julia, self-oscillating

ABSTRACT

Self-oscillation modes in control systems of data transmission networks negatively affect the characteristics of these networks. To investigate the self-oscillation mode for systems with a control module, the analytical model of the active queue management module was developed. The problem of verification of the obtained theoretical results arises in the study. Previously, a software system was developed for software emulation of the router. However, its use has caused some difficulties. Alternatively, the simulation model of network with active queue management module was developed. The paper describes a software package for verifying theoretical calculations based on the NS-2 simulation system. For illustration, a numerical example is given.

INTRODUCTION

When modeling technical systems, there is often a question of verification of results. Either there is no access to data on the functioning of such systems, or the acquisition of data is associated with large resource and time costs. But some simulation experiments can be seen as a solution to this problem.

The problem of the occurrence of self-oscillatory regime in systems with control is considered (see Lautenschlaeger and Francini (2015)). In particular, the active queue management modules with RED-like algorithms were studied. Based on the theoretical model (see Misra et al. (1999); Kulyabov et al. (2018)) of the functioning of the RED module, the parameters of self-oscillating regimes were investigated. However, for completeness of the study, the verification of the results is necessary. We have developed the installation for a full-scale experiment on the basis of emulation of images of network equipment (see Velieva et al. (2015)). However, this approach required some extra resources, which were not at authors disposal. So, as a substitute, it was suggested to use the simulation model.

In this paper, the simulation model based on the NS-2 network protocol simulation tool is described. This approach proved to be more flexible in comparison with the full-scale experiment on the basis of emulation of network operating systems.

RED ADAPTIVE CONGESTION CONTROL MECHANISM

The RED algorithm (see Adams (2013); Kushwaha and Shwer (2013); Kushwaha and Gupta (2014)) uses a weighted queue length as a factor determining the probability of packets drop. As the average queue length grows, the probability of

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

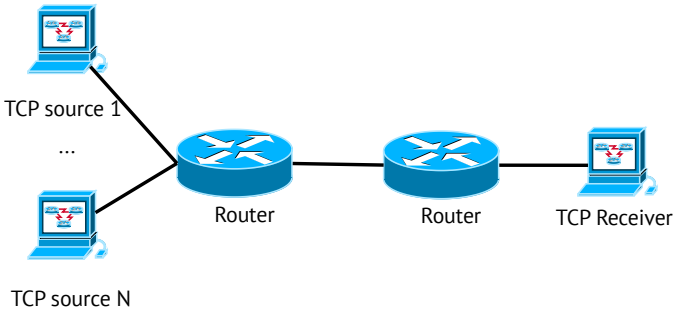


Fig. 1. Dumbbell topology

packets drop also increases. The algorithm uses two threshold values of the average queue length to control the drop function.

$$p(\hat{Q}) = \begin{cases} 0, & 0 < \hat{Q} \leqslant Q_{\min}, \\ \dfrac{\hat{Q} - Q_{\min}}{Q_{\max} - Q_{\min}} p_{\max}, & Q_{\min} < \hat{Q} \leqslant Q_{\max}, \\ 1, & \hat{Q} > Q_{\max}. \end{cases}$$

Here $p(\hat{Q})$ is the packets drop function (drop probability), $\hat{Q}$ is the exponentially-weighted moving average of the queue size average, $Q_{\min}$ and $Q_{\max}$ are the thresholds for the weighted average of the queue length, $p_{\max}$ is the the maximum level of the packets drop.

The RED algorithm is quite effective due to simplicity of its implementation in the network hardware, but it has a number of drawbacks. In particular, for some parameters values there is a steady oscillatory mode in the system, which negatively affects the quality of service (QoS) indicators (see Jenkins (2013); Ren et al. (2005); Lautenschlaeger and Francini (2015)). Unfortunately there are no clear selection criteria for RED parameters values, at which the system does not enter in self-oscillating mode.

SIMULATION MODEL

The full-scale experiment often involves certain difficulties. For example, the real equipment is not always available. Also the use of a virtual stand is associated with high demands on computer equipment (see Velieva et al. (2015)). In addition, since the simulation takes place in real time, the whole process is extremely long.

To save resources and time, simulation tools are usually used. The package ns2 (see Issariyakul and Hossain (2012); Altman and Jiménez (2012)) is a tool for network protocols simulating. This package was created as a reference modeling tool, so it is often used as an alternative to the full-scale experiment.

For an imitation experiment, we will use the so-called dumbbell topology (see Fig. 1). Additional TCP sessions are emulated by addition of extra sources.

The program for ns2 is written in TCL language (see Welch and Jones (2003); Nadkarni (2017)).

First, we need to create a simulator object. Let's set the experiment time.

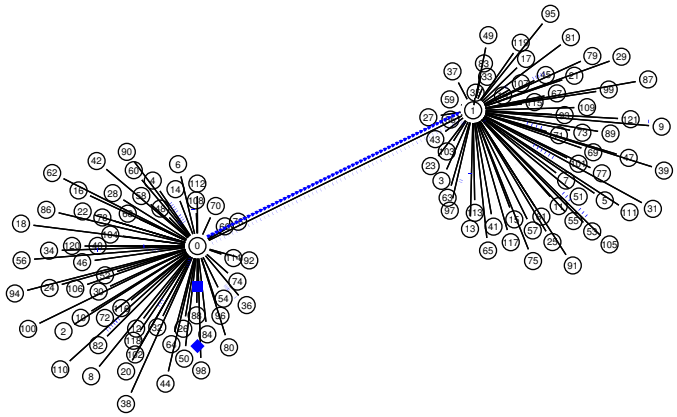


Fig. 2. Visualization of the simulation. Packets drop is shown

```
set ns [new Simulator]

set sduration 20
set simTime 20
```

We may write the data for the nam visualization tool (see Fig. 2). In the final version of the script, we will disable this feature to save resources.

```
# set nf [open out.nam w]
# $ns namtrace-all $nf
```

Let's set the number of TCP sessions (sources).

```
set numSrc 60
```

The current experiment is connected with the study of thresholds values influence on the occurrence of self-oscillation mode. Therefore, the threshold values are set as arguments.

```
if { $argc == 2 } {
  puts "[lindex $argv 0] [lindex $argv 1]"
  # thresh_ <=> Q_min
  # maxthresh_ <=> Q_max
  Queue/RED set thresh_ [lindex $argv 0]
  Queue/RED set maxthresh_ [lindex $argv 1]
}
```

In ns2, there are implementations of three varieties of the RED discipline: the original RED Floyd and Jacobson (1993) algorithm, the ARED algorithm, and the Gentle RED algorithm. The settings of the algorithms are controlled by parameters.

- `bytes_`: turns on (`true`) or turns off (`false`) the mode *byte mode*, in which the size of the package affects the probability of their tagging to drop;
- `Queue-in-bytes_`: if the value of the parameter is set to `true`, then the average queue length will be measured in bits. Also, the `thresh_` and `maxthres_` will be measured by the calculated average packet size (`mean_pktsize_`). The default value is `false`;
- `thres_`: the minimum queue length threshold $q_{\min}$;

- `maxthres_`: the maximum queue length threshold $q_{\max}$;
- `mean_pktsize_`: an approximate estimation of the packet size in bytes. The default value is 500;
- `q_weight_`: the w_q weighting factor is used in calculating of the average queue length;
- `wate_`: this option allows to maintain the interval between drops packets if its valueis set as a `true`;
- `linterm_`: the inverse of the parameter $p_{\max}$. The default is 10;
- `setbit_`: takes the value `false`, if RED discards marked packets. If the value is set as `|true|`, a congestion bit is added to the marked packets;
- `drop-tail_`: when the value is true and the buffer is overflowed then the active queue management algorithm switches to the Drop Tail algorithm.

The default values for the parameters `q_weight_`, `maxthresh_` and `thres_` are 0.002, 15 and 5 respectively.

```
Queue/RED set q_weight_ 0.002
# Queue/RED set drop_tail_ true
Queue/RED set setbit_ false
Queue/RED set bytes_ false
Queue/RED set queue_in_bytes_ false
Queue/RED set gentle_ false
Queue/RED set mean_pktsize_ 1000
Queue/RED set cur_max_p_ 0.1
```

Two nodes that will play the role of routers are created.

```
set R1 [$ns node]
set R2 [$ns node]
```

Using a loop, we create the nodes that will simulate a TCP session.

```
# create the tcp/ftp src nodes
for {set i 1} {$i<=$numSrc} { incr i } {
    # Create node
    set n($i) [$ns node]
    # Create link
    $ns duplex-link $n($i) $R1 100Mb 20ms DropTail
    # Create TCP agent on node n($i)
    set tcp($i) [new Agent/TCP/Reno]

    $tcp($i) set window_ 32
    $tcp($i) set fid_ 2
    $tcp($i) set paketSize_ 1000
    $tcp($i) set class_ 1

    $ns attach-agent $n($i) $tcp($i)
    # FTP
    set ftp($i) [new Application/FTP]
    $ftp($i) attach-agent $tcp($i)
    $ftp($i) set type_ FTP

    # Create sink
    set s($i) [$ns node]
    # Create link
    $ns duplex-link $s($i) $R2 100Mb 20ms DropTail
    # Create sink agent on node s($i)
    set sink($i) [new Agent/TCPSink]
    $ns attach-agent $s($i) $sink($i)

    # Connect n($i) and s($i)
    $ns connect $tcp($i) $sink($i)

}
```

The queue is connected to the link between the routers. Since we are only interested in traffic in the forward direction, the discipline Drop Tail is set to the link in the opposite direction.

```
set flink [$ns simplex-link $R1 $R2 15Mb 35ms RED]
$ns simplex-link $R2 $R1 15Mb 35ms DropTail
$ns queue-limit $R1 $R2 300
```

One of the most important objects in ns2 is the queue monitor. It allows to gather information not only about the length of the queue, but also about arriving, departing and dropped packets.

```
set qmon [$ns monitor-queue $R1 $R2 [open qm.tr ↩
    w] 0.01]
[$ns link $R1 $R2] queue-sample-timeout
```

To monitor the parameters of the RED queue (for example, between the nodes $n2$ and $n3$), the following lines of code should be added:

```
set redq [[$ns link $n2 $n3] queue]
set traceq [open red-queue.tr w]
$redq trace curq_
$redq trace ave_
$redq attach $traceq
```

Here `curq_`is the current size of the queue, `ave_` is the average queue size. As a result, the output file consisting of three columns is obtained. The first column contains the flag `Q` (the current queue size) or `a` (average queue size). The other columns are the time and value of the observed parameter.

The size of the windowWe will be under control.

```
# Obtaining TCP CWND Window information

# plotWindow(tcpSource file k): Write CWND of k ↩
    tcpSources in file
# The output format is as follows:
# TIME Win_flow1 Win_flow2 Win_flow3 ... ↩
    Win_flowN
proc plotWindow {tcpSource file k} {
  global ns numSrc

  set time 0.03
  set now [$ns now]
  set cwnd [$tcpSource set cwnd_]
```

```
  if {$k == 1} {
    puts -nonewline $file "$now \t $cwnd \t"
  } else {
    if {$k < $numSrc } {
      puts -nonewline $file "$cwnd \t"
    }
  }

  if { $k == $numSrc } {
    puts -nonewline $file "$cwnd \n"
  }
  if { $k == $numSrc } {
    puts -nonewline $file "$cwnd \n"
  }
  $ns at [expr $now+$time] "plotWindow    ↩
      $tcpSource $file $k"
}

# Start plotWindow() for all tcp sources
# Output to stdout
for {set j 1} {$j<=$numSrc} { incr j } {
    $ns at 0.1 "plotWindow $tcp($j) stdout $j"
}
```

The process of simulation is started:

```
proc finish {} {
    exit 0
}

for {set i 1} {$i<=$numSrc} {incr i} {
    $ns at 0.0 "$ftp($i) start"
    $ns at $simTime "$ftp($i) stop"
    # $ns at $simTime "calc_throughput $tcpsrc($j)    ↩
      $j $simTime"
}

$ns at $simTime "finish"
$ns run
```

If the file with the model is named as `red.tcl`. Then, in order to run the simulation, the `ns red.tcl` command should be executed.

PROCESSING OF THE SIMULATION RESULTS

After the simulation experiment a large amount of raw data is obtained and it is necessary to process this data.

By using the output data the parameters of self-oscillations may be obtained. Here are the fragments of the program in the Julia language (see Joshi and Lakhanpal (2017)), in which the spectral portrait of the self-oscillatory mode is constructed on the basis of the Fast Fourier Transform algorithm (see Rao et al. (2010)).

The file being processed is passed as an argument.

```
#!/usr/bin/env julia
if length(ARGS) > 0
  try
    global const window_size = readcsv(ARGS[1])[:, 2]
    global const count = readcsv(ARGS[1])[:, 1]
  catch err
    if isa(err, SystemError)
      error("File $(ARGS[1]) not found")
    else
      throw(err)
    end
  end
end
```

Actually, the spectral analysis is carried out on the basis of a Fast Fourier Transform algorithm.

```
delta = count[2] .- count[1]
Fd = 1.0 ./ delta
fftCount = length(window_size)
X = fft(window_size[1:512])
amplitude_spectrum = 2 .* abs.(X) / 512.0
amplitude_spectrum[1] = amplitude_spectrum[1] ./ 2.0
frequency = collect(0:Fd./512.0:Fd./2 - 1.0./512.0)
```

As a result the amplitude and frequency of the first harmonic of self-oscillations are obtained.

```
max_amp=findmax(amplitude_spectrum,1)
println(max_amp[1][1],",",frequency[max_amp[2]][1])
```

In addition, we may derive the point values of the spectrum in order to build a graph on it later.

```
for (f, A) in zip(frequency, amplitude_spectrum)
  println(f, ",", A)
end
```

As in the current experiment the dependence of self-oscillations on the threshold values of the RED algorithm is investigated, we will generate files with different threshold values. This data will be used in carrying out the simulation.

```
#!/usr/bin/env python3
import itertools
import numpy as np

DIR = 'parameters'
Q_START = 10
Q_STOP = 90
Q_DELTA = 1

Q = np.arange(Q_START, Q_STOP+1, Q_DELTA)
gen = (p for p in itertools.product(Q, Q) if p[0]    ↩
    < p[1])
for i, p in enumerate(gen):
    with open("./{0}/{1:04d}".format(DIR, i),    ↩
      mode='w', encoding='utf-8') as f:
        f.write("{0[0]} {0[1]}".format(p))
```

In order to collect all the elements together, we will use the *Snakemake* assembly system [https://snakemake.readthedocs.io]. This system is ideologically similar to the Make assembly

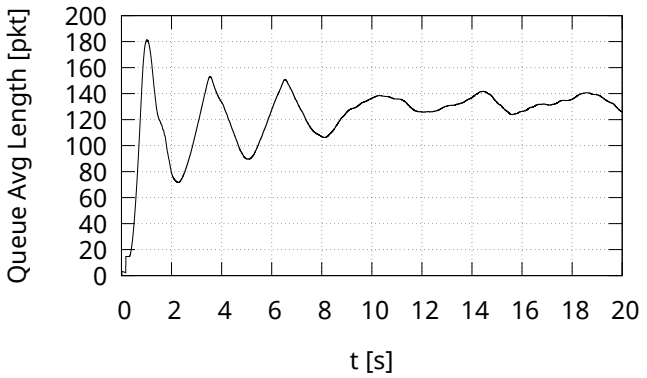


Fig. 3. Average queue length at link capacity $C = 5$ Mbps

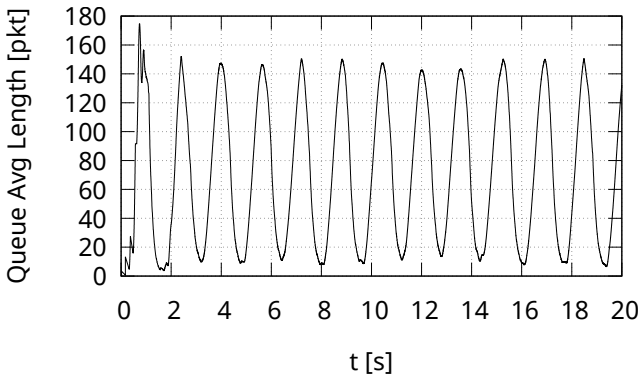


Fig. 4. Average queue length at link capacity $C = 20$ Mbps

system. However, it is not aimed at assembling software, but for reproducible and scalable data analyses. The syntax of *Snakemake* language is similar to the *Python* language.

```
rule all:
  input:
    ['spectrum/{0}'.format(f) for f in
      os.listdir("./parameters")]
  output:
    touch(".status")

rule fft:
  input:
    "parameters/{file}"
  output:
    "4fft/{file}"
  shell:
    "ns red.tcl `cat {input}` > {output}"

rule spectrum:
  input:
    rules.fft.output
  output:
    "spectrum/{file}"
  shell:
    "julia spectrum.jl {input} > {output}"
```

The resulting set of programs can be parallelized according to the SPMD ideology (single program, multiple data) (see Darema (2001)).

SIMULATION EXPERIMENT

As an illustration, we give a concrete example. Let's set the following parameters of the RED algorithm: the number of sessions $N = 60$, round-trip time $T_p = 0.075$ s, thresholds $Q_{\min} = 75$ packages and $Q_{\max} = 150$ packets, drop probability $p = 0.1$, parameter $w_q = 0.002$.

In the study examines the impact of the parameters on the character of the self-oscillation mode. Let us investigate the dependence of self-oscillation on the link capacity C.

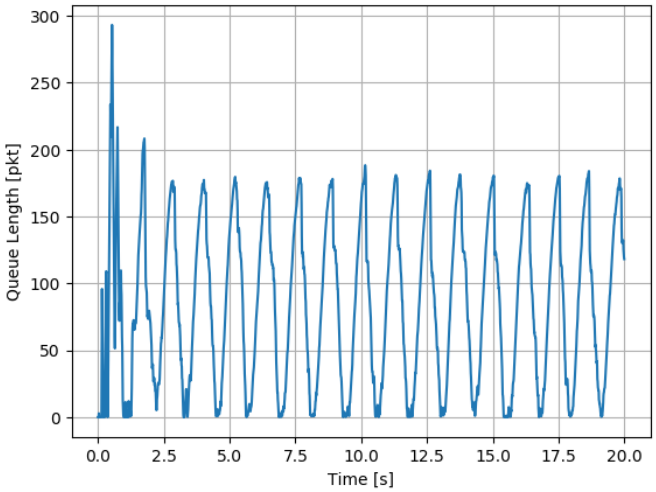


Fig. 5. Instantaneous queue length at link capacity $C = 20$ Mbps

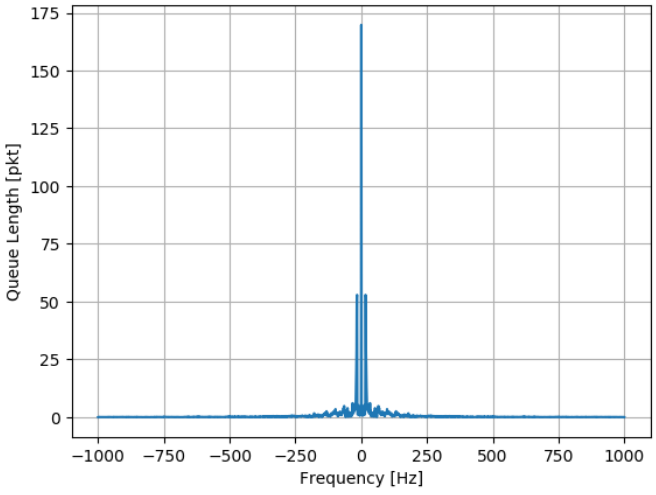


Fig. 6. Spectrum of self-oscillations of instantaneous queue length at link capacity $C = 20$ Mbps

The Fig. 3 and Fig. 4 show the behavior of the average queue length for link capacity $C = 5$ Mbps and $C = 20$ Mbps. In the second case clearly shows the presence of the self-oscillation mode. Theoretically obtained characteristic of this mode: oscillation frequency $\nu = 0.6$ Hz, oscillation amplitude $A = 150$ packets. In the spectral study of the results of the simulation, we obtained the following characteristics: the frequency of self-oscillations $\nu = 0.5$ Hz, the amplitude of the oscillations $A = 169$ packets (see Fig. 5 and Fig. 6). As can be seen, the theoretical and experimental values are very close. Thus, our program complex can serve the purposes of verification of theoretical studies of the self-oscillatory regime in control systems.

CONCLUSION

The authors have developed the set of programs for simulation experiment in order to investigate the self-oscillation mode of control systems and to verify theoretical results. It is assumed that this experiment will confirm the analytical model of the active queue management module with the RED-like algorithm proposed by the authors.

ACKNOWLEDGMENT

The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by Russian Foundation for Basic Research (RFBR) according to the research project No 16-07-00556. The computations were carried out on the Felix computational cluster (RUDN University, Moscow, Russia) and on the HybriLIT heterogeneous cluster (Multifunctional center for data storage, processing, and analysis at the Joint Institute for Nuclear Research, Dubna, Russia).

REFERENCES

Adams, R. (2013), Active Queue Management: A Survey, *IEEE Communications Surveys & Tutorials* 15(3), 1425–1476.

Altman, E. and Jiménez, T. (2012), NS Simulator for Beginners, *Synthesis Lectures on Communication Networks* 5(1), 1–184.

Darema, F. (2001), The SPMD Model: Past, Present and Future, pp. 1–1.

Floyd, S. and Jacobson, V. (1993), Random Early Detection Gateways for Congestion Avoidance, *IEEE/ACM Transactions on Networking* 1(4), 397–413.

Issariyakul, T. and Hossain, E. (2012), *Introduction to Network Simulator NS2*, Springer US, Boston, MA.

Jenkins, A. (2013), Self-Oscillation, *Physics Reports* 525(2), 167–222.

Joshi, A. and Lakhanpal, R. (2017), *Learning Julia*, Packt Publishing.

Kulyabov, D. S., Korolkova, A. V., Velieva, T. R., Eferina, E. G. and Sevastianov, L. A. (2018), The Methodology of Studying of Active Traffic Management Module Self-oscillation Regime, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak and J. Kacprzyk, eds, DepCoS-RELCOMEX 2017. Advances in Intelligent Systems and Computing, Vol. 582 of *Advances in Intelligent Systems and Computing*, Springer International Publishing, Cham, pp. 215–224.

Kushwaha, V. and Gupta, R. (2014), Congestion Control for High-Speed Wired Network: A Systematic Literature Review, *Journal of Network and Computer Applications* 45, 62–78.

Kushwaha, V. and Shwer, R. (2013), A Review of Router based Congestion Control Algorithms, *International Journal of Computer Network and Information Security* 6(1), 1–10.

Lautenschlaeger, W. and Francini, A. (2015), Global Synchronization Protection for Bandwidth Sharing TCP Flows in High-Speed Links, Proc. 16-th International Conference on High Performance Switching and Routing, IEEE HPSR 2015, Budapest, Hungary.

Misra, V., Gong, W.-B. and Towsley, D. (1999), Stochastic Differential Equation Modeling and Analysis of TCP-Windowsize Behavior, *Proceedings of PERFORMANCE* 99.

Nadkarni, A. P. (2017), *The Tcl Programming Language: A Comprehensive Guide*, CreateSpace Independent Publishing Platform.

Rao, K. R., Kim, D. N. and Hwang, J. J. (2010), *Fast Fourier Transform - Algorithms and Applications*, Signals and Communication Technology, Springer.

Ren, F., Lin, C. and Wei, B. (2005), A Nonlinear Control Theoretic Analysis to TCP-RED System, *Computer Networks* 49(4), 580–592.

Velieva, T. R., Korolkova, A. V. and Kulyabov, D. S. (2015), Designing Installations for Verification of the Model of Active Queue Management Discipline RED in the GNS3, 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE Computer Society, pp. 570–577.

Welch, B. and Jones, K. (2003), *Practical Programming in Tcl and Tk*, 4th edn, Prentice Hall.

AUTHOR BIOGRAPHIES

TATYANA R. VELIEVA postgaduate student in Peoples' Friendship University of Russia. Her current research activity focuses on mathematical modeling. Her email address is velieva_tr@rudn.university.

ANNA V. KOROLKOVA received her Ph.D. in Mathematics in 2010. Since then, she has worked as associate professor in RUDN University (Peoples' Friendship University of Russia). Her current research activity focuses on mathematical modeling. Her email address is korolkova_av@rudn.university.

MIGRAN N. GEVORKYAN received his Ph.D. in Mathematics in 2013. Since then, he has worked as associate professor in RUDN University (Peoples' Friendship University of Russia). His current research activity focuses on mathematical modeling. His email address is gevorkyan_mn@rudn.university.

IVAN S. ZARYADOV received his Ph.D. in Mathematics in 2010. Since then, he has worked as associate professor in

RUDN University (Peoples' Friendship University of Russia). His current research activity focuses on probability theory. His email address is zaryadov_is@rudn.university.

DMITRY S. KULYABOV received his Ph.D. in Physics in 2000. Since then, he has worked as associate professor in RUDN University (Peoples' Friendship University of Russia). His current research activity focuses on mathematical modeling. His email address is kulyabov_ds@rudn.university.

SERGEY A. VASILYEV received his Ph.D. in Physics in 2003. Since then, he has worked as associate professor in RUDN University (Peoples' Friendship University of Russia). His current research activity focuses on mathematical modeling. His email address is vasilyev_sa@rudn.university.

SIMULATION OF THE LIMITED RESOURCES QUEUING SYSTEM FOR PERFORMANCE ANALYSIS OF WIRELESS NETWORKS

Eduard Sopin*†, Kirill Ageev*, Sergey Shorgin†

*Peoples' Friendship University of Russia (RUDN University)
Applied Informatics and Probability Theory Department
Miklukho-Maklaya St 6, Moscow, 117198, Russian Federation
{sopin_es, ageev_ka}@rudn.university

†Institute of Informatics Problems, Federal Research Center "Computer Science and Control
Vavilova str., Moscow, 119333, Russian Federation
sshorgin@ipiran.ru

KEYWORDS
Queuing system, limited resources, random requirements, performance analysis, simulation, M2M traffic, LTE.

ABSTRACT

Queuing systems with limited resources, in which customers require a device and a certain amount of limited resources for the duration of their service, proved their effectiveness in the performance analysis of modern wireless networks. However, the application of the queuing systems leads to complex computations. In this paper, we develop the simulation tool for the limited resources queuing systems and apply it to the analysis of M2M traffic characteristics in a LTE network cell.

I. INTRODUCTION

Queuing systems with limited resources are very promising in the performance analysis of the contemporary wireless networks. The key advantage of them is possibility to capture peculiarities of different radio resources allocation schemes by definition of appropriate cumulative distribution function (CDF) of resource requirements.

All analytical results for probability measures of resource queuing systems were obtained under assumption of Poisson [5,8] or state-dependent Poisson arrivals [6]. However, analytical formulas for probabilistic characteristics are too complex to be used directly due to multiple convolutions of the resource requirements CDF. In [7], we derived the recurrent algorithm for evaluation of stationary measures for the case of discrete resource requirements and proposed the sampling approach for continuous resources. However, the complexity of the calculations is still high and the algorithms are only applicable under assumption of Poisson arrivals. Hence, the development of a tool for simulation of queuing systems with limited resources became a very important task.

In the paper, we describe the developed simulation tool and use it to evaluate performance measures of M2M traffic in a LTE network cell. By the way, we provide comparison of calculations accuracy with the methods, proposed in [7].

The rest of the paper is organized as follows. Section II describes the queuing system with limited resources and provides a short summary of analytical results for it, while section III gives a brief description of the simulation tool. In Sections IV, the developed simulation tool is used for the analysis of LTE cell characteristics, simulation results are presented. Section V concludes the paper.

II. QUEUING SYSTEM WITH LIMITED RESOURCES

We consider a multiserver queueing system with N servers, in which arriving customer occupies not only a server, but also a volume of limited resources. The total volume of resources in the system is R and volumes of customers' resource requirements are independent identically distributed random variables with CDF $F(x)$. Customers arrive according to a Poisson process with intensity λ and the serving times have exponential distribution with rate μ. We assume here that resource requirements are independent of arrival and serving processes.

Let $\xi(t)$ be the number of customers in the system at moment $t>0$ and $\gamma(t)=\left(\gamma_1(t),....,\gamma_{\xi(t)}(t)\right)$ - the vector of occupied resources by each customer. The system behavior is described by the stochastic process $X(t)=\left(\xi(t),\gamma(t)\right)$ over the set of states $S=\left\{(n,r_1,...,r_n):0\le n\le N, r_i\ge 0, \sum_{i=1}^{n} r_i \le R\right\}$. Figure 1 shows the scheme of the queuing system.

Let t_i be the moment of arrival of the *i*-th customer. If upon arrival of the *i*-th customer with resource requirements r_i, the system does not have free servers ($\xi(t_i)=N$) or there is not enough unoccupied resources to meet the resource requirements, i.e. $r_i > R-\sum_{j=1}^{\xi(t_i)} \gamma_i(t_i)$,

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

then the customer is lost. If $\xi(t_i) < N$ and $r_i \le R - \sum_{j=1}^{\xi(t_i)} \gamma_i(t_i)$, then the customer is accepted and occupies r_i resources. Upon the departure of a customer, it releases the server and all resources that were occupied by it.

In [5], the described system was analyzed and formulas for stationary probabilities were obtained. Stationary probabilities

$$q_0 = \lim_{t\to\infty} P\{\xi(t) = 0\}, \quad (1)$$

$$q_n(x) = \lim_{t\to\infty} P\left\{\xi(t) = n, \sum_{i=1}^{n} \gamma_i(t) \le x\right\}, \quad (2)$$

are given by formulas (3) and (4):

$$q_0 = \left(1 + \sum_{n=1}^{N} \frac{\rho^n}{n!} F^{(n)}(R)\right)^{-1}, \quad (3)$$

$$q_n(x) = q_0 \frac{\rho^n}{n!} F^{(n)}(x), 1 \le n \le N, 0 \le x \le R, \quad (4)$$

where $F^{(n)}(x)$ is the *n*-fold convolution of resource requirements CDF $F(x)$ and $\rho = \lambda / \mu$ is the offered load.

The main characteristics of interest in the model are blocking probability *B* and the average volume of occupied resources *b*. In [5], the formulas for the characteristics were derived:

$$B = 1 - q_0 \sum_{n=0}^{N-1} \frac{\rho^n}{n!} F^{(n+1)}(R), \quad (5)$$

$$b = q_0 \sum_{n=1}^{N} b_n \frac{\rho^n}{n!}, \quad (6)$$

where $b_n = \int_0^R x F^{(n)}(dx)$.

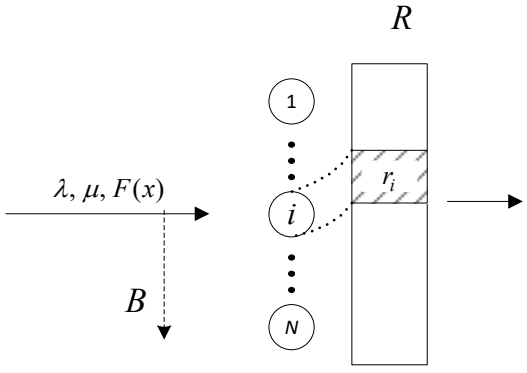


Figure 1. Resource queueing system scheme

The main challenge in evaluating blocking probability and average volume of occupied resources is calculating *n*-fold convolutions of $F(x)$. For some special types of $F(x)$, its convolutions may be evaluated analytically and then characteristics of interest are calculated easily according to formulas (5) and (6). But in most cases, numerical calculation of convolutions is inevitable. For that reason, we developed the simulation tool for limited resources queuing systems.

III. SIMULATION TOOL DESCRIPTION

In this section, we describe the developed simulation tool. Figure 2 shows block-diagram of the simulation tool.

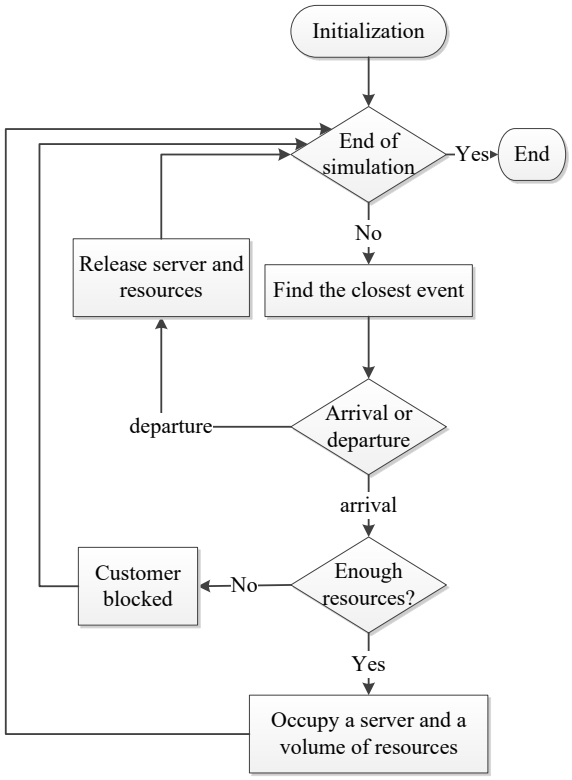


Figure 2. Block-diagram of the simulation tool.

1) Initialization: initial parameters setup. Define two types of events: arrival and departure of customers. Set *current_time*, *total_occupied_resources* and all components of *N*-dimensional vector *resources* to zero, define next *arrival_time* according to the Poisson distribution with rate λ and set all elements of *N*-dimensional vector *departure_time* to infinity. Define *total_number_of_customers* in the simulation session.
2) Start simulation cycle. Check for the end of simulation. If not, find the closest event, set *current_time=event_time*.
   a. If the closest event is arrival, then go to 3.
   b. If the closest event is departure, then go to 5.
3) Set *resource_requirements* of the customer according to $F(x)$, check whether the customer is accepted or not.
   a. If volume of unoccupied resources is greater or equal to *resource_requirements* and there is a free server in the system, then customer is accepted, go to 4.

b. If volume of unoccupied resources is less than *resource_requirements* or there are no free servers in the system, then the customer is blocked. Set new *arrival_time,* update statistics and go to 2.

4) Define a server to serve the customer, set the corresponding element of *departure_time* vector according to the exponential distribution with rate μ and set corresponding element of *resources* vector to *resource_requirements*. Increase *occupied_resources* by *resource_requirements* and set new *arrival_time*. Update statistics and go to 2.
5) Decrease *total_occupied_resources* by corresponding element of *resources* vector and set then set this element to zero. Moreover, set the corresponding element of *departure_time* to infinity, update statistics and go to 2.

The simulation session continues until the number of arrived customers exceeds *total_number_of_customers*.

IV. SIMULATION RESULTS

In this section, we use the developed simulation tool for the analysis of machine-to-machine (M2M) traffic [4] characteristics in a LTE cell. Table 1 shows radio propagation and load parameters of the base station [1].

Table 1: Numerical example data

Parameter	Value
L	100 m
ω	10 MHz
N	1000
$p_{\max}$	0.00398 W
r_0	100 Kbit/s
λ	[5 – 200] 1/s
μ	1 s
N_0	10^{-9} W
G	197.43
κ	5

First three parameters in table 1 define cell characteristics. Here L is the range of the base station, ω is the frequency bandwidth used for M2M traffic and N is the maximum number of active devices. M2M devices are characterized by the maximum transmit power $p_{\max}$, required bitrate for data transmission r_0, the rate of device activation λ and the mean duration of data transmission μ^{-1}. Final block of in table 1 define signal propagation parameters: N_0 is the noise power, G is the propagation constant and κ is the propagation exponent.

Earlier in [7], we derived CDF of resource requirements assuming that Full Power scheduler is used on the base station:

$$F(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{L^2}\left(\dfrac{Gp_{\max}}{N_0}\right)^{2/\kappa}\left(e^{\frac{r_0}{x\omega}}-1\right)^{-2/\kappa}, & x\in(0,\varphi];\\ 1, & x>\varphi,\end{cases} \quad (7)$$

where $\varphi=\dfrac{r_0}{\omega\ln\left(\dfrac{Gp_{\max}}{N_0L^{\kappa}}+1\right)}$. The corresponding probability density function (PDF) $f(x)$ is given by

$$f(x)=\begin{cases}0, & x\notin(0,\varphi]\\ \dfrac{2r_0}{L^2\omega\kappa}\left(\dfrac{Gp_{\max}}{N_0}\right)^{2/\kappa}\dfrac{e^{\frac{r_0}{x\omega}}}{x^2}\cdot\left(e^{\frac{r_0}{x\omega}}-1\right)^{-\frac{2+\kappa}{\kappa}}, & \\ & x\in(0,\varphi].\end{cases} \quad (8)$$

Here volume of resources means share of a time slot that is needed to guarantee the bitrate r_0 [2, 3]. Thus, the total amount of resources in the system is R=1.

For the generation of random numbers with CDF $F(x)$ one needs the inverse function $F^{-1}(x)$:

$$F^{-1}(x)=\frac{\frac{r_0}{\omega}}{\ln\left(1+\dfrac{Gp_{\max}}{N_0}\left(L^2x\right)^{-\kappa/2}\right)}. \quad (9)$$

The simulation results were compared with calculations from [7], where performance measures of the queueing system were obtained using sampling of the CDF (7) and applying recurrence algorithm for discrete resource requirements (figures 3 and 4).

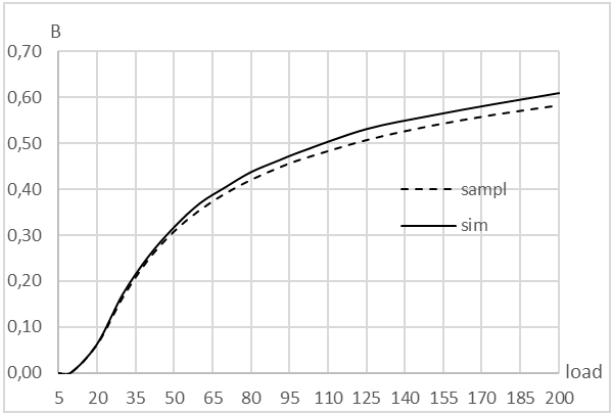


Figure 3. Blocking probability *B*.

As it is seen from figure 4, the blocking probability growth slows down with the increase of the load. The effect can be explained by different resource requirements of M2M sessions. Indeed, with the growth of the load, the system accepts mainly low-requirement sessions and blocks resource-greedy sessions. Both

(figure 3 and figure 4) show us that results of samplification and simulations are rather close.

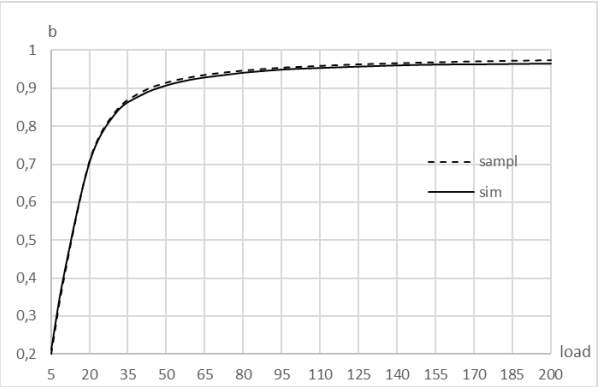


Figure 4. Average volume of occupied resources *b*.

Table 2 shows relative error of the simulations compared to the samplification method for different values of offered load. One can see that the relative error of average volume of occupied resources is nearly 1%, but when it comes to blocking probability the relative error rises to nearly 4%.

Table 2. Relative errors for *B* and *b* for various ρ .

ρ	**Relative error *B* (%)**	**Relative error *b* (%)**
5	0,354	4,641
10	9,294	1,843
20	2,038	0,642
30	4,284	0,742
40	2,613	1,056
50	2,976	0,968
60	4,117	0,860
70	3,431	0,893
80	4,061	0,757
90	3,573	0,687
100	3,686	0,689
125	4,747	0,797
150	4,127	0,776
175	4,139	0,981
200	4,602	1,085

V. CONCLUSION

In the paper, we developed the simulation tool for the queueing systems with limited resources and random resource requirements. The tool allows to evaluate performance measures of modern wireless networks with any custom radio resource scheduler.

In our future work, we plan to improve the condition for the end of simulation, so that the simulations continue until the relative error does not exceed some predefined level.

ACKNOWLEDGEMENTS

The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 16-37-60103 and No. 16-07-00766.

We thank Prof. Valeriy Naumov for the methodic assistance and scientific guidance in the preparation of this paper.

REFERENCES

[1] 3GPP TS 36.300: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN): overall description (Release 13).

[2] Galinina, O, Andreev, S, Turlikov, A, Koucheryavy, Y Optimizing Energy Efficiency of a Multi-Radio Mobile Device in Heterogeneous Beyond-4G Networks, Performance Evaluation, vol.78, 2014, pp. 18-41.

[3] Gudkova I., Samouylov K., Buturlin I., Borodakiy V., Gerasimenko M., Galinina O., Andreev S., Analyzing impacts of coexistence between M2M and H2H communication on 3GPP LTE system, Lecture Notes in Computer Science, vol.8458, 2014, pp. 162-174.

[4] Jabbar A.I.A., Fawaz Y. Abdullah Long Term Evolution (LTE) Scheduling Algorithms in Wireless Sensor Networks (WSN), International Journal of Computer Applications, vol. 121, 2015.

[5] Naumov, V.A., Samuilov, K.E., Samuilov, A.K., On the total amount of resources occupied by serviced customers, Autom Remote Control, 2016, volume 77, issue 8, pp 1419–1427.

[6] Naumov V., Samouylov K. Analysis of multi-resource loss system with state dependent arrival and service rates, Probability in the Engineering and Informational Sciences, vol. 31, issue. 4 (G-Networks and their Applications), 2017, pp. 413-419.

[7] Sopin E.S., Gaidamaka Yu.V., Markova E.V., Gudkova I.A. Performance analysis of M2M traffic in LTE network using queuing systems with random resource requirements, Automatic Control and Computer Science, in print.

[8] Tikhonenko O., Generalized Erlang problem for service systems with finite total capacity. Problems of Information Transmission, 41 (3), 2005, pp. 243–253.

AUTHOR BIOGRAPHIES

EDUARD SOPIN received his B.Sc. and M.Sc. degrees in applied mathematics from the Peoples' Friendship University of Russia (RUDN University) in 2008 and 2010, respectively. In 2013, he received his PhD degree in applied mathematics and computer science. Since 2009, Eduard Sopin works at the Telecommunication Systems Department of RUDN University, now he is an associate professor at the Department of Applied Probability and Informatics of RUDN University. His current research interests lie in the area of performance analysis of modern wireless networks and cloud/fog

computing. His e-mail address is sopin_es@rudn.university

KIRILL AGEEV was born in Aqtobe, Kazakhstan and went to the RUDN University of Russia, where he studied computer scince and obtained his degree in 2017. He is now studing as a PhD student of department of telecommunications and informatics. His e-mail address is: ageev_ka@rudn_university.

SERGEY SHORGIN received his Ph.D. degree from the Moscow State University in 1979 and Doctor of Sciences degree in 1997. Since 1992 works at Institute of Informatics Problems, Federal Research Center "Computer Science and Control", where he became an associate director in 1999. An executive editor of "Informatics and Applications" scientific journal. His current research interests are Probabilities theory, complex systems modeling, actuarial mathematics and mathematics of finance. His e-mail address is: sshorgin@ipiran.ru

AUTHOR INDEX