5-7 July 2017

Zilina - Slovakia

Proceedings of

The International Conference on

Information and Digital Technologies 2017

IEEE ESRA CERES

The International Conference on

Information and Digital Technologies 2017

5–7 JULY 2017
ZILINA, SLOVAKIA

ISBN PREPRINT
ISSN PREPRINT
IEEE Catalog Number CFP17CDT-USB

The International Conference on Information and Digital Technologies 2017

IEEE Catalog Number CFP17CDT-USB
ISBN PREPRINT
ISSN PREPRINT

PROGRAM COMMITTEE

Chair
Viergever Max, Netherlands

Co-Chairs:
Zaitseva Elena, Slovakia
Pancerz Krzysztof, Poland

Ablameyko Serge, Belarus
Androulidakis Iosif, Greece
Astola Jaakko, Finland
Aven Terje, Norway
Barach Paul, USA
Baraldi Piero, Italy
Beer Michael, United Kingdom
Berenguer Christophe, France
Bris Radim, Czech Republic
Bourek Ales, Czech Republic
Cepin Marko, Slovenia
Cimrak Ivan, Slovakia
Coolen Frank, United Kingdom
Czapp Stanislaw, Poland
Dao Phan, Vietnam
Deserno Thomas, Germany
Di Maio Francesco, Italy
Drozd Alexander, Ukraine
Filatova Daria, Poland
Fiser Petr, Czech Republic
Frenkel Ilia, Israel
Fukushi Masaru, Japan
Grondzak Karol, Slovakia
Herrler Andreas, Netherlands
Huang Hong-Zhong, China
Kameyama Michitaka, Japan
Kerntopf, Pawel, Poland
Kharchenko Vyacheslav, Ukraine
Kolowrocki Krzysztof, Poland
Krsak Emil, Slovakia
Levashenko Vitaly, Slovakia
Levitin Gregory, Israel
Lukac Martin, Japan
Majernik Jaroslav, Slovakia
Matiasko Karol, Slovakia
Moraga Claudio, Germany
Paprzycki Marcin, Poland
Pastor Luis, Spain
Pattinson Colin, United Kingdom
Pedroni Nicola, France
Perkowski Marek, USA
Podofillini Luca, Switzerland
Sasao Tsutomu, Japan
Schmidt Jan, Czech Republic
Simic Zdenko, Netherlands
Singh Suraj B., India
Soda Paolo, Italy
Soszynska-Budny Joanna, Poland
Stankevich Sergey, Ukraine
Stankovic Radomir, Serbia
Steinbach Bernd, Germany
Subbotin Sergey, Ukraine
Vintr Zdenek, Czech Republic
Xie Min, Hong Kong
Yanfu Li, France
Zhao Qiangfu, Japan
Zio Enrico, France / Italy

Organizing Committee

Chair: Kostolny Jozef, Slovakia
Chovancova Olga, Slovakia
Endersova Brita, Slovakia
Ilovska Anna, Slovakia
Kvet Michal, Slovakia
Kvet Marek, Slovakia
Kvassay Miroslav, Slovakia
Ponomarenko Vladimir, Slovakia
Rabcan Jan, Slovakia
Rusnak Patrik, Slovakia
Vaclavkova Monika, Slovakia
Varga Michal, Slovakia

CONTENTS

INTRODUCTION

Dear participants,

It is a great pleasure to welcome you to International Conference on Information and Digital Technologies (IDT 2017). The Conference is organized and hosted by the Faculty of Management Science and Informatics of the University of Žilina. The IDT history goes back to the Conference on Digital Technologies (DT) that was founded at the Faculty of Electrical Engineering of our University. The DT allows growing of new conference with widened scope that is started in year 2015 as IDT. IDT 2017 provides a forum for presentation and discussion of scientific contributions covering the theories and methods in the field of information and digital technologies, and their applications to a wide range of industrial, civil and social sectors and problem areas. IDT 2017 is also an opportunity for researchers, practitioners, academics and engineers to meet, exchange ideas, and gain insights from each other. IDT 2017 offers a multidisciplinary platform to address the technological, societal and financial aspects of information systems.

The conference program is divided into some workshops that cover numerous aspects of information and digital technologies:

- Workshop on Biomedical Technologies,
- Workshop on Reliability Technologies,
- Workshop on New Frontier Information Digital Technology,
- Workshop on Dynamical Systems and Real World Applications
- Workshop on CERES: Modern Experience on Young Researchers Organization.

We would like to thank colleagues who organized these workshops.

Initially, more than 130 papers were submitted for the conference. The final number of 73 papers to accepted is the result of a rigid reviewing procedure performed by reviewers from all over the world. The workshop chairs mainly organized the review process and the process was made by a large number of reviewers, which are gratefully acknowledged for their contributions to the improvement of quality of the accepted papers. At least two anonymous reviewers reviewed each paper in order to ensure fair and high-quality reviews. In addition to regular sessions, IDT 2017 offers distinguished keynote lectures. We thank the IDT 2017 Invited Speakers for offering their unique perspectives on information technologies at the Conference.

In addition, we are hosting the Workshop on CERES. Participants of this workshop are PhD students and young researchers dominantly. This workshop was formed under project TEMPUS CERES (Centers of Excellence for young RESearchers, reg.no. 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES) and hold the second time in the framework of the Conference. We gratefully acknowledge the Faculty of Management Science and Informatics of the University of Žilina, European Reliability and Safety Association (ESRA) and The Czechoslovakia section of IEEE for the sponsoring, organizational and technical support.

We also thank all the contributed paper authors for their submissions and presentations.

July 2017
Organization team of IDT 2017

Evaluation of Healthcare Enterprise Information Systems: A Structural Equation Model

Ahed Abugabah
College of Technological Innovation, Zayed University, UAE
Email: ahed.abugabah@zu.ac.ae

Abstract: *The aim of this study is to evaluate Healthcare Enterprise Information Systems from both system and user perspectives. The study used a structural equation modelling to analyse the impacts of Healthcare Enterprise Information Systems on employees' performance and health care service provided by healthcare organizations. This was measured using several dimensional factors including, system quality, service quality and information quality. The study used a questionnaire survey to gather data from healthcare organizations and their staff in different departments and at different levels. The study findings indicated that Hospital Enterprise Information Systems (HEIS) helped improve service quality by providing complete and accurate information related to patients and patient service processes.*

Keywords: *Healthcare Enterprise Information Systems, Information Systems Models, Healthcare service, System Quality.*

I. Introduction

Information systems have significant potential to improve patient service, health care efficiency, patient satisfaction and health care process. For example, physician systems can help reduce the serious medication error. Over the past years, substantial initiatives were proposed in health care industry to prompt the adoption and implementation of healthcare information systems and their applications [1]. As a consequence, health care organizations made huge investments and spent enormous amounts of money and resources in shifting to or adopting Health care Enterprise Information Systems with the hope of improving organizational performance and patient service, and operational efficient [2].

General speaking, it is believed that the use of health care Information Systems provides great opportunities for improving the healthcare process, service quality and the efficiency health care operations. Recent research in health care industry reported that if these organizations do not adopt relevant and advanced applications of healthcare information systems, they will become ineffective and face many operational and functional deficiencies [1,2]. From practical perspective, health care practitioners and managers continue their quest for more efficient techniques and methods of gathering patient information to meeting patient requirements and work needs. These requirements place increased demands on healthcare information systems and their applications [4,5,6]. This also makes it interesting to evaluate the actual benefits of the Health care Enterprise Information Systems (HEIS) and their role in health care organizations [3].

Consequently, researchers as well as stakeholders of health care organizations seem to be interested in evaluating the outcomes of HEIS projects implementations and analysing the level to which HEIS applications are able to meet the varied needs of groups and individual in their organizations [3]. Fortunately, there has been a number of Information Systems (IS) models and frameworks available for researchers in this endeavour [2,6].

This study therefore attempts to evaluate the (HEIS) and assess their impacts on health care service quality and staff performance from both system and user perspectives. The study uses a structural equation modelling to analyse the impacts of healthcare information systems on employees and organizational performance and healthcare services provided by hospitals. The study is a part of a large research project investigating information systems in health care organizations at different levels and from different user perspectives.

II. Literature Review

As an attempt to build a rigor foundation for this research, prior IS and heath care informatics research were reviewed with a focus on IS models and IS evaluation frameworks research in health care context. This provided a theoretical basis for the main factors investigated in this study and their dimensions and measurements. This also helped avoid the limitations of previous evaluation frameworks and make improvements to the current study framework [2,5].

Most existing evaluations of HEIS focused on either technical issues or behaviour issue leading to improper evaluation and or poor evaluation results. Ammenwerth et al. [3] stated that, despite the growing body of research literature on healthcare information systems evaluation, there have been many concerned issues in conducting a system evaluation. These issues include the complexity of the systems themselves, the evaluation goals and the lack of clarity [12, 13, 14]. Whilst many frameworks proposed for HIS may have been complementary to one another because they focus on different aspects of evaluation,

however none of them provided explicit evaluation categories for the evaluators, including technical, behavioral and systematic categories [11]. In addition, one of the major shortcomings in previous HEIS frameworks used in previous studies is the lack of a proper reporting format for HEIS evaluation. The lack of consideration in addressing the what, why, who, how and when of the HEIS evaluation is evident [5,6,7, 19]. Whilst previous frameworks complement each other, it appears that caution should be exercised in applying the frameworks based on their shortcomings [10].

A. Research Model

In order to build a rigor research framework, the research model was built based on previous IS models. Namely, the IS success model, the Task-Technology Fit (TTF) model and Technology Acceptance Model (TAM) [8,9,10]. These models have been used widely in information technology/systems fields to evaluate numerous types of IS applications in various business environments, and have been proven to provide valid findings [13, 14, 15, 18].

Notwithstanding the expansion of frequent frameworks for HEIS evaluation, there appears to be a lack of consensus about what aspects to evaluate, when to conduct the evaluation, and how to carry out such evaluation [7, 8, 9, 16]. In addition to who are the main stakeholders of the evaluation as well as why carrying that evaluation, meaning the main the purpose of the evaluation must be clearly identified.

III. Research Model and Hypotheses

That is, jointing these models together will provide a more comprehensive and solid research framework that can examine many factors including the HEIS and user related factors, as illustrated in Figure 1. Besides, the integrated model will help overcome the limitations of each model standing alone [6,7,8]. The study model along with the hypotheses examined in this study are illustrated in figure 1

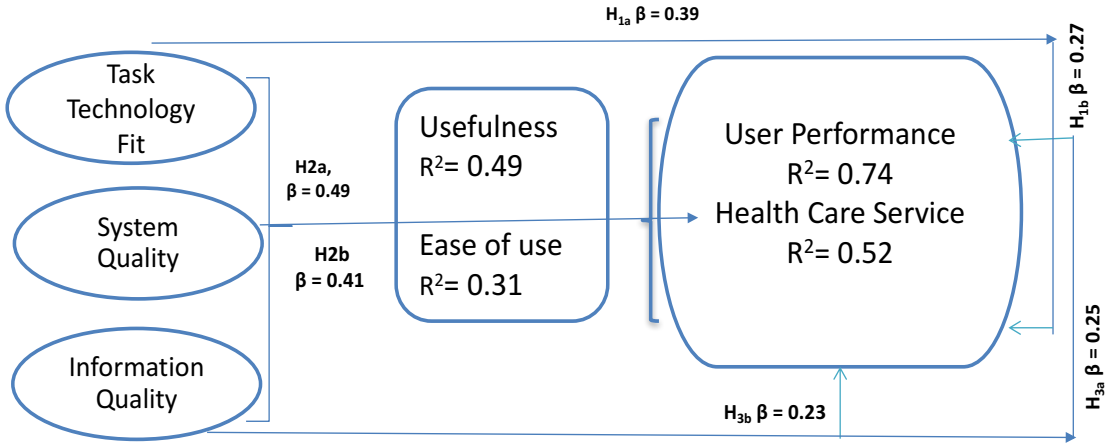


Figure 1. The Study Model

.

B. Hypotheses

This research examined the following hypotheses, which are derived from the integrated models and based on an extensive review of the IS in health care

- H1a: TTF is associated with user performance and health care service quality in health care organizations.
- H1b: TTF is associated with health care service in health care organizations.
- H2a: system quality of HEIS is associated with user performance in health care organizations
- H2b: system quality of HEIS is associated with health care service in health care organizations.
- H3a: Information quality of HEIS is associated with user performance and health care service quality in health care organizations.
- H3b: Information quality of HEIS is associated with health care service in health care organizations.

IV. Methodology

The results reported in this study represent a part of larger research project investigating HEIS. The study utilized a survey questionnaire to collect data from different types of users including decision makers and IT managers in healthcare organizations. The questionnaire was constructed based on an extensive review of the literature in the areas of information technology and health care information systems [3,7,8,9, 17].

The questionnaire first was pre-tested with 5 users and academicians in order to improve the face validity. The study was conducted in five hospitals that had implemented the HEIS more than two years ago in UAE. A total of 197 questionnaires were analysed after data filtration and cleaning as summarized in Table 1.

Table 1. Sample Characteristics (N=197).

Factor	Classification	No	%
Gender	Male	104	52.8%
	Female	93	47.2%
Qualification	Vocational	11	5.6%
	Bachelor	166	84.3%
	Postgraduate	20	10.1%
Position	Physician	97	49.2%
	Nursing	52	26.4%
	Administrative	21	10.7%
	IT staff	27	13.7%

A. Reliability and Validity

The validity of the study model was examined using item reliability, discriminant validity and convergent validity. The reliability of the measurement items was confirmed by calculating factor loadings. A factor loading greater than 0.5 is considered as an acceptable threshold for item reliability [11,4]. As presented in Table 2 , the factor loadings for all items exceeded 0.5, indicating acceptable level of reliability. In relation to construct validity, the Average Variance Extracted (AVE) was used to measure construct validity [4]. An AVE must be higher than 0.5 to ensure that variance captured by the construct is greater than the measurement error. As shown in Table 2 all values of the AVEs were above 0.60, indicating that construct validity was satisfied.

Table 2. Correlations and AVE values.

Factors	AVE	TTF	SQ	IQ	UP	HS	Alphaα
TTF	0.89	(0.94)					0.84
IQ	0.83	0.78	(0.91)				0.87
SQ	0.87	0.63	0.59	(0.93)			0.81
UP	0.89	0.52	0.39	0.47	(0.94)		0.85
HSQ	0.72	0.48	0.29	0.59	0.41	(0.85)	0.82

Table 3. Goodness-of-fit

Criteria/Indices	Recommended Value	Measurement Model	Structural Model
Chi-square (χ2)	- - -	307	301
Degree of Freedom	- - -	192	195
χ2/df	>2	1.60	1.54
GFI	>0.90	0.94	0.93
NFI	>0.90	0.91	0.92
NNFI	>0.90	0.93	0.93
CFI	>0.90	0.92	0.92
RMSEA	>0.08	0.81	0.81

Furthermore, discriminant validity was tested for each construct using the square root of its AVE, which must be greater than the correlation coefficient with other constructs in the model. As shown in Table 2 all square root values are higher than the correlation coefficient of all constructs demonstrating thus a high level of discriminant validity.

V. Analysis and Findings

A Structural Equation Modelling (SEM) was used to test the research structural model. The structural model was examined by using the goodness-of-fit (GFI) of the hypothesized research model [4], the 2/df, Confirmatory Factor Analysis (CFI), and the Root Mean Square Error of Approximation (RMSEA) were used to test the goodness-of-fit of the structural model. , the goodness-of-fit indices for the structural model were x2/df=1.54, GFI=0.93, CFI=0.93, Normed Fit Index (NFI) =0.92, and RMSEA=0.81, indicating that the model provides a good fit with the data

Table 4. Results of hypotheses tests

Hypothesis	Path Coefficient β	t-Value	Support
H1aTTF→ performance	0.39	6.8	Yes
H1b TTF→ Health care service	0.27	3.5	Yes
H2a System Quality → performance	0.49	6.3	Yes
H2b System Quality → Health care service	0.41	3.0	Yes
H3a Information Quality → performance	0.25	4.5	Yes
H3b Information Quality → Health care service	0.23	3.4	Yes

A Structural Equation Modelling (SEM) was used to test the research structural model. The path coefficients and t-test values were used to examine the relationships between all study variables.

The findings indicated a significant support for all hypotheses. The path coefficient between all study factors indicated positive relationships significantly. As shown in table 4 P<0.05. Task technology fit significantly affects both health care service and user performance, where (β=0.27, t=3.5, and (β=0.39, t=6.85) respectively. This indicates that whenever HEIS provides the appropriate level of fitness with staff needs and work requirements, health care staff tend to realize more system benefits and significant impacts on both performance and healthcare service quality. That is, having a HEIS with the right outputs on time will lead to enhanced patient's services as it will save staff time and expedite service process with more accuracy and less errors. Thus, both H1a and H1b are supported as listed in Table 4.

The findings also indicated that the SQ of HEIS was the most powerful variable that affects performance and service quality (β=0.41, t=3.0, and β=0.49, t=6.39, p<0.05) respectively. Hence, both H2a and H2b were significantly supported. Similarly, the findings indicated a positive effect of IQ of HEIS on both service quality and User Performance (UP) β=0.23, t=3.4 and, (β=0.25, t=4.5 p<0.05) respectively and therefore, H3a and H3b are

supported. Unexpectedly, the study revealed no significant impact of IQ on Perceived Ease of Use (PEOU). However, other factors such as system quality showed relatively significant but small impact on both PU and PEOU, as shown in figure 1. To sum up, the findings indicated that HEIS plays a critical role in helping health care staff performing their job and tasks more effectively, leading to improved health care services.

VI. CONCLUSIONS AND IMPLICATIONS

This study aimed to evaluate HEIS from user perspective in order to identify main benefits and impacts of the HEIS on users and service quality. The main conclusions of the study bring into our attention some important highlights on critical factors relevant to the evaluation of HEIS at both system and user levels.

Evaluation of HEIS is in itself a complex and multidimensional task. A proper framework then must address the what, why, who, how and when of the HEIS evaluation is evident and provide appropriate solutions from health care context. Some observations have been clearly made in this study. These include identifying a need for an evaluation framework that clearly answers the aforementioned questions raised by many researchers. An evaluation framework must consider also contextual factors when evaluating health care information systems. In addition to the need for a better approach for identifying the right system stakeholders, instead of using simple classifications.

The finding of the study indicated that system incompatibility leads probably to disrupted system interaction to system users including medical staff and administrative staff. Consequently, information and data needed in some certain circumstances are inaccessible and/or unreadable, causing serious difficulty for doctors to and managers to make decision related to patient diagnosis or treatment. Poor system functionality is also attributed to different factors, such as administration disputes and missing user and task requirements. Therefore, technology fit between the system functionality and user needs and task requirements is a priority that should be considered by both system designers and health care management.

Furthermore, system quality of HEIS is associated with ease of use and staff performance leading to better services quality. The study revealed that when the HEIS is designed to be easy to use and user friendly, better system usage will occur leading to more positive impacts on both user tasks and service quality provided to patients. This can happen because of time saving in answering questions and obtaining relevant and complete information when needed. Relevant training on the system itself also helps users to become competent and utilize the system functionality more effectively.

Information quality is measured from systems outputs and user interface displays in various forms such as patient report, record, and prescription. Information quality in many cases is subjected to user perspective on information characteristics such as accuracy, consistency and completeness. Poor quality of information retrieved from the HEIS will lead to medical errors and task mistakes, leading consequently to poor performance and poor services quality.

Overall, the potential of HEIS system impact is influenced by system ease to use and system quality, which in return affect the level of system use, leading to major effects on patient services and user performance

REFERENCES

[1] A. Andargoli, , D. Rajendran and A Sohal, Health information systems evaluation frameworks: A systematic review. International Journal of Medical Informatics, (97), January 2017; 195-209

[2] A.Azizi, R. Abdolkhani, The case study of effect of hospital information system in improvement of Razi hospital performance, J. Jundishapour 2 (2011) 185–190.

[3] Alfarraj, O and Abugabah, A (2017). Extending Information System Models to the Health Care Context: An Empirical Study and Experience from Developing Countries, The International Arab Journal of Information Technology. 14 (2), p159-167.

[4] Abugabah, A and Alfarraj, O, 2015. Issues to Consider in Designing Health Care Information Systems: A User-centred Design Approach, electronic Journal of Health Informatics, 9(1), pp1-15.

[5] Abugabah, A Sansogni, L and Alfarraj, O., 2015. Evaluating the impact of ERP systems, International Journal of Information and Learning Technology, 32 (1), pp. 45-64

[6] Ahed Abugabah, Louis Sanzogni (2014). Exploring factors affecting end-user performance of information systems: the International Journal for Infonomics, 7 (3/4), 956-973.

[7] E. Ammenwerth, N. de Keizer, An inventory of evaluation studies of information technology in health care: trends in evaluation research 1982–2002, Methods Inf. Med. 44 (2005) 44–56.

[8] J. Hair, B. Babin, A. Money and P. Samouel, P, Essential of business research methods. John Wiley & Sons: United States of America, 2003.

[9] J. Sligo, R. Gauld, V. Roberts and L. Villa, International Journal of Medical Informatics, (97), January 2017; 86-97. A literature review for large-scale health information system project planning, implementation and evaluation

[10] L. Ahmadian, S. Nejad and R. Khajouei, Evaluation methods used on health information systems (HISs) in Iran and the effects of HISs on Iranian healthcare: A systematic review, International Journal of Medical Informatics, 84, (6), 2015; 444-453

[11] L. Lapointe, M. Mignerat and I. Vedel , The IT productivity paradox in health: A stakeholder's perspective. International Journal of Medical Informatics. 2011. 80 (2):102-115.

[12] L. Politi, S. Codish, I. Sagy and L. Fink, Use patterns of health information exchange systems and admission decisions: Reductionistic and configurational approaches, International Journal of Medical Informatics, 84, (12), 2015; 1029-1038

[13] M. Beuscart-Zéphir, S. Pelayoa and S. Bernonvillea, Example of a human factors engineering approach to a medication administration work system: potential impact on patient safety. International Journal of Medical Informatics. 2010. 79(4): E43-E57.

[14] M. Lluch, Healthcare professionals' organizational barriers to health information technologies: A literature review. International Journal of Medical Informatics. 2011. 80(12): 849–862

[15] M.M. Yusof, J. Kuljis, A. Papazafeiropoulou, L.K. Stergioulas, An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit), Int. J. Med. Inf. 77 (2008) 386–398.

[16] P. Sockolow, P. Crawford, H. Lehmann, Health services research evaluation principles, Methods Inf. Med. 51 (2012) 122–130.

[17] R. Khajouei, A.A. Azizi, A. Atashi, Usability evaluation of an emergency information system: a heuristic evaluation, J. Health Admin. 16 (2013) 61–72.

[18] S. Martikainen, M. Korpela and T. Tiihonen. User participation in healthcare IT development: A developers' viewpoint in Finland: International Journal of Medical Informatics, 2014. 83 (3); p 189–200

[19] T. Teixeira, C. Ferreira, B. Santos, User-centered requirements engineering in health information systems: Computer Methods and Programs in Biomedicine. 2012. 106 (3):60

Discrete-Continual Method of Analysis of the Coupled System “Plate - Soil Foundation” in Context of Microseismic and Gravitational Processes in Foundation

Prof., Dr.Sc. Pavel Akimov
Russian Academy of Architecture and Construction Sciences, Moscow, Russia;
Scientific Research Center “StaDyO”, Moscow, Russia;
Department of Applied Mathematics,
National Research Moscow State University of Civil Engineering, Moscow, Russia;
e-mail: pavel.akimov@gmail.com

Prof., Dr.Sc. Marina L. Mozgaleva
Department of Applied Mathematics,
National Research Moscow State University of Civil Engineering, Moscow, Russia
e-mail: marina.mozgaleva@gmail.com

***Abstract*—This paper is devoted to discrete-continual method of analysis of coupled system "plate-soil foundation" in context of microseismic and gravitational processes in foundation. This method is called discrete-continual because it presupposes finite element approximation with respect to space coordinates while corresponding problem remains continual in time. This reasonable combination of numerical and analytical approaches offers numerous advantages and provides a high degree of accuracy. Time discretization is not required and therefore this appreciable fact in its turn leads to considerable reduction in computing time and number of operations. Due to specificity of initial data standard step-by-step methods couldn’t meet the challenge for the considering problem. Brief introduction to the problems of geological efficiency of microseismic processes in foundations is provided. Original formulation is based on elastic theory and linear creep theory. Special soil model allowing propagation of plastic deformations and nonlinear model of soil foundation under cyclic loadings are used. Being foreground subject of the paper, discrete-continual method is considered in the aspects of theoretical basics, numerical implementation and software packages. Numerical example is presented as well.**

Keywords—discrete-continual method, semianalytical methods, finite element approximation, microseismic processes in foundation, gravitational processes in foundation

I. Introduction

Discrete-continual methods of structural analysis for static problems have been considered in [1-4]. However, these methods embraces much more wide-ranging field of application. This paper in particular is devoted to discrete-continual method for the problem of analysis of coupled system “plate-soil foundation” in context of microseismic and gravitational processes in foundation. The distinctive method is discrete-continual because it presupposes finite element approximation with respect to space coordinates while problem remains continual in time. This reasonable combination of numerical and analytical approaches offers numerous advantages providing a high degree of accuracy. We should stress that time discretization is not required and therefore this appreciable fact in its turn leads to considerable reduction in computing time and number of operations. Due to specificity of initial data standard step-by-step methods couldn’t meet the challenge for the considering problem. Finally note that starting point for the presenting method has been given by formulation of the problem presented in [5].

II. Problems of Geological Efficiency of Microseismic Processes in Fgoundations

Problems of geological efficiency of microseismic processes in foundations have been poorly studied field of structural mechanics and soil mechanics until the last decade of 20th century. However, in our days they are problems of present interest especially in congested areas, industrial zones, megapolises, foundations of heavy buildings including nuclear power stations [6]. Bad cases of this geological efficiency include segmentation of geological covering, nonuniform soil consolidation, disturbances in the natural flow of groundwater and storm water, propagation of karst processes etc. Major problems of geoecology including questions from rock mechanics, geology, hydrogeology, geochemistry dealing with safety of construction are considered in [6]. The aim of this paper is to present results of particular geoecological problem closely coordinated with mathematical modelling, full-scale models, tests and verifications of geoecological efficiency of man-caused vibro-seismic processes in foundations.

Corresponding instrumental and theoretical research has shown that one of the main reasons of accelerated propagation of above mentioned processes deals with nonuniform microcyclic fatigue of soil foundation due to multiyear exposure to natural and man-caused microseismic and

microgravity loads. In accordance with results of statistical analysis soil foundations in various districts of Moscow undergo more than 10 billions cycles of microloadings for a period of 10 years. Researches indicated substantial increase of geoecological efficiency in cases with severe heterogeneity of fine structure of geological section of foundation or nonuniform vibro-seismic excitation. These reasons lead to decrease in geotechnic safety, banks, nonuniform subsidence of buildings and line network. So unpleasant problems of present-day cities escalates both in seismoactive and seismopassive areas all over the world if buildings have high-Q geodynamic resonances [7, 8], which concentrate energy of microseismic excitation of foundations [9] and capable to increase its intensity more that 75 times [10].

Efficiency and risks of effects caused by microseismic loadings on cyclic fatigue of heterogeneous soil foundations remain rather difficult for design estimations. Thus, instrumental verification of proposed theory and algorithms for analysis of geoecological effectiveness of vibro-seismic excitation of heterogeneous soil foundations has been carried out [5]. This theory allows investigator to specify risky zones with respect to negligible consolidation ("fatigue") of soil foundation and take into account modal geoecological efficiency for various types of seismic waves including longitudinal waves, transverse waves, surface waves etc. Models from [5,11,12] are used for describing soil foundation.

III. Formulations of problems

A. Physical parameters of materials

Analysis of the real coupled system "plate-soil foundation" and corresponding model is under consideration.

Material of the plate is concrete. Let E_p be the modulus of elasticity of concrete; ν_p is the corresponding Poisson's ratio; ρ_p is the density of material.

As for soil foundation both pattern and practical design diagrams are considered. Pattern ones presuppose homogeneous soil foundation while practical ones deal with heterogeneous foundations with multilayer structures. Particularly let E_i, ν_i, ρ_i be the modulus of elasticity, the Poisson's ratio and the density of material of the i-th layer.

B. Loads and effects

Within static analysis concentrated loads applied to the plate with values P, $2P$ and distributed uniform load with value q are used. Besides, we optionally take into account soil dead weight. As for dynamic analysis, we normally take $T = 1-20$ years as a design period for dynamic analysis. Applied loads are varied in accordance with the formula

$$P(t) = P \cdot K(t), \quad 0 \le t \le T, \tag{1}$$

where

$$K(t) = (g + \Delta g^{ms} + \Delta g^{gr}) / g; \tag{2}$$

$$\Delta g^{ms} = \Delta g^{ms}_{\max} \sin(2\pi t / t_{ms}); \tag{3}$$

$$\Delta g^{gr} = \Delta g^{gr}_{\max} \sin(2\pi t / t_{gr}); \tag{4}$$

Δg^{ms} is the microseismic component of increment of free fall acceleration $g = 9.8\ \mathrm{m/s^2}$; Δg^{gr} is the gravity component of increment of free fall acceleration; $\Delta g^{ms}_{\max}$, $\Delta g^{gr}_{\max}$ are corresponding peak values; t_{ms} is the period of microseismic oscillations; t_{gr} is the period of gravity oscillations. According to research [5] we have $t_{ms} = 0.1$ s, $t_{gr} = 20$ min.

C. Design diagrams

We have considered more than twenty design diagrams: homogeneous structure, horizontal planar multilayer structure, inclined multilayer structure, wedge-shaped structure, folded-plate structure, synclinal structure, antisynclinal structure, heterogeneous structure (Fig. 1, for instance).

D. Design models

Two-dimensional problem of elasticity is used for initial analysis of each design diagram.

Let us consider linear creep theory as a design model at the distinctive case. Let T be design period for formulation allowing soil creep. Let ε_{ij} be deformation (strain) in x_i-axis direction, perpendicular to x_j-axis direction. Let σ_{ij} be stress in x_i-axis direction, perpendicular to x_j-axis direction.

Basic formulas for deformations (strains) of plate and soil foundation have the following form:

$$\varepsilon_{11}(t) = \frac{1}{E_0}[\sigma_{11}(t) - \nu\sigma_{22}(t) + \int_0^t [\sigma_{11}(\tau) - \nu\sigma_{22}(\tau)]L(t-\tau)d\tau]; \tag{5}$$

$$\varepsilon_{22}(t) = \frac{1}{E_0}[\sigma_{22}(t) - \nu\sigma_{11}(t) + \int_0^t [\sigma_{22}(\tau) - \nu\sigma_{11}(\tau)]L(t-\tau)d\tau]; \tag{6}$$

$$\varepsilon_{12}(t) = \frac{2(1+\nu)}{E_0}\left\{\sigma_{12}(t) + \int_0^t \sigma_{12}(\tau)L(t-\tau)d\tau\right\}, \tag{7}$$

where E_0 is the elastic modulus of elasticity; ν is corresponding Poisson's ratio;

$$L(t-\tau) = \delta \exp[-\delta_1(t-\tau)]. \tag{8}$$

Basic formulas for stresses in plate and soil foundation have the following form:

$$\sigma_{11}(t) = \frac{E_0}{1-\nu^2}[\varepsilon_{11}(t) + \nu\varepsilon_{22}(t) - \int_0^t [\varepsilon_{11}(\tau) + \nu\varepsilon_{22}(\tau)]R(t-\tau)d\tau]; \tag{9}$$

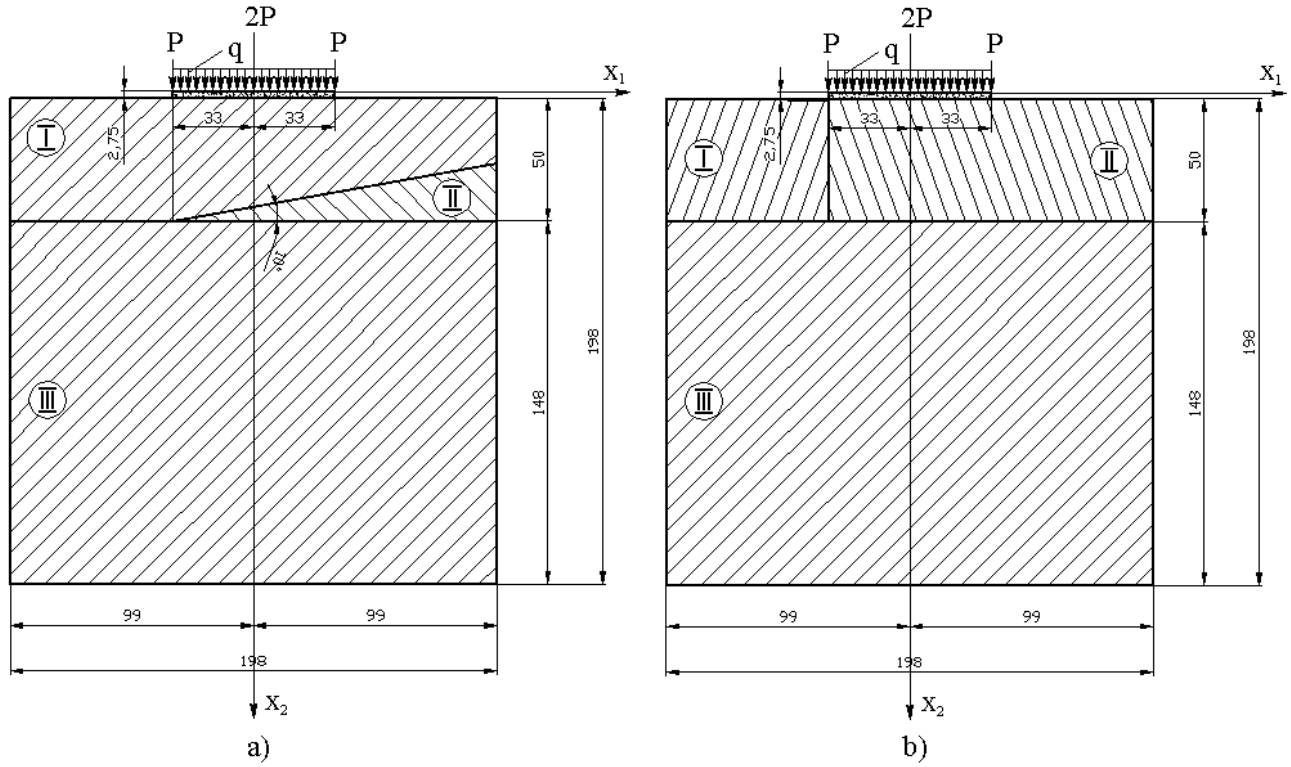


Figure 1. Design diagrams: wedge-shaped structure (a); folded-plate structure (b).

$$\sigma_{22}(t)=\frac{E_0}{1-\nu^2}[\varepsilon_{22}(t)+\nu\varepsilon_{11}(t)- -\int_0^t[\varepsilon_{22}(\tau)+\nu\varepsilon_{11}(\tau)]R(t-\tau)d\tau]; \quad (10)$$

$$\sigma_{12}(t)=\frac{E_0}{1-\nu^2}\left\{\varepsilon_{12}(t)-\int_0^t\varepsilon_{12}(\tau)R(t-\tau)d\tau\right\}; \quad (11)$$

$$R(t-\tau)=\delta\exp[-(\delta+\delta_1)(t-\tau)]. \quad (12)$$

Here δ,δ_1 are parameters of creep of plate. According to RIEM research we have $\delta=0.07605$ 1/day; $\delta_1=0.03$ 1/day for plate and $\delta=0.001-0.02333$ 1/day; $\delta_1=0.1813-0.255$ 1/day for soil foundation.

Soil model from [17-19] is used to describe behavior of the foundation. Generally we applied simplified formulas for determination of soil deformations recommended by [5]:

$$\bar{\varepsilon}=\bar{\varepsilon}_0+\frac{1}{E_0}\Delta\bar{\sigma}_0+[\bar{\varepsilon}^p(n^*)-\bar{\varepsilon}_0]_-(1-\exp(-\beta n)), \quad (13)$$

where

$$[\bar{\varepsilon}^p(n^*)-\bar{\varepsilon}_0]_-=\begin{cases}\bar{\varepsilon}^p(n^*)-\bar{\varepsilon}_0, & \bar{\varepsilon}^p(n^*)-\bar{\varepsilon}_0<0\\ 0, & \bar{\varepsilon}^p(n^*)-\bar{\varepsilon}_0\ge 0;\end{cases} \quad (14)$$

$$\bar{\varepsilon}_0=[\varepsilon_{11}^0\ \ \varepsilon_{22}^0\ \ \varepsilon_{12}^0]^{\mathrm{T}};\quad \bar{\sigma}_0=[\sigma_{11}^0\ \ \sigma_{22}^0\ \ \sigma_{12}^0]^{\mathrm{T}}; \quad (15)$$

ε_{ij}^0 are strain components; σ_{ij}^0 are stress components; $i=1,2;\ j=1,2$; E_0 is the elastic modulus of elasticity; $\Delta\bar{\sigma}_0$ is the increment of stress components, caused by variation Δg of free fall acceleration $g=9.8\ \mathrm{m/s^2}$;

$$\Delta\bar{\sigma}_0=k\bar{\sigma}_0\approx 0.1\bar{\sigma}_0;\ \ k=1-(g+\Delta g)/g. \quad (16)$$

$\bar{\varepsilon}^p(n^*)$ is the deformation corresponding to complete stabilization. It is defined equal to 0.1 componentwise for all formations. n^* is the number of loading cycles till stabilization in accordance with [5]; β is the soil parameter,

$$\beta=10/n^*. \quad (17)$$

Cyclic loadings of soil foundations cause additional settlements and banks leading to abnormal working conditions or sometimes to emergency conditions [11-15]. This problem is habitual for nuclear, thermal and water power stations causing cyclic low-frequency effects in soil foundations of these buildings. Their possible sources are gravidinamic processes in topsoil. There are no methods at present for quantitative estimates of parameters (such as amplitude and frequency) of gravidinamic effects in foundations of heavy buildings. In this connection we may only specify certain disturbances on the boundaries of considering domain of soil body or uniformly distributed loads throughout soil body. This approach allows investigator forecasting of additional

deformations in arbitrary point of foundation on the assumption of stabilized stress condition under soil dead weight and building dead weight. Initial stress condition is apparently heterogeneous with respect to depth and strike of soil body as well as in contact zone between bedplate and soil foundation. Thus, additional deformations caused by cyclic effects are different in different points of soil foundation due to their substantial dependence from initial stress conditions (i.e. from mean stress $\sigma(x_1, x_2, x_3)$, shearing-stress intensity $\tau_i(x_1, x_2, x_3)$, proximity to limit state τ_i / τ_i^* in the considering point of soil foundation, amplitude and frequency of cyclic loading). Solution of this complicated geomechanical problem is in account with stress analysis of soil foundation subjected to static loads and soil dead weight on the one hand and additional stress analysis caused by cyclic effects throughout soil body on the other hand. Corresponding changes in strain-stress distribution cause changes in stress condition of reinforced concrete bedplate of structure. Therefore we have transformations in diagrams of reaction pressure, moments etc. Stress and strain parameters of soil foundation under static and subsequent cyclic loadings are required for solving problems. It is rather complex problem in itself. However, these parameters have been determined in [5,11,12]. Corresponding equations of soil state subjected to cyclic loadings are used in distinctive paper. Formulation of corresponding problem was given by [5].

Let σ_1, σ_2 be normal stresses in x_1 and x_2-axis directions. Consequently we have

$$\sigma_3 = \nu(\sigma_1 + \sigma_2). \tag{18}$$

Let $\varepsilon_1, \varepsilon_2$ be linear deformations in x_1 and x_2-axis directions and

$$\varepsilon_3 = 0. \tag{19}$$

Average stress is defined according to formula

$$\sigma_{av} = (\sigma_1 + \sigma_2 + \sigma_3)/3. \tag{20}$$

Plot $E(\sigma)$ in considering nonlinear soil model is presented by Figure 2. We have

$$E(\sigma) = E_\infty[1 - \exp(-c\sigma_{av})]; \tag{21}$$

$$c = 10/\sigma_{av}^{\lim}; \tag{22}$$

$\sigma_{av}^{\lim}$ is the limiting value of σ_{av} corresponding to stabilization of modulus of elasticity.

The following formula allowing this factor is recommended in [5,11,12]:

$$E(n) = [E_\infty - E(\sigma)][1 - \exp(-kn)]; \tag{23}$$

$$k = 10/n^*; \tag{24}$$

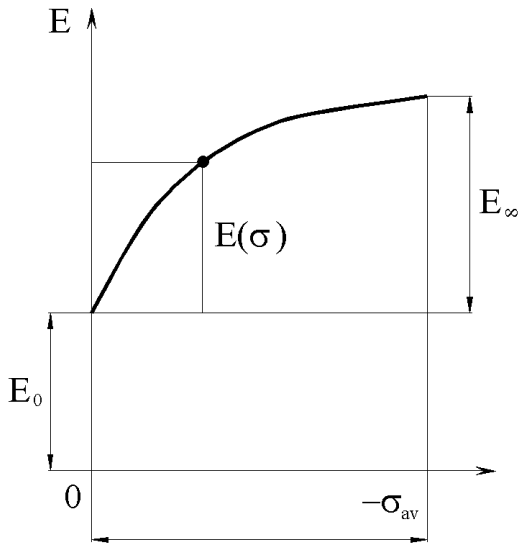


Figure 2. Plot $E(\sigma)$ in nonlinear soil model.

where n^* is the number of cycles corresponding to stabilization of modulus of elasticity.

Multifactor formula for modulus of elasticity has form

$$E = E_0 + E_\infty[1 - \exp(-c\sigma_{av})] + [E_\infty - E(\sigma)][1 - \exp(-kn)]. \tag{25}$$

In accordance with recommendations from [5] we don't consider tensile normal stress components in soil foundation. Otherwise

$$\sigma_1 \le 0; \quad \sigma_2 \le 0. \tag{26}$$

Nevertheless we don't somehow restrict shearing-stress components $\sigma_{12} = \sigma_{21}$.

IV. Discrete-continual method of analysis of coupled system "plate – soil foundation" in context of microseismic and gravitational processes in foundation

A. Basic foundation

Analysis of the real coupled system "plate-soil foundation" and corresponding model is under consideration.

Resolving equations of the considering problems have the following form (in displacements):

$$B^*(\theta J)D_0 B\bar{u} - B^*(\theta J)RD_0 B\bar{u} + C\bar{u} = \bar{F}, \tag{27}$$

where Ω is the given domain occupied by structure; $\partial\Omega$ is its boundary; $\theta = \theta(x)$ is the characteristic function of the domain Ω; R is the integral operator of creep theory set forth below; J is the corresponding Jacobian; E is the modulus of elasticity; ν is the Poisson's ratio; $\bar{u}$ is the displacement vector; B is the Cauchy matrix; D_0 is the matrix of elastic

parameters in case of plane strain; $\overline{F}$ is the body force vector; $\overline{F}_{stat}$ is the constant (long-term) component of body force vector; $\overline{F}_{harm}$ is the variable (periodic) component of body force vector;

$$\overline{u} = [\, u_1 \quad u_2 \,]^{\mathrm{T}}; \quad B = \begin{bmatrix} \partial_1 & 0 \\ 0 & \partial_2 \\ \partial_1 & \partial_2 \end{bmatrix}; \quad \partial_1 = \frac{\partial}{\partial x_1}; \quad \partial_2 = \frac{\partial}{\partial x_2}; \quad (28)$$

$$D_0 = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1 & \frac{\nu}{1-\nu} & 0 \\ \frac{\nu}{1-\nu} & 1 & 0 \\ 0 & 0 & \frac{1-2\nu}{2(1-\nu)} \end{bmatrix}; \quad (29)$$

$$\overline{F} = [\, F_1 \quad F_2 \,]^T; \quad \overline{F} = \overline{F}_{stat} + \overline{F}_{harm}; \quad (30)$$

$$\overline{F}_{stat} = \begin{bmatrix} F_{stat,1} \\ F_{stat,2} \end{bmatrix}; \quad \overline{F}_{harm} = \begin{bmatrix} F_{harm,1} \\ F_{harm,2} \end{bmatrix}. \quad (31)$$

We find it useful to consider analysis of system "plate-soil foundation" under long-term and periodic loadings separately.

B. Analysis of coupled system "plate – soil foundation" subjected to long-term loadings

Resolving equations have the form

$$B^*(\theta J) D_0 B\overline{u} - B^*(\theta J) R D_0 B\overline{u} + C\overline{u} = \overline{F}_{const}. \quad (32)$$

The reader will have no difficulty in showing that

$$R(x) = \int_{-\infty}^{0} K(x, t-\tau) d\tau; \quad (33)$$

$$K(x, t-\tau) = \delta(x) \exp[-(\delta(x) + \delta_1(x))(t-\tau)]; \quad (34)$$

Operator (33) takes each function $C = C(x)$ to

$$R(x)C = C(x) \int_{-\infty}^{0} K(x, t-\tau) d\tau \quad (35)$$

or after integration

$$R(x) = \delta(x) / (\delta(x) + \delta_1(x)). \quad (36)$$

Combining (36) and (32), we obtain

$$B^*(\theta J)[D_0(1-R)]B\overline{u} + C\overline{u} = \overline{F}_{const}$$

and consequently

$$B^*(\theta J) D_1 B\overline{u} + C\overline{u} = \overline{F}_{const}, \quad (37)$$

Where D_1 is specially modified matrix of elastic parameters,

$$D_1 = D_0(1-R). \quad (38)$$

C. Analysis of coupled system "plate – soil foundation" subjected to periodic loadings

With regard to original formulation we have

$$\overline{F}_{harm} = \overline{F}_{ms} + \overline{F}_{gr}. \quad (39)$$

Components from (39) are given by formulas

$$F_{ms} = \overline{A}_{ms} \sin(\beta_{ms} t); \quad F_{gr} = \overline{A}_{gr} \sin(\beta_{gr} t), \quad (40)$$

$$\overline{A}_{ms} = \frac{\Delta g^{ms}_{\max}}{g} \overline{F}; \quad \overline{A}_{gr} = \frac{\Delta g^{gr}_{\max}}{g} \overline{F}; \quad (41)$$

$$\beta_{ms} = 2\pi / t_{ms}; \quad \beta_{gr} = 2\pi / t_{gr}. \quad (42)$$

Generalizing (40)-(42), we consider analysis of system subjected to load

$$\overline{F}_{harm} = F \sin(\beta t). \quad (43)$$

In this connection we have

$$B^*(\theta J) D_0 B\overline{u} - B^*(\theta J) R D_0 B\overline{u} + C\overline{u} = \overline{F}_{harm}. \quad (44)$$

Operator (33) takes each function $f(x,\tau)$ to

$$Rf(x,\tau) = \int_{-\infty}^{0} K(x, t-\tau) f(x,\tau) d\tau. \quad (45)$$

By definition, put

$$L = B^*(\theta J) D_0 B - B^*(\theta J) R D_0 B + C. \quad (46)$$

From (43), (44) and (46) we get the following formulation

$$L\overline{u} = \overline{F} \sin(\beta t). \quad (47)$$

Let us consider auxiliary problem

$$L\overline{v} = \overline{F} \cos(\beta t). \quad (48)$$

Multiplying both sides of (47) by imaginary unit $\tilde{i}$ and summing with (48), we get

$$L(\overline{v} + i\overline{u}) = \overline{F}[\cos(\beta t) + \tilde{i} \sin(\beta t)].$$

By definition, put

$$\overline{w} = \overline{v} + \tilde{i}\,\overline{u}. \quad (49)$$

The result is

$$L\overline{w} = \overline{F}[\cos(\beta t) + \tilde{i}\sin(\beta t)].$$

Taking into account well-known Euler formula

$$\exp(\tilde{i}\beta t) = \cos(\beta t) + \tilde{i}\sin(\beta t)$$

we finally obtain

$$L\overline{w} = \overline{F}\exp(\tilde{i}\beta t). \quad (50)$$

We may define the solution of the considering problem in form

$$\overline{w} = \overline{w}_0 \exp(\tilde{i}\beta t). \quad (51)$$

Let us demonstrable consistency of this step. Indeed we have

$$B^*(\theta J)[\exp(\tilde{i}\beta t) - \mathrm{Re}\, xp(\tilde{i}\beta t)]D_0 B\overline{w} + \\ + C\overline{w}\exp(\tilde{i}\beta t) = \overline{F}\exp(\tilde{i}\beta t). \quad (52)$$

It can be shown in the usual way that

$$\mathrm{Re}\, xp(\tilde{i}\beta t) = \int_{-\infty}^{0} \exp[-\alpha(t-\tau)]\exp(\tilde{i}\beta t)d\tau = \delta\gamma\exp(\tilde{i}\beta t); \quad (53)$$

$$\alpha = \delta + \delta_1; \quad \gamma = (\alpha - \tilde{i}\beta)/(\alpha^2 + \beta^2) \quad (54)$$

or in other words

$$\gamma = \tilde{\alpha} - \tilde{i}\tilde{\beta}; \quad \tilde{\alpha} = \frac{\alpha}{\alpha^2+\beta^2}; \quad \tilde{\beta} = \frac{\beta}{\alpha^2+\beta^2}. \quad (55)$$

Transforming (52) we get

$$B^*(\theta J)[1-\gamma]D_0 B\overline{w} + C\overline{w} = \overline{F}. \quad (56)$$

Let us consider real and imaginary parts of coefficient matrix and right-side vector of the system

$$[B^*(\theta J)(1-\tilde{\alpha})D_0 B + iB^*(\theta J)\tilde{\beta}D_0 B]\overline{w} + C\overline{w} = \overline{F}. \quad (57)$$

By definition, put

$$A = B^*(\theta J)(1-\tilde{\alpha})D_0 B; \quad S = B^*(\theta J)\tilde{\beta}D_0 B. \quad (58)$$

It follows that

$$(A + iS)\overline{w} + C\overline{w} = \overline{F}. \quad (59)$$

Solution of the problem (59) has the form

$$\overline{w} = \overline{w}_0\exp(\tilde{i}\beta t) = (\overline{u}_0 + \overline{v}_0)\exp(\tilde{i}\beta t) = \\ = v_0\cos(\beta t) + \tilde{i}v_0\sin(\beta t) + \tilde{i}u_0\cos(\beta t) - u_0\sin(\beta t). \quad (60)$$

Solution of (47) in its turn

$$u = u_0\cos(\beta t) + v_0\sin(\beta t). \quad (61)$$

After computing of displacements we may find corresponding stress and strain components in accordance with (5)-(12):

$$\overline{\sigma} = [\cos(\beta t) - R\cos(\beta t)]\overline{\sigma}_u + \\ + [\sin(\beta t) - R\sin(\beta t)]\overline{\sigma}_v, \quad (62)$$

where $\overline{\sigma}$ is the vector containing "elastic" stress components,

$$\overline{\sigma} = [\,\sigma_{11}\ \ \sigma_{22}\ \ \sigma_{12}\,]^{\mathrm{T}}; \quad (63)$$

$$\overline{\sigma}_u = DB\overline{u}_0; \quad \overline{\sigma}_v = DB\overline{v}_0. \quad (64)$$

D. Analysis of coupled system "plate – soil foundation" subjected to long-term and periodic loadings

Taking into account principle of superposition and presented algorithms we consider three problems: analysis of the system subjected to one long-term loading and two periodic loadings. Results are summed up in the end.

V. Investigation of Nonlinear Soil Models into Discrete-Continual Method of Analysis

Special numerical procedures describing soil properties are introduced within proposed discrete-continual method of analysis of coupled system "plate-soil foundation" in context of microseismic and gravitational processes in foundation. In other words we have to modify corresponding approaches considered in paragraph III.

Suppose deformations all over domain satisfies to the condition

$$\overline{\varepsilon}^{\,p}(n^*) - \overline{\varepsilon}_0 < 0, \quad (65)$$

Combining (65) and (13) we obtain

$$\overline{\varepsilon} = \overline{\varepsilon}_0 + \frac{1}{E_0}\Delta\overline{\sigma}_0 + [\overline{\varepsilon}^{\,p}(n^*) - \overline{\varepsilon}_0](1 - \exp(-\beta n)). \quad (66)$$

Using (16) we get

$$\overline{\varepsilon} = (1-\alpha)\overline{\varepsilon}_0 + \frac{0.1}{E_0}\overline{\sigma}_0 + \alpha\overline{\varepsilon}^{\,p}, \quad \alpha = 1 - \exp(-\beta n). \quad (67)$$

As is well known

$$\bar{\sigma}_0 = D\bar{\varepsilon}_0, \quad D = \begin{bmatrix} \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda+2\mu & 0 \\ 0 & 0 & 2\mu \end{bmatrix}, \tag{68}$$

$$\bar{\varepsilon} = B\bar{u}, \quad B = \begin{bmatrix} \partial_1 & 0 \\ 0 & \partial_2 \\ \partial_1 & \partial_2 \end{bmatrix}. \tag{69}$$

D is the matrix of elastic parameters in case of plane strain; λ, μ are Lame coefficients of soil foundation.

We obviously have

$$\bar{\varepsilon}_0 = D^{-1}\bar{\sigma}_0 \tag{70}$$

and consequently

$$\bar{\varepsilon} = \left[(1-\alpha)D^{-1} + \frac{0.1}{E_0}I\right]\bar{\sigma}_0 + \alpha\bar{\varepsilon}^p. \tag{71}$$

Here I is the identity matrix of corresponding order. It remains to note that

$$\bar{\sigma}_0 = \left[(1-\alpha)D^{-1} + \frac{0.1}{E_0}I\right]^{-1}(\bar{\varepsilon} - \alpha\bar{\varepsilon}^p) \tag{72}$$

or in other words

$$\bar{\sigma}_0 = \tilde{D}(\bar{\varepsilon} - \alpha\bar{\varepsilon}^p), \ \tilde{D} = \left[(1-\alpha)D^{-1} + \frac{0.1}{E_0}I\right]^{-1} \tag{73}$$

Equations of equilibrium have form

$$B^*\bar{\sigma}_0 = \bar{F}, \tag{74}$$

where $\bar{F}$ is the body force vector. Thus

$$B^*\tilde{D}(\bar{\varepsilon} - \alpha\bar{\varepsilon}^p) = \bar{F},$$

and it is clear that

$$B^*\tilde{D}B\bar{u} = \bar{F} + B^T\tilde{D}\alpha\bar{\varepsilon}^p. \tag{75}$$

We have problem of functional minimization (with respect to $\bar{u}$) in the final analysis

$$\Phi(\bar{u}) = \frac{1}{2}\int_\Omega (\tilde{D}(\bar{\varepsilon} - \alpha\bar{\varepsilon}^p), \bar{\varepsilon})\mathrm{dx} - \int_\Omega (\bar{F}, \bar{u})\mathrm{dx} \ \rightarrow \ \min. \tag{76}$$

We see that as stated above original problem can be handled by proposed method if condition (65) is satisfied. Otherwise there are points with

$$\bar{\varepsilon}^p(n^*) - \bar{\varepsilon}_0 > 0. \tag{77}$$

Special iteration method has been developed for this situation. Let $\bar{u}^{(0)}$ be solution of (76) or equations set

$$L_0\bar{u}^{(0)} = \bar{F}; \quad L_0 = B^*\tilde{D}(\bar{\varepsilon} - \alpha\bar{\varepsilon}^p). \tag{78}$$

Basic recurrence formula has the form

$$\bar{u}^{(k+1)} = \bar{u}^{(k)} - L_0^{-1}(B^*\bar{\sigma}^{(k)} - \bar{F}), \tag{79}$$

where

$$\bar{\sigma}^{(k)} = \tilde{D}(\bar{\varepsilon}^{(k)} - \alpha\bar{\varepsilon}^{(p)}); \quad \bar{\varepsilon}^{(k)} = B\bar{u}^{(k)}; \tag{80}$$

$\bar{u}^{(k)}$, $\bar{\varepsilon}^{(k)}$, $\bar{\sigma}^{(k)}$ denote displacements, deformations and stresses at iteration number k.

Convergence condition has the form

$$\| L_0\bar{u}^{(k)} - \bar{F} \| < \gamma \| \bar{F} \|. \tag{81}$$

Symbol $\|\cdot\|$ denotes corresponding vector norm; γ is a small factor, for instance $\gamma = 0.1$ (degree of accuracy).

Now we introduce method for analysis of nonlinear problems of theory of elasticity. It is based on corresponding variation formulation

$$\Phi(\bar{u}) = \frac{1}{2}\iint_\Omega (\bar{\sigma}, \bar{\varepsilon})dx - \iint_\Omega (\bar{F}, \bar{u})dx \rightarrow \min, \tag{82}$$

where $\Phi(u)$ is the energy functional; $\bar{u}$ is the displacement vector; $\bar{F}$ is the body force vector; $\bar{\sigma} = [\sigma_{11} \ \sigma_{22} \ \sigma_{12}]^T$ is the vector containing stress components; $\bar{\varepsilon} = [\varepsilon_{11} \ \varepsilon_{22} \ \varepsilon_{12}]^T$ is the vector containing strain components; Ω is the considering domain.

Original domain Ω is divided into quadrangular finite elements. We use two-index numbering for nodes and elements of the constructed mesh.

In compliance with introduced finite element model [16] we have

$$\Phi(\bar{u}) = \sum_{i_1}\sum_{i_2}\Phi_{i_1 i_2}(\bar{u}); \quad \Phi(\bar{u}) = \iint_{finite\,element\,number\ i}\Phi_{i_1 i_2}(\bar{u}). \tag{83}$$

Local coordinate system (Ot_1 and Ot_2, $t_1, t_2 \in [0,1]$) is introduced in every finite element $(x_1, x_2) \Rightarrow (i, t_1, t_2)$ (Fig. 3). Renumbering of nodes of element is performed $i, j \Rightarrow 1,1$; $i+1, j \Rightarrow 2,1$; $i, j+1 \Rightarrow 1,2$; $i+1, j+1 \Rightarrow 2,2$.

We use bilinear approximation for unknown vector function $\bar{u}$ and coordinate function $\bar{x}$ within each element

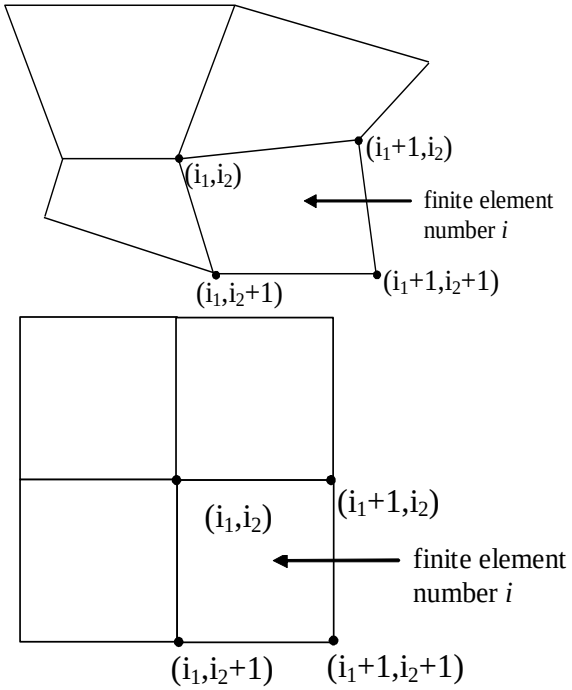


Figure 3. Discretization of the given domain: finite elements, numbering of nodes and elements.

$$u_k(t_1,t_2) = u_k(i) + t_1\Delta_1 u_k(i) + \\ + t_2\Delta_2 u_k(i) + t_1t_2\Delta_{12}u_k(i), \quad k=1,2; \tag{84}$$

$$x_k(t_1,t_2) = x_k(i) + t_1\Delta_1 x_k(i) + \\ + t_2\Delta_2 x_k(i) + t_1t_2\Delta_{12}x_k(i), \quad k=1,2. \tag{85}$$

Here $x(i)$ is a coordinate of node number $i=(i_1,i_2)$; $f(i)$ is a value of function f,

$$\Delta_1 f(i) = f(i_1+1,i_2) - f(i); \quad \Delta_2 f(i) = f(i_1,i_2+1) - f(i); \\ \Delta_{12} f(i) = f(i_1+1,i_2+1) - f(i_1+1,i_2) - \\ - f(i_1,i_2+1) + f(i); \tag{86}$$

t_s, $0 \le t_s \le 1$, $s=1,2$ are local coordinates.

Thus we have finite element mesh topologically equivalent to rectangular [17]. Corresponding finite elements are unit squares. Basic unknowns for elastic problems are vectors $u_k(i)$ containing nodal displacements. Global vector of nodal displacements is denoted by $\overline{U}_{gl}$. It is computed as a result of the following iterative process:

$$\overline{U}_{gl}^{k+1} = \overline{U}_{gl}^{k} - A_k^{-1}\overline{F}^k \quad , \quad k=0,1,... \\ \overline{U}_{gl}^{0} \equiv 0 \tag{87}$$

A_k is the matrix with elements A_{ij}^k given by

$$A_{ij}^k = \frac{1}{\alpha_1\alpha_2}[\Phi(\overline{U}_{gl}^k + \alpha_1\bar{e}_i + \alpha_2\bar{e}_j) - \\ - \Phi(\overline{U}_{gl}^k + \alpha_1\bar{e}_i) - \Phi(\overline{U}_{gl}^k + \alpha_2\bar{e}_j) + \Phi(\overline{U}_{gl}^k)]; \tag{88}$$

$\overline{F}^k$ is the nullity vector with elements F_i^k,

$$F_i^k = \frac{1}{\alpha_1}[\Phi(\overline{U}_{gl}^k + \frac{\alpha_1}{2}\bar{e}_i) - \Phi(\overline{U}_{gl}^k - \frac{\alpha_1}{2}\bar{e}_i)]; \tag{89}$$

$\bar{e}_i$ is the unit vector with elements $\bar{e}_i(j) = \delta_{i,j}$; $\delta_{i,j}$ is the Chronicler's symbol; $\alpha_1 > 0$; $\alpha_2 > 0$ is defined as percent of the next iteration; i, j are indexes of global numeration.

Criterion of truth for solution computed by iterative method has the form (for instance $\varepsilon_{\max} = 0.1$)

$$\left\|\overline{F}^k\right\| / \left\|\overline{F}^0\right\| < \varepsilon = \varepsilon_{\max}. \tag{90}$$

VI. Computer Realization of Discrete-Continual Method of Analysis in Structural Dynamics

Discrete-continual method, considering in the distinctive paper, have been realized in software package DCMDyn. The main purpose of this software is analysis of coupled system "plate-soil foundation" in context of microseismic and gravitational processes in foundation.

Programming environment is Intel Parallel Studio XE 2017 (Fortran Programming Language). Program is designed for Microsoft Windows 7/8/8.1/10.

VII. Numerical Example

Let us consider analysis of the real system "plate-soil foundation" and corresponding model in context of heterogeneous creep and pulsating loads.

Material of the plate is concrete. Let $E_0 = 2.6\cdot10^5 kg/cm^2$ be the modulus of elasticity; $\nu_0 = \nu = 0.16$ is the Poisson's ratio.

Parameters of soil foundation are presented in tables I, II.

TABLE I. Physical Parameters of Soil.

Number of soil type	$E_0, kg/sm^2$	ν
1	690	0.15
2	1500	0.30
3	2050	0.30
4	350	0.35
5	550	0.35

TABLE II. Parameters of Creep

Material	δ, day^{-1}	δ_1, day^{-1}
concrete (plate)	0.07605	0.03
1st soil type	0.0049	0.236
2nd soil type	0.0051	0.225
3rd soil type	0.0125	0.205
4th soil type	0.0025	0.251
5th soil type	0.001	0.255

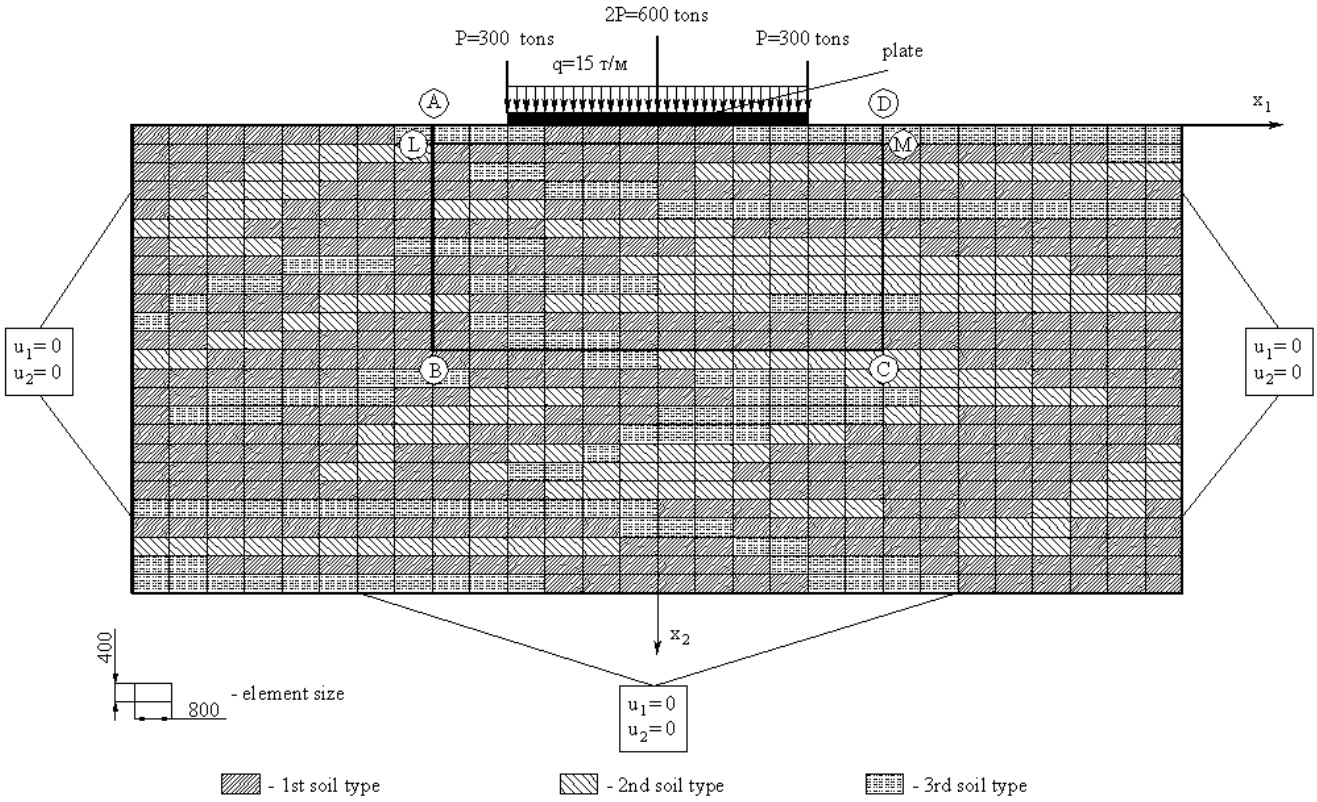


Figure 4. Loading diagram of the real system “plate-soil foundation”.

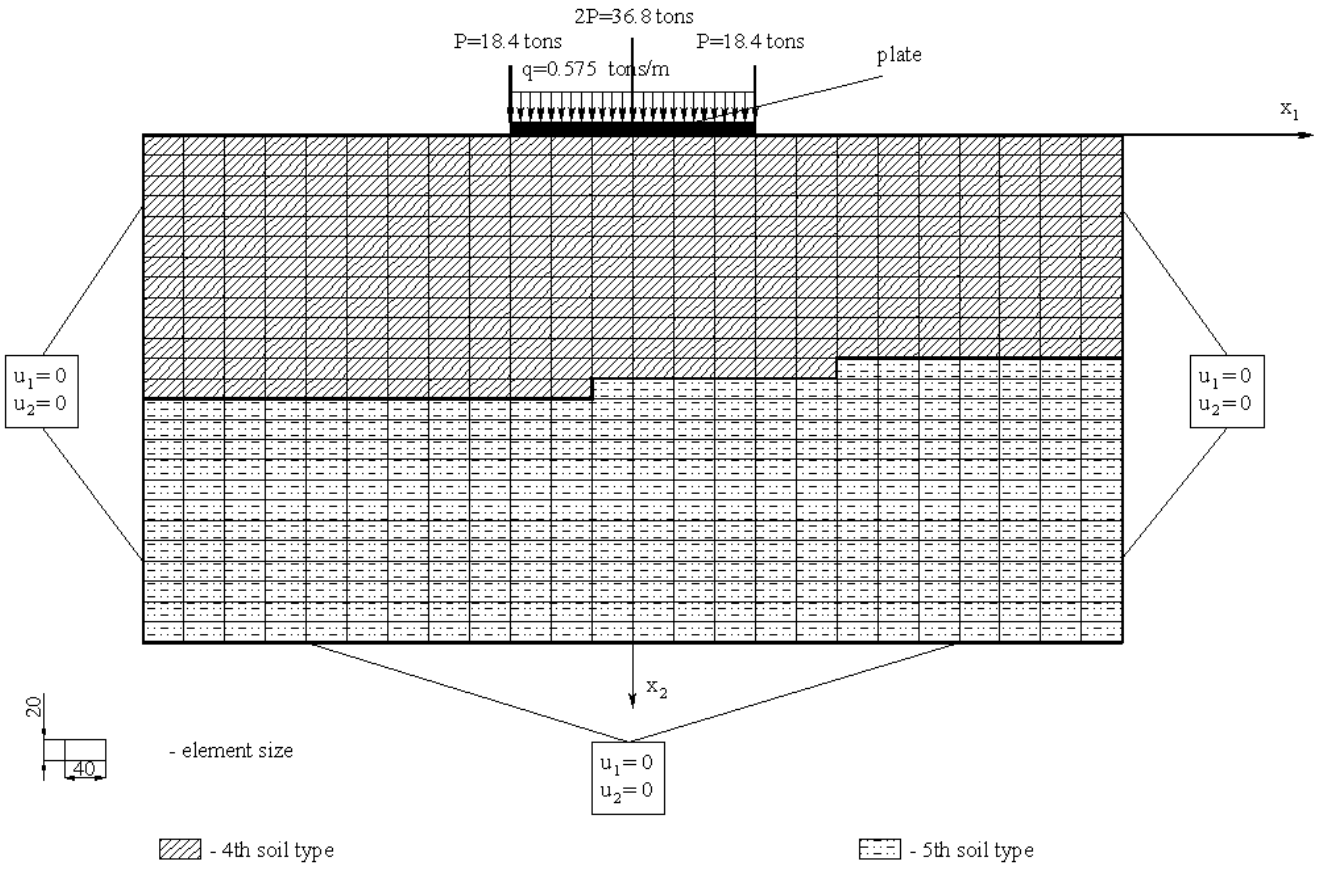


Figure 5. Loading diagram of the model.

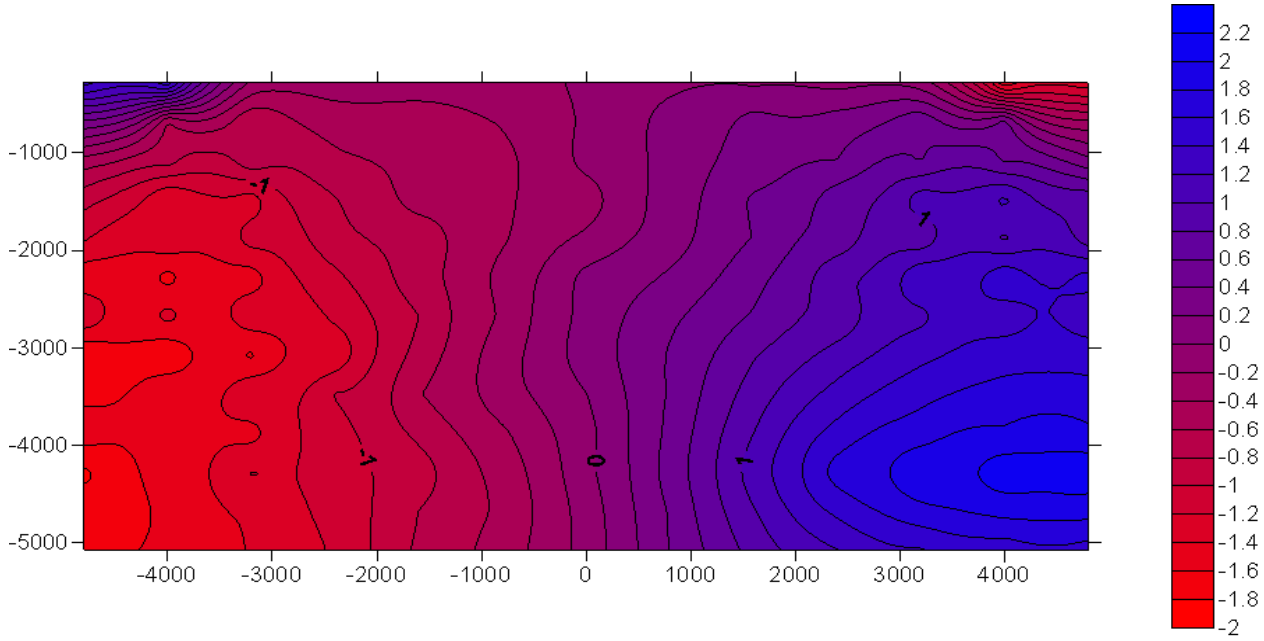


Figure 6. Field of displacements u_1 (area ABCD).

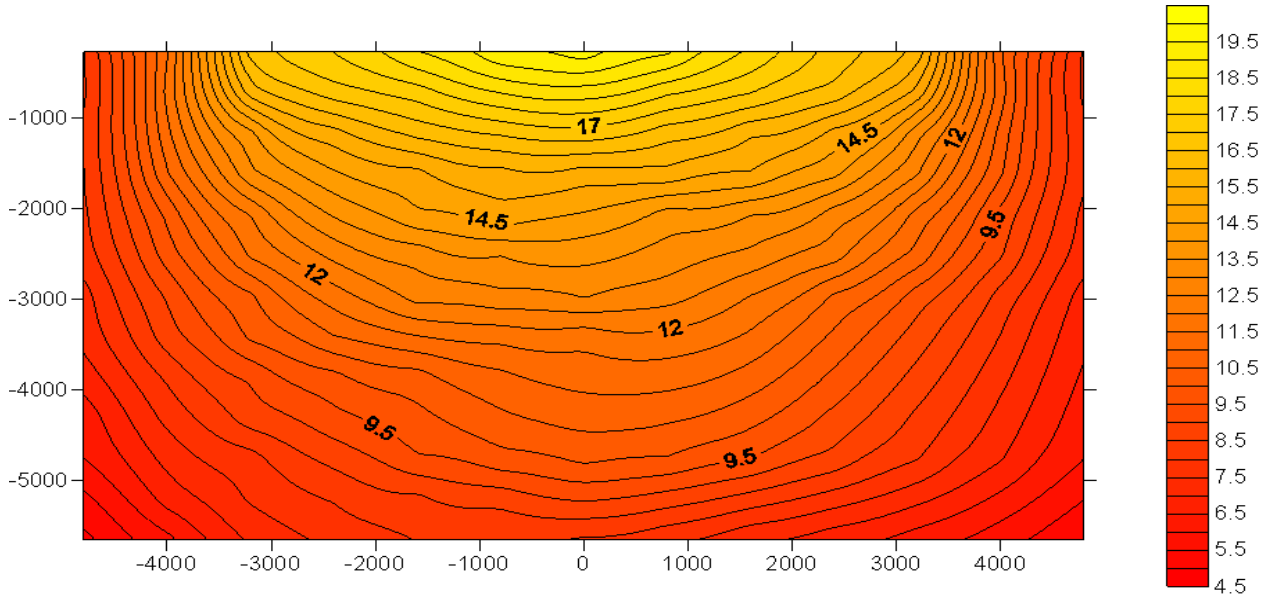


Figure 7. Field of displacements u_2 (area ABCD).

We take into account concentrated loads $P = 300$ tons, $2P = 600$ tons and distributed load $q = 15$ tons/m.

Parameters for dynamic analysis are

$$\Delta g^{ms}_{\max} = 1.2 \text{ m/s}^2;\ \Delta g^{ms}_{gr} = 0.75 \text{ m/s}^2;\ t_{ms} = 0.1\text{s};\ t_{gr} = 0.5\text{s};$$
$$0 \le t \le 10 \text{ years}.$$

Loading diagrams of the real system "plate – soil foundation" and corresponding model are given by Figures 4, 5. Displacements at the domain boundary equated to zero.

Original domain Ω is embordered by extended one ω of rectangular shape [17]. It is optimally approximated by mesh, topologically equivalent to rectangular. Corresponding key features includes regular numeration of nodes and therefore convenient mathematical formulas, effective computational schemes and algorithms, simple data processing and so on [5]. Contour maps and selective graphs of horizontal and vertical displacements are given by Figures 6-9 (units are centimeters). Corresponding contour maps and graphs for the model are given by Figures 10,11,12 (Contour map of stress component σ_{22} is given by Figures 12 (units for σ_{22} are $\kappa N/cm^2$)).

We use software package DCMDyn for the analysis of numerous problems. In accordance with analytical researches in quasi-static cases we can conclude that for some geological heterogeneities of soil foundation irregularities in slumps of buildings may exceed 20%.

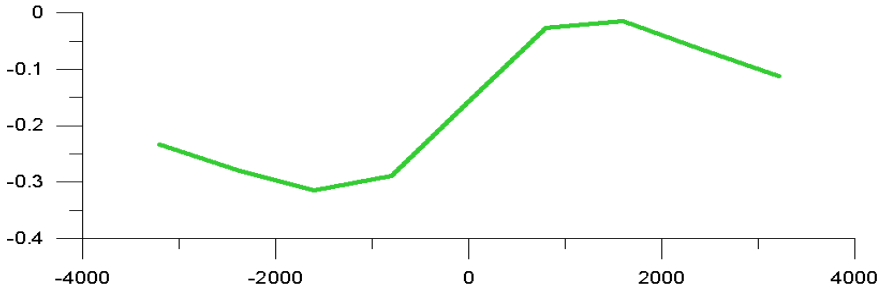


Figure 8. Plot of displacements u_1 at the contact of soil and plate (vertical axis – displacements, horizontal axis – coordinates).

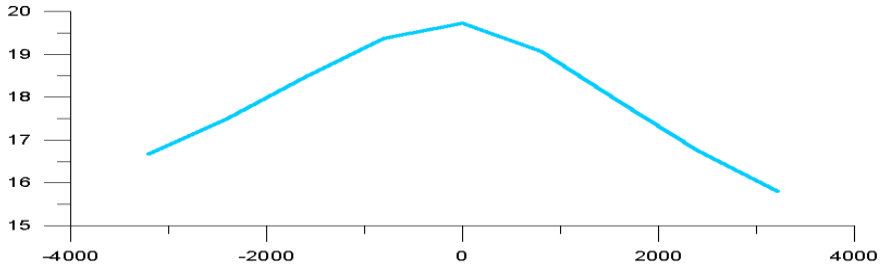


Figure 9. Plot of displacements u_2 at the contact of soil and plate (vertical axis – displacements, horizontal axis – coordinates).

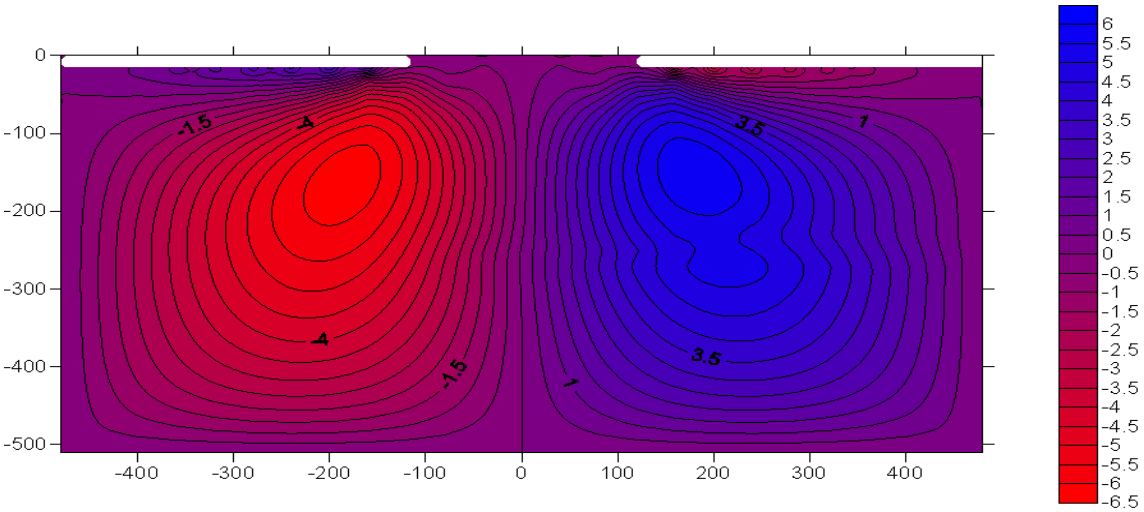


Figure 10. Field of displacements u_1 in the model.

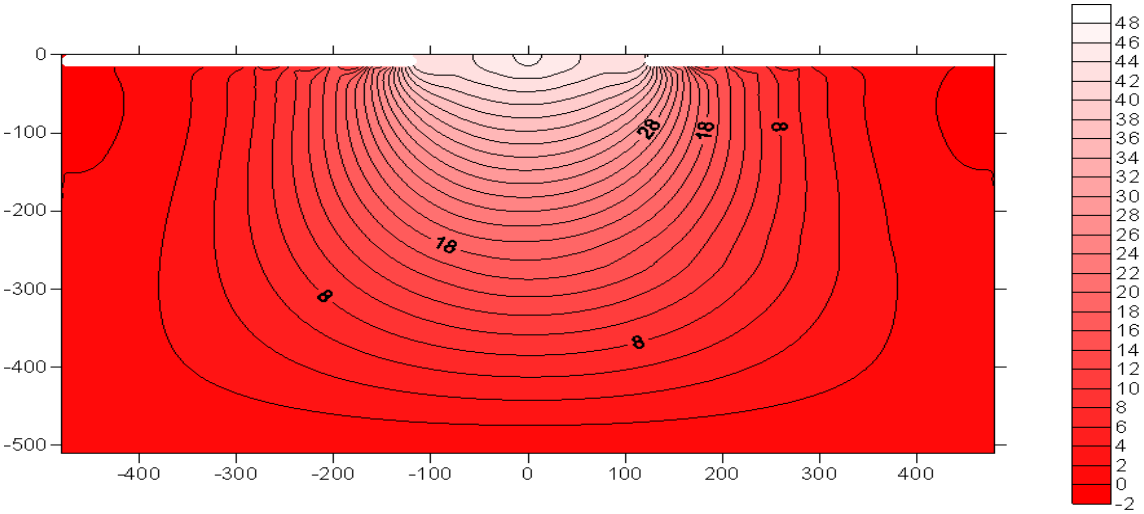


Figure 11. Field of displacements u_2 in the model.

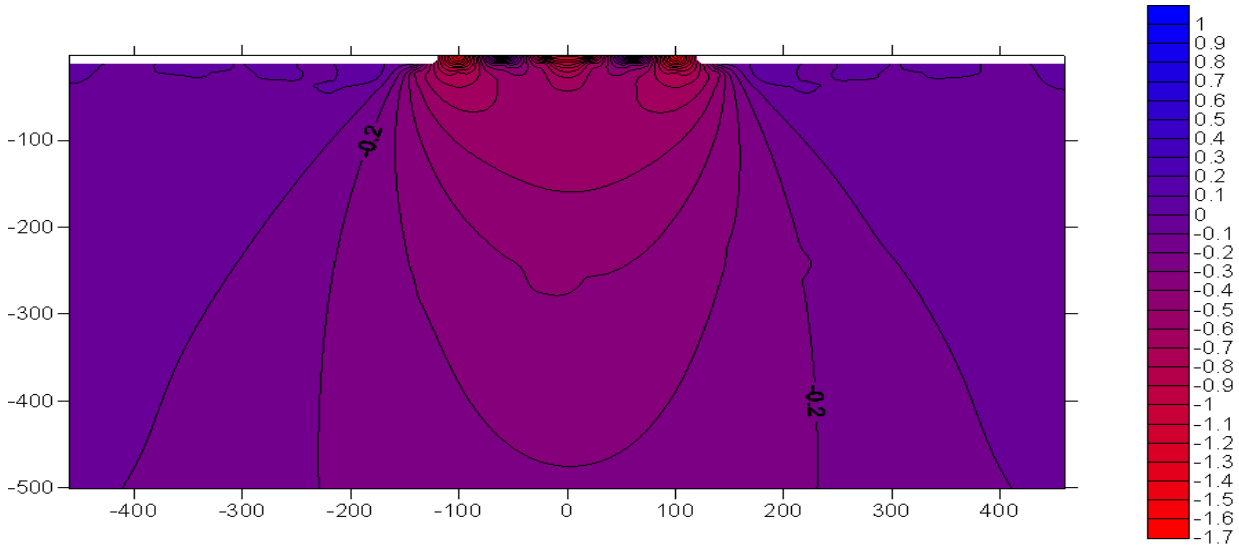


Figure 12. Field of stresses σ_{22} in the model.

VIII. Conslusions

Thus, discrete-continual method of analysis of the coupled system "plate – soil foundation" in context of microseismic and gravitational processes in foundation is presented in the distinctive paper. This method presupposes finite element approximation with respect to corresponding space coordinates while problem remains continual in time. In addition to increasing accuracy, the absence of time discretization significantly reduces the amount of computation and provides correct solution of the problem with the rapidly changing (rapidly oscillating) nature of applied loads. Considering method was implemented in the software package DCMDyn and verified on a large number of test, model and practical tasks, one of the examples of analysis is given in the paper [18]. The reliability and validity of the results are based on the rigor of the mathematical apparatus used, comparison of results of analysis with experimental data and results obtained by verified software packages of industrial type.

References

[1] P.A. Akimov, "Correct Discrete-Continual Finite Element Method of Structural Analysis Based on Precise Analytical Solutions of Resulting Multipoint Boundary Problems for Systems of Ordinary Differential Equations", Applied Mechanics and Materials, Vols. 204-208, pp. 4502-4505, 2012.

[2] P.A. Akimov, A.M. Belostosky, V.N. Sidorov, M.L. Mozgaleva and O.A. Negrozov, "Application of Discrete-Continual Finite Element Method for Global and Local Analysis of Multilevel Systems. // Applied Mechanics and Materials", AIP Conference Proceedings, 1623, 3, pp. 3-6, 2014;

[3] P.A. Akimov and M.L. Mozgaleva, "Method of Extended Domain and General Principles of Mesh Approximation for Boundary Problems of Structural Analysis", Applied Mechanics and Materials, Vols. 580-583, pp. 2898-2902, 2014.

[4] P.A. Akimov, M.L. Mozgaleva and O.A. Negrozov, "Advanced Wavelet-Based Multilevel Discrete-Continual Finite Element Method for Three-Dimensional Local Structural Analysis", ACSR-Advances in Computer Science Research, Vol. 18, pp. 713-716, 2015.

[5] V.N. Savostyanov, M.S. Khlystunov, A.B. Zolotov and P.A. Akimov, "About One Discrete-Continual Method of Strctural Dynamics. Part 1: Formulation of the Problem", International Journal for Computational Civil and Structural Engineering, Vol. 6, Issue 1&2, pp. 188-196, 2010 (in Russian).

[6] M.S. Khlistunov and Z.G. Mogiluk, "Geological efficiency of microseismic processes in foundations with inclined boundaries of layers", Problems of Applied Mathematics and Computational Mechanics, Number 6, MSUCE, pp. 330-340, 2003 (in Russian).

[7] M.S. Khlistunov, "Ballistic and float resonances of structures", Proceedings of Colloquium of MSUCE, Moscow, MSUCE, 1999 (in Russian).

[8] M.S. Khlistunov, "Geodynamic stability of geological foundations", Earthquake-resistant Construction, Number 4, Moscow, 2001 (in Russian).

[9] E.A. Voznesensky, "Earthquakes and soil dynamics", Educational Journal of Soros, Number 2. Moscow, pp. 101-108, 1998 (in Russian).

[10] S.I. Zavalishin and M.S. Khlistunov, "Graviseismic resonances of structures", Earthquake-resistant Construction, Number 3, Moscow, 2000 (in Russian).

[11] Z.G. Ter-Martirosyan, Predictions of mechanical processes in multiphase soil bodies. Moscow, Nedra, 1986 (in Russian).

[12] Z.G. Ter-Martirosyan, Rheological parameters of soil and analysis of foundations. Moscow, Stroyizdat, 1992 (in Russian).

[13] A.K. Bugrov, R.M. Narbut and V.P. Sipidin, "Soil researches in triaxial cell", Leningrad, Stroyizdat, 1987, 185 pages (in Russian).

[14] M.N. Goldshteyn, V.Ya. Khain and N.S. Bogolubov, "Experimental and laboratorial tests of vibrocreep of sandy soil", Soil foundation and soil mechanics, Number 1, 1974.

[15] S.S. Vyalov, Rheological basics of soil mechanics. Moscow, 1978 (in Russian).

[16] O.C. Zienkiewicz, R.L. Taylor and J.Z. Zhu, The Finite Element Method: Its Basis and Fundamentals. Butterworth-Heinemann, Sixth edition, 2005.

[17] P.A. Akimov and M.L. Mozgaleva, "Correct Wavelet-based Multilevel Discrete-Continual Methods for Local Solution of Boundary Problems of Structural Analysis", Applied Mechanics and Materials, Vols. 353-356, pp. 3224-3227, 2013.

[18] A.B. Zolotov, P.A. Akimov and M.L. Mozgaleva, "About one discrete-continual method of strctural dynamics. Part 2: Theoretical foundation of method", International Journal for Computational Civil and Structural Engineering, Volume 6, Issue 1&2, pp. 120-127, 2010 (in Russian).

Semianalytical Solution of Multipoint Boundary Problems of Structural Analysis with the Use of Combined Application of Finite Element Method and Discrete-Continual Finite Element Method

Prof., Dr.Sc. Pavel Akimov
Russian Academy of Architecture and Construction Sciences, Moscow, Russia;
Scientific Research Center "StaDyO", Moscow, Russia;
Department of Applied Mathematics,
National Research Moscow State University of Civil Engineering, Moscow, Russia;
e-mail: pavel.akimov@gmail.com

Oleg A. Negrozov
Department of Applied Mathematics,
National Research Moscow State University of Civil Engineering, Moscow, Russia;
Russian Academy of Architecture and Construction Sciences, Moscow, Russia;
e-mail: genromgsu@gmail.com

***Abstract*—Development, research and verification of correct mathematical models and methods of structural mechanics are the most important aspects of ensuring safety of buildings and complexes. The distinctive paper is devoted to semianalytical solution of multipoint boundary problems of structural analysis with the use of combined application of finite element method and discrete-continual finite element method. Structures containing parts (subdomains) with regular (in particular, constant or piecewise constant) physical and geometrical parameters in some dimension are under consideration. Operational formulations of two-dimensional and three-dimensional problems of structural mechanics with the use of so-called method of extended domain, corresponding numerical implementations (including construction of discrete (finite element) and discrete-continual approximation models for subdomains) and numerical examples are presented.**

Keywords—discrete-continual finite element method, finite element method, semianalytical solution, multipoint boundary problems, structural analysis

I. INTRODUCTION

As is well known, problems of structural analysis normally lead to complex two-dimensional and three-dimensional systems of equations, and solutions of these systems can be obtained only by numerical approaches [1-3]. Generally it is impossible to find the correct analytical (exact) solution [4, 5] while corresponding experimental researches are more expensive and non-complete. In this connection, the idea of development of semianalytical methods, combining the qualitative properties of analytic solutions with the generality of numerical methods, is quite natural.

Semianalytical methods of structural analysis have a long history. We should mention here in particular such traditional methods as finite strip method (semi-analytical, analytical and numerical variants) [6, 7], Kantorovich-Vlassov's method, method of lines etc. The finite strip method (FSM) is not so universal, powerful and versatile than the finite element method (FEM). However, for instance, for linear elastic analysis of long thin-walled structures the FSM is more effective, simple and economical than the traditional FEM since the 2D problem is reduced to 1D problem. Besides, different versions of the method have different disadvantages (nonuniversality i.e. practical impossibility for complex cases of boundary conditions or presence of nonlinearities; problems dealing with allowance for abrupt changes in the stiffnesses; actual absence of exact analytical solutions etc) [7].

Recent developments in computer industry and mathematics resulted in emerging of new branches in computational mechanics under the name of discrete-continual methods [5, 8-18]. These methods allow obtaining solutions in correct analytical form, which is more efficient, especially in the areas of boundary effects (results have normally fast variational behaviour in there areas). Therefor accuracy and convergence of solutions obtained by numerical methods are often heavily dependent on the type of selected shape (basis) functions for approximating of unknowns and the number of discretization nodes (elements) [1, 19]. Convergence in the areas of boundary effects, located near the concentered forces and stress concentrations (i.e. in the most critical areas) are very slow for numerical approaches and weakly depends on the number of elements. Even, for example, if the convergence for displacements is relatively high, it is much smaller for internal strains and stresses. Besides, preliminary simplified analytical research of complicated problem can be very useful for understanding of solution properties or behaviour of structure. Applications of discrete-continual methods provides substantial reduction of size of problems, especially for multidimensional analysis [8]. Semianalytical formulations are contemporary mathematical models which are currently becoming available for computer realization [20].

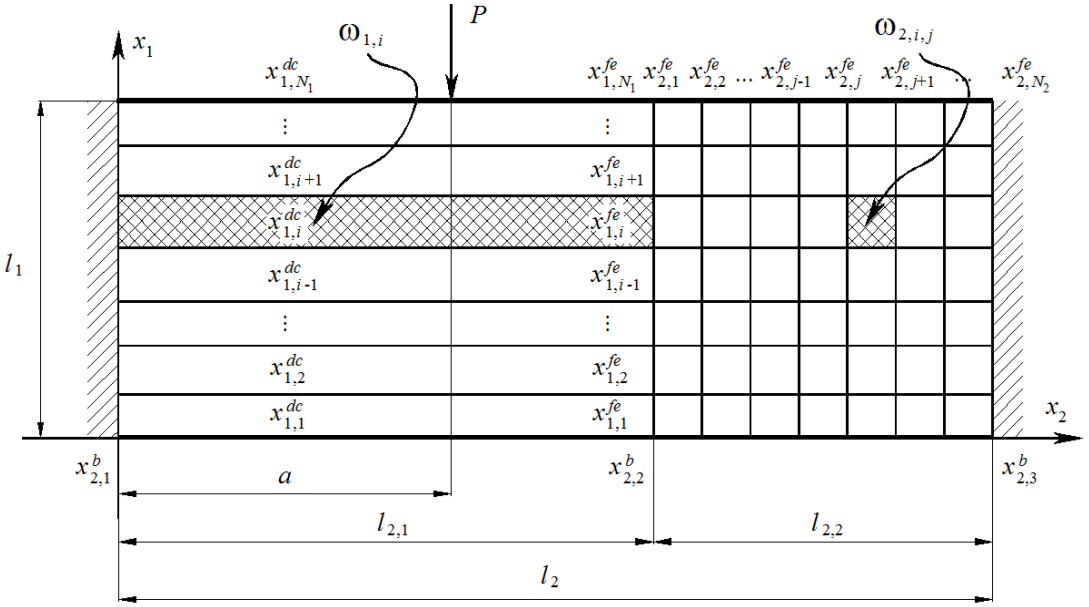


Figure 1. Considering structure (deep beam).

This paper is devoted to so-called semianalytical structural analysis, based on combined application of FEM [1, 19] and discrete-continual finite element method (DCFEM) [5, 8-18]. The field of application of DCFEM comprises structures with regular (constant or piecewise constant) physical and geometrical parameters in some dimension ("basic" dimension). DCFEM presupposes finite element approximation for non-basic dimensions of extended domain while in basic dimension problem remains continual. Thus, DCFEM provides correct analytical solution of the considering problem along basic direction of structure and finite element approximation along other directions. As a result, more accurate solutions can be obtained, especially in areas of boundary effect. Another feature of application of DCFEM is the absence of limitations on lengths of structures and in this context DCFEM is peculiarly relevant.

Regularity of physical and geometrical parameters in one dimension (so-called "basic dimension (direction)) exists in various problems of structural analysis and corresponding mathematical models. We should mention here in particular such objects as beams, strip foundations, thin-walled bars, deep beams, plates, shells, high-rise buildings, extensional buildings, pipelines, rails, dams and others. Correct analytical solution in corresponding constructive parts is apparently preferable in all aspects for qualitative analysis of calculation data. It allows investigator to consider boundary effects when some components of solution are rapidly varying functions. Due to the abrupt decrease inside of mesh elements in many cases their rate of change can't be adequately considered by conventional numerical methods while analytics enables study.

Boundary problems of two-dimensional and three-dimensional theory of elasticity are under consideration. In accordance with the method of extended domain [21], the given domain (occupied by structure) is embordered by extended one of arbitrary shape, particularly elementary. Corresponding discrete and discrete-continual approximation models for subdomains and coupled multilevel approximation model for extended domain are under consideration. Resultant system of linear algebraic equations are constructed. Brief information about software and verification samples are given.

II. TWO-DIMENSIONAL PROBLEMS OF STRUCTURAL ANALYSIS

A. *Formulation of the problem and notation system*

Let us consider problem of static structural analysis of deep beam loaded by concentrated force with hinged ends (cross-sections) along basic dimension (Fig. 1). Some elements of notation system are presented at Fig. 1 as well.

Let Ω be the domain occupied by structure, $\Omega = \{(x_1, x_2):\ 0 < x_1 < l_1,\ 0 < x_2 < l_2\}$, where $\Omega = \Omega_1 \cup \Omega_2$ and $\Omega_k = \{(x_1, x_2):\ 0 < x_1 < l_1,\ x^b_{2,k} < x_2 < x^b_{2,k+1}\}$, $k = 1, 2$; x_1, x_2 are coordinates (x_2 corresponds to basic dimension); $x^b_{2,1} = 0$, $x^b_{2,2} = l_{2,1}$, $x^b_{2,3} = l_{2,1} + l_{2,2} = l_2$ are coordinates of boundary points (cross-sections) along basic dimension; Ω_1 and Ω_2 are subdomains of Ω; ω_1 and ω_2 are extended subdomains, embordering subdomains $\Omega_1 \subset \omega_1$ and $\Omega_2 \subset \omega_2$; $\omega = \omega_1 \cup \omega_2$; $x^{dc}_{1,i}$, $i = 1, 2, ..., N^{dc}_1$ are coordinates (along x_1) of nodes (nodal lines) of discrete-continual finite elements (these elements are used for approximation of domain ω_1); $(N^{dc}_1 - 1)$ is the number of discrete-continual finite elements; $x^{fe}_{1,i}$, $i = 1, 2, ..., N^{fe}_1$; $x^{fe}_{2,j}$, $j = 1, 2, ..., N^{fe}_2$ are coordinates (along x_1, x_2) of nodes of finite elements (these elements are used for approximation of domain ω_2); $(N^{fe}_1 - 1)$ is the number of finite elements along coordinate x_1; $(N^{fe}_2 - 1)$ is the number of finite elements along x_2.

Two-index notation system is used for numbering of discrete-continual finite elements. Typical number has the form (k,i), where k is the number of subdomain, i is the number of element (along x_1). Three-index system is used for numbering of finite elements. Typical number has the form (k,i,j), where k is the number of subdomain, i and j are numbers of element (along x_1 and x_2). Let us consider the following simplest case:

$$N_1^{fe} = N_1^{dc} = N_1;\ x_{1,i}^{fe} = x_{1,i}^{dc} = x_{1,i},\ i = 1, 2, ..., N_1.$$

B. Discrete-continual approximation model for subdomain

Discrete-continual approximation model is used for two-dimensional problems. It presupposes mesh approximation for non-basic dimension of extended domain (along x_1) while in the basic dimension (along x_2) problem remains continual. Thus extended subdomain ω_1 is divided into discrete-continual finite elements

$$\omega_1 = \bigcup_{i=1}^{N_1-1} \omega_{1,i}; \quad (1)$$

$$\omega_{1,i} = \{(x_1, x_2):\ x_{1,i} < x_1 < x_{1,i+1},\ x_{2,1}^b < x_2 < x_{2,2}^b\}. \quad (2)$$

Lame constants for discrete-continual finite element are defined by formulas:

$$\overline{\lambda}_{1,i} = \theta_{1,i}\lambda; \quad \overline{\mu}_{1,i} = \theta_{1,i}\mu; \quad \theta_{1,i} = \begin{cases} 1, & \omega_{1,i} \subset \Omega_1; \\ 0, & \omega_{1,i} \not\subset \Omega_1, \end{cases} \quad (3)$$

where $\theta_{1,i}$ is the characteristic function of element $\omega_{1,i}$ (it is defined in accordance with the method of extended domain (henceforward corresponding details are presented in [22])).

Basic nodal unknown functions are displacement components $u_1^{(1)}, u_2^{(1)}$ and their derivatives $v_1^{(1)}, v_2^{(1)}$ with respect to x_2 (superscript hereinafter corresponds to the number of considered subdomain i.e. ω_1). Thus for node $(1,i)$ we have the following unknown functions: $u_1^{(1,i)}, u_2^{(1,i)}$ and $v_1^{(1,i)}, v_2^{(1,i)}$.

Linear approximation is used for unknown functions within discrete-continual finite element (it should be noted, that in the general case, other types of approximation can also be used).

DCFEM is reduced at some stage to the solution of systems of $4N_1$ first-order ordinary differential equations:

$$\overline{U}_1'(x_2) = A_1\overline{U}_1(x_2) + \overline{\overline{R}}_1(x_2), \quad (4)$$

where $\overline{U}_1(x_2)$ is the global vector of nodal unknown functions (subscript corresponds to the number of subdomain ω_1),

$$\overline{U}_1 = \overline{U}_1(x_2) = [\,(\overline{u}_1)^T \quad (\overline{v}_1)^T\,]^T; \quad (5)$$

$$\overline{u}_1 = \overline{u}_1(x_2) = [\,(\overline{u}_n^{(1,1)})^T \quad (\overline{u}_n^{(1,2)})^T \quad ... \quad (\overline{u}_n^{(1,N_1)})^T\,]^T;$$

$$\overline{v}_1 = \overline{v}_1(x_2) = [\,(\overline{v}_n^{(1,1)})^T \quad (\overline{v}_n^{(1,2)})^T \quad ... \quad (\overline{v}_n^{(1,N_1)})^T\,]^T; \quad (6)$$

$$\overline{u}_n^{(1,i)} = \overline{u}_n^{(1,i)}(x_2) = [\,u_1^{(1,i)} \quad u_2^{(1,i)}\,]^T;$$

$$\overline{v}_n^{(1,i)} = \overline{v}_n^{(1,i)}(x_2) = [\,v_1^{(1,i)} \quad v_2^{(1,i)}\,]^T; \quad (7)$$

A_1 is the global matrix of coefficients of order $4N_1$; $\overline{\overline{R}}_1(x_2)$ is the global right-side vector of order $4N_1$.

Correct analytical solution of (4) is defined by formula

$$\overline{U}_1(x_2) = E_1(x_2)\overline{C}_1 + \overline{S}_1(x_2), \quad (8)$$

where $\overline{C}_1$ is the vector of constants of order $4N_1$;

$$E_1(x_2) = \varepsilon_1(x_2 - x_{2,1}^b) - \varepsilon_1(x_2 - x_{2,2}^b);$$

$$\overline{S}_1(x_2) = \varepsilon_1(x_2) * \overline{\overline{R}}_1(x_2); \quad (9)$$

$\varepsilon_1(x_2)$ is the fundamental matrix-function of system (4) (it is constructed in the special form, convenient for problems of structural mechanics [5]); $*$ is a convolution notation.

C. Discrete (finite element) approximation model for subdomain

Discrete (finite element) approximation model for the considering two-dimensional problems presupposes finite element approximation along x_1 and x_2. Thus extended subdomain ω_2 is divided into finite elements

$$\omega_2 = \bigcup_{i=1}^{N_1-1}\bigcup_{j=1}^{N_2-1} \omega_{2,i,j}; \quad (10)$$

$$\omega_{2,i,j} = \{(x_1, x_2):\ x_{1,i} < x_1 < x_{1,i+1},\ x_{2,j}^{fe} < x_2 < x_{2,j+1}^{fe}\}. \quad (11)$$

Lame constants for finite element are defined by formulas:

$$\overline{\lambda}_{2,i,j} = \theta_{2,i,j}\lambda; \quad \overline{\mu}_{2,i,j} = \theta_{2,i,j}\mu; \quad \theta_{2,i,j} = \begin{cases} 1, & \omega_{2,i,j} \subset \Omega_2; \\ 0, & \omega_{2,i,j} \not\subset \Omega_2, \end{cases} \quad (12)$$

where $\theta_{2,i,j}$ is the characteristic function of element $\omega_{2,i,j}$ (it is defined in accordance with the method of extended domain).

Basic nodal unknowns are displacement components $u_1^{(2)}, u_2^{(2)}$ (superscript hereinafter corresponds to the number of subdomain i.e. ω_2). Thus for node $(2,i,j)$ we have the following unknowns: $u_1^{(2,i,j)}, u_2^{(2,i,j)}$.

Bilinear approximation of unknowns is used within finite element (conventional plane rectangular 4-node finite element of two-dimensional problem of elasticity theory is used in the simplest case).

As is well known, FEM is reduced to the solution of systems of $2N_1N_2$ linear algebraic equations:

TABLE I. ALGORITHM OF CONSTRUCTION OF MATRIX B_1^+ (ALL OTHER ELEMENTS OF MATRIX B_1^+ ARE EQUAL TO ZERO).

Numbers (indexes) of elements	Element value	Corresponding boundary condition
$(2i-1, 2i-1),\ \ i = 1, 2, ..., N_1$	1	The first equation from (17)
$(2i, 2i),\ \ i = 1, 2, ..., N_1$	1	The second equation from (17)

$$K_2 \overline{U}_2 = \overline{R}_2, \tag{13}$$

where K_2 is the global stiffness matrix of order $2N_1N_2$; $\overline{R}_2$ is the global right-side (load) vector of order $2N_1N_2$; $\overline{U}_2$ is the global vector of nodal unknowns (subscript corresponds to the number of subdomain ω_2),

$$\begin{aligned}\overline{U}_2 = [\,&(\overline{u}_n^{(2,1,1)})^T \quad (\overline{u}_n^{(2,2,1)})^T \quad ... \quad (\overline{u}_n^{(2,N_1,1)})^T \\ &... \quad (\overline{u}_n^{(2,1,2)})^T \quad (\overline{u}_n^{(2,2,2)})^T \quad ... \quad (\overline{u}_n^{(2,N_1,2)})^T \quad ... \\ &... \quad (\overline{u}_n^{(2,1,N_2)})^T \quad (\overline{u}_n^{(2,2,N_2)})^T \quad ... \quad (\overline{u}_n^{(2,N_1,N_2)})^T \,]^T;\end{aligned} \tag{14}$$

$$\overline{u}_n^{(2,i,j)} = [\,u_1^{(2,i,j)} \quad u_2^{(2,i,j)}\,]^T,\ \ i = 1, 2, ..., N_1,\ \ j = 1, 2, ..., N_2; \tag{15}$$

D. Multilevel approximation model for domain

System (13) can be rewritten for all nodes with indexes $1 < j < N_2$ (i.e. $x_{2,2}^b < x_2 < x_{2,3}^b$) in the following form (the resolving system of $2N_1(N_2-2)$ linear algebraic equations is obtained):

$$\widetilde{K}_2 \overline{U}_2 = \overline{\overline{R}}_2, \tag{16}$$

where $\widetilde{K}_2$ is the reduced global stiffness matrix of size $[2N_1(N_2-2)] \times [2N_1N_2]$; $\overline{\overline{R}}_2$ is the reduced right-side vector of order $2N_1(N_2-2)$.

Boundary conditions at section $x_2 = x_{2,1}^b$ (hinged edge) have the form ($2N_1$ equations):

$$\begin{aligned}u_1^{(1,i)}(x_{2,1}^b + 0) = 0,\ \ i = 1, 2, ..., N_1;\\ u_2^{(1,i)}(x_{2,1}^b + 0) = 0,\ \ i = 1, 2, ..., N_1.\end{aligned} \tag{17}$$

Equations (17) can be rewritten in a matrix form:

$$B_1^+ \overline{U}_1(x_{2,1}^b + 0) = \overline{g}_1^+, \tag{18}$$

where B_1^+ is the matrix of boundary conditions of size $2N_1 \times 4N_1$ (Table 1); $\overline{g}_1^+$ is the zero vector of order $2N_1$ (i.e. $\overline{g}_1^+ = 0$).

Combining (8) and (18), we obtain

$$B_1^+ E_1(x_{2,1}^b + 0)\overline{C}_1 = \overline{g}_1^+ - B_1^+ \overline{S}_1(x_{2,1}^b + 0) \quad \text{or} \quad Q_1 \overline{C}_1 = \overline{G}_1, \tag{19}$$

where Q_1 is the matrix of size $2N_1 \times 4N_1$; $\overline{G}_1$ is the vector of order $2N_1$;

$$Q_1 = B_1^+ E_1(x_{2,1}^b + 0); \quad \overline{G}_1 = \overline{g}_1^+ - B_1^+ \overline{S}_1(x_{2,1}^b + 0). \tag{20}$$

Boundary conditions at section $x_2 = x_{2,2}^b$ (perfect contact) have the form ($4N_1$ equations):

$$u_1^{(1,i)}(x_{2,2}^b - 0) = u_1^{(2,i,j)},\ \ i = 1, 2, ..., N_1,\ \ j = 1; \tag{21}$$

$$u_2^{(1,i)}(x_{2,2}^b - 0) = u_2^{(2,i,j)},\ \ i = 1, 2, ..., N_1,\ \ j = 1; \tag{21}$$

$$\sigma_{1,2}^{(1,i)}(x_{2,2}^b - 0) = \sigma_{1,2}^{(2,i,j)},\ \ i = 1, 2, ..., N_1,\ \ j = 1; \tag{22}$$

$$\sigma_{2,2}^{(1,i)}(x_{2,2}^b - 0) = \sigma_{2,2}^{(2,i,j)},\ \ i = 1, 2, ..., N_1,\ \ j = 1; \tag{22}$$

where $\sigma_{1,2}^{(1,i)}(x_2)$ and $\sigma_{2,2}^{(1,i)}(x_2)$ are nodal functions (after corresponding averaging) of stress components $\sigma_{1,2}(x_2)$ and $\sigma_{2,2}(x_2)$ for discrete-continual finite element $(1, i)$; $\sigma_{1,2}^{(2,i,j)}$ and $\sigma_{2,2}^{(2,i,j)}$ are nodal stress components $\sigma_{1,2}$ and $\sigma_{2,2}$ (after corresponding averaging) for finite element $(2, i, j)$; $j = 1$.

Equations (21) and (22) can be rewritten in a matrix form:

$$B_2^- \overline{U}_1(x_{2,2}^b - 0) = B_2^+ \overline{U}_2, \tag{23}$$

where B_2^- is the matrix of boundary conditions of size $4N_1 \times 4N_1$ (it can be constructed in accordance with the algorithm presented at [16]); B_2^+ is the matrix of boundary conditions of size $4N_1 \times 2N_1N_2$ (it can be constructed in accordance with so-called method of basis variations [22]).

Combing (8) and (22), we obtain

$$B_2^- E_1(x_{2,2}^b - 0)\overline{C}_1 - B_2^+ \overline{U}_2 = -B_2^- \overline{S}_1(x_{2,2}^b - 0) \quad \text{or}$$

$$Q_{2,1} \overline{C}_1 + Q_{2,2} \overline{U}_2 = \overline{G}_2, \tag{24}$$

where $Q_{2,1}$ is the matrix of size $4N_1 \times 4N_1$; $Q_{2,2}$ is the matrix of size $4N_1 \times 2N_1N_2$; $\overline{G}_2$ is the vector of order $4N_1$,

$$Q_{2,1} = B_2^- E_1(x_{2,2}^b - 0); \quad Q_{2,2} = -B_2^+; \quad \overline{G}_2 = -B_2^- \overline{S}_1(x_{2,2}^b - 0). \tag{25}$$

Boundary conditions at section $x_2 = x_{2,3}^b$ (hinged edge) have the form ($2N_1$ equations; $j = N_2$):

$$u_1^{(2,i,j)} = 0, \quad i = 1, 2, ..., N_1, \quad j = N_2;$$
$$u_2^{(2,i,j)} = 0, \quad i = 1, 2, ..., N_1, \quad j = N_2. \quad (26)$$

Equations (26) can be rewritten in a matrix form:

$$B_3^- \overline{U}_2 = \overline{g}_3^-, \quad (27)$$

where B_3^- is the matrix of boundary conditions of size $2N_1 \times 2N_1N_2$ (it can be constructed in accordance with the algorithm presented at [16]); $\overline{g}_3^-$ is the zero vector of order $2N_1$ (i.e. $\overline{g}_3^- = 0$).

Thus, the total number of equations is equal to $2N_1(N_2 - 2) + 2N_1 + 4N_1 + 2N_1 = 2N_1N_2 + 4N_1$ (i.e. $4N_1$ components of vector $\overline{C}_1$ and $2N_1N_2$ components of nodal displacements $u_1^{(2,i,j)}, u_2^{(2,i,j)}$, $i = 1, 2, ..., N_1$, $j = 1, 2, ..., N_2$). Corresponding coupled system of $2N_1N_2 + 4N_1$ linear algebraic equations with $2N_1N_2 + 4N_1$ unknowns has the form:

$$\begin{bmatrix} Q_1 & 0 \\ Q_{2,1} & Q_{2,2} \\ 0 & \tilde{\tilde{K}}_2 \\ 0 & B_3^- \end{bmatrix} \begin{bmatrix} \overline{C}_1 \\ \overline{U}_2 \end{bmatrix} = \begin{bmatrix} \overline{G}_1 \\ \overline{G}_2 \\ \tilde{\tilde{R}}_2 \\ \overline{g}_3^- \end{bmatrix}. \quad (28)$$

Conditions (27) can be taken into account automatically within construction of the global stiffness matrix and the global right-side vector corresponding to ω_2. The result is

$$\begin{bmatrix} Q_1 & 0 \\ Q_{2,1} & Q_{2,2} \\ 0 & \tilde{\tilde{K}}_2 \end{bmatrix} \begin{bmatrix} \overline{C}_1 \\ \overline{U}_2 \end{bmatrix} = \begin{bmatrix} \overline{G}_1 \\ \overline{G}_2 \\ \tilde{\tilde{R}}_2 \end{bmatrix}, \quad (29)$$

where $\tilde{\tilde{K}}_2$ is the corresponding reduced global stiffness matrix of size $[2N_1(N_2 - 1)] \times [2N_1N_2]$; $\tilde{\tilde{R}}_2$ is the corresponding reduced global right-side vector of order $2N_1(N_2 - 1)$.

Strain and stress components are computed according to well-known formulas after solving of system (29).

E. Software

We should stress that all methods and algorithms considered in this paragraph have been realized in proprietary software. The main purpose of software system CSASA2D (DCFEM + FEM) is semianalytical structural analysis (static analysis of two-dimensional structures within two-dimensional theory of elasticity), based on combined application of FEM and DCFEM. Programming environment is Microsoft Visual Studio 2013 Community and Intel Parallel Studio 2017XE with Intel MKL Library [23]. This software system is designed for Microsoft Windows 8.1/10.

III. THREE-DIMENSIONAL PROBLEMS OF STRUCTURAL ANALYSIS

A. Formulation of the problem and notation system

Let us consider the problem of static analysis of three-dimensional structure (Fig. 2) loaded by concentrated force with hinged ends (cross-sections) along basic dimension. Some elements of notation system are presented at Fig. 2.

Let Ω be the domain occupied by structure, $\Omega = \{(x_1, x_2) : 0 < x_1 < l_1,\ 0 < x_2 < l_2,\ 0 < x_3 < l_3\}$, where $\Omega = \Omega_1 \cup \Omega_2$ and $\Omega_k = \{(x_1, x_2, x_3) : 0 < x_1 < l_1,\ 0 < x_2 < l_2,\ x_{3,k}^b < x_3 < x_{3,k+1}^b\}$, $k = 1, 2$; x_1, x_2, x_3 are coordinates (x_3 corresponds to basic dimension); $x_{3,1}^b = 0$, $x_{3,2}^b = l_{3,1}$, $x_{3,3}^b = l_{3,1} + l_{3,2} = l_3$ are coordinates of corresponding boundary points (cross-sections) along basic dimension; Ω_1 and Ω_2 are subdomains of Ω; ω_1 and ω_2 are extended subdomains, embordering subdomains $\Omega_1 \subset \omega_1$ and $\Omega_2 \subset \omega_2$; $\omega = \omega_1 \cup \omega_2$; $x_{1,i,j}^{dc}$, $x_{2,i,j}^{dc}$, $i = 1, 2, ..., N_1^{dc}$, $j = 1, 2, ..., N_2^{dc}$ are coordinates (along x_1 and x_2) of nodes (nodal lines) of discrete-continual finite elements (discrete-continual finite elements are used for approximation of domain ω_1); $(N_1^{dc} - 1)$ is the number of discrete-continual finite elements along coordinate x_1; $(N_2^{dc} - 1)$ is the number of discrete-continual finite elements along coordinate x_2; $x_{1,i,j,r}^{fe}$, $x_{2,i,j,r}^{fe}$ $x_{3,i,j,r}^{fe}$, $i = 1, 2, ..., N_1^{fe}$, $j = 1, 2, ..., N_2^{fe}$, $r = 1, 2, ..., N_3^{fe}$ are coordinates (along x_1, x_2 and x_3) of nodes of finite elements (finite elements are used for approximation of domain ω_2); $(N_1^{fe} - 1)$ is the number of finite elements along coordinate x_1; $(N_2^{fe} - 1)$ is the number of finite elements along coordinates x_2; $(N_3^{fe} - 1)$ is the number of finite elements along coordinates x_3.

Let us consider the case of rectangular mesh approximation of domain ω_2 (Fig. 2):

$$x_{1,i,j,r}^{fe} = x_{1,i}^{fe}, \quad x_{21,i,j,r}^{fe} = x_{2,j}^{fe}, \quad x_{31,i,j,r}^{fe} = x_{3,r}^{fe},$$
$$i = 1, 2, ..., N_1^{fe}, \quad j = 1, 2, ..., N_2^{fe}, \quad r = 1, 2, ..., N_3^{fe}. \quad (30)$$

Three-index notation system is used for numbering of discrete-continual finite elements. Typical number has the form (k, i, j), where k is the number of subdomain, i and j are numbers of element (along x_1 and x_2). Four-index system is used for numbering of finite elements. Typical number has the form (k, i, j, r), where k is the number of subdomain, i, j and k are numbers of element (along x_1, x_2 and x_3). Let us consider the following case: $N_1^{fe} = N_1^{dc} = N_1$ and $N_2^{fe} = N_2^{dc} = N_2$, $x_{q,i,j}^{dc} = x_{q,i,j,r}^{fe}$, $i = 1, 2, ..., N_1$, $j = 1, 2, ..., N_2$, $r = 1, 2, ..., N_3^{fe}$, $q = 1, 2, 3$.

Let us consider the simplest case:

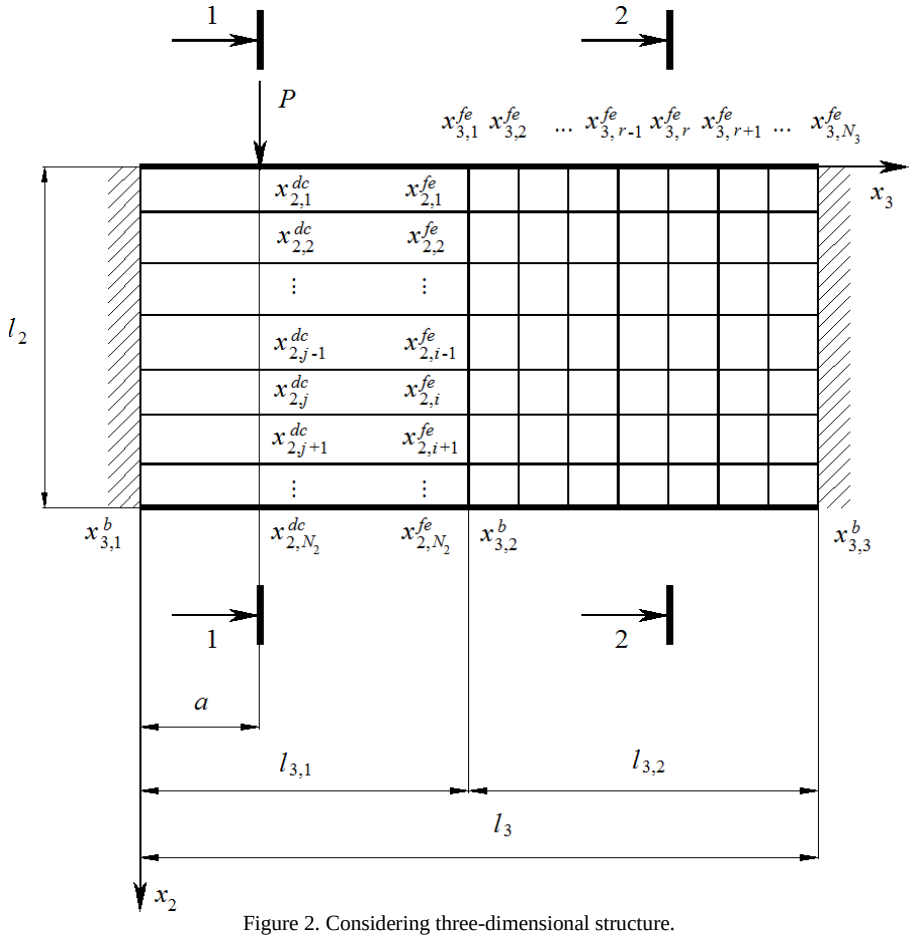


Figure 2. Considering three-dimensional structure.

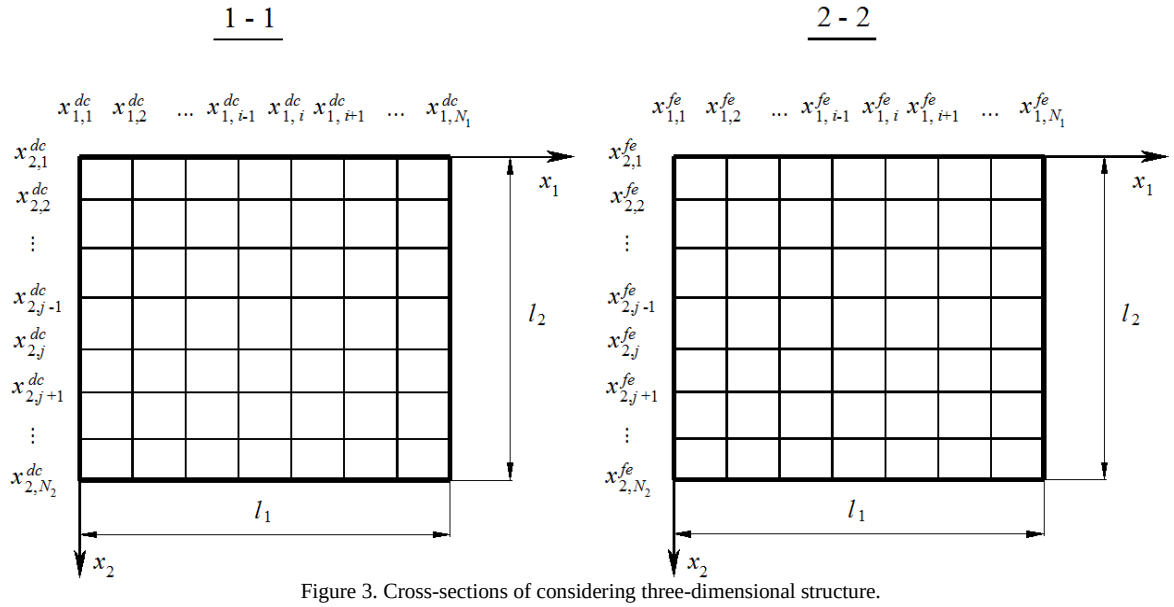


Figure 3. Cross-sections of considering three-dimensional structure.

$$x_{q,i,j} = x^{dc}_{q,i,j} = x^{fe}_{q,i,j,r}, \quad i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2, \quad r = 1, 2, ..., N^{fe}_3, \quad q = 1, 2, 3; \tag{31}$$

$$x_{1,i} = x_{1,i,j}, \quad i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2;$$

$$x_{2,j} = x_{2,i,j}, \quad i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2. \tag{32}$$

B. *Discrete-continual approximation model for subdomain*

Discrete-continual approximation model is used for three-dimensional problems. It presupposes mesh approximation for non-basic dimensions of extended domain (along x_1 and x_2) while in the basic dimension (along x_3) problem remains

continual. Thus extended subdomain ω_1 is divided into discrete-continual finite elements

$$\omega_1 = \bigcup_{i=1}^{N_1-1} \bigcup_{j=1}^{N_2-1} \omega_{1,i,j} ; \quad (33)$$

$$\omega_{1,i,j} = \{ (x_1, x_2, x_3) : \; x_{1,i} < x_1 < x_{1,i+1}, \; x_{2,j} < x_2 < x_{2,j+1}, \; x_{3,1}^b < x_3 < x_{3,2}^b \}. \quad (34)$$

Lame constants for discrete-continual finite element are defined by formulas:

$$\overline{\lambda}_{1,i,j} = \theta_{1,i,j} \lambda; \quad \overline{\mu}_{1,i,j} = \theta_{1,i,j} \mu ; \quad \theta_{1,i,j} = \begin{cases} 1, & \omega_{1,i,j} \subset \Omega_1; \\ 0, & \omega_{1,i,j} \not\subset \Omega_1, \end{cases} \quad (35)$$

where $\theta_{1,i,j}$ is the characteristic function of element $\omega_{1,i,j}$ (it is defined in accordance with the method of extended domain).

Basic nodal unknown functions are displacement components $u_1^{(1)}, u_2^{(1)}, u_3^{(1)}$ and their derivatives $v_1^{(1)}, v_2^{(1)}, v_3^{(1)}$ with respect to x_3 (superscript hereinafter corresponds to the number of considered subdomain i.e. ω_1). Thus for node $(1, i, j)$ we have the following unknown functions: $u_1^{(1,i,j)}, u_2^{(1,i,j)}, u_3^{(1,i,j)}$, $v_1^{(1,i,j)}, v_2^{(1,i,j)}, v_3^{(1,i,j)}$. Bilinear approximation is used for unknown functions within element.

DCFEM is reduced at some stage to the solution of systems of $6N_1N_2$ first-order ordinary differential equations:

$$\overline{U}_1'(x_3) = A_1 \overline{U}_1(x_3) + \overline{\overline{R}}_1(x_3), \quad (36)$$

where $\overline{U}_1(x_2)$ is the global vector of nodal unknown functions (subscript corresponds to the number of subdomain ω_1),

$$\overline{U}_1 = \overline{U}_1(x_2) = [(\overline{u}_1)^T \quad (\overline{v}_1)^T]^T ; \quad (37)$$

$$\overline{u}_1 = \overline{u}_1(x_3) = [(\overline{u}_n^{(1,1,1)})^T \quad (\overline{u}_n^{(1,2,1)})^T \quad ... \quad (\overline{u}_n^{(1,N_1,1)})^T \quad \dots \quad (\overline{u}_n^{(1,1,2)})^T \quad (\overline{u}_n^{(1,2,2)})^T \quad ... \quad (\overline{u}_n^{(1,N_1,2)})^T \quad \dots \quad \dots \quad (\overline{u}_n^{(1,1,N_2)})^T \quad (\overline{u}_n^{(1,2,N_2)})^T \quad ... \quad (\overline{u}_n^{(1,N_1,N_2)})^T]^T ; \quad (38)$$

$$\overline{v}_1 = \overline{v}_1(x_3) = [(\overline{v}_n^{(1,1,1)})^T \quad (\overline{v}_n^{(1,2,1)})^T \quad ... \quad (\overline{v}_n^{(1,N_1,1)})^T \quad \dots \quad (\overline{v}_n^{(1,1,2)})^T \quad (\overline{v}_n^{(1,2,2)})^T \quad ... \quad (\overline{v}_n^{(1,N_1,2)})^T \quad \dots \quad \dots \quad (\overline{v}_n^{(1,1,N_2)})^T \quad (\overline{v}_n^{(1,2,N_2)})^T \quad ... \quad (\overline{v}_n^{(1,N_1,N_2)})^T]^T ; \quad (39)$$

$$\overline{u}_n^{(1,i,j)} = \overline{u}_n^{(1,i,j)}(x_3) = [u_1^{(1,i,j)} \quad u_2^{(1,i,j)} \quad u_3^{(1,i,j)}]^T ;$$

$$\overline{v}_n^{(1,i,j)} = \overline{v}_n^{(1,i,j)}(x_3) = [v_1^{(1,i,j)} \quad v_2^{(1,i,j)} \quad v_3^{(1,i,j)}]^T ; \quad (40)$$

A_1 is the global matrix of coefficients of order $6N_1N_2$; $\overline{\overline{R}}_1(x_2)$ is the right-side vector of order $6N_1N_2$.

Correct analytical solution of (36) is defined by formula

$$\overline{U}_1(x_3) = E_1(x_3) \overline{C}_1 + \overline{S}_1(x_3), \quad (41)$$

where * is convolution notation; $\overline{C}_1$ is the vector of constants of order $6N_1N_2$;

$$E_1(x_3) = \varepsilon_1(x_3 - x_{3,1}^b) - \varepsilon_1(x_3 - x_{3,2}^b) ;$$

$$\overline{S}_1(x_3) = \varepsilon_1(x_3) * \overline{\overline{R}}_1(x_3) ; \quad (42)$$

$\varepsilon_1(x_3)$ is the fundamental matrix-function of system (4), which is constructed in the special form convenient for problems of structural mechanics [5].

C. Discrete (finite element) approximation model for subdomain

Discrete (finite element) approximation model for the considering three-dimensional problems presupposes finite element approximation along x_1, x_2 and x_3. Thus extended subdomain ω_2 is divided into finite elements

$$\omega_2 = \bigcup_{i=1}^{N_1-1} \bigcup_{j=1}^{N_2-1} \bigcup_{r=1}^{N_3-1} \omega_{2,i,j,r} ; \quad (43)$$

$$\omega_{2,i,j,r} = \{ (x_1, x_2) : \; x_{1,i} < x_1 < x_{1,i+1}, \; x_{2,j} < x_2 < x_{2,j+1}, \; x_{3,r}^{fe} < x_3 < x_{3,r+1}^{fe} \}. \quad (44)$$

Lame constants for finite element are defined by formulas:

$$\overline{\lambda}_{2,i,j,r} = \theta_{2,i,j,r} \lambda; \quad \overline{\mu}_{2,i,j,r} = \theta_{2,i,j,r} \mu ; \quad (45)$$

$$\theta_{2,i,j,r} = \begin{cases} 1, & \omega_{2,i,j,r} \subset \Omega_2; \\ 0, & \omega_{2,i,j,r} \not\subset \Omega_2, \end{cases} \quad (46)$$

where $\theta_{2,i,j,r}$ is the characteristic function of element $\omega_{2,i,j,r}$ (it is defined in accordance with the method of extended domain).

Basic nodal unknowns are displacement components $u_1^{(2)}, u_2^{(2)}, u_3^{(2)}$ (superscript hereinafter corresponds to the number of considered subdomain i.e. ω_2). Thus for node $(2, i, j, r)$ we have the following unknowns: $u_1^{(2,i,j,r)}, u_2^{(2,i,j,r)}, u_3^{(2,i,j,r)}$. Bilinear approximation of unknowns is used within finite element (conventional three-dimensional parallelepipedic 8-node finite element of three-dimensional problem of elasticity theory).

As is well known, FEM is reduced to the solution of systems of $3N_1N_2N_3$ linear algebraic equations:

$$K_2 \overline{U}_2 = \overline{R}_2, \quad (47)$$

where K_2 is the global stiffness matrix of order $3N_1N_2N_3$; $\overline{R}_2$ is the global right-side (load) vector of order $3N_1N_2N_3$; $\overline{U}_2$ is the global vector of nodal unknowns (subscript corresponds to the number of subdomain ω_2),

$$\{\overline{U}_2\}_{i_g} = u_q^{(k,i,j,r)}; \quad (48)$$

i_g is the global index of element of vector $\overline{U}_2$; k, i, j, r, q are corresponding local indexes,

$$k = 2; \quad r = \left[\frac{i_g}{3N_1N_2}\right] + 1; \quad j = \left[\frac{i_g - 3(r-1)N_1N_2}{3N_1}\right] + 1;$$
$$i = \left[\frac{i_g - 3(r-1)N_1N_2 - 3(j-1)N_1}{3}\right] + 1; \quad (49)$$
$$q = i_g - 3(r-1)N_1N_2 - 3(j-1)N_1 - 3i. \quad (50)$$

D. Multilevel approximation model for domain

System (36) can be rewritten for all nodes with indexes $1 < r < N_3$ (i.e. $x_{3,2}^b < x_3 < x_{3,3}^b$) in the following form (the resolving system of $2N_1(N_2 - 2)$ linear algebraic equations):

$$\widetilde{K}_2 \overline{U}_2 = \overline{\overline{R}}_2, \quad (51)$$

where $\widetilde{K}_2$ is the reduced global stiffness matrix of size $[3N_1N_2(N_3 - 2)] \times [3N_1N_2N_3]$; $\overline{\overline{R}}_2$ is the reduced right-side vector of order $3N_1N_2(N_3 - 2)$.

Boundary conditions at section $x_3 = x_{3,1}^b$ (hinged edge) have the form ($3N_1N_2$ equations):

$$u_1^{(1,i,j)}(x_{3,1}^b + 0) = 0, \quad u_2^{(1,i,j)}(x_{3,1}^b + 0) = 0,$$
$$u_3^{(1,i,j)}(x_{3,1}^b + 0) = 0, \quad i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2. \quad (52)$$

Equations (52) can be rewritten in a matrix form:

$$B_1^+ \overline{U}_1(x_{2,1}^b + 0) = \overline{g}_1^+, \quad (53)$$

where B_1^+ is the matrix of boundary conditions of size $3N_1N_2 \times 6N_1N_2$ with elements defined by formula

$$\{B_1^+\}_{p,q} = \delta_{p,q}, \quad p = 1, 2, ..., 3N_1N_2, \quad q = 1, 2, ..., 6N_1N_2; \quad (54)$$

$\overline{g}_1^+$ is the zero vector of order $3N_1N_2$ (i.e. $\overline{g}_1^+ = 0$); $\delta_{p,q}$ is the Kronecker delta.

Combining (41) and (53), we get

$$B_1^+ E_1(x_{3,1}^b + 0)\overline{C}_1 = \overline{g}_1^+ - B_1^+ \overline{S}_1(x_{3,1}^b + 0) \quad \text{or} \quad Q_1 \overline{C}_1 = \overline{G}_1, \quad (55)$$

where Q_1 is the matrix of size $3N_1N_2 \times 6N_1N_2$; $\overline{G}_1$ is the vector of order $3N_1N_2$;

$$Q_1 = B_1^+ E_1(x_{3,1}^b + 0); \quad \overline{G}_1 = \overline{g}_1^+ - B_1^+ \overline{S}_1(x_{3,1}^b + 0). \quad (56)$$

Boundary conditions at section $x_3 = x_{3,2}^b$ (perfect contact) have the form ($6N_1N_2$ equations):

$$u_q^{(1,i,j)}(x_{3,2}^b - 0) = u_q^{(2,i,j,r)}, \quad i = 1, 2, ..., N_1,$$
$$j = 1, 2, ..., N_2, \quad r = 1, \quad q = 1, 2, 3; \quad (57)$$
$$\sigma_{1,3}^{(1,i,j)}(x_{3,2}^b - 0) = \sigma_{1,3}^{(2,i,j,r)}, \quad \sigma_{2,3}^{(1,i,j)}(x_{3,2}^b - 0) = \sigma_{2,3}^{(2,i,j,r)},$$
$$\sigma_{3,3}^{(1,i,j)}(x_{3,2}^b - 0) = \sigma_{3,3}^{(2,i,j,r)}, \quad (58)$$
$$i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2, \quad r = 1;$$

where $\sigma_{1,3}^{(1,i,j)}(x_3)$, $\sigma_{1,3}^{(2,i,j,r)}$, $\sigma_{3,3}^{(1,i,j)}(x_3)$ are nodal functions (after corresponding averaging) of stress components $\sigma_{1,3}(x_3)$, $\sigma_{2,3}(x_3)$, $\sigma_{3,3}(x_3)$ for element $(1, i, j)$; $\sigma_{1,3}^{(2,i,j,r)}, \sigma_{2,3}^{(2,i,j,r)}, \sigma_{3,3}^{(2,i,j,r)}$ are nodal stress components $\sigma_{1,3}$, $\sigma_{2,3}$, $\sigma_{3,3}$ (after corresponding averaging) for finite element $(2, i, j, r)$; $r = 1$.

Equations (57) and (58) can be rewritten in a matrix form:

$$B_2^- \overline{U}_1(x_{3,2}^b - 0) = B_2^+ \overline{U}_2, \quad (59)$$

where B_2^- is the matrix of boundary conditions of size $6N_1N_2 \times 6N_1N_2$ (it can be constructed in accordance with so-called method of basis variations [22]); B_2^+ is the matrix of boundary conditions of size $6N_1N_2 \times 3N_1N_2N_3$ (it can be constructed in accordance with the method of basis variations).

Combining (12) and (30) we get

$$B_2^- E_1(x_{3,2}^b - 0)\overline{C}_1 - B_2^+ \overline{U}_2 = -B_2^- \overline{S}_1(x_{3,2}^b - 0) \quad \text{or}$$
$$Q_{2,1}\overline{C}_1 + Q_{2,2}\overline{U}_2 = \overline{G}_2, \quad (60)$$

$Q_{2,1}$ is the matrix of size $6N_1N_2 \times 6N_1N_2$; $Q_{2,2}$ is the matrix of size $6N_1N_2 \times 3N_1N_2N_3$; $\overline{G}_2$ is the vector of order $6N_1N_2$,

$$Q_{2,1} = B_2^- E_1(x_{3,2}^b - 0); \quad Q_{2,2} = -B_2^+; \quad \overline{G}_2 = -B_2^- \overline{S}_1(x_{3,2}^b - 0). \quad (61)$$

Boundary conditions at section $x_3 = x_{3,3}^b$ (hinged edge) have the form ($3N_1N_2$ equations):

$$u_1^{(2,i,j,r)} = 0, \quad u_2^{(2,i,j,r)} = 0, \quad u_3^{(2,i,j,r)} = 0,$$
$$i = 1, 2, ..., N_1, \quad j = 1, 2, ..., N_2, \quad r = N_3. \quad (62)$$

Equations (62) can be rewritten in a matrix form:

$$B_3^- \overline{U}_2 = \overline{g}_3^-, \quad (63)$$

where B_3^- is the matrix of boundary conditions of size $3N_1N_2 \times 3N_1N_2N_3$ with elements defined by formula

$$\{B_3^-\}_{p,q} = \delta_{p,q}, \quad p = 1, 2, ..., 3N_1N_2, \quad q = 1, 2, ..., 3N_1N_2N_3; \quad (64)$$

$\bar{g}_3^-$ is the zero vector of order $3N_1N_2$ (i.e. $\bar{g}_3^- = 0$).

Thus, the total number of equation is equal to $3N_1N_2N_3 + 6N_1N_2$. Corresponding coupled system of $3N_1N_2N_3 + 6N_1N_2$ linear algebraic equations with $3N_1N_2N_3 + 6N_1N_2$ unknowns has the form:

$$\begin{bmatrix} Q_1 & 0 \\ Q_{2,1} & Q_{2,2} \\ 0 & \tilde{K}_2 \\ 0 & B_3^- \end{bmatrix} \begin{bmatrix} \bar{C}_1 \\ \bar{U}_2 \end{bmatrix} = \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \\ \tilde{\bar{R}}_2 \\ \bar{g}_3^- \end{bmatrix}. \quad (65)$$

It should be noted that boundary conditions (63) can be taken into account automatically within construction of the global stiffness matrix and the global right-side vector corresponding to subdomain ω_2. The result is

$$\begin{bmatrix} Q_1 & 0 \\ Q_{2,1} & Q_{2,2} \\ 0 & \tilde{\tilde{K}}_2 \end{bmatrix} \begin{bmatrix} \bar{C}_1 \\ \bar{U}_2 \end{bmatrix} = \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \\ \tilde{\bar{R}}_2 \end{bmatrix}, \quad (66)$$

where $\tilde{\tilde{K}}_2$ is the corresponding reduced global stiffness matrix of size $[3N_1N_2(N_3 - 1)] \times [3N_1N_2N_3]$; $\tilde{\bar{R}}_2$ is the corresponding reduced global right-side vector of order $3N_1N_2(N_3 - 1)$.

Strain and stress components are computed according to well-known formulas after solving of system (67).

E. Software

We should stress that methods and algorithms considered in this paragraph have been realized in proprietary software. The main purpose of software system CSASA3D (DCFEM + FEM) is semianalytical structural analysis (static analysis of three-dimensional structures within three-dimensional theory of elasticity), based on combined application of FEM and DCFEM. Programming environment is Microsoft Visual Studio 2013 Community and Intel Parallel Studio 2017XE with Intel MKL Library [23]. Software system is designed for Microsoft Windows 8.1/10.

IV. VERIFICATION SAMPLES

Let us consider verification sample dealing with static analysis of three-dimensional structure loaded with two uniform loads over x_1 ($P = 100$ kN) with hinged ends (cross-sections) along the basic dimension (Fig. 4). It should be noted that for verification purposes, symmetric problem of static structural analysis is considered. Besides, one half of this structure is modelled with the use of DCFEM and the other half of structure is modelled with the use of FEM. This may explain some asymmetry in the results of analysis.

Domain, occupied by structure, is divided into two subdomains. The right subdomain is approximated by finite elements (within FEM) and the left subdomain is approximated by discrete-continual finite elements (within DCFEM). Modulus of elasticity of material of structure is equal to $E = 3000$ kN/sm2. Poisson's ratio of material of structure is equal to $\nu = 0.16$. The material of the structure is linearly elastic and homogeneous throughout the domain (such case is considered for the sake of simplicity and clarity). Some results of structural analysis are presented at Fig. 5-8.

ANSYS Mechanical 15.0 [24] was used for verification purposes. We can conclude, that results of analysis obtained by the ANSYS Mechanical and CSASA2D generally agree well with each other. According to expectations, DCFEM is more effective in the most critical, vital, potentially dangerous areas of structure in terms of fracture (areas of the so-called edge effects), where some components of solution are rapidly changing functions and their rate of change in many cases can't be adequately taken into account by the standard FEM [1,19].

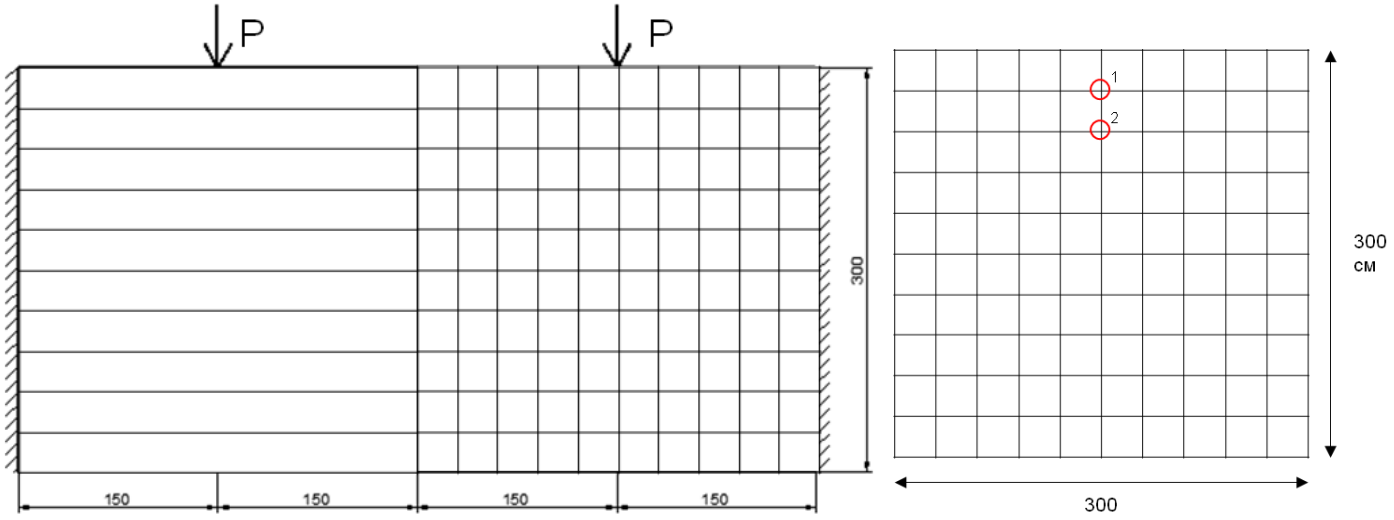


Figure 4. Verification sample and sections for verification.

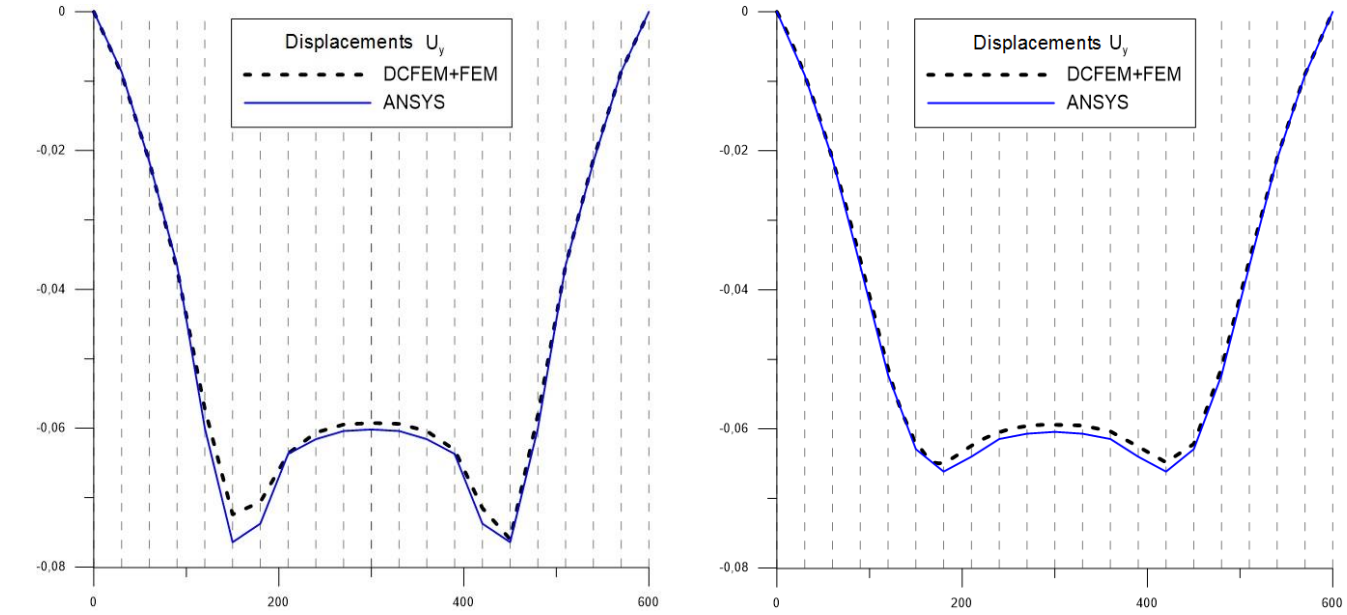


Figure 5. Comparison of results, obtained by ANSYS Mechanical and CSASA2D (DCFEM + FEM): distribution of nodal displacements along section x_2 (cm).

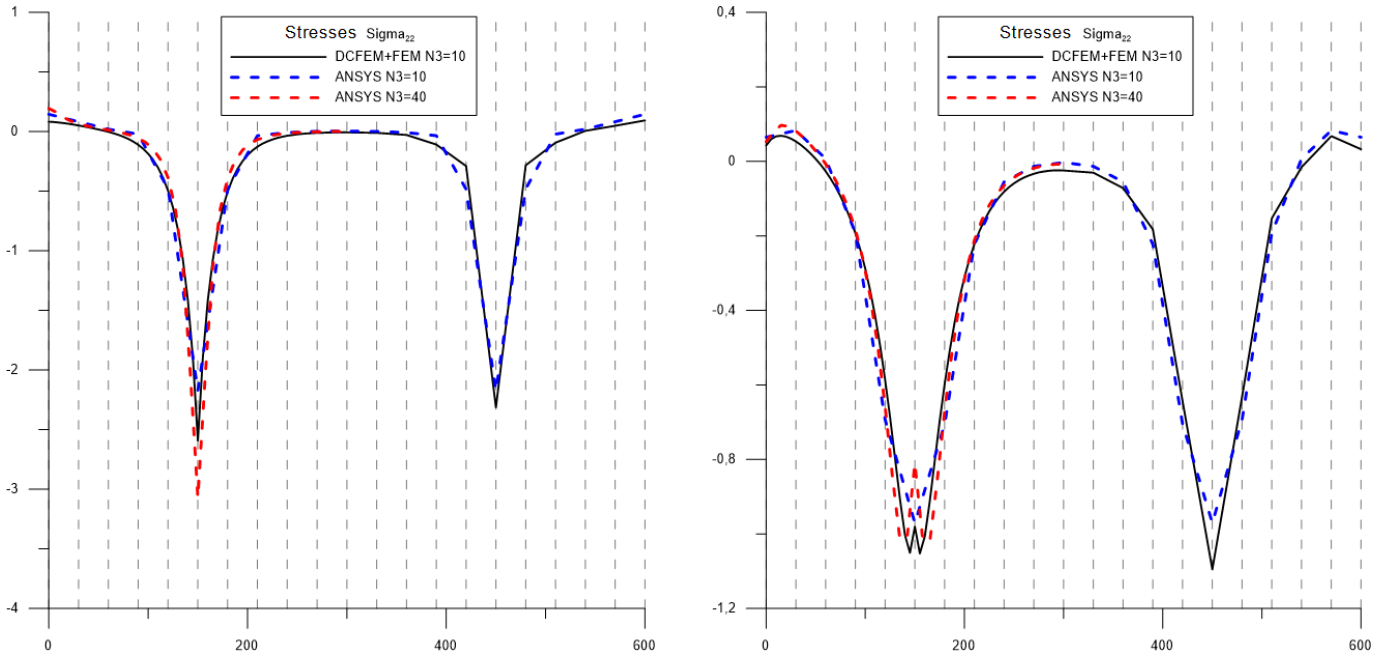


Figure 6. Comparison of results, obtained by ANSYS Mechanical and CSASA2D (DCFEM + FEM): distribution of nodal stresses $\sigma_{2,2}$ (kN/cm2)

Generally it should be noted that structural discontinuities in the analytical direction can be taken into account, by addition of new appropriate boundary condition at the relevant section.

V. Conclusions and an Outlook on Future Work

DCFEM has completely computer-oriented algorithm, computational stability, optimal conditionality of resultant systems and applicable for the various loads at arbitrary points or areas of the considering structure. This method allows increasing the accuracy of the solution and provides reduction of the computational efforts [8]. Moreover, DCFEM is reliable basis for their further development and practical application for modelling of behavior of a wide class of structures, by "integrating" this method as effective alternative modules into advanced finite / superelement software systems, which are used for analysis of typical and unique building structures, buildings and facilities at the stages of their design and operation (structural health monitoring).

Main advantages of combined application of FEM and DCFEM include wide range of applicability of this approach; high accuracy of determination of parameters of stress-strain state in construction parts with regular physical and geometric characteristics (parameters) along one of the directions; orientation to implementation in the computational software systems of industrial type.

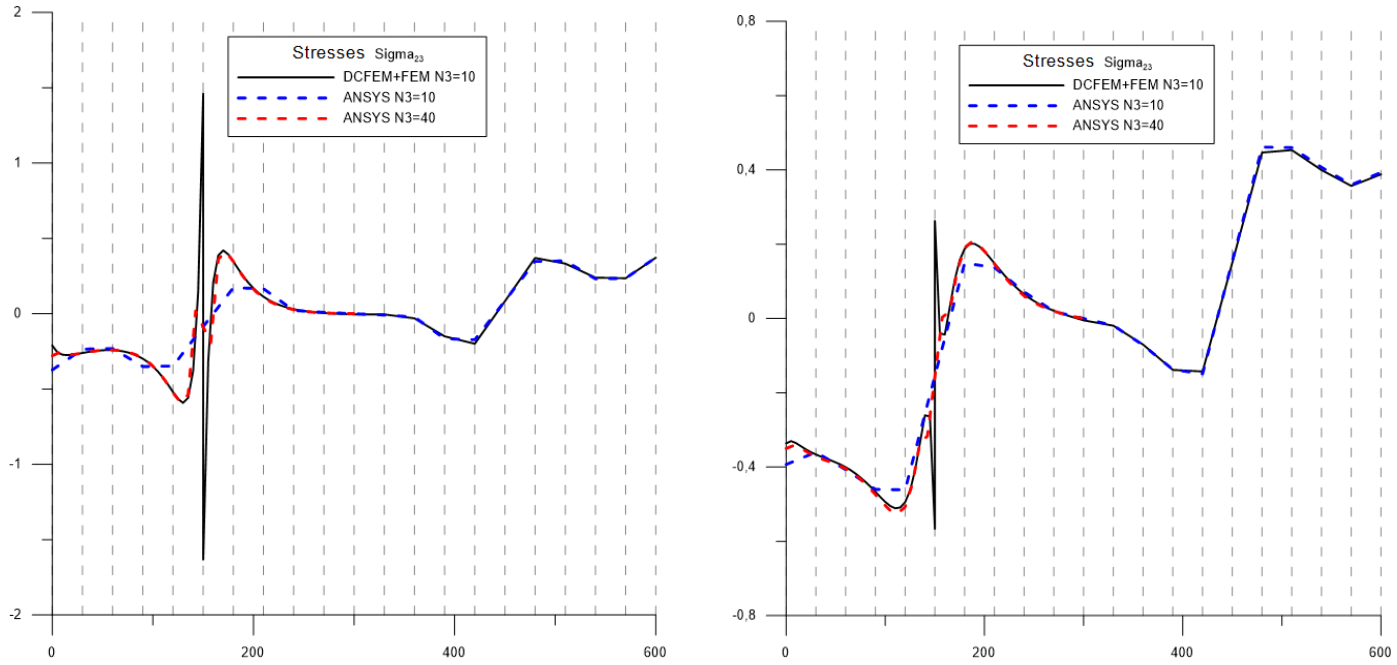


Figure 7. Comparison of results, obtained by ANSYS Mechanical and CSASA2D (DCFEM + FEM): distribution of nodal stresses $\sigma_{2,3}$ (kN/cm2).

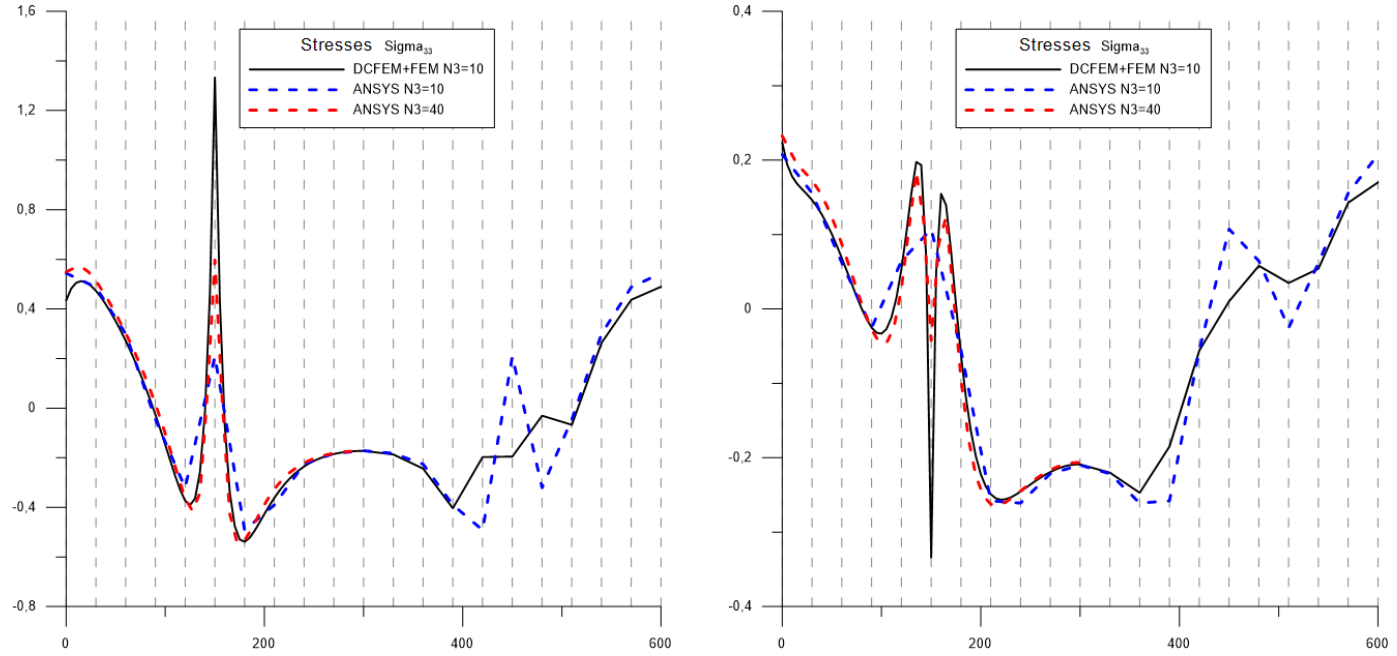


Figure 8. Comparison of results, obtained by ANSYS Mechanical and CSASA2D (DCFEM + FEM): distribution of nodal stresses $\sigma_{3,3}$ (kN/cm2).

Vital tendency of future development of DCFEM is associated with the use of multigrid approach for semianalytical structural analysis [9]. Since the pioneering work of R.P. Fedorenko [25], multigrid literature has grown at an astonishing rate.

Multigrid methods are among the most efficient iterative algorithms for solving the linear algebraic systems associated with elliptic partial differential equations [26-34]. As is known, the algorithm requires a series of problems be "solved" on a hierarchy of grids with different mesh sizes. For many problems, it is possible to prove that its execution time is asymptotically optimal. The niche of multigrid algorithms is large-scale problems where this asymptotic performance is critical. The need for high-resolution partial differential equations (PDE) simulations has motivated the parallelization of multigrid algorithms. However, these algorithms require to handle the operations on several coarser levels where the communication costs are higher than the computation costs.

The main steps in development of multigrid methods will include: coarsening the fine grid (corresponding fine grid operator can be generated by finite element application), choosing grid transfer operators to move between meshes (i.e.,

the restriction and prolongation operators), determining the coarse mesh discretization matrices and finally development of appropriate smoothers. Development of effective multigrid methods often boils down to striking a good balance between setup times, convergence rates, and cost per iteration. These features in turn depend on operator complexity, coarsening rates, and smoother effectiveness.

ACKNOWLEDGMENT

The Reported study was funded by Government Program of the Russian Federation "Development of science and technology" (2013-2020) within Program of Fundamental Researches of Ministry of Construction, Housing and Utilities of the Russian Federation and Russian Academy of Architecture and Construction Sciences, the Research Project 7.1.1".

REFERENCES

[1] O.C. Zienkiewicz, R.L. Taylor and D.D. Fox, The Finite Element Method for Solid and Structural Mechanics. Butterworth-Heinemann, 2013.

[2] R. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems. Classics in Applied Mathematics, SIAM, 2007.

[3] S.A. Sauter and C. Schwab, Boundary element methods. Springer Berlin Heidelberg, 2011.

[4] J. Kevorkian, Partial differential equations: Analytical solution techniques. Vol. 35, Springer, 2000.

[5] P.A. Akimov, "Correct Discrete-Continual Finite Element Method of Structural Analysis Based on Precise Analytical Solutions of Resulting Multipoint Boundary Problems for Systems of Ordinary Differential Equations", Applied Mechanics and Materials, Vols. 204-208, pp. 4502-4505, 2012.

[6] Y.K. Cheung, Finite Strip Method in Structural Analysis. Pergamon Press, Oxford – New York – Toronto – Sydney – Paris – Frankfurt, 1976.

[7] C.T. Christov and L.B. Petrova, "Comparison of Some Variants of the Finite Strip Method for Analysis of Complex Shell Structures", Internationales Kolloquium über Anwendungen der Informatik und Mathematik in Architektur und Bauwesen, IKM, Weimar, Bauhaus-Universitat, Vol. 15, 2000.

[8] P.A. Akimov, M. Aslami, M.L. Mozgaleva and Zh.I. Mskhalaya, "Semianalytical Analysis of Shear Walls With the Use of Discrete-Continual Finite Element Method. Part 1: Mathematical Foundations", MATEC Web Conf., Vol. 86, 2016.

[9] P.A. Akimov, M. Aslami and M.L. Mozgaleva, "Semianalytical Analysis of Shear Walls With the Use of Discrete-Continual Finite Element Method. Part 2: Numerical Examples, Future Development", MATEC Web Conf., Vol. 86, 2016.

[10] P.A. Akimov, M.L. Mozgaleva, Mojtaba Aslami and O.A. Negrozov, "About Verification of Discrete-Continual Finite Element Method of Structural Analysis. Part 1: Two-Dimensional Problems", Procedia Engineering, Vol. 91, pp. 2-7, 2014.

[11] P.A. Akimov, M.L. Mozgaleva and O.A. Negrozov, "About Verification of Discrete-Continual Finite Element Method for Two-Dimensional Problems of Structural Analysis. Part 1: Deep Beam with Constant Physical and Geometrical Parameters Along Basic Direction", Advanced Materials Research, Vols. 1025-1026, pp. 89-94, 2014.

[12] P.A. Akimov, M.L. Mozgaleva and O.A. Negrozov, "About Verification of Discrete-Continual Finite Element Method for Two-Dimensional Problems of Structural Analysis. Part 2: Deep Beam with Piecewise Constant Physical and Geometrical Parameters Along Basic Direction", Advanced Materials Research, Vols. 1025-1026, pp. 95-103, 2014.

[13] P.A. Akimov, M.L. Mozgaleva and V.N. Sidorov, "About Verification of Discrete-Continual Finite Element Method of Structural Analysis. Part 2: Three-Dimensional Problems", Procedia Engineering, Vol. 91, pp.14-19, 2014.

[14] P.A. Akimov and O.A. Negrozov, "On the Use of Discrete-continual Finite Element Method with Unstructured Meshes", Procedia Engineering, Vol. 111, pp. 8-13, 2015.

[15] P.A. Akimov and O.A. Negrozov, "On the Use of Discrete-continual Finite Elements with Triangular Cross-section for Semianalytical Structural Analysis", Procedia Engineering, Vol. 111, pp. 14-19, 2015.

[16] P.A. Akimov and O.A. Negrozov, "Semianalytical Structural Analysis Based on Combined Application of Finite Element Method and Discrete-continual Finite Element Method Part 1: Two-Dimensional Theory of Elasticity", Procedia Engineering, Vol. 153, pp. 8-15, 2016.

[17] P.A. Akimov and O.A. Negrozov, "Semianalytical Structural Analysis Based on Combined Application of Finite Element Method and Discrete-continual Finite Element Method Part 2: Three-Dimensional Theory of Elasticity", Procedia Engineering, Vol. 153, pp. 16-23, 2016.

[18] O.A. Negrozov, P.A. Akimov and I.Yu. Lantsova, "Semianalytical Structural Analysis Based on Combined Application of Finite Element Method and Discrete-continual Finite Element Method Part 4: Verification Samples", Procedia Engineering, Vol. 153, pp. 926-932, 2016.

[19] O.C. Zienkiewicz, R.L. Taylor and J.Z. Zhu, The Finite Element Method: Its Basis and Fundamentals. Butterworth-Heinemann, Sixth edition, 2005.

[20] W.X. Zhong and X.X. Zhong, "Computational structural mechanics, optimal control and semi-analytical method for PDEs", Computers & Structures, Vol. 37(6) , pp. 993-1004, 1990.

[21] P.A. Akimov and M.L. Mozgaleva, "Method of Extended Domain and General Principles of Mesh Approximation for Boundary Problems of Structural Analysis", Applied Mechanics and Materials, Vols. 580-583, pp. 2898-2902, 2014.

[22] M.L. Mozgaleva, P.A. Akimov, Z.I. Mskhalaya and V.A. Kharitonov, "About Method of Basis (Local) Variations and its Applications in Structural Analysis", Procedia Engineering, Vol. 153, pp. 501-508, 2016.

[23] R.J. Hanson and T. Hopkins, Numerical Computing With Modern Fortran (Applied Mathematics). SIAM-Society for Industrial and Applied Mathematics, 2013.

[24] ANSYS 15.0. User's Guide. Canonsburg. 2014.

[25] R.P. Fedorenko, "A Relaxation Method for Solving Elliptic Difference Equations", USSR Computational Math. and Math. Phys., Vol. 1(5) , pp. 1092-1096, 1962.

[26] H. Boffy and C.H. Venner, "Multigrid Solution of the 3D stress field in strongly heterogeneous materials", Vol. 74, pp. 121-129, 2014.

[27] W.L. Briggs, V.E. Henson and S. McCormick, A multigrid tutorial. SIAM, 2000.

[28] A.K. Garg, I. Kumar, S. Bansal and I. Goyal, "Multigrid Approach for Solving Elliptic Type Partial Differential Equations", International Journal of Science and Research, Vol. 3, Issue 4, pp. 473-475, 2014.

[29] H. Gua, J. Rethorea, M.-C. Baiettoa, P. Sainsota, P. Lecomte-Grosbrasb, C.H. Vennerc and A.A. Lubrechta, "An efficient MultiGrid solver for the 3D simulation of composite materials", Computational Materials Science, Vol. 112, Part A, pp. 230-237, 2016.

[30] W. Hackbusch, Multigrid methods and applications. Vol. 4 of Computational Mathematics, Springer – Verlag, 1985.

[31] K.S. Kang, "Scalable implementation of the parallel multigrid method on massively parallel computers", Computers & Mathematics with Applications, Vol. 70, Issue 11, pp. 2701-2708, 2015.

[32] C. Rodrigo, F.J. Gaspar and F.J. Lisbona, "Multigrid methods on semi-structured grids", Archives of Computational Methods in Engineering, Vol. 19(4) , pp. 499-538, 2012.

[33] I.M. Smith, Programming the Finite Element Method. John Wiley & Sons Ltd, 2004.

[34] U. Trottenberg, C. Oosterlee and A. Schuller, Multigrid. Academic Press, 2001.

Practical Implementation of High Power and Efficiency Dc-dc Full-Bridge PWM Boost Converter

Sofia Alexandrova
Nikolay Nikolaev
Olga Slita
Dept. of Control Systems and Informatics
ITMO University
Saint-Petersburg, Russia
alexandrova_sophie@mail.ru

Andrey Baev
Michail Goncharenko
Control Devices Development Sector
JSC Research Institute of Fine Mechanics
Saint-Petersburg, Russia
baevap.niitm@gmail.com

***Abstract*—Design and simulation problems of high power full-bridge boost converter with 175…320 VDC supply voltage are considered. The converter under investigation consists of a full-bridge inverter, a boost high-frequency transformer, a diode rectifier connected to a capacitive filter and an active load. Additional inductance, connected in series with the transformers primary winding, is brought in the converters structure to achieve soft commutation of power switches and limitation of the current switched by them, in order to improve the reliability of the device and increase its efficiency of energy conversion. Selection of the additional inductance value is an important task, because too much of it could not allow to provide load power requirements, and too small of it could bring about defects of expensive power semiconductor elements. The choice of additional inductance is also complicated by the difficulty of measuring the transformer leakage inductance with sufficient accuracy. This problem is solved using the proposed method of selection the additional inductance value, based on an analysis of the mathematical model and on an analytical description of the output inverter current curve. We also propose increasing of energy transformation efficiency by variation of the PWM carrier frequency. The curves that measured on real device 100 kW (175 ... 320V / 610V) show correctness of the model and the proposed method of selection of the carrier frequency and the additional inductance value.**

Keywords—full-bridge inverter; boost converter; soft commutation; phase-shift control; transformer leakage inductance; variable carrier frequency.

I. INTRODUCTION

Power electronics development opens up prospects of energy converters design with high efficiency of semiconductor components usage and simultaneous improvement of weight-and-size indices of devices and cost reducing. But wide range of components, topologies and control schemes complicates the choice of an optimal topology of power cascade with reliable performance of the designed device. Simulation of impulse systems can solve this problem partly.

Scheme of full-bridge inverter with boost transformer is implemented for low input and high output voltage at high power and frequency. The advantage of this scheme compared to buck-boost converter is galvanic isolation, and compared to half-bridge is half the value of switching current. But introduction of transformer complicates the converter analysis due to its non-ideality: it is rather difficult to estimate with sufficient accuracy the value of the transformer leakage inductance which brings significant changes to performance characteristics of the bridge inverter. At the same time, for the chosen topology “soft” commutation of the power switches could be ensured: zero voltage switching (ZVS) mode of the power switches for full supply voltage range, and zero current switching (ZCS) for values closed to the minimum of supply voltage [1].

This type of converters is implemented in various industries, such as oil plants, marine power systems and widely applied frequency converters with DC input voltage. There are different approaches to improving the scheme topology of ZVS full-bridge converters by additional components introduction such as auxiliary inductances, diodes, serially connected transformers [2-8], but selection of these components and analysis of their characteristics requires time and increases the cost of the device.

In this paper we consider the full-bridge converter which topology consists of minimal quantity of elements and includes a full-bridge IGBT inverter, a boost transformer and a diode rectifier. The transformer leakage inductance value might not be enough for sot commutation mode ensuring and to limit the switching currents. In this case it is necessary to include additional inductance which can be connected both to primary and secondary transformer winding. Determination of the inductance value is a very important problem as its large values can lead to increasing of losses. Some researchers suggest determining its value from conditions based on unknown values: drain-to-source capacitance of MOSFET or collector-to-emitter capacitance of IGBT [6, 7, 10].

There are also might be difficult to determine leakage inductance of real transformer windings as its value depends on frequency, core design, mutual position of primary and secondary windings and number of turns. There are many

This work was supported by the Government of the Russian Federation (grant 074-U01), by the Ministry of Education and Science of Russian Federation (project 14.Z50.31.0031).

approaches to determine these values [11-13] - either experimental or theoretical investigations. Experimental approaches require taking into account ratio of measuring device accuracy and the measured value. Theoretical approaches require knowledge of design parameters values, but some of them might be unknown.

In [14] approaches to efficiency rise by PWM frequency variation are considered for the case when load varies in the range of 5-100%, and an adaptive control system realization is described in detail. In this paper analysis of the current curve is presented which allows determining frequency for different values of supply voltage.

In this paper an approach for selection of an additional inductance is proposed: maximum value is obtained by output inverter current curve analytical description taking into consideration variation of PWM frequency depending on input voltage.

The reminder of the paper is organized as follows. Section "Problem Statement" contains the converter functional diagram and its main characteristics. Also the problem of selection an additional inductance value is stated. A model and a simplified equivalent diagram of the investigated converter which allow analytical estimation of the additional inductance maximum value and PWC frequency calculation are received in section "Main Result". Section "Example" contains data from the real device with power 100 kW (input 175-320 VDC; output 610V/164 A) and a numerical example which confirms correctness of the model and proposed method of the additional inductance value selection.

II. Problem Statement

Let us list the main parameters of the designed boost voltage converter: supply voltage is 175-320 V, rated supply voltage 250-280V, output stabilized voltage is 610 V. Maximum current switched by IGBT (two intelligent power modules are used PM800DV1B060 Mitsubishi Electric Semiconductor [15]) is 1250 A. Minimal output power is 100 kW. Parameters of the device power elements: transformer ratio is 1:6, capacitor bank capacity is 9900 μF, optimal commutation frequencies up to 10 kHz.

Converter considered in this paper consists of a full-bridge inverter loaded with power transformer which is connected to a rectifier. Such a system should maintain predetermined average voltage (610 V) on the output of diode rectifier and filter by feedback control. Output voltage should be constant despite the changes of the input voltage. Selected structure of the converter includes transformer leakage inductance which value is impossible to be determined exactly. This fact complicates the choice of the additional inductance.

Functional chart of the converter power part is shown in Fig.1. The following notations are used: *DC* is DC voltage source, *VT1-VT4* are IGBT transistors of the full-bridge inverter, *VD1-VD4* are diodes, *L* is additional inductivity, *T* is high frequency boost transformer, C_f is a capacitive filter, R_l is an active load resistance.

Let us state the problem of obtaining a computational model of the considered boost converter (see Fig.1) and output inverter current curve analytical description to determine the maximum value of additional inductance for the ensuring soft commutation mode and to limit switching current and optimize energy transformation.

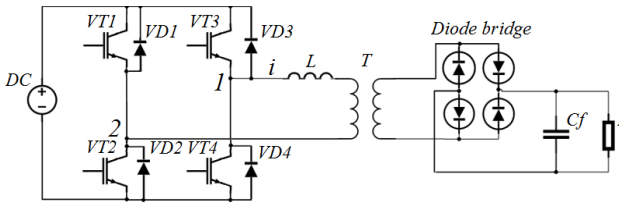


Figure 1. Functional chart of the converter

III. Main Result

Consider model of the boost converter of 100 kW power with phase-shift switching algorithm of IGBT transistors.

The computational model of this system was created with Power Elements toolbox of Matlab Simulink and is shown in Fig. 2. This model includes the subsystem "Bridge Inverter" shown in Fig 3.

In Fig. 2 we use the following notations: L is inductor of the primary winding of the transformer, C_f is capacitive filter, R_1 is active load resistance (chosen as 3.6 Ω which matches 100 kW); transformer is given as "Linear Transformer" element with rated power, frequency and voltages of primary and secondary windings. Resistance of the primary circuit is set very small, not equal to zero; its leakage inductance is 1 μH. Resistance R_m of the magnetizing circuit is set very large. Other windings parameters are set equal to zero. Value of the additional inductance is set approximately 3 μH.

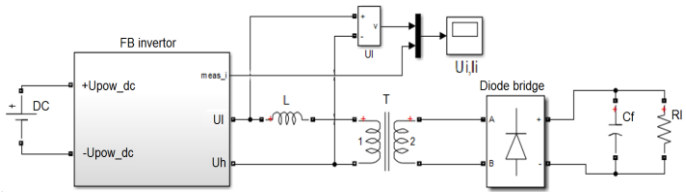


Figure 2. The computational model of the DC-DC converter.

Fig. 4 shows multipolar rectangular pulses of the inverter output voltage with the same duration $0.5T_{PWM}K_{PWM}$ and amplitude 230 V equal to inverter supply voltage, where T_{PWM} is period of PWM carrier frequency, K_{PWM} is PWM coefficient. Fig.4 also shows control signals *S1-S4* of the transistors *VT1-VT4* respectively. It also shows there is a time delay between transistors control signals which ensures soft commutation of switches [7], switching at zero voltage. Control signals *S3* and *S4* are shifted with respect to *S1* and S2 that is phase-shift

control is realized and pre-assigned stabilized voltage on the converter output is obtained. Fig. 5 shows current and voltage (between nodes 1 and 2 in Fig.1) curves on the output of the full-bridge inverter, PWM frequency is 7.5 kHz.

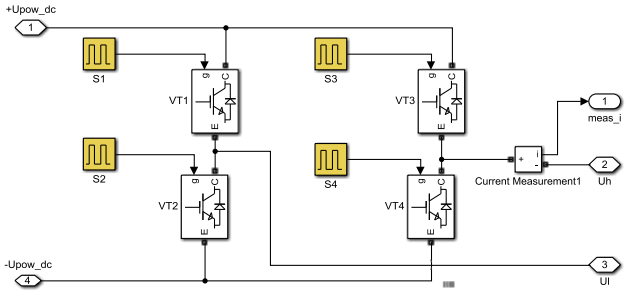


Figure 3. The subsystem "Bridge Inverter".

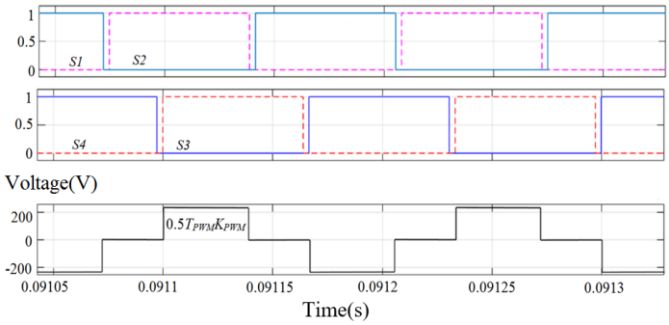


Figure 4. Control signals and output inverter voltage.

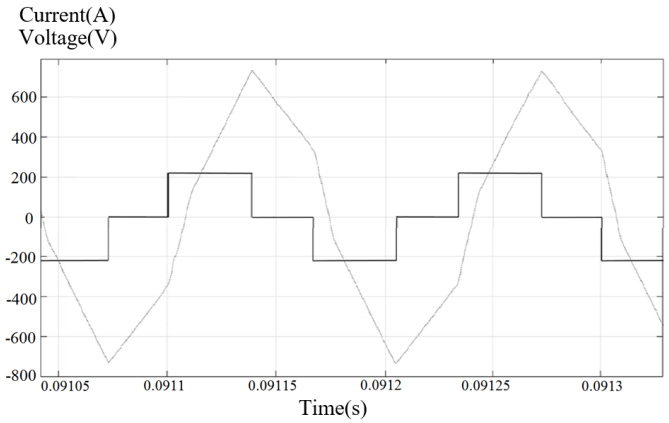


Figure 5. Output inverter current and voltage

Fig.5 illustrates the following transistors commutation law: transistors *VT1*, *VT4* are switched on the first specific current plot area (dashed lines in Fig.4). Current from supply source flows through transistor *VT1*, inductance and the transformer primary winding (let us suppose that polarity is positive) then trough transistor *VT4* to the supply source. Then *VT1* is switched off and *VT2* is switched on after time delay. During the time delay *VD2* conducts current. After switching on of *VT2* short circuit of transformer is retained (*VT2* and *VD2)*, the second specific current plot area appears and current may fall down to zero for small values of K_{PWM}. Then *VT4* is switched off, *VT3* is switched on after time delay, continuity of current in case it has not fallen to zero in ensured by diode *VD3*. Polarity of voltage of the inverter output is changed (*VT2*, *VT3* are on). If current has not fallen to zero it is conducted by *VD2*, *VD3*, and it flows into supply source. Simultaneously negative polarity current grows in the load by *VT2*, *VT3*. Total current will fall down faster until it goes through zero, after that it growth (negative polarity) slows. If current falls to zero before the transistors switching, then negative polarity current growth will be observed after the switching thanks to pair *VT2*, *VT3*. Then transients repeat with negative polarity current.

Consider the following substitution connection of the boost converter [16].

In Fig. 6 we use the following notations: L, R are inductance and resistance of the additional inductor; L_1, R_1 are leakage inductance and resistance of the primary transformer winding; L_m, L_2, R_2 are magnetizing, the secondary leakage inductance and winding resistance modified to the primary side; C_f, R_l – capacitance bank capacity and load resistance modified to the primary side.

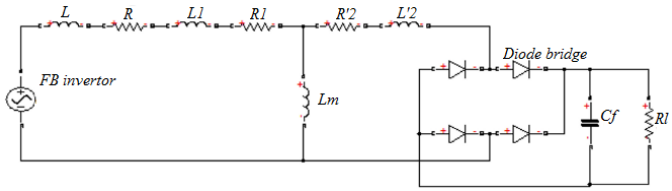


Figure 6. The equivalent circuit.

For influence analysis of the output filter capacity value on the primary winding current we use simulations with filters of different capacities. Output current and voltage plots of the inverter are shown in Fig.7 (I_{i1} corresponds to capacity of 9900 μF of capacitors bank, I_{i2} corresponds to capacity of 1100 μF).

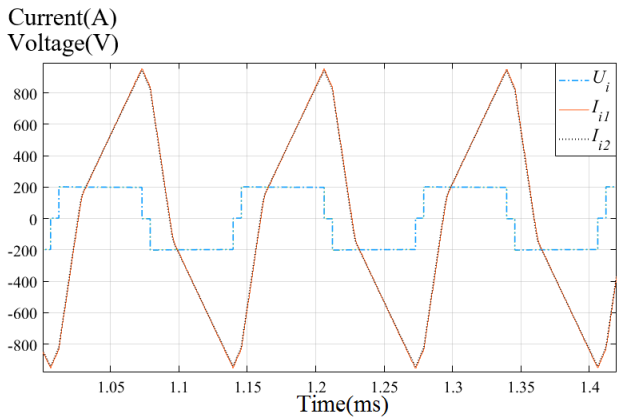


Figure 7. Output current and voltage plots of the inverter with different capacities.

Fig. 7 shows that current curves of primary winding with capacity of 9900 μF and with capacity of 1100 μF of

capacitors bank are nearly do not differ from each other. So we can conclude that due to high value of capacity a simplified substitution connection can be considered for qualitative and approximate quantitative description of the output inverter current as relation $(\omega C_f)^{-1} \ll R_l$ is correct. Fig. 8 shows simplified substitution connection where capacitor bank is substituted with ideal voltage source [17]. At short-circuiting of secondary winding through ideal voltage source relatively low current of transverse branch of the substitution transformer connection is not taken into account at calculations, so we can neglect magnetizing inductance. Then simplified substitution connection of transformer can be represented as *RL* circuit with parameters determined by parameters of transverse branch of the substitution connection shown in Fig. 1.

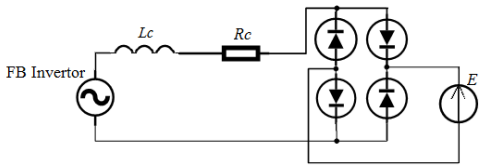


Figure 8. The simplified substitution connection.

Let us suppose that in fig.8 the inductance is $L_c = L + L_1 + L'_2$ and the resistance is $R_C = R + R_1 + R'_2$. Respect to the fact that *RL* time constant is greater than the carrier cycle, i.e. $L_c / R_c \gg T_{PWM}$ ($R_C \approx 0$), exponent transient responses could be changed by linear processes [18], so when a positive voltage pulse appears at the inverter output, the output current will grow linearly according to the law:

$$i(t) = U_{sv}(t - t_d) / L_c. \quad (1)$$

Due to Fig. 5 and Fig. 7 show steady-state processes, the beginning of positive current growth and the voltage jump are not synchronized and t_d is time delay (U_{sv} is supply voltage).

Then the current will go up until the end of the voltage pulse, the maximum value of it will be equal to:

$$I_{\max} = U_{sv}(T_{PWM} K_{PWM} - t_d) / L_c. \quad (2)$$

At the end of the positive voltage pulse, the current will go down linearly according to the law, in which the reference time is beginning of zero-level of voltage:

$$i(t) = I_{\max}(1 - R_c t / L_c). \quad (3)$$

After the occurrence of a negative voltage pulse, the rate of current decay increases:

$$i(t) = I_{\max}(1 - R_c t / L_c) - U_{sv}(t - t_p) / L_c, \quad (4)$$

where t_p is the time from reaching the maximum current to occurrence of the negative inverter output voltage.

It should be noted that in this section of the quasi-transient process a direction of an output inverter current does not coincide with the direction of the power source, i.e. the inverter supplies stored in the inductances energy to the power source. And the output inverter current lags behind the output inverter voltage.

The current will continue to go down until it becomes negative, the rectifier diodes will commutate, the polarity of the secondary winding voltage will change to the opposite, and the law of current variation will change to:

$$i(t) = -U_{sv}\, t / L_c, \quad (5)$$

where time *t* is measured from zero-crossing of current. The further process is identical to that described above, only current values are negative.

The qualitative description of the inverter output current curve requires knowledge of the value of the total inductance. Accurate calculation of the inductance value is impeded by its dependence on many plant parameters, which are adjusted during the debugging of the device. However, when some of parameters of the converter, the transformer and the capacitor bank are fixed, estimation technique the maximum permissible value of the total inductance L_c could be really effective.

Estimation of the total inductance of the *RL* circuit can be obtained from (2):

$$L_c = U_{sv}(T_{PWM} K_{PWM} - t_d) / I_{\max}. \quad (6)$$

The maximum value of L_c could be estimated from the fact that at pre-assigned values of T_{PWM}, minimum K_{PWM} and maximum supply voltage U_{SV} the transistor current should reach the possible maximum value, which is most often specified.

Fig. 9 (U_{SV} =320 V, K_{PWM} =0.17) shows that for low value of K_{PWM} time delay is approximately zero.

Then, according to the formula (6) and taking into account the fact that for a low value of K_{PWM} $t_d = 0$:

$$L_c < U_{sv\_\max}(T_{PWM} K_{PWM}) / I_{\max}. \quad (7)$$

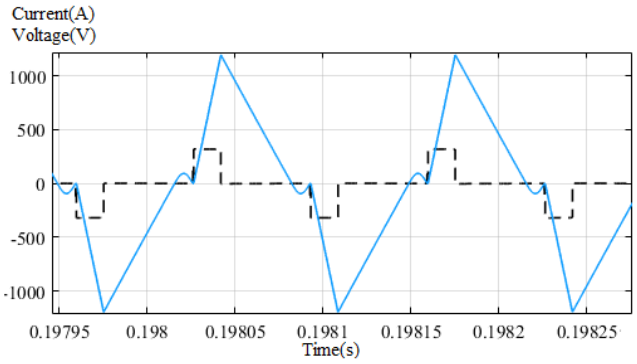


Figure 9. Output inverter current and voltage at U_{SV} =320 V, K_{PWM} =0.17.

Let us write down method of selection the additional inductance value [19].

Step 1. The minimum value of the additional inductance is chosen by using the simulation: the maximum switched current value by the IGBT does not exceed a specified value with respect to given PWM carrier frequency, the maximum supply voltage and the minimum value of K_{PWM}.

Step 2. According to (7) find the necessary and sufficient value of the total inductance L_C .

Step 3. Calculate the transformer leakage inductance by using known techniques or from the (6) by using the experimental data without the additional inductance, estimate the value of the total inductance L_C approximately.

Step 4. Having received the values of the transformer leakage inductance and L_C , estimate the value of the additional inductance.

Step 5. In practice, it should be taken into account that K_{PWM} varies with voltage supply with regard to specified power roughly linearly. Therefore, the maximum total inductance should be set slightly less than the found one.

This technique is enough to specify parameters of the model. But because of the difficulty of measuring the transformer leakage inductance with sufficient accuracy, the maximum value of the additional inductance should be selected based on the experimental data.

To study the prototype, an additional inductance should not be exceeding value $0.6L_C$.

Step 6. If it is necessary, estimate the total resistance R_C from the decreasing current section from expression (3) substituting a found value of L_C .

After estimation of L_C and R_C values it is possible to continue with analysis of thermal processes for the purpose of the transformer efficiency rise by variation of PWM carrier frequency for a prescribed range of supply voltage. Thermal processes calculation of power transistors should be carried out for average rectified current while energy balance of the system is calculated by average current of the inverter supply circuit.

As the third specific current section is absent for low values of K_{PWM} , then it is possible to estimate average output inverter current. From (1) for a known value of maximum current an expression for the first time section can be obtained in the form

$$t_1 = I_{\max} L_c / U_{sv}. \quad (8)$$

From (2) we have an expression for the second time section

$$t_2 = L_c / R_c . \quad (9)$$

Average output inverter current can be calculated if we divide sum of the two triangles' squares formed by currents of the first and the second sectors and the time axis by $T_{PWM} / 2$,

$$I_{av} = I_{\max} (L_c / R_c + I_{\max} L_c / U_{sv}) / T_{PWM}. \quad (10)$$

It is more difficult to calculate average output inverter current for larger values of K_{PWM}, when we have three sectors of current. From (1)-(5) it is possible to obtain moments of time t_d and t_p . But approximate estimation of average output inverter current is rather simple

$$I_{av} = I_{\max} / 2 . \quad (11)$$

The value of the average current is necessary for thermal calculation of power units.

It is important to estimate switching power losses of transistor modules. A known expression for transistors [17] can be used for this purpose

$$P_{SWC} = f_{sw}(E_{on} + E_{off}) \quad (12)$$

where P_{SWC} is switching power losses in W, f_{sw} is PWM carrier frequency in Hz, E_{on} , E_{off} are energies released at switching on and off of the transistor module for corresponding switching currents.

It should be noted that transistors of the module switch on at zero current and switch off at the maximum current and zero collector voltage.

For PM800 transistors [15]: for switched current $I_s = 1250$ A from plots we find E_{on} (for zero current) is about 5 mJ, E_{off} is about 50 mJ. For PWM carrier frequency of 10 kHz switching losses of the module transistor are $P_{SWC} = 10000 \cdot 55/(1000) = 550$ W, for 7.25 kHz are 400 W. So we can conclude that optimal PWM frequency is 7.25 kHz.

Increase of K_{PWM} at increase of inverter supply voltage is achievable for fixed $I_{\max}$, L_C, R_c only by increase of PWM carrier frequency which leads to increase of switching losses. So minimal operating frequency of the transformer and its characteristics which influence on L_C are very important.

The minimum value of the additional inductance is chosen using the simulation: the maximum switched current value by the IGBT does not exceed a specified value with respect to the minimum PWM carrier frequency, the maximum supply voltage and the minimum value K_{PWM}. As average value of rated operating voltage (265 V) is 1.2 times less than maximum supply voltage then required PWM carrier frequency for predetermined $I_{\max\_z} = 1250$ A is approximately 1.2 times more ≈ 8.7 kHz.

For reduction of switching losses in prototype boost converter which is a part of frequency converter adaptive control is used, carrier frequency is changed in the interval 7.25…10 kHz depending on supply voltage 175...320 V: frequency increases when K_{PWM} increases. Frequency converter is designed for power supply and control of an asynchronous motor of a compressor station drive by JSC Research Institute of Fine Mechanics (prototype is shown in Fig. 10). Boost converter is intended to perform in a small and confined space, so the main problems of its practical implementation are ensuring the load with the required power, reducing heat losses and high reliability. The prototype efficiency is 0.95. If we use equation (11) it is possible to make thermal calculation which proves appropriateness of the frequency variation.

Figure 10. Frequency converter (boost converter is on the left)

IV. EXAMPLE

The transformer leakage inductance of the research prototype is known (it is conditioned by an elaboration of the transformer construction) and approximately equal to 1 μH. The minimum value of additional inductance is determined by simulation and equal to 2 μH. For specified characteristics $U_{sv\_\max} = 320$ V, $I_{\max\_z} = 1250$ A, the calculation should be carried out for maximum PWM frequency that is for $T_{PWM} = 10^{-4}$ s as K_{PWM} varies with supply voltage with regard to specified power linearly, i.e. K_{PWM} depends on relation $0.5 \cdot U_{sv\_min} / U_{sv\_max}$ and for (7) it could be defined as $K_{PWM} \approx 0.28$. Then $L_{C\max} = 6.94$ μH. With respect to step 5 the maximum value of additional inductance is $L_{\max} = 4.2$ μH.

The experimental data of 100kW industrial prototype with specified characteristics and the additional inductance is about $L = 3$ μH are provided below in Fig.11: (a) the full-bridge inverter output voltage U_i (mirrored) and current I_i (190V/div and 600A/div) $U_{sv} = 187$ V, $K_{PWM} = 0.35$, $f_{sw} = 10$ kHz; (b,c) the full-bridge inverter output voltage U_i and current I_i for a low value K_{PWM} (190V/div and 1200 A/div) U_{SV}=320 V, $K_{PWM} = 0.18$, $f_{sw} = 7.25$ kHz. It is clear that simulation results in Fig.6 and Fig.9 and experimental results in Fig.11 are similar.

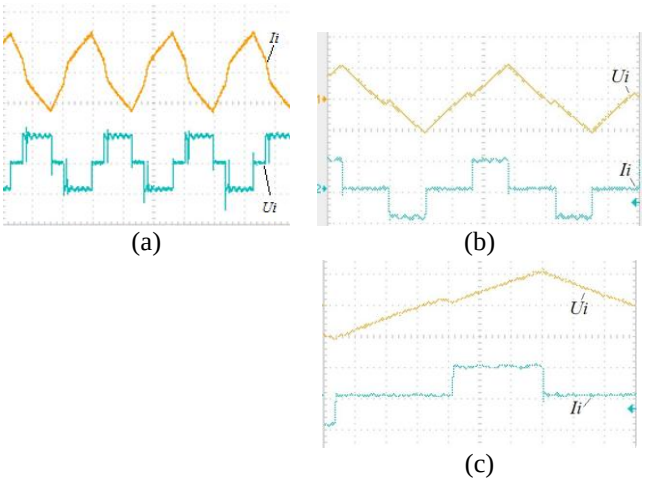


Figure 11. Experimental data.

Fig.11 illustrates that power switches work in the soft commutation mode, and it is possible to observe ZVS and ZCS for low value of K_{PWM}.

REFERENCES

[1] Baei M., Narimani M., Moschopoulos G. A. "New ZVS-PWM Full Bridge Boost Converter," Journal of Power Electronics. vol.14, 2014, pp.1-12.

[2] Koo G. B., Moon G. W., Youn M. J. "Analysis and design of phase shift full bridge converter with series-connected two transformers," IEEE Trans. Power Electron., vol. 19, no. 2, 2004, pp. 411-419.

[3] Jang Y., Jovanovic M. M. "A new family of full-bridge ZVS converter," IEEE Trans. Power Electron., vol. 19, no. 3, 2004, pp. 701-708.

[4] Jain P.K., Kang W., Soin H., Xi Y. "Analysis and design considerations of a load an line independent zero voltage switching full bridge DC/DC converter topology," IEEE Trans. Power Electron., vol. 17, no. 5, 2002, pp. 649-657.

[5] Jang Y., Jovanovic M. M. "A new PWM ZVS full-bridge converter," IEEE Trans. Power Electron.," vol. 22, no. 3, 2007, pp. 987-994.

[6] Jeon S.J., Cho G. H. “A Zero-Voltage and Zero-Current Switching Full Bridge DC-DC Converter with Transformer Isolation,” IEEE Trans. Power Electron., vol. 16, no. 5, 2001, pp. 573-580.

[7] Sabate J.A, Vlatkovic Y., Ridel R.B., Lee F.C., Cho B. “Design considerations for high-voltage high-power full-bridge zero-voltage-switched PWM converter,” IEEE Applied Power Electronics Conference Proceedings 1990, pp.275-284.

[8] Delbin J.V., Rajaram M. “A new soft switched full bridge converter with voltage-doubler-type rectifier for high voltage applications,” International Journal of Computer and Electrical Engineering, vol.4, no. 2, 2012, pp.246-251.

[9] Zhu L. “A novel soft-commutating isolated boost full-bridge ZVS-PWM dc-dc converter for biderictional high power applications,” IEEE Trans. Power Electron., vol.21, no. 2, 2006, pp. 422-429.

[10] Chen B.Y, Lai Y.S. “Switching control technique of phase-shift-controlled full-bridge converte to improve efficiency under light-load and standby conditions without additional auxiliary components” IEEE Trans. Power Electron., vol.25, no. 4, 2010, pp. 1001-1012.

[11] Petrov R. “Optimum design of a high-power, high-frequency transformer,” IEEE Trans. Power Electron., 1996, pp.33-42.

[12] Hurley W.G., Wilcox D.J. “Calculation of leakage inductance in transformer windings,” IEEE Trans. Power Electron., vol.9, no. 1, 1994, pp. 121-126.

[13] Erickson R. W., Maksimovic D. “A multiple-winding magnetics model having directly measurable parameters,” PESC 98 Record. 29th Annual IEEE Power Electronics Specialists Conference , vol.2, 1998, pp. 1472-1478.

[14] Zhao L., Li H., Liu Y., Li Z. “High efficiency variable-frequency full-bridge converter with a load adaptive control method based on the loss model,” Energies, 2015, pp. 2647-2673.

[15] Application manual. Intelligent power module PM800DV1B060 (http://www.mitsubishichips.com).

[16] Blache F., Pierre K., Cogitore B. “Stray capacitances of two winding transformers: equivalent circuit, measurements, calculation and lowering,” Conference Record of the 1994 IEEE Industry Applications Society Annual Meeting, vol.2, 1994, pp. 1211-1217.

[17] Erickson R., Maksimovic D. Fundamentals of power electronics, Norwell, Mass: Kluwel Academic. 2001. 885 p.

[18] Charles K. A., Sadiku M. N. O. Fundamentals of electric circuits— 4th ed. McGraw-Hill. 2009, 1056 p.

[19] Alexandrova S., Baev A., Goncharenko M., Nikolaev N., Slita O. “Method of additional inductance selection for full-bridge boost converter,” Proceedings of the International Scientific Conference on Physics and Control (Physcon 2017, Florence, Italy), (*to be published*).

Methods of exploring the Red Blood Cells rotation during the simulations in devices with periodic topology*

Hynek Bachratý, Kristína Kovalčíková, Katarína Bachratá and Martin Slavík
Cell-in-fluid Research Group, http://cell-in-fluid.fri.uniza.sk
Faculty of Management Science and Informatics, University of Žilina
Žilina, Slovakia
hynek.bachraty@fri.uniza.sk, kristina.kovalcikova@fri.uniza.sk

***Abstract*—Progress in the development of simulation tools makes it possible to conduct experiments with blood flow containing a large number of solid objects, typically red blood cells (RBC). In our work [2] we designed and verified a set of statistical methods allowing to either confirm or deny the compliance of the behavior of such simulations and real experiments. Rotation (angular speed) of RBC during their movement has proven to be a highly sensitive and precise tool to this exploration. In our article, we compare and confirm the compliance of several methods of measuring and detecting this parameter and its sensitivity to changes in the duration of the experiment. We use Discrete Fourier Transform (DFT) method and rank correlation methods as a key mathematical tool.**

Keywords—numerical simulation; RBC rotation; DFT; periodical behavior

I. INTRODUCTION

The research group Cell-in-fluid´s long-term goal has been to implement a detailed simulation of blood flow (its liquid component and its individual solid particles) in artificial microfluidic devices. The specific goal is mainly to optimize the shape and topology of the devices for separating circulating tumor cells (CTC) from blood samples. However, the model can be developed and used in several other fields, too. There is more detailed information on the model used in [1], [4], and on some of its possible applications e. g. in [3], [5], [6].

The main aim of the simulation experiments is to substitute the real experiments which are not possible to conduct in the required scope due to technical, time and financial reasons. Naturally, for the simulations to be meaningful, it is necessary to continuously make sure the simulation models and algorithms are precise, true and consistent with the reality. This is most frequently achieved by comparing the conduct of simulations to real experiments. In our area of research several successful comparisons have been done, however, almost all of them have concerned the behavior and attributes of the individual elastic objects, most frequently RBC, in certain specific situations. As the research is being developed and extended, and the computing options of the simulations increase, there are more possibilities and needs to verify the quality and trueness of the simulation models with the experiments with the flow of tens to hundreds of objects in larger devices. In addition to the comparison with the real experiments, this field also offers the verification possibility by comparing the results of identically designed in-silico experiments. Even though they are carried out by means of various simulation tools, comparing them is technically easier to perform. In the end, to verify the inner consistency it is also useful to compare various simulations within one model, within which we verify and compare their expected behavior.

In all cases, our current goal is to find and verify methods and methodology which will offer an appropriate answer to the question whether the compared experiments are identical or not. The nature of the matter makes us concentrate on the characteristics of the behavior, respectively the movement, of the particles, while the starting parameters of the device and the flow within are identical.

In the article [2], (simulation experiments in Fig. 1) we have designed and verified an extensive set of statistical data and methods of processing them which enable to compare and evaluate the behavior compliance of a large group of objects in a liquid flow. One group of data is acquired by continuously monitoring the movement of all individual objects during the entire experiment, which provides a continuous record of the position and velocity (in some special positions) of each of them. It is possible to acquire such data by processing a continuous video-record of the laboratory experiment. The second group is based on the data of so-called simulation snapshots (in the laboratory experiment, from its static image) which capture the positions and skews of all objects. Intentionally, we have designed the data as relatively simple, to make it easy and natural to acquire them as the simulation algorithms outputs, or by processing the image records of a real experiment. In spite of that, the data describe precisely enough the dynamics of the movement of tens to hundreds of objects during the experiment, stochasticity of the experiment, the mutual interactions of the objects and the device etc. The amount of data acquired is so large that it makes it impossible to compare them to the similar data set from the compared experiment. Therefore, the next step is processing them into a form enabling to evaluate and quantify the rate of their compliance with each other. To do so, we have designed statistical methods substantially reducing the data into the contents and forms suitable to be compared, and which,

* Supported by the Slovak Research and Development Agency [contract number APVV-15-0751]

moreover, we are able to verify using the standard statistical compliance tests.

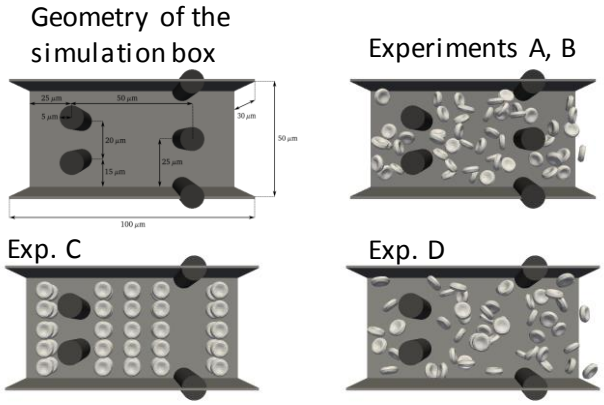


Figure 1. Series of comparing simulation experiments from the article [2]. With the experiments A and B, the identical device and randomly generated two different seedings of 50 RBCs were used. With the experiment C, while maintaining the shape of the device, regular seeding of 50 RBCs was used. The experiment D started similarly to A, B, from a random seeding of 50 RBCs, however, its topology was changed by leaving out one obstacle. All statistical methods designed confirmed the expected statistical compliance of the experiments A and B, and identified the distinctness (denied the compliance) with the experiments C and D, as well as their mutual compliance. The compliance tests were carried out by means of the standard Kolmogor – Smirnov test. Experimental setup of the simulations is as follows: All experiments were performed in the simulation software ESPResSo with usage of Object-in-fluid framework for modelling cells and LB for modelling fluid. As a model of each of the RBCs we had chosen a surface with 141 nodes and the dimensions of 7.82μm x 7.82μm x 2.56μm. The elastic coefficients of the mesh were *ks*=0.0044, *kb*=0.0715, *kal*=0.005, *kag*=1, *kv*=1.25. The simulation mass of every cell was 0.059574468085106386 pg. The interactions among the cells were modeled by membrane collision interaction with the parameter of *mc_K*=0.005, *mc_n*=2.0, *mc_cut*=0.5. The interactions between the cells and surface of the channel were ensured by soft sphere interactrion with the parameters of *soft_K*=0.00035, *soft_n*=1.0, *soft_cut*=0.5. The fluid was discretized by a three-dimensional regular lattice with the space step 1mm. The fluid viscosity was 1.5 mPa.s and the density was 1000 kg/m^3, representing blood plasma at 20 °C. The interaction parameter between fluid and object called friction was 0:025865531914893616. Different external volume forces were used to induce the flow in the x-direction.

Having verified the basic correctness of the proposed methods, we continue to explore the attributes of some of them in more details. In this article, we want to focus on analyzing RBC rotation (angular speed) while monitoring its movement continuously. It is a significant characteristic, since the course of the rotations of the individual RBC reflects the influence of the mutual interactions of the individual RBCs, their collisions with the boundaries of the device and with the obstacles inside, as well as the reactions to the diverse velocities of the liquid inside the device. Thus, by comparing the course of RBC rotations in the individual simulation experiments we can compare the quality, or verify the correctness and consistency of modelling the interactions in the individual models. Example of rotation as measured in [2] is illustrated in Fig. 2. When comparing the experiments based on the characteristics acquired from RBC rotation in our work [2] we obtained the most precise results confirming the expected compliance, respectively differences, of the individual simulations. To measure the rotation, we designed and used a simplified method based on the data easy to obtain and measure. In this contribution, we would like to verify the compliance of this method with the physical definition of rotation, as well as with other possible and suitable ways of measuring it.

II. Methods of Measuring RBC Rotation

We have found out several ways of evaluating the rotation of a cell in the simulation. In order to compare the different approaches to each other, we are going to apply all of them to the same dataset. The dataset concerns one of the cells in the numerical experiment A (Fig. 1 and Fig. 3). The results, however, have been verified by similar comparisons for the other cells of the simulation.

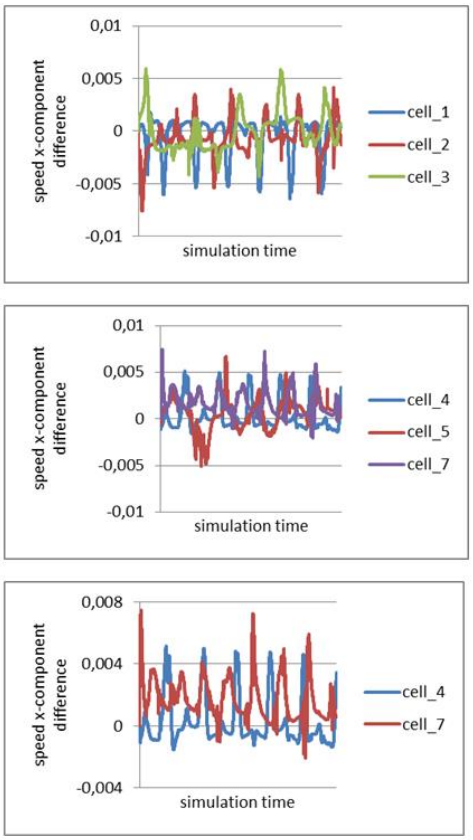


Figure 2. Presentation of simplified data describing the rotation of several RBCs (see part II. A). Vertical movement was neglected, only the horizontal components of the cell displacement are considered. Most frequently, the rotation is caused by slowing down the parts of a cell surface due to mutual collisions and collisions with the device, further on by the uneven velocity profile of the device flow, and, to a certain extent, also by the deformation changes of the cell shape. Within each graph of RBC rotation, we measured its maximum, minimum and overall rotation rate for further statistical processing. Another significant characteristics of periodic devices is also the number of changes in spin (rotation direction), and mainly the general periodicity characteristics of rotation course.

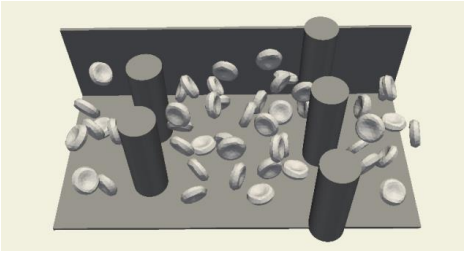


Figure 3. Disposition of the numerical simulation used for the evaluation of the different approaches. The left-to-right main direction of the flow corresponds to the positive direction of axis x, the perpendicular horizontal direction to axis y. The vertical depth of the device corresponding to axis z is usually omitted.

The records of laboratory experiments are often two-dimensional top-view videos. Therefore, if we want to compare those results to numerical simulations, the estimation of a cell-rotation has to be reduced to two dimensions as well. That is why we neglect one dimension from the calculation of the angular speed. (The second reason is the general theoretical, as well as practical complexity of measuring the rotations in a three-dimensional space.) In this article we further work with the rotation reduction into the horizontal plane *xy*, nevertheless, from the viewpoint of calculation, the reduction into the coordinate planes *xy* or *yz* is also possible and equivalent.

The approaches that are susceptible to being used for the calculation of the angular speed from a video recording must require minimum input data. The data should be easy to obtain from a post-processing of a sequence of pictures. However, they should produce results sufficiently similar to the reference one which is calculated from a set of precise data obtained from the simulation. Even obtaining such a precise data from numerical simulation happen to be impossible – it depends on the software running the simulation. Without possibility of the software modifying, we often have to be happy with more simple data outputs.

Let us note in advance that our goal has been to focus mainly on exploring and comparing the periodical properties of RBCs rotation. That enables us to do the measuring neglecting e. g. the issue of the physical units used but more precisely to monitor the rotation spin and its alterations. Thus, it will result in each graph of the angular speed issued from the different approach going to be treated with Discrete Fourier Transformation (DFT).

A. ***Simplified method*** *of rotation measuring: difference in the velocity x-component*

The above mentioned initial simplified measuring of RBC rotation used in [2] is based on the standard data output of our model, which is taken after every 200 simulation steps which corresponds to the time period of 40 microseconds and RBC shift by cca 0.5 micrometers. With on-going monitoring of RBC movement, we obtain the information on the position of its center and 6 extremal points in the direction of the coordinate axes. They are 6 selected points of the cell-surface triangulation which, viewed in the direction of axis *x*, represent its ´first´ and ´last´, ´most right´ and ´most left´, ´upper´ and ´lower´ points. For each of the points, we know the current values of its 3 velocity components in the axial directions. Consequently, we calculated the horizontal rotation of RBC as the difference of *x*-components of the velocities in *y*-extreme points (Fig. 4). The sign of this difference corresponds to the spin of the rotation. We verified the inner consistency of the method by confirming the expected compliance with the dual manner of measuring the difference of the velocity y-components in x-extreme points. Some uncertainty and therefore a need to verify the method thoroughly originate from the fact that already with the following measuring g after next 200 simulation steps, other points can be, and usually are, evaluated as extreme and used for the calculation.

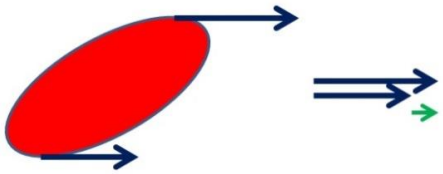


Figure 4. For Simplified method, the measure of the rotation is calculated as a difference between the x-coordinates of the velocity of two opposite extreme points (two points with the biggest and the lowest y-coordinate)

B. ***Extended method*** *of rotation measuring: difference in the position x-component*

The mentioned shortcoming of simplified method can be rectified by complementing the simulation program by supplementary calculation, in order to get more detailed output. This amendment, however, adds a burden on the course of the simulation, and therefore should not remain its standard part.

The modification consists in tracking the six extremal points and evaluating their position and velocity after another 100 steps of the simulation. This allows to monitor their actual movement and rotation. On the other hand, that means that measures are not taken every 200 steps, but twice as often. We have designed a method of measuring the rotation, which is based on collected data. Even though the method is approximative, it is fast and simple enough. Here we used as a measure of the rotation the difference between the *x*-coordinates of the ´left´ extreme point in two consecutive moments (Fig. 5). As we stipulate below, in spite of the unusual nature of the calculation, further reduced for monitoring only one of the extremal points, the method proved to be surprisingly accurate. The theoretical explanation points out in realizing the fact that considering the distance of the tracked point from the center of the rotation quasi constant, the angular speed can be calculated as a multiple of the x-coordinate of the velocity. And this one is as well a multiple of the difference between the x-coordinates of the tracked point in two subsequent measures.

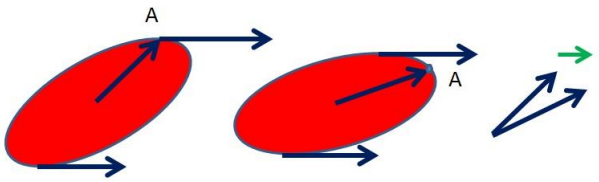


Figure 5. For Extended method, the considered measure of the rotation is difference between the x-coordinates of the ´left´ extreme point in two consecutive moments

C. ***Reference method**: Exact evaluation of the angular speed*

In order to verify the correctness of the previously mentioned methods, a more precise calculation, considering the exact physical definition of the angular speed, was provided with those additional data.

In general, there are two possible options of calculating the angular speed properly. The first one takes into account the vector of the velocity of a point (v) and its distance from the center of the rotation (r). Then the angular speed is obtained by dividing the velocity v by the distance r. The second option takes into account the position of the tracked point in two consecutive moments. The angle between the two position vectors divided by the time lapse between the two moments gives the value of the angular speed of the tracked point (Fig. 6).

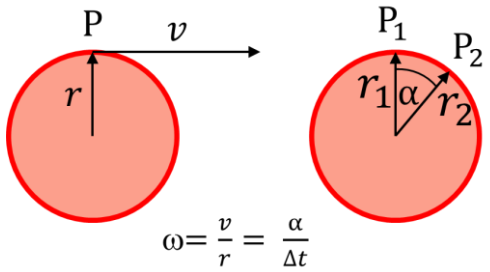


Figure 6. Two different ways of obtaining the precise value of the angular speed. Left: the position vector of the tracked point, regarding the center of rotation, and the instantaneous velocity of the tracked point. Right: positions of the tracked point in two different moments. The angle between the two position vectors and the time between the two moments are used to calculate the angular speed.

The reference approach calculates the angular speed by using the second option. Every 200th step of the simulation the point with the most important x-coordinate is identified on the cell, and its position (regarding the center of the mass) is recorded. The principle is illustrated in Fig. 7. One hundred steps later its new position is recorded. Knowing the two positions and the duration of the 100 steps of simulation, the precise value of angular speed can be evaluated. Even though the 3-dimensional vector can be evaluated this way, only 2 dimensions are considered, as explained before.

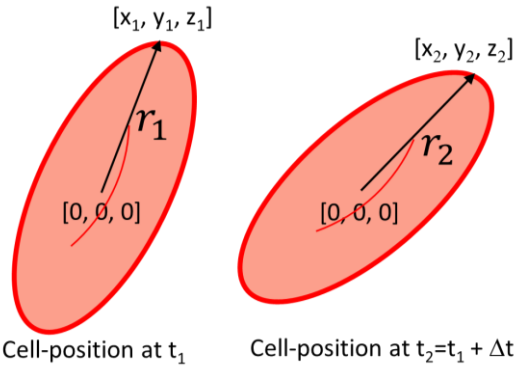


Figure 7. Tracking a fixed point on the cell, in order to evaluate the angular speed.

To calculate the angle between the two vectors we use equation (1) issued from the scalar product of two vectors. To evaluate the sense of the rotation we used the vector product of the two vectors (2). Combination of the two equations gives us whole information about the angular speed (3).

$$\cos(\alpha) = \frac{x_1 * x_2 + y_1 * y_2}{r_1 * r_2} \quad (1)$$

$$sign(\omega) = x_1 * y_2 - x_2 * y_1 \quad (2)$$

$$\omega = \frac{\alpha}{\Delta t} * sign(\omega) \quad (3)$$

Where r_1 and r_2 are two position vectors of the tracked point on the cell at time t_1 and t_2; α is the angle between these two vectors; x_1, x_2, y_1 and y_2 are coordinates of these vectors; ω is the angular speed of the cell and Δt is the time difference between t_1 and t_2 (see Fig. 7).

The exact value of the angular speed of the cell during the simulation is presented in Figure 8.

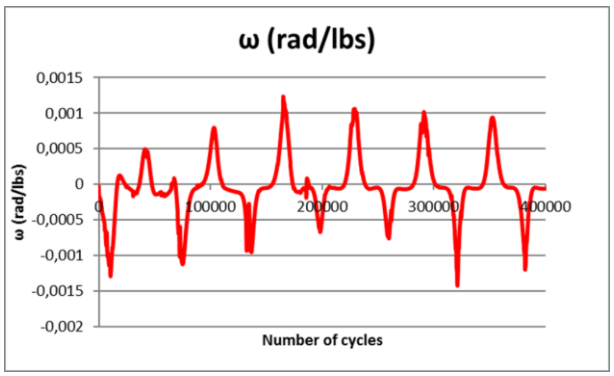


Figure 8. Example of angular speed for a cell from simulation

The disadvantage of this approach is its demand for the input data. Tracking a single point on the cell from a video recording is difficult, even more so if the tracked point is "rolled out of the view". Nevertheless, this approach gives us the exact value of the angular speed, and the required data can be obtained from any numerical simulation.

D. ***Video method***: *Evaluation of the angular speed using the circumscribed box around a cell*

Until now, we have presented methods of evaluation of the rotation from numerical simulations. This allows us to compare the numerical experiments between each other. To evaluate the rotation from a video recording, a different approach is required. As the video is a sequence of snapshots, and data about cells are coming mostly from their post-processing, the required data should be obtainable from static pictures. So, the inputs for the video method are not velocities, but only a sequence of simple positions of certain points of the cell.

An example of a set of data easily obtainable from a video recording can be demonstrated by a circumscribed box around the cell (Fig. 9). Such a box can be drawn for a cell in individual snapshots of a video recording. Then the foursome of the tangent points between the cell and the box (or x and y coordinate extremal points, as they were called in part II.A) are identified, and their positions are recorded.

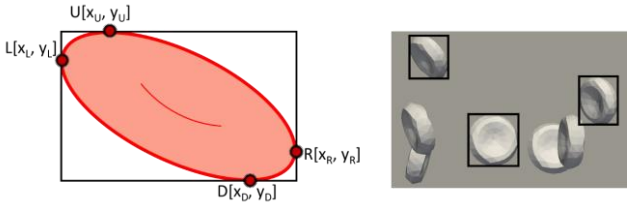


Figure 9. Circumscribed box around a cell in the simulation. The foursome of the tangent points is denoted as L, R, U and D. This labeling is defined by their position in the picture – Left, Right, Up and Down. The four points correspond to the first, last, the most left and the most right point, as defined in the section 2.A.

For our purposes, such a box was made for the tracked cell every 200 steps of the simulation. The positions of the foursome points were recorded. The center of the mass of the cell was identified as an average position of these points. Then the coordinates of the farther of the extreme points (always regarding the center of the mass) were used to calculate the angular speed. The principle presented in Figure 5 was used, by considering the sequence of the coordinates of the tangent point as the coordinates of one fixed tracked point.

This approach has several weaknesses. The first one is lack of accuracy. The coordinates of the tangent point in two consecutive moments do not correspond exactly to the coordinates of a fixed point (Fig. 10). However, this inaccuracy can be diminished by taking frequent records. This way the tangent extreme points in two consecutive snapshots are almost superposed.

There is also a reason why we should use the farther one of the points to calculate the angular speed. Using the closer point can lead to getting a false value, and even the direction of the rotation can be miscalculated. The explication is shown in Fig. 10.

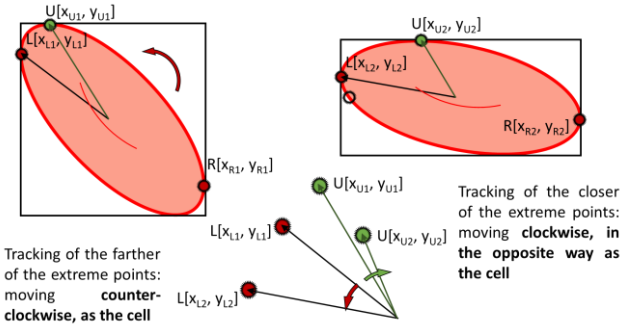


Figure 10. Inaccuracy of the approach – in the second snapshot the position of the tangent point is not the same as the position of the tracked point from the first snapshot The further point should be used to evaluate the angular speed of the cell.

In Fig 11a, we present a graph of the angular speed obtained by using the explained algorithm, without any extra treatment. We can state a significant amount of the noise in the graph. The origin of this noise is probably an occasional sudden rotation of the cell in the third direction, which is neglected in our approach.

The advantage is that this noise is not systematic, so it is possible to treat the data and to filter it out. This filtering consists of suppression of all measures, in which the distance of the tracked point from the center of the mass differs more than 0,1% from the previous measure. The graph of the treated angular speed is shown in Fig. 11b, and it is this one which is kept as a result of the video-method. There are still some rare peaks but this time they are rare enough, they even do not disturb the DFT processing, as shown further. The treatment of the peaks can be pushed even further, so the totality of the peaks is eliminated. But this time the graph is not as smooth as before (Fig. 11c). The principal frequencies calculated with DFT remain comparable to the reference approach.

A

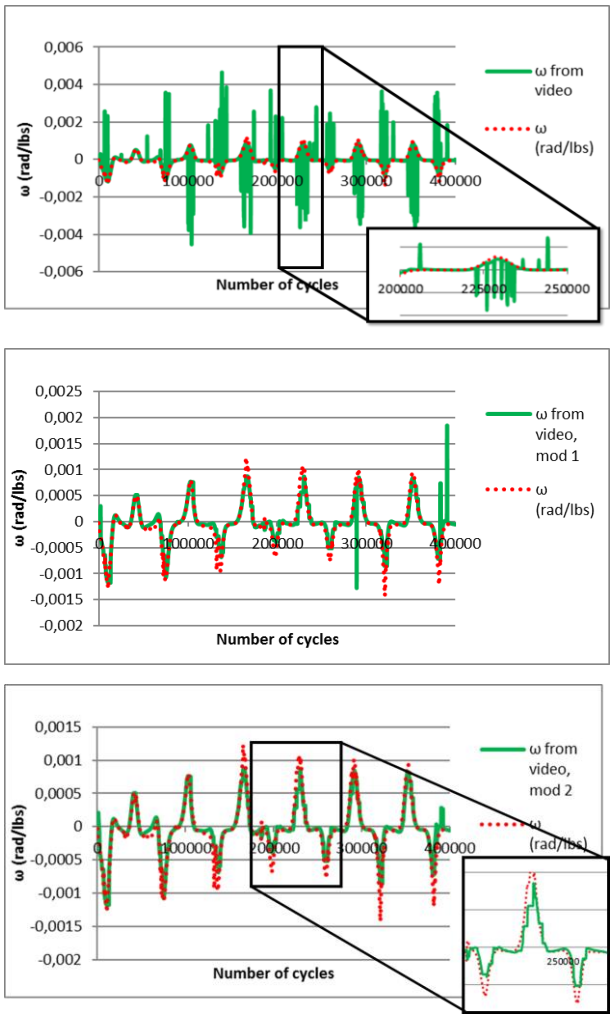


Figure 11. A) Graph of the angular speed derived from video-like data, without any treatment, green full line. The real angular speed is provided as well, in red dotted line. B) Graph of the angular speed derived from video-like data, after the application of a cleansing procedure. C) Graph of the angular speed issued from video-like data, after suppression of all the peaks

III. Comparison of different methods of measurement using DFT

In the previous parts, II.A, II.B, and II.C, we have described the step-by-step formation of the simplified, extended and reference methods of measuring RBC rotation. Of course, after acquiring all three datasets for the monitored cell, we, with great expectations and excitement, immediately compared their graphs. They confirmed the expected differences in the range of the values measured, and, at the same time, surprisingly similar periodic behavior (Fig. 12).

At closer inspection, however, we can see certain differences, caused e. g. by certain measurement errors. Confirming the compliance of the oscillations only by observing the graphs (which are, in addition, presented in various scales) is not a sufficiently reliable method. Therefore, to perform a further analysis of the data, we have used Discrete Fourier Transform, which happens to be an optimal tool to analyze the periodic behavior of processes. Our analysis also included later acquired data originating from the proposed measurement of the rotation based on processing the video-record of the experiment, pursuant part II.D.

Figure 12. View and comparisons of the graphs of the data measured by the individual methods. Regarding the differences in the scales and units of ampliudes (compare the values on y-axis), presenting the data in one common graph proves itself impractical.

Having processed all four rotation courses by means of DFT, we defined and sorted in descending order 25 of their most significant periods for each of them, based on the amplitude spectrum values (Tab. I).

TABLE I. Rank of 25 Most Important Frequencies Found by the Four Various Methods

Rnk.	Simpl	Ext	Ref	Video
1	6	6	6	6
2	7	7	19	19
3	19	19	7	7
4	12	12	12	12
5	1	1	1	8
6	20	20	20	5
7	5	5	32	32
8	8	8	5	1
9	32	32	8	11
10	0	0	0	20
11	17	17	17	18
12	25	25	18	25
13	16	16	4	17
14	3	3	25	4
15	18	18	16	21
16	11	11	3	24
17	4	4	11	10
18	10	10	21	16
19	33	21	10	9
20	21	15	33	2
21	15	33	9	3
22	13	13	24	0
23	28	28	15	13
24	2	2	2	27
25	27	27	45	15

The 25-period range corresponded to the selection within the scope from 100% to 10% of the amplitude maximum value. According to our experience, this range is sufficient to capture all substantial periodic characteristics of the rotation. When using DFT for interpolation of functions, periods with smaller amplitude value slightly modify the shape of its course, but do not improve the description of its periodicity. In signal analysis theory, the range of periods is considered as highly sufficient, if it covers more then 95% of signal energy, our ranges cover over 97%. The Tab III shows the sequence of the dominant periods for all methods. Here we can already observe a

similarity, even compliance, of the periodic behavior measured by the individual methods at a more transparent, numerical level. But we wanted to use a standard statistical method to confirm the assumption that the sequences and significance of the individual periods are similar enough to entitle us to talk about measuring the same periodic behavior. That is why we did not use the relatively complicated methods of verifying the spectral coherence. For the comparison, we used Kendall´s tau coefficient as a well-established rank correlation method.

Nevertheless, it was necessary to modify it modestly for our needs. In the four considered lists of dominant periods, there were all together 28 different values. Therefore, we filled in the missing ones into each sequence, and we put them in their particular order into the 26th to 28th positions. Consequently, we did the Kendall´s tau calculations for the individual pairs of sequences, testing the standard hypothesis that such sequences were statistically independent. The Tab. II shows the input values of the period ranks for the individual measuring methods and Tab. III the Kendall´s tau values for their pairs.

TABLE II. UNION OF THE TOP 25 FREQUENCIES, USED FOR EVALUATION OF THE KENDALL'S TAU VALUES

per.	Rnk.Sim.	Rnk.Ext	Rnk.Ref.	Rnk.Vid
6	1	1	1	1
7	2	2	3	3
19	3	3	2	2
12	4	4	4	4
1	5	5	5	8
20	6	6	6	10
5	7	7	8	6
8	8	8	9	5
32	9	9	7	7
0	10	10	10	22
17	11	11	11	13
25	12	12	14	12
16	13	13	15	18
3	14	14	16	21
18	15	15	12	11
11	16	16	17	9
4	17	17	13	14
10	18	18	19	17
33	19	21	20	27
21	20	19	18	15
15	21	20	23	25
13	22	22	26	23
28	23	23	27	26
2	24	24	24	20
27	25	25	28	24
24	26	26	22	16
9	27	27	21	19
45	28	28	25	28

TABLE III. RESULTS OF THE COMPARISON OF DIFFERENT METHODS BY KENDALL'S TAU CALCULATION

alpha 1% c.v. 0,312

Tau	Simpl	Ext	Ref	Video
Simpl	1	0,989	0,857	0,646
Ext	0,989	1	0,857	0,656
Ref	0,857	0,857	1	0,714
Video	0,646	0,656	0,714	1

alpha 10% c.v. 0,180

We can see that all values are close to the value 1 (1 is a result for two identical sequences) and, more importantly, they disclaim clearly the hypothesis on the independence of these sequences. The red color in the tables indicate that the hypothesis of their statistical independence has to be rejected. At the strict significance level $\alpha = 0.01$, it is enough for the Kendall´s tau value to exceed the value of 0.312 (for $\alpha = 0.1$ the value of 0.180). Therefore, we may conclude that in this case all of the 4 above presented methods of measuring RBC rotation are equal from the viewpoint of characterizing their periodic attributes. This fact has not only been confirmed by comparing the graphs of their courses but also by using the standard methods of exploring periodic processes and by using a reliable statistical method of comparing them.

Let us remind our original intention to use the RBC rotation course as one of the important characteristics to evaluate the compliance of experiments in microfluidic devices with periodical topology. The presented results enable us to use for this purpose (according to the current possibilities of the experiments carried out) any of these 4 measuring methods.

IV. APPLICATION OF DFT TO ANALYSE INFLUENCE OF SIMULATION NUMERICAL ONSET

In all the cases mentioned above, the evaluation of the angular speed was done for the entire simulation. The DFT considered the totality of the calculated angular speed as well. We did not consider the possible difference of the signal at the beginning, caused by the numerical onset of the simulation. This is not something that disturbs the pertinency of our research about the methods of acquisition of the cell rotation. However, it is something that should be considered in comparing the numerical simulation with a laboratory experiment. In this chapter, we present a preliminary analysis which help us to investigate the importance of the difference.

There are two kinds of numerical onset to be considered. The first one, the "strictly numerical onset", concerns only the flow of the fluid inside the simulation box. The second one, the "cells behavior onset", concerns the more complex behavior of the cells. Event if the fluid is flowing in steady conditions, the cells have to start to rotate and to advance inside of the simulation box and to find their proper way to move forward. Our aim is to find out what is the impact of the "strictly numerical onset" to the "cells behavior onset". To make such an evaluation, the DFT was used to determine principal

frequencies of some subsequences of the rotation graphs, and the above Kendall-tau principle was used to compare their frequency contents.

At first, the length of the numerical onset of the fluid flow simulation was evaluated in the model. We have measured fluid velocities in 10 selected points in the simulation box. The values of the velocities get stabilized after approximately 400 cycles. The total length of the simulation is about 400 000 cycles, so it means that the numerical onset takes approximately 0.1 % of measured data. The Fig. 13 illustrates the evolution of the x-coordinate of the velocity in 9 monitored points.

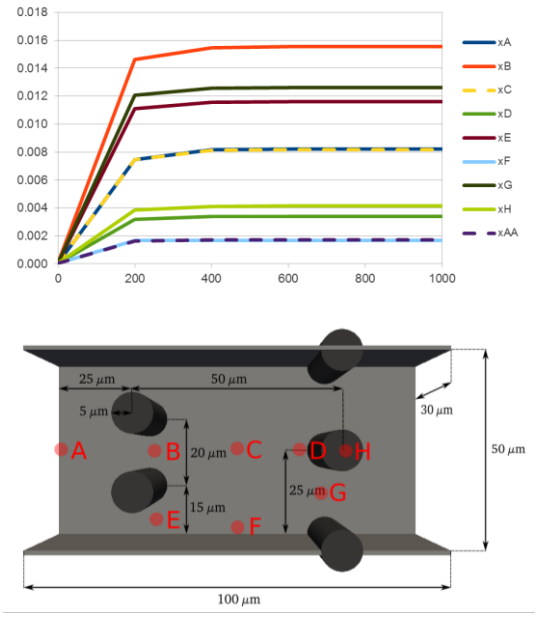


Figure 13. Graph of the x-coordinate of the velocity in several points in the model. The point H is positioned straight after the obstacle in direction of the flow. The point AA is not in the picture, but it has the same X and Y coordinates as the point A, and it is closer to the top boundary of the model.

After that, we examined the periodical behavior of the cells rotation, in order to know whether this short numerical onset of the flow has a more important impact on the cells. The aim was to do a preliminary analysis to see, whether the periodic behavior of the cells can be influenced by the numerical onset of the simulation. To do so, we compared several sub-sequences of the angular speed graph to each other. We provided such a comparison for 3 different cells.

For each cell, two different sets of subsequences were compared. Firstly, we split the graph of the rotation into two parts, one part containing 0-50% of the graph, second part containing 50-100% of the graph. Then we compared the results of the DFT for the whole course of the rotation (0-100% of the graph) with the results of the DFT of the two bisections. The position of the three parts is presented in Fig. 14.

The next set consists of seven sub-sequences with the same length (25% of the total length of the graph). The precise position of each of them is presented in Fig. 15. The idea was to observe the progressive evolution of the periodical behavior of the rotation in the overlapping intervals.

Each subsequence was treated with DFT. Afterwards the frequencies were sorted out according to their amplitudes. Then the list of several most important frequencies for each subsequence was compared with the other lists of other sequences, using the Kendall's Tau method mentioned in III.

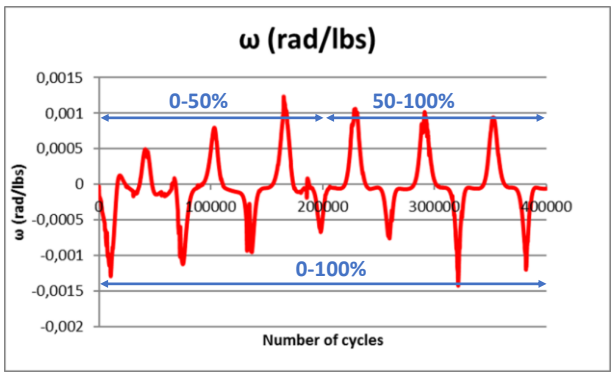


Figure 14. Position of the three parts of the graph, compared with each other.

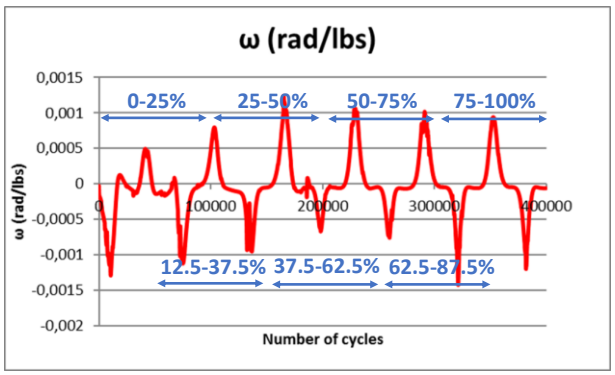


Figure 15. Position of the seven parts of the graph, compared with each other. Each part has a length of 25%.

For each 25% subsequence, the list of at most 10 important frequencies was considered. The list was shorter in case that the amplitude of the frequencies in the tail was inferior than 10% of the most important frequency in the list. After that, we used the union of all frequencies from the 7 lists, and joined the missing ones to the lists where they did not appear before.

For the two 50% and the 100% sequence of frequencies, the algorithm was similar, but in this case the list of at most 15 important frequencies was considered.

Comparing of two lists of the most important frequencies by Kendall's Tau method determines whether they are statistically independent. If the calculated Kendall's Tau value is superior to the critical value of chosen alpha, the two compared lists cannot be interpreted as independent. It means that the two lists of the frequency rank correlate sufficiently

with each other, and the intervals have similar frequency content.

The critical value is dependent on the length of the compared lists. As it appears to be slightly different for each cell, the values differ as well. Moreover, there are different critical values for 1% significance level of the statistical independence, and for 10% significance level of statistical independence.

The Tab. IV to IX show the results of the comparison between the different subsequences. The color code represents whether the values of calculated tau exceeds (red) or not (green) the critical value. The upper right corner constitutes the evaluation of the independence for 1% significance level, the lower left corner constitutes the evaluation of the independence for 10% significance level.

TABLE IV. COMPARISON OF THREE SECTIONS OF SIGNAL OF LENGTH 50% OF SIGNAL AND 100% OF SIGNAL, CELL A. ALPHA FOR 1%: 0.426, ALPHA FOR 10%: 0.250

alpha 1% c.v. 0,426

Tau	0-100	0-50	50-100
0-100	1	0.676	0.529
0-50	0.676	1	0.265
50-100	0.529	0.265	1

alpha 10% c.v. 0,250

TABLE V. COMPARISON OF THREE SECTIONS OF SIGNAL OF LENGTH 50% OF SIGNAL AND 100% OF SIGNAL, CELL B. ALPHA FOR 1%: 0.412, ALPHA FOR 10%: 0.242

alpha 1% c.v. 0,412

Tau	0-100	0-50	50-100
0-100	1	0.621	0.216
0-50	0.621	1	-0.085
50-100	0.216	-0.085	1

alpha 10% c.v. 0,242

TABLE VI. COMPARISON OF THREE SECTIONS OF SIGNAL OF LENGTH 50% OF SIGNAL AND 100% OF SIGNAL, CELL C. ALPHA FOR 1%: 0.467, ALPHA FOR 10%: 276

alpha 1% c.v. 0,467

Tau	0-100	0-50	50-100
0-100	1	0.6	0.448
0-50	0.6	1	0.238
50-100	0.448	0.238	1

alpha 10% c.v. 0,276

TABLE VII. COMPARISON OF SEVEN SECTIONS OF SIGNAL OF LENGTH 25% OF SIGNAL, CELL A. ALPHA FOR 1%: 0.513, ALPHA FOR 10%: 0.308

alpha 1% c.v. 0,513

Tau	0-25	12.5-37.5	25-50	37.5-62.5	50-75	62.5-87.5	75-100
0-25	1	0.615	0.282	0.667	0.590	0.333	0.564
12.5-37.5	0.615	1	0.564	0.590	0.744	0.462	0.538
25-50	0.282	0.564	1	0.410	0.513	0.333	0.410
37.5-62.5	0.667	0.590	0.410	1	0.821	0.615	0.641
50-75	0.590	0.744	0.513	0.821	1	0.615	0.641
62.5-87.5	0.333	0.462	0.333	0.615	0.615	1	0.513
75-100	0.564	0.538	0.410	0.641	0.641	0.513	1

alpha 10% c.v. 0,308

TABLE VIII. COMPARISON OF SEVEN SECTIONS OF SIGNAL OF LENGTH 25% OF SIGNAL, CELL B. ALPHA FOR 1%: 0.473, ALPHA FOR 10%: 0.275

alpha 1% c.v. 0,473

Tau	0-25	12.5-37.5	25-50	37.5-62.5	50-75	62.5-87.5	75-100
0-25	1	0.231	0.385	0.582	0.714	0.516	0.582
12.5-37.5	0.231	1	0.604	0.319	0.319	0.604	0.538
25-50	0.385	0.604	1	0.538	0.495	0.429	0.407
37.5-62.5	0.582	0.319	0.538	1	0.868	0.582	0.604
50-75	0.714	0.319	0.495	0.868	1	0.670	0.692
62.5-87.5	0.516	0.604	0.429	0.582	0.670	1	0.890
75-100	0.582	0.538	0.407	0.604	0.692	0.890	1

alpha 10% c.v. 0,275

TABLE IX. COMPARISON OF SEVEN SECTIONS OF SIGNAL OF LENGTH 25% OF SIGNAL, CELL C. ALPHA FOR 1%: 0.545, ALPHA FOR 10%: 0.303

alpha 1% c.v. 0,545

Tau	0-25	12.5-37.5	25-50	37.5-62.5	50-75	62.5-87.5	75-100
0-25	1	0.394	0.636	0.636	0.636	0.727	0.606
12.5-37.5	0.394	1	0.636	0.576	0.576	0.485	0.606
25-50	0.636	0.636	1	0.576	0.636	0.485	0.606
37.5-62.5	0.636	0.576	0.576	1	0.576	0.667	0.848
50-75	0.636	0.576	0.636	0.576	1	0.848	0.667
62.5-87.5	0.727	0.485	0.485	0.667	0.848	1	0.576
75-100	0.606	0.606	0.606	0.848	0.667	0.576	1

alpha 10% c.v. 0,303

We can notice that for each of the 3 cells, the comparison of the sequence 0-50% and the sequence 50-100% points out a significant difference between the two parts, so they can be considered as independent. To understand better the origin of such a difference, we have to investigate the tables in more details.

The table of the comparison of the 25% subsequences can be divided in 3 principal parts, as illustrated in Fig. 16. The left part contains the results of the comparison between the subsequences of the first half of the graph. If the results in this part are below the critical value of tau, it means that the frequencies in the first half of the rotation graph are changing during the advancement of the simulation, and so this part cannot be considered as stabilized. The upper right part contains the results of the comparison between the subsequences from the first half and the subsequences from the second half of the graph. If the values in this part are below the critical value of tau, it means, that the first and second half of the graph do not have correlated frequency content. It can be interpreted as a fact that the rotation does not evolve in the similar way in the first and in the second half of the graph. The last part of the table, the right lower one, contains the results of the comparison between the subsequences of the second half of the graph. If the results are above the critical tau value, it can be interpreted as that the second half of the graph is stabilized, and the rotation of the cell does not evolve significantly any more. We can observe such a conduct for all of the three cells.

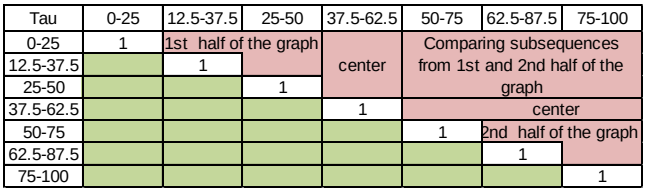


Figure 16. Decomposition of the table into parts showing comparison of different type of subsequences.

On the other hand, even if the statistical tests demonstrate the difference between the first and the second half of the rotation course, we should keep in mind the severity of the Kendall's Tau rank correlation analysis. If we concentrate only to investigation whether the set of the top 15 frequencies (out of all 900) is the same in the three lists, even for the worst cell we can state a considerable match between the first and the second half of the rotation graph (Tab. X).

TABLE X. MATCH OF THE FREQUENCIES IN THE SET OF THE TOP 15 FREQUENCIES FOR EACH LIST, CELL B.

Rnk.	0-100%	0-50%	50-100%
1	6	18	6
2	0	6	0
3	18	8	18
4	8	14	30
5	14	0	42
6	20	20	24
7	12	16	12
8	16	28	4
9	4	22	20
10	2	2	2
11	30	26	36
12	32	4	16
13	26	32	22
14	22	12	32
15	24	30	14

To make a general conclusion on the numerical simulation onset impact, we should examine more than 3 objects (which will be a subject for further research). Even though we can already observe possible tendencies of behavior of the cells during the simulation.

Each of the three cells manifests a significant difference between the first and the second half of the rotation course, and a kind of instability which occurs in the first half and disappears in the second one. That could be explained by the rearranging of the cells in the beginning of the simulation. As the initial seeding was set randomly, the cells must find a more stabilized order of flow at the beginning. The initial disorder could correspond to the non-stabilized part of the rotation graph. However, this hypothesis has to be confirmed by investigation of the bigger ensemble of the cells in more experiments.

If this hypothesis is truthful, the beginning of the simulation should be omitted for the examination of the cells behavior in the numerical experiment. Therefore, to keep the length of the pertinent data, the simulation time should be prolonged. The exact length of this onset can be dependent on the number of the cells in the simulation or the topology of the simulation box. This will be a subject of further research. In any case, the DFT and the Kendall's Tau rank correlation analysis appears to be a convenient mathematical and statistical tool to evaluate the extent of the numerical onset.

V. CONCLUSION

The aim of our article was to propose a methodology for accurate statistical examination of periodic properties of RBC rotation in simulation and real experiments. Based on our research objective, we are working with simulations in microfluidic devices with periodic topology. The periodic behavior of rotation (and other characteristics) of the cells is typical for such devices and it helps us to verifying the consistency and credibility of the simulation model. The precise RBC rotation measuring is difficult and it demands a considerable amount of data from the simulation or from video. That's why we have proposed a statistical methodology to verify the consistency of several approximate and exact ways of calculating the rotation in the first part of the article. Applying the proposed procedure to calibration simulation data confirmed the utility of the methodology. It follows that the results of different methods match and therefore it is possible to replace accurate but demanding RBC rotation measurement methods by simpler ones. Verification of this conclusion using the proposed methodology will be the subject of further studies.

The second part of the paper proposes a method to study the effect of the simulation numerical onset to RBC rotation periodic properties. Testing of the method on calibration experiments confirmed its efficiency. The method detected certain differences in RBC behavior at the individual proportions of the simulation. The proposed methodology is currently intensively used for a detailed analysis of an extensive data from the simulation experiment.

REFERENCES

[1] K. Bachratá and H. Bachratý, "On modeling blood flow in microfluidic devices," in ELEKTRO 2014: 10th International Conference, IEEE, May 2014, pp. 518-521

[2] K. Bachratá, H. Bachratý, and M. Slavík, "Statistics for comparison of simulations and experiments of flow of blood cells," International conference Experimental Fluid Mechanics, Mariánské Lázne 2016, *efm.kez.tul.cz*

[3] M. Bušík, I. Jančigová, R. Tóthová, and I. Cimrák, "Simulation study of rare cell trajectories and capture rate in periodic obstacle arrays," Journal of Computational Science, November 2016, pp. 370-376

[4] I. Cimrák, M. Gusenbauer, and T. Schrefl, “Modelling and simulation of processes in microfluidic devices for biomedical applications,” Computers and Mathematics with Applications, 2012, Vol 64(3), pp. 278-288

[5] I. Cimrák, K. Bachratá, H. Bachratý, I. Jančigová, R. Tóthová, M. Bušík, M. Slavík, and M. Gusenbauer, “Object-in-fluid framework in modeling of blood flow in microfluidic channels,” Comunications, Scientific Letters of the University of Zilina, 2016, vol. 18/1a, pp. 13-20

[6] M. Gusenbauer, G. Mazza, M. Brandl, T. Schrefl, R. Tóthová, I. Jančigová, and I. Cimrák, “Sensing platform for computational and experimental analysis of blood cell mechanical stress and activation in microfluidics,” in Procedia Engineering: Proceedings of the 30th anniversary Eurosensors Conference, September 2016, Vol. 168, pp. 1390-1393

Influence of CRF Value for Compression Efficiency

Juraj Bienik
Department of Multimedia and Information-Communication Technologies
Faculty of Electrical Engineering, University of Zilina
Zilina, Slovakia
juraj.bienik@fel.uniza.sk

Miroslav Uhrina
Department of Multimedia and Information-Communication Technologies
Faculty of Electrical Engineering, University of Zilina
Zilina, Slovakia
miroslav.uhrina@fel.uniza.sk

Peter Kortis
Department of Multimedia and Information-Communication Technologies
Faculty of Electrical Engineering, University of Zilina
Zilina, Slovakia
peter.kortis@fel.uniza.sk

***Abstract*—This paper deals with the dependence of Constant Rate Factor (CRF) on the bitrate for most commercial used compression standards H.264 and H.265. The measurements were done for eight different video sequences with high resolutions (HD, Full HD, Ultra HD) depending on content. The results indicated biggest differences in the coding efficiency in CRF range from 10 to 20 and from 40 to 51 by both mentioned codecs.**

Keywords— Constant Rate Factor, bitrate, H.264, H.265

I. INTRODUCTION

The multimedia services have become a global trend in the last few years, while the demand for these services has been constantly increasing. This trend has impact to many areas of the multimedia industry, including the services for video content distribution. The customers permanently place higher and higher demands on multimedia applications developers to create more innovative and creative products and the consumers market is forcing service providers to provide these services.

The users of the services need to meet their expectations. The degree of satisfaction of these needs is called quality. The users of video services require sharp and realistic image and continuous video flow. The fair view and sharpness of the image is dependent on the picture parameters as a resolution and a color depth. The video continuity is dependent on the reliability of the transmission path and the size of the cache. The parameter which has the most significant impact on the reliability in case of digital video transmission is the packet loss. Each end customer, as an assessor of service quality, takes into account all the aspects mentioned upper.

Not many studies and publications deal with this kind of theme [1]. The goal of this publication is to express the relationship between the CRF values and the bitrate.

II. TEST SEQUENCES

In this experiment eight video test sequences were used. These sequences are part of database [2]. The next paragraphs contain brief description of used sequences.

- **Bund Nightscape** – city night shot. The scene is time lapsed, the dynamic segments of scene are moving cars and walkers on the curb, static segments are represented by urban buildings. The camera captures scene form static position (Fig.1).

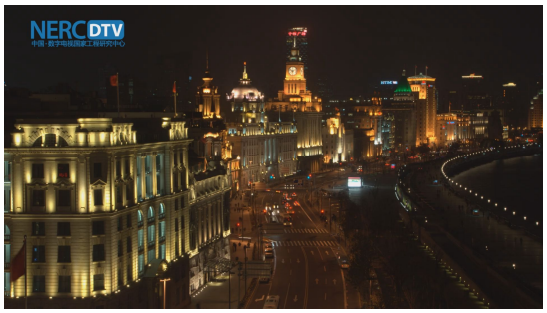


Figure 1. Bund Nightscape

- **Campfire Party** – night scene close to the fire. In the front of the image is flaming bonfire (the fast change of temporal and luminance information). In the background of the image is a group of slightly static people. At the end of the sequence, the camera zooms on the group of people (Fig.2).

Figure 2. Campfire Party

- **Construction Field** – shot on the construction site, where the static background is represented by buildings under construction, dynamic objects are represented by construction vehicles (excavator) and

walking workers. The slow-motion scene captured statically (Fig.3).

Figure 3. Construction Field

- **Fountains** – the daily shot on the city fountain. The foreground consists of squirting water (a lot of edges in the picture), the background is static formed by trees and the buildings. The capturing is static, scene with low dynamic of motion (Fig.4).

Figure 4. Fountains

- **Marathon** – marathon competition. The runners are multiple moving objects with moderate dynamic, the background is static road. The camera capturing is static from high point of view (Fig.5).

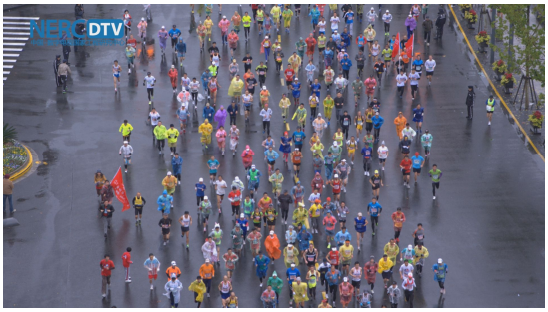


Figure 5. Marathon

- **Runners** – the running challenge, but in contrast to "marathon scene" there are fewer runners. The camera is static, located in the front of the runners slightly angled to the side (higher spatial information). Scene is relatively dynamic (Fig.6).

Figure 6. Runners

- **Tall buildings** – the shot on the modern city. The static objects are skyscrapers, river and the urban infrastructure; the slow-motion objects are represented by city traffic. The camera is moving slowly form the left to the right side. The scene is characteristic with the change of spatial and temporal information (Fig.7).

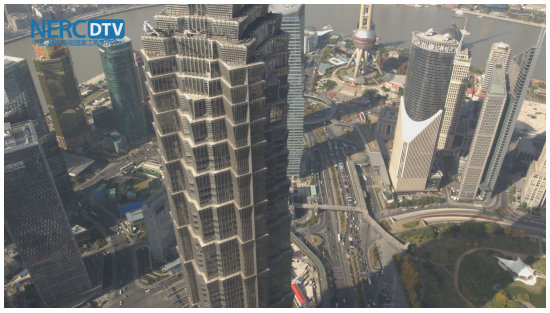


Figure 7. Tall Buildings

- **Wood** – the forest scenery. The shot on the trees in the forest (captured objects are static), the motion of the camera is from the left to the right side and the motion is accelerating in the sequence. Relatively high value of the spatial and temporal information (Fig.8).

Figure 8. Wood

III. VIDEO PROCESSING AND DATA ANALYZE

In our experiments eight input video sequences in the YUV 4:2:0 chroma subsampled format were used. Each sequence was encoded into H.264 [3] and H.265 [4] compression standards and three different resolutions (HD, FHD and UHD) using the compression tool ffmpeg [5]. The GoP size was set to half of the framerate (M=3, N=15). The Constant Rate Factor value was sequentially set to 0, 10, 18, 20, 23, 25, 28, 40, 45

and 51. Finally, the bitrate value of each compressed sequence using the MediaInfo 0.7.94 application [6] was obtained. The whole structure of the video process and data analyze is shown in the fig. 9. Spatial and temporal information of used scenes were calculated using the Mitsu tool [7] and are shown in the fig. 10. The obtained results of above mentioned process are summarized in the tables I. and II. and in the figures 11-19. Table III. shows how many times is H.265 more effective compared to H.264 for each resolution and each type of scene.

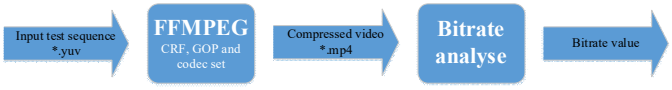


Figure 9. Video processing and data analyze

Figure 10. Spatial and teporal information of used scenes

TABLE I. BITRATES OF ENCODED SCENES OF H.264 COMPRESSION STANDARD FOR HIGH DEFINITION

CRF	BN [Mbps]	CP [Mbps]	CF [Mbps]	F [Mbps]	M [Mbps]	R [Mbps]	TB [Mbps]	W [Mbps]
0	78.50	104.0	95.40	114.0	135.0	117.0	104.0	121.0
10	17.10	43.00	17.40	43.50	47.40	36.40	25.40	38.90
12	11.20	32.60	12.20	34.80	35.10	28.30	19.30	30.10
14	8.15	24.30	8.867	28.00	26.00	22.10	14.70	23.20
16	6.263	18.200	6.629	22.20	19.60	17.50	11.30	17.90
18	4.910	13.80	5.122	17.30	15.00	14.00	8.765	13.8
20	3.983	10.60	4.031	13.20	11.50	11.20	6.844	10.6
23	2.874	7.387	2.864	8.381	7.785	7.944	4.751	6.986
25	2.318	5.825	2.292	5.980	6.026	6.290	3.750	5.289
28	1.681	4.086	1.646	3.439	4.109	4.399	2.643	3.578
40	0.436	1.167	0.389	0.233	0.848	0.964	0.598	0.809
45	0.241	0.691	0.207	0.097	0.448	0.517	0.317	0.441
51	0.116	0.363	0.099	0.053	0.229	0.255	0.149	0.253

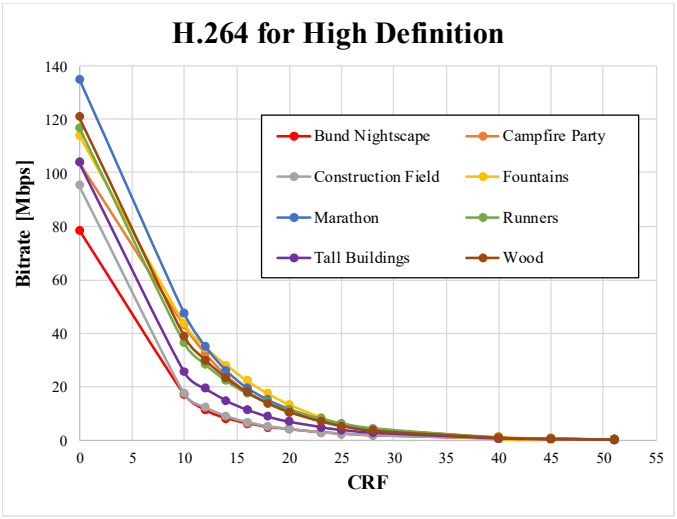


Figure 11. CRF to Bitrate for HD H.264

TABLE II. BITRATES OF ENCODED SCENES OF H.265 COMPRESSION STANDARD FOR HIGH DEFINITION

CRF	BN [Mbps]	CP [Mbps]	CF [Mbps]	F [Mbps]	M [Mbps]	R [Mbps]	TB [Mbps]	W [Mbps]
0	54.30	83.80	55.30	88.20	104.0	79.30	58.30	79.60
10	9.779	29.40	10.80	33.90	36.90	28.00	16.50	28.80
12	7.369	22.50	7.871	27.40	28.60	22.40	12.60	22.60
14	5.759	17.00	5.778	21.90	21.80	17.80	9.695	17.60
16	4.582	12.80	4.398	17.20	16.40	14.20	7.484	13.50
18	3.688	9.790	3.481	13.30	12.60	11.30	5.809	10.20
20	2.987	7.549	2.797	9.964	9.661	8.848	4.542	7.590
23	2.179	5.179	2.019	6.190	6.449	6.103	3.206	4.864
25	1.760	4.030	1.627	4.369	4.892	4.746	2.560	3.727
28	1.271	2.796	1.157	2.468	3.215	3.251	1.817	2.609
40	0.325	0.574	0.252	0.147	0.485	0.644	0.380	0.542
45	0.176	0.285	0.127	0.065	0.195	0.312	0.184	0.255
51	0.101	0.250	0.069	0.048	0.157	0.213	0.100	0.160

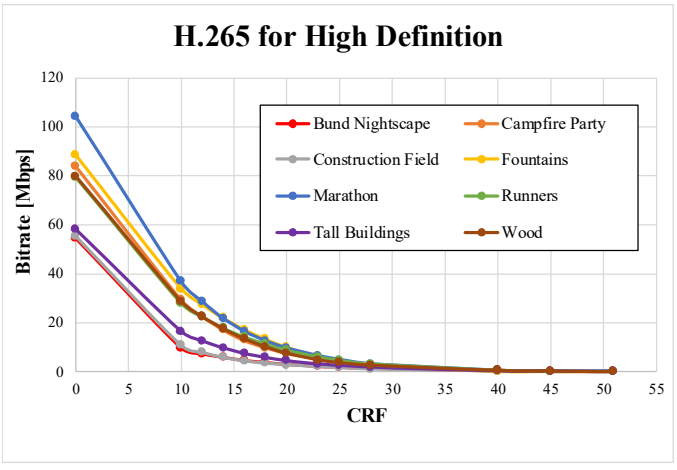


Figure 12. CRF to Bitrate for HD H.265

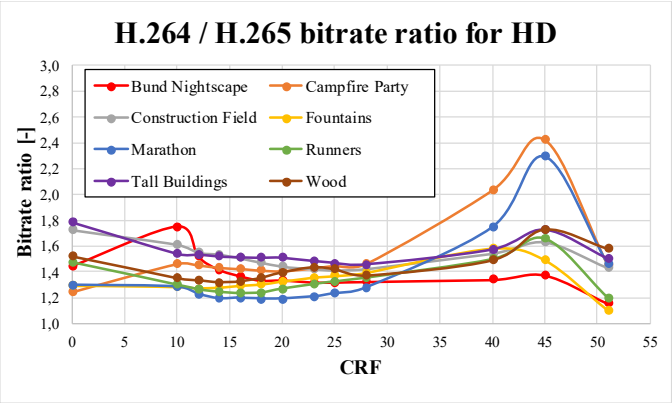


Figure 13. H.264 / H.265 bitrate ratio for HD

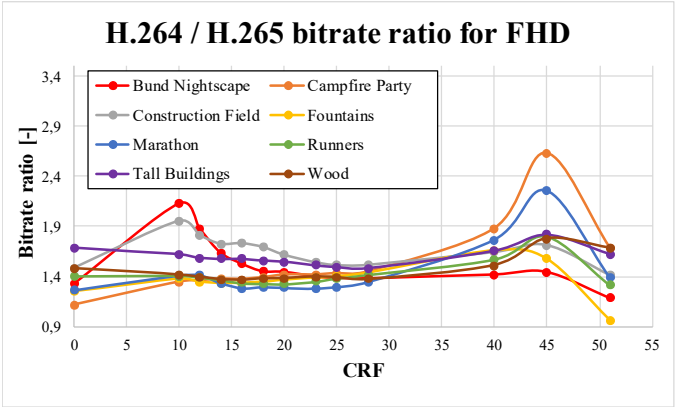


Figure 16. H.264 / H.265 bitrate ratio for FHD

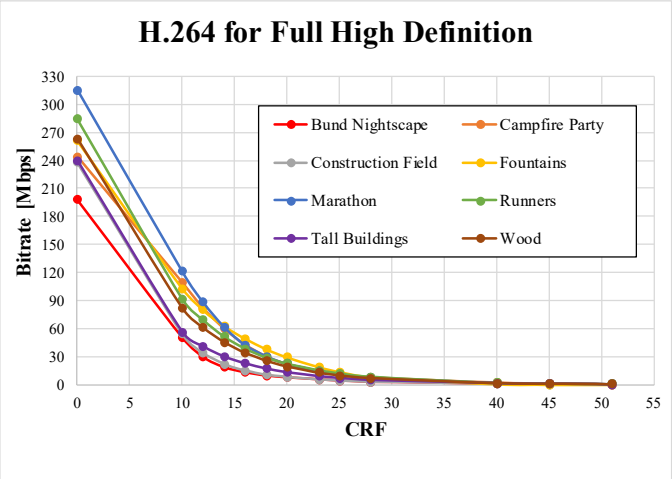


Figure 14. CRF to Bitrate for FHD H.264

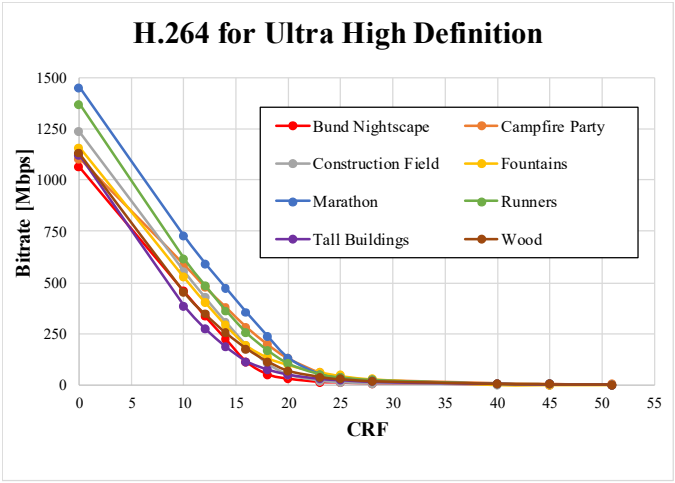


Figure 17. CRF to Bitrate for UHD H.264

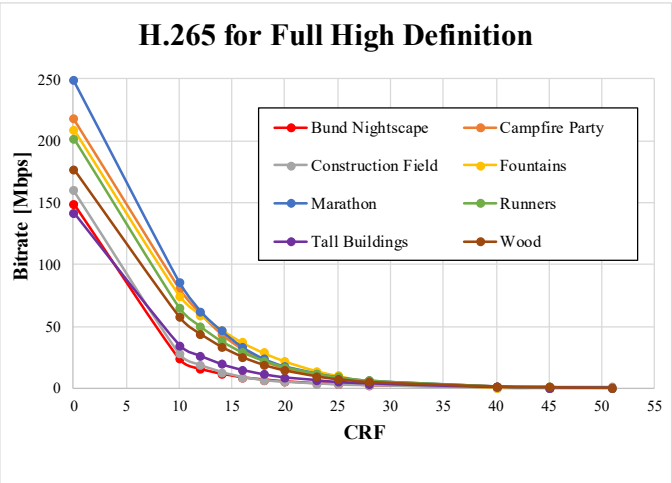


Figure 15. CRF to Bitrate for FHD H.265

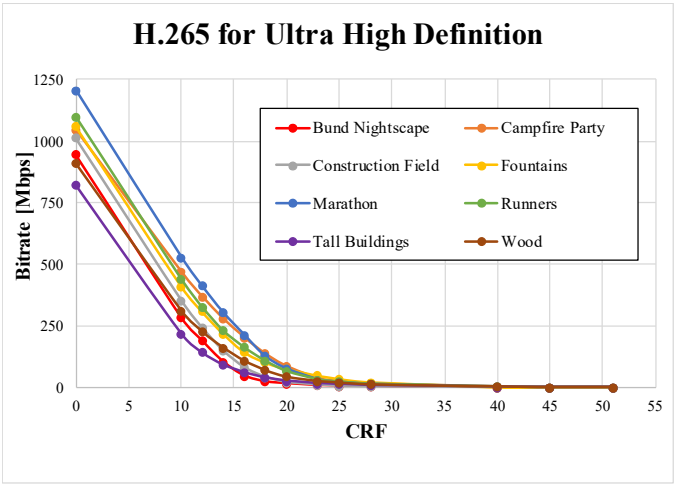


Figure 18. CRF to Bitrate for UHD H.265

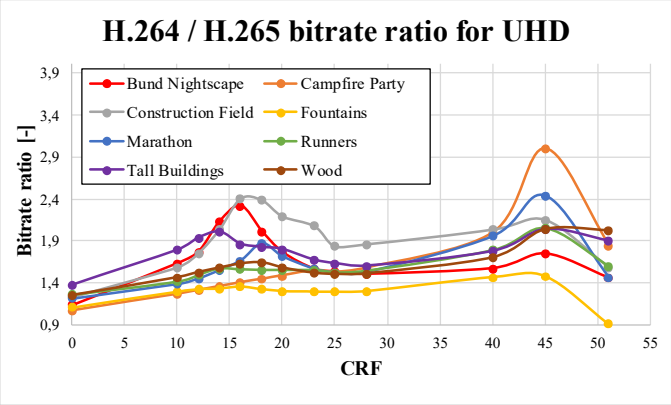


Figure 19. H.264 / H.265 bitrate ratio for UHD

TABLE III. EFFICIENCY COMPARISON OF H.264 AND H.265 COMPRESSION STANDARD (H.265 IS REFERENCE)

	BN	CP	CF	F	M	R	TB	W	Average
HD	1.384	1.543	1.515	1.333	1.369	1.335	1.546	1.433	**1.432**
FHD	1.514	1.530	1.648	1.369	1.434	1.416	1.599	1.460	**1.496**
UHD	1.590	1.925	1.279	1.630	1.565	1.779	1.601	1.590	**1.633**

According to the plots and tables some conclusions can be stated:

1) Compression efficiency of H.265 outperformed H.264 compression standard. This fact is valid for all used resolutions (HD. FHD. UHD) and CRF values in range from 0 to 45; except scene Fountains for CRF = 51 for FHD and UHD resolution (see fig. 16. and 19.). Since sane values of CRF are in range from 18 to 28 (default value is 23), the compression efficiency of H.265 always outperformed H.264. CRF value 51 corresponds to bitrate value for UHD resolution only 669 kbps. This bitrate is unusable for a video distribution with UHD resolution.
2) Coding efficiency depends on the content of test sequence. The scenes Construction Field, Campfire Party and Tall Buildings indicate higher compression efficiency; the lowest efficiency was by the Fountain scene (see table III). The compression efficiency depends on many factors including spatial (SI) and time information (TI). Unfortunately, the SI and TI parameters are not so dominant. Therefore, it is not possible to estimate the mapping to compression efficiency. The compression efficiency for scenes Fountain and Bund Nightscape is different, but SI and TI parameters are similar.
3) Compression effectivity grows with rising resolution. This fact is clearly demonstrated by average value shown in the last column in the table III. The dependence of compression efficiency on resolution is nonlinear, while the codec efficiency increment grows slower than linearly.

IV. CONCLUSION

This paper dealt with dependency of CRF on the bitrate for most used compression standards (H.264 and H.265). The measurement was done for eight types of video sequences with HD, Full HD and Ultra HD video resolutions with different spatial and temporal information. The results showed that the biggest differences in coding efficiency between H.264 and H.265 is by setting CFR value close to 18 and 45, where H.265 significantly outperforming H.264 in all resolutions.

ACKNOWLEDGMENT

This paper is supported by the following project: University Science Park of the University of Zilina – II. phase (ITMS: 313011D13) supported by the Operational Programme Research and Innovation funded by the European Regional Development Fund.

REFERENCES

[1] W. Robitza. "CRF Guide". Avalaible at: http://slhck.info/video/2017/02/24/crf-guide.html

[2] L. Song1. X. Tang. W. Zhang. X. Yang and P. Xi. "The SJTU 4K Video Sequence Dataset". 5th International Workshop on Quality of Multimedia Experience (QoMEX 2013). Klagenfurt. Austria. 3rd-5th July 2013. pp. 34-35. ISBN: 978-1-4799-0738-0.

[3] Richardson. E.G Iain. The H.264 Advanced Video Compression Standard. 2nd ed. John Wiley and Sons Ltd.. 2003. 316 pages. ISBN 978-0-470-51692-8.

[4] G. J. Sullivan. J.-R. Ohm. W.-J. Han and T. Wiegand. "Overview of the High Efficiency Video Coding (HEVC) IEEE Transactions on Circuits and Systems for Video Technology. bol. 22. no. 12. 28th December 2012. ISSN: 1558-2205.

[5] FFmpeg tool [online]. Available at: https://www.ffmpeg.org/.

[6] MediaInfo tool [online]. Avalaible at: https://mediaarea.net/sk/MediaInfo.

[7] P. Romaniak. L. Janowski. M. Leszczuk and Z. Papir. "Perceptual Quality Assessment for H.264/AVC Compression." IEEE Consumer Communications and Networking Conference (CCNC). 14.-17.1.2012. Las Vegas. NV. USA. pp. 597-602. ISSN: 2331-9852.

Adoption model of m-government services

Renata Bilkova, Anna Kralova
Department of System Engineering and Informatics
University of Pardubice
Pardubice, Czech Republic
renata.bilkova@upce.cz

***Abstract* - E-government is an important tool for public sector transformation. The number of e-services available on government websites continues to grow by leaps and bounds. The usage of mobile technologies in the public administration not only provides an alternative communication and services channel for citizens, but also allows to go further. M-government is emerging as the next big wave for ICT use in the public sector.**

The aim of this article is to find out what are the factors that affect user acceptance of mobile e-government services or how can we motivate the citizens to utilize services. This paper applies the technology acceptance model to explore users' requirements for the adoption and usage of e-government services.

Keywords - e-government; m-government; technology acceptance model

I. INTRODUCTION

Recently the possibilities of electronic communication have significantly improved on both sides - the citizens and the authorities and institutions of the public administration. The inclusion of electronic services in the public sphere enables more effective government, improving the quality of services for the private and business sector, ultimately, increase business support and increase the competitiveness of enterprises in foreign markets. The concept of e-Government also includes the possibility of the opposite approach: the citizen or a company's fielding individual agenda when it suits them - online way and, therefore, at a distance, from the comfort of their home or from their office, faster and more efficiently with lower costs on both sides.

E-government is a government existing on the web and characterized by functions such as information dissemination (government to users), communication, transaction, interoperability (vertical and horizontal integration) as well as participation [10]. The next step in the process of approximation government services to private and business sector is using mobile technologies. Mobile communication technologies are a key catalyst for transformational change of functionalities of governments. Mobile phones have emerged from being a luxury product to a mass necessity. The integration of internet and mobile devices have opened the doors of real mobility and 24×7 services. Mobile devices, equipped with various features and functions, are rapidly changing human social interaction, resulting in both new challenges and new opportunities. These special features of mobile phones have attracted many people to use it. Mobile services have huge potential to be one of the government's most effective tools, to govern, control, and administer community requirements and justice. In order for the governments to offer acceptable and attainable mobile services, these services have to be used by citizens. [3]

Mobile government services are a subset of e-government. M-government has advanced the dynamic nature of e-government and also created certain channels which either are not available or would be problematic for e-government. This has resulted in offering more dynamic and versatile methods for citizens to access certain government services.

M-government implies the use of mobile technologies (e.g., smartphones and tablets) in e-government. M-government enables location-based services [6], which are "personalized services delivered to a mobile device user at a remote location, so citizens can get immediate access to certain government information and services on an anywhere-anytime basis. The government can use the scalable and swift wireless channels to send time-sensitive information such as terror and severe weather alerts to citizens quickly and directly" [20]. M-government services require another type of data, namely geo-referenced information (data with spatial attributes) [16] and devices which are equipped with a Global Positioning System (GPS) and also Web 2.0 technology enable the efficient and effective information sharing, peer creation, and collective deliberation [13].

Since one of m-government´s ultimate goals is to provide better services for citizens, a suggestion, that the application of m-government should also be examined from users´ viewpoint. Therefore, the objective of this paper is to investigate the users´ requirements for the adoption of m-government services and positive factors that can contribute to an understanding of citizen intention to adopt these services.

II. MATERIAL AND METHODS

M-government with integrated and customized services represents the advanced e-government services. Implementation and successive upgrading of the m-government follow certain paths, levels of maturity, stages, or phases. Different countries implementing e-government in their ICT framework certainly have different missions and objectives; however, the gradual development of an e-government system in any country follows some unique levels of service maturity for evolution [17]. Each of the service levels represents a different service pattern, different levels of technological sophistication, different stakeholder orientation, different types of interaction, different security requirements, and different

reengineering processes. Table I. shows the basic level model of e-government services.

TABLE I. OVERVIEW OF BASIC MATURITY LEVEL MODELS OF E-GOVERNMENT SERVICES

Source / Level	Access	Interact	Transaction	Integration	Customization
World Bank	Publish	Interact	Transaction		
IBM			Automate Enhance	Integrate	On Demand
Gartner	Presence	Interaction	Transaction	Transformation	
OSN	Emerging	Enhanced	Transactional	Connected	
UN/ASPA	Emerging Enhanced	Interactive	Transactional	Seamless	
Layne & Lee	Cataloguing		Transaction	Vertical integration Horizontal integration	
Siau&Long	Web presence	Interaction	Transaction	Transformation	E-Democracy
Capgemini	Information	One-way interaction	Two-way interaction	Transaction	Targetisation

Sources: [2], [17], [21], [14], [5], [21]

All of these models expect the advanced services requiring progressive technology, which will provide e-government to their users - citizens and businesses. Especially citizens' behaviour, in terms of adopting a new technology-driven system, is a very complex and robust subject. The problem with measuring of the success of e-government is described for example in [10]. Understanding and estimating the effect of citizens' adopting criteria, which leads to successful implementation of m-government, would have important managerial implications.

The concept of connected government looks towards technology as a strategic tool and an enabler for public service transformation, innovation and productivity growth.

M-government as a part of e-government is not only adoption of mobile technologies, it is also providing new choices for communication and access channels to citizens, including new social media applications and options [15]. So, as well as e-government services, m-government services can be divided into four levels:

- Information applications - created purely to provide information. This is a one-sided presentation of an office's information resource. Examples include mobile apps that replicate the content of the authority web site, or an services to find the nearest authority office.
- Interactive services - can provide some more or less personalized service for the user. An example is signing into waiting list or preparation of the claims based on fulfilled data about current life situation.
- Transactional services and applications for electronic submission - allow users to prepare, fill in and trustfully submit a form with any request or submission. Examples are applications for filling out forms or for filling in and submitting a tax return.
- Governance and citizen engagement – services and applications for mutual communication between citizens and government. An example might be an application for reporting the current traffic situation or all the tools of social participation created in the traditional internet apps for social networking.

A. The Technology Acceptance Model

To design and to deliver m-government services, authorities should consider the expectations and the perceptions of citizens toward using the services.

There are several studies dealing with adoption of m-government [3], [9], [11], [18] which indicates that whether or not citizens adopt m-government services is influenced by the following beliefs:

- perceived ease of use; efficiency in time and distance; value for money; convenience; availability of device and infrastructure; usefulness; responsiveness; relevance, quality and reliability of information; risk to user privacy; reliability of the mobile network and the SMS-based system; risk to money; compatibility;
- trust in the mobile service technology, in the government and perceived quality of public services;
- self-efficacy in using mobile technology.

The Technology Acceptance Model (TAM) model was chosen for the purpose of this study as one of the best-describing models from theory of information systems, which explains how users accept and use new technologies. TAM is one of the most important expansions of the original Ajzen&Fishbein's Theory of Reasoned Action (TRA) [1]. Both models - TRA and TAM - contain elements strongly affected by behaviour. The TAM model shows that if a user meets new technologies, there are a number of factors affecting decisions about how and when to use them. Many input factors of the TRA model are replaced by two factors [7] only:

- perceived usefulness - the degree to which a person believes that using a particular system would enhance his or her job performance,
- perceived ease of use – that is the degree to which a person believes that using a particular system would be free of physical and mental efforts.

The proposed m-government services adoption model is based on the original TAM model, which we transform in the first step into a simple CLD diagram using a feedback loop – Figure 1.

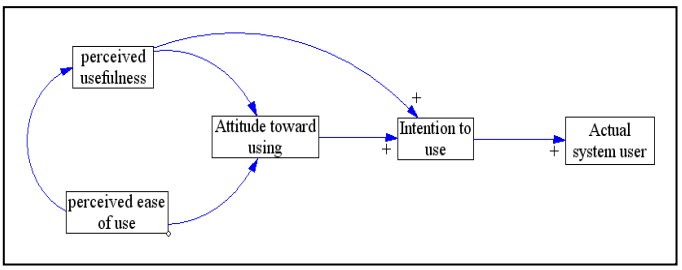


Figure 1. Transformation of the original TAM model into a CLD diagram *(modified by [8])*

III. Results and Discussion

The aim of public administration should be to offer services in such forms and quality that customers expect, which they are used to from, for example e-commerce services. The user's satisfaction with the service is an important indicator for determining whether users would return and continue in using the provided online service. The user's satisfaction is one of the subjective qualitative metrics dealing with issues that are more general and related to perceptions and attitudes (e.g. perceived ease of use, usability, etc.).

A. Criteria for the assessment of user satisfaction in m-government services

To determine a set of criteria evaluating the users' satisfaction with provided mobile services is quite a challenging task. It is hard to decide what individual users consider to be important, what their initial knowledge about the use of the Internet is, or what experience with the public administration services they have. Verdegem's study [22] summarized the basic indicators that have a fundamental impact on the users' satisfaction into three groups: accessibility (the service must be easily available without complicated searching), ease of use, and efficiency when users preferred lowering of the administrative burden. Tinholt [18] develops evaluating of the efficiency into criterion of the efficiency to meet the expected benefits. In contrast, Gouscos et al. [8] add to these subjective criteria an objective criterion of time and error rate. And even though these indicators belong rather to the lower layers of the e-services quality evaluation, as shown in the previous charts, these factor significantly influence the perception of the system reliability which has a direct impact on satisfaction.

The objective is therefore to define the key benefits of electronic public administration services, which are expected by the user based on their previous experience, for example with e-commerce services, and which therefore appear as fundamental evaluation criteria of their satisfaction. The essential expected benefits of m-government services based on previous models and studies mentioned above include the following user requirements:

- simpler services – the perception of the service simplicity is mainly based on:
  - ease of use - service usability evaluation,
  - services availability - whether and what additional documents, information or materials (including. electronic signature) are required for the use of the service,
  - technical support - whether there is for example hotline, FAQ, etc..,
- faster services – the speed of service depends on:
  - speed of service search
  - speed of data entry / sending electronic reports - how much time is required for complete service processing,
  - speed of system response,
- more efficient services - efficiency perception is particularly based on:
  - usability of services - what is the portion of users able to use the e-service,
  - quality of information - whether the website offering the service gives all the information necessary for its use,
  - error rates of services - frequency of errors during data entry/ sending electronic reports.

B. Factors affecting user's satisfaction with eGovernment services

All of these criteria mentioned above can usually be monitored using several rating systems and evaluation logic, for example through user questionnaires. Continuity of individual factors affecting the user's satisfaction with public e-services is expressed in Figure 2 with Basic casual loop diagram representing "mind maps" for visual mapping of interrelations between various factors that enter the system.

The proposed diagram of m-government services is based on e-government Adoption Model (GAM) [17] and on the model evaluating user satisfaction with the use of e-commerce services [4], because as also the user of m-government services is the customer for public administration and expects a certain quality, which they are used to from the commercial sector. The website quality a can significantly affect the satisfaction and it is one of the factors influencing the intention to purchase goods. The chart presents a position of website quality and value of provided information which certainly influence the overall evaluation of the service. Another depicted factor - service availability – affects interest in offered services and the number of new people interested in using this services. Also feedback captured in the chart explains how positive assessment affects the number of new users or regular users of these services.

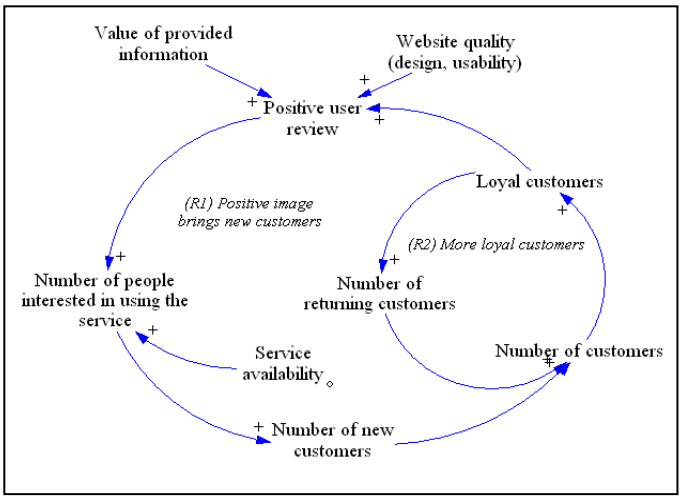


Figure 2. Impact of user satisfaction on the number of service users

The following model Figure 3 is extended by a new variable – perceived reliability. Besides objective factors such as system response speed or service error rate perceived reality of the system is also affected by the subjective expectations on service quality. It is a very subjective viewpoint influenced by previous experience with the use of other electronic services in the business sector. In recent years, it has been possible to observe a gradual decrease of the satisfaction with the offered services caused by rising expectations from offering new ICT technologies [18]. At the same time, we can see that high expectations with the lack of quality (information, website, services) can reduce user's satisfaction. The main problem is that a reduction in only one of these aspects of quality can reduce overall user's satisfaction.

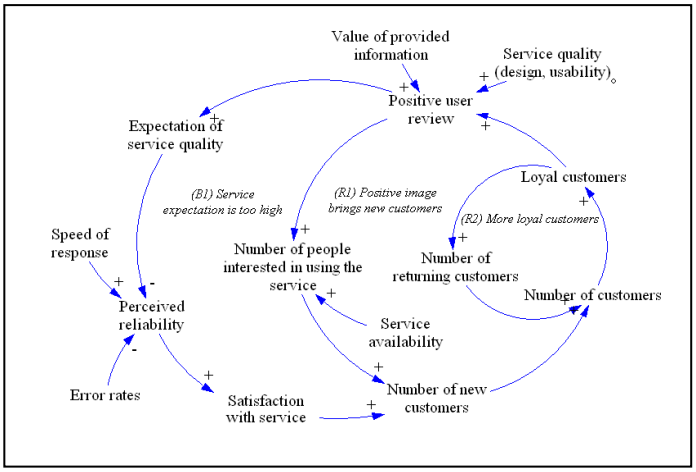


Figure 3. Impact of service expectation on user satisfaction

IV. CONCLUSION

Obviously, one research study such as this one cannot cover all topics in m-government user acceptance. This paper proposes an m-government adoption model based on original TAM model extended by factors affecting m-government customers´ behavior. Hence, there is a room to pursue further research, which should be conducted with more dimensions, depending on the research goals.

The outcome model of this study can help the government officials to understand the needs of the users and to implement the m-government services that will be more urgent and efficient.

REFERENCES

[1] AJZEN, I., FISHBEIN, M. 1980. Understanding attitudes and predicting social behaviour, Englewood Cliffs, NJ: Prentice-Hall

[2] AL-HASHMI, A.; DAREM, A. B. 2008. Understanding phases of eGovernment project. New Delhi: Retrieved from http://www. csi-sigegov. org/emerging_pdf/17_152-157. pdf

[3] ALTHUNIBAT A., ALRAWASHDEH T. A., MUHAIRAT M. 2014. The Acceptance of Using M-government Services in Jordan, Information Technology: New Generations (ITNG), 2014 11th International Conference on, Las Vegas, NV, 2014, pp. 643-644.

[4] BILKOVA, R., KOPACKOVA, H. Enhancing e-commerce by website quality. Recent Advances in Business Administration, Marketing and Economics, vol. 12, September 2013, pp. 40-48, ISSN 2227-460X.

[5] CAPGEMINI. Method paper 2010: Preparing the 9th Benchmark Measurement, European Commission, Directorate General for Information Society and Media, Brussels, 2010.

[6] CARROLL, J. M., GANOE, C. H. Supporting community with location-sensitive mobile applications. In M. Foth (Ed.), Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, 2009, pp. 339- 352. Hershey, PA: IGI Global

[7] DAVIS, F. A technology acceptance model for empirically testing new end-user informatics systems: theory and results. MIT Sloan School of Management, Cambridge, MA.

[8] GOUSCOS, D., KALIKAKIS, M. et al. A general model of performance and quality for one-stop eGovernment service offerings, Government Information Quarterly, Volume 24, Issue 4, October 2007, pp. 860-885, ISSN 0740-624X

[9] HUNG S.Y., CHANG C.M., KUO S.R. User acceptance of mobile e-government services: An empirical study, Government Information Quarterly, Volume 30, Issue 1, January 2013, pp. 33-44, ISSN 0740-624X

[10] KOPÁČKOVÁ, H. The issue of measuring e-government success in context of the Initiative 202020. In Proceedings of the 21th International Conference Current Trends in Public Sector Research. Brno: Masarykova univerzita, 2017, pp. 41-49. ISBN 978-80-210-8448-3.

[11] LIU, Y. et al. An empirical investigation of mobile government adoption in rural China: A case study in Zhejiang province. Government Ingormation Quarterly 31, 2014, pp. 432-442.

[12] MOON, M. The evolution of e-government among municipalities: Rhetoric or reality? Public Administration Review, 2002, 62(4), pp. 424-433.

[13] NAM T. Suggesting frameworks of citizen-sourcing via Government 2.0, Government Information Quarterly, Volume 29, Issue 1, January 2012, pp. 12-20, ISSN 0740-624X

[14] OECD eGovernment studies. eGovernment for better government. Paříž : OECD Publishing, 2005. 205 s. ISBN: 92-64-01833-6.

[15] OECD/International Telecommunication Union. M-Government: Mobile Technologies for Responsive Governments and Connected Societies, OECD Publishing, 2011.

[16] REUVER, M., STEIN, S., HAMPE, J. F. 2013. From e-participation to mobile participation: Designing a service platform and business model for mobile participation. Information Polity, 18(1), pp. 57-73.

[17] SHAREEF M.A., KUMAR V., KUMAR A., DWIVEDI Y. K. E-Government Adoption Model (GAM): Differing service maturity levels, Government Information Quarterly, Volume 28, Issue 1, January 2011, pp. 17-35, ISSN 0740-624X

[18] SULTANA, R., AHLAN, A.R., HABIBULLAH MD. A comprehensive adoption model of m-government services among citizens in developing countries. Journal of Theoreticel and Applied Information Technology, Vol.90, No.1, 2016. ISSN: 1992-8645.

[19] TINHOLT, D., LINDEN, N. Public services online: Assessing user centric eGovernment performance in Europe – eGovernment Benchmark 2012. Capgemini, European Union, 2013, ISBN: 978-92-79-29949-0

[20] TRIMI, S., SHENG, H. Emerging trends in m-government. Communications of the ACM, 2008, 51(5), pp. 53-58

[21] UNITED NATIONS: United Nations eGovernment Survey 2016. New York: United Nations, 2016. ISBN 978-92-1-123205-9

[22] VERDEGEM, P., VERLEYE, G. User-centered eGovernment in practice: A comprehensive model for measuring user satisfaction, Government Information Quarterly, Volume 26, Issue 3, July 2009, pp. 487-497, ISSN 0740-624X.

Detecting Compromised Accounts on the Pokec Online Social Network

Jan Bohacik[1], Antonin Fuchs[2], Miroslav Benedikovic[3]
[1]Department of Informatics, University of Zilina, Univerzitna 8215/1, 010 26 Zilina, Slovakia
[2,1]Data Mining Group, Pokec Team, Ringier Axel Springer Slovakia, a. s., Murgasova 2/243, 010 01 Zilina, Slovakia
[3]Department of Software Technologies, University of Pardubice, Studentska 95, 532 10 Pardubice, Czech Republic
[1]Jan.Bohacik@fri.uniza.sk (corresponding author), [2]fuchs@rubicon.sk, [3]Miroslav.Benedikovic@upce.cz

Abstract—**Online social networks have billions of users worldwide when combined and they still keep increasing this amount. Their users typically develop trust relationships with the accounts of other users. But large numbers of users and potential gains from abuses of the trust relationships have attracted the attention of cyber-criminals. Therefore, it is important to stop accounts from being compromised by these criminals. In this paper, an anomaly model trained on the previous login data of users is applied to detection of compromised accounts. The login data comes from the Pokec online social network, which is the largest community in Slovakia where people can meet others and talk to their friends. The anomaly model watches sudden changes in the behavior of a user trying to log in to her or his account. A change in the behavior can indicate an attempt from someone else to compromise the account of the user. The efficiency of the anomaly model is validated with computation of measures such as sensitivity, specificity and overall accuracy. Achieved results are promising with a real potential to detect compromised accounts.**

Keywords—compromised account; anomaly model; login data; online social network

I. INTRODUCTION

Online social networks are online platforms used by people to build social networks or social relations with those who share similar interests, activities, backgrounds or real-life connections [8]. They have become increasingly popular and time spent on the networks varies by country, but many countries average more than two hours per day per user [13]. People use them to share knowledge, opinions, and experiences; seek information and resources; and expand personal connections [14]. The most popular online social network in the world is Facebook with more than 1.8 billion active users [11]. Other popular networks include QZone with 632 million active users, Twitter with 317 million active users, LinkedIn with 106 million active users and VKontakte with 90 million active users. In Slovakia, the largest community where people can meet others and talk to their friends is the Pokec online social network. There are about 5.4 million people in Slovakia including children [10] and the network is visited daily by more than 400 thousand people. Overtime, the users of a social network build trust relationships with other people, friends, colleagues etc., which helps them to express their identity and gain social validation, find like-minded people and communicate. Unfortunately, potential gains from abuses of these trust relationships have attracted the attention of cyber-criminals and their malicious activities. Information access and interaction is based on trust and users typically share a substantial amount of personal information with their friends [12]. Depending on the network, this information may be public or not. Some users also accept any friendship just to gain popularity and thus expose themselves to potential attacks. There are also users who do not realize these risks even if they are well aware of e-mail spams for example.

As a consequence, researches are analyzing malicious activities on online social networks actively and several approaches have been proposed [1], [2], [3], [4], [5], [9]. They are specialized in several areas such as detection of fake accounts [2], [5], detection of regular compromised accounts [3], [9] and detection of high profile compromised accounts [4]. Phishing website detection is also relevant [1]. Fake accounts are accounts created for the purpose of spreading malicious content (spam, phishing attacks etc.) or increasing the visibility in the social network (appearing high in forum posts, manipulating votes, etc.). The method in [2] uses social graph properties to rank users according to their perceived likelihood of being fake. It is supposed for this ranking that fake accounts mostly have relationships with other fake ones and that a group of fake accounts randomly tries to befriend a group of interconnected real accounts. Large-scale spam campaigns transmitted via the wall messages of users with the use of fake accounts are detected in [5]. Wall posts are modelled as nodes in a large graph and edges are built when two posts are similar. It is supposed that a single account has a limited number of wall posts and that messages in a single campaign are sent within a relatively short time window. As a response to detection of fake accounts, cyber-criminals started compromising accounts because it is harder to label this as malicious activity. Compromised accounts are existing legitimate accounts which have been taken over by an attacker. The approach in [3] uses anomaly detection based on computation of value frequencies for messages to identify accounts that experience a sudden change in behavior. This is combined with identification of account groups which experience similar changes within a short period of time. It is assumed here that these changes are the result of a malicious campaign. In paper [9], a behavioral feature metrics is proposed for watching the behavioral profile of an individual user. This allows to distinguish one user from another. According to [4], it is useful to distinguish between regular and high-profile accounts in online social networks. High-profile accounts have large social circles such as followers and at the same time they are likely trustworthy to many of them. Typically, high-profile

This work was supported by Ringier Axel Springer Slovakia, a. s., project no. 64866: "Phishing – Identification of Compromised Accounts with Machine Learning Tools" and by faculty research grant FVG/2/2017: "User-centred Approach for Decision Support Systems" of the Faculty of Management Science and Informatics, University of Zilina, Slovakia.

accounts are owned by newspapers or popular brand names. Compromised high-profile accounts can cause huge damages within a very short time and so it is very useful to keep watching them automatically with a computer system. These accounts often have well-defined behavioral profiles and so it is possible to find out if they are compromised just with sudden behavioral changes. In general, compromised accounts have other users logging in them through login usernames and passwords which are often found with a phishing website. It is therefore important to detect phishing websites. The author of [1] proposes a data mining fuzzy-based classification system for phishing website detection where a layered fuzzy structure is constructed for all gathered and extracted phishing website features and patterns. In this paper, we develop a system for detection of compromised accounts on the basis of sudden changes in login behavior where the changes are measured through computation of frequencies. A login of some user produces one set of values and this set is compared with previous sets of values for the user so that an anomaly is detected and an anomaly score is computed. If the anomaly score is higher than a given threshold, the account of the user is considered to be compromised.

The paper is organized as follows. Section II describes the login data taken from the Pokec online social network. The anomaly model used for detection of compromised accounts is presented in Section III. Section IV contains experimental results achieved with the model. Conclusions are in Section V.

TABLE I. LOGIN TABLE

Column	Data Type	Description
ts	bigint	Timestamp of the login.
iduser	bigint	Identifier of the particular account owner.
ip	bigint	Internet Protocol address of the device.
status	integer	Value 1 for a successful login and value 0 for a failed login.
hashpswd	varchar	Result of the hash function applied to the password used by the user logging in the account.
medium	varchar	Medium used by the user logging in the account - defined as either a PC browser (*desktop*), a mobile phone browser (*mpokec*), some version of the provided application running on the Android mobile operating system (*appAndroid*) or some version of the application running on the iOS mobile operating system (*appIOS*).
useragent	varchar	User-Agent string in HTTP.
ldate	varchar	Date of the login in format yyyy-mm-dd.

II. LOGIN DATA

Employed login data is represented as an SQL table which has columns *ts*, *iduser*, *ip*, *status*, *hashpswd*, *medium*, *useragent*, *ldate* and 3805474 rows. The data contains logins for 2806 accounts chosen by Pokec Team from the Pokec online social network. Among these 2806 accounts, there are 2794 normal accounts and 12 compromised accounts. Compromised accounts have logins from users who are not the owners of these accounts. Particular columns, their SQL data types and descriptions are in Table I. Column *ts* identifies when the login recorded in a row happened through date and the time of the day. Column *iduser* stores a unique identifier for any recorded action of an account in the social network. Column *ip* keeps the IP address of the device used for logging in the account. Column *status* stores if the attempt to log in the account was successful. This means that some rows of the SQL table contain information about unsuccessful logins. Column *hashpswd* keeps the password used for logging in the account. The password is hashed so that it cannot be read or recreated easily and so that the hashed password is always the same for some particular password. Column *medium* contains information about the device used for logging in the account. Column *ldate* contains the date of the login in a row and it is there for optimization purposes.

TABLE II. SAMPLE MADE-UP LOGINS

Attribute/Login	$\boldsymbol{l_1}$	$\boldsymbol{l_2}$
TS (B_1)	1451994221467	1464184238003
IDUser (B_2)	54473838	62473886
IP (B_3)	2548884151	2548884122
Status (B_4)	1	1
HashPswd (B_5)	*8a5468294a43dba13f45*	*cdb932605b94652a09ab*
Medium (B_6)	*desktop*	*appAndroid*
UserAgent (B_7)	*Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36*	*Pokec.sk/2.3.2 (Android 5.1.1; Sony D2303)*
LDate (B_8)	2016-01-05	2016-05-25

Suppose a set $\boldsymbol{L}$ representing all logins is defined. An individual login $\boldsymbol{l} \in \boldsymbol{L}$ corresponds to a row in the SQL table and cardinality of $\boldsymbol{L}$ $\#(\boldsymbol{L}) = 3805474$. Attribute $TS = B_1$ is defined for column *ts*. The particular value for attribute *TS* and login $\boldsymbol{l} \in \boldsymbol{L}$ can be obtained as $TS(\boldsymbol{l}) = B_1(\boldsymbol{l})$. Attribute $IDUser = B_2$ is defined for column *iduser*. The particular value for attribute *IDUser* and login $\boldsymbol{l} \in \boldsymbol{L}$ can be obtained as $IDUser(\boldsymbol{l}) = B_2(\boldsymbol{l})$. Similarly, attribute $IP=B_3(\boldsymbol{l})$/ $Status=B_4(\boldsymbol{l})$/ $HashPswd=B_5(\boldsymbol{l})$/ $Medium=B_6(\boldsymbol{l})$/ $UserAgent=B_7(\boldsymbol{l})$ / $LDate=B_8(\boldsymbol{l})$ is defined for column *ip* / *status* / *hashpswd* / *medium* / *useragent* / *ldate*, respectively. The particular value for *IP*/*Status*/*HashPswd*/*Medium*/*UserAgent*/*LDate* and login $\boldsymbol{l} \in \boldsymbol{L}$ is obtained as $IP(\boldsymbol{l}) = B_3(\boldsymbol{l})$/ $Status(\boldsymbol{l}) = B_4(\boldsymbol{l})$/$HashPswd(\boldsymbol{l}) = B_5(\boldsymbol{l})$/$Medium(\boldsymbol{l}) = B_6(\boldsymbol{l})$/$UserAgent(\boldsymbol{l}) = B_7(\boldsymbol{l})$/$LDate(\boldsymbol{l}) = B_8(\boldsymbol{l})$. A sample of logins $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ can be found in Table II as an example. It contains made-up values so that there are no privacy issues, however, the values resemble real values in the login data. $TS(\boldsymbol{l_1})$ = 1451994221467, $IDUser(\boldsymbol{l_1})$ = 54473838, $IP(\boldsymbol{l_1})$ = 2548884151, $Status(\boldsymbol{l_1})$ = 1, $HashPswd(\boldsymbol{l_1})$ = *8a5468294a43dba13f45*, $Medium(\boldsymbol{l_1})$ = *desktop*, $UserAgent(\boldsymbol{l_1})$ = *Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36*, $LDate(\boldsymbol{l_1})$ = 2016-01-05. Similarly, $TS(\boldsymbol{l_2})$ = 1464184238003,

IDUser ($\boldsymbol{l}_2$) = 62473886, *IP*($\boldsymbol{l}_2$) = 2548884122, *Status*($\boldsymbol{l}_2$) = 1, *HashPswd*($\boldsymbol{l}_2$) = *cdb932605b94652a09ab*, *Medium*($\boldsymbol{l}_2$) = *appAndroid*, *UserAgent*($\boldsymbol{l}_2$) = *Pokec.sk/2.3.2 (Android 5.1.1; Sony D2303)*, *LDate*($\boldsymbol{l}_2$) = 2016-05-25.

TABLE III. SQL TABLE OF FREQUENCIES FOR USERAGENT

iduser	*useragent*	*freq_useragent*
54473838	*Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36*	96
54473838	*Mozilla/5.0 (Linux; U; Android 4.1.1; sk-sk; SonyC1505 Build/11.3.A.2.33) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30*	156
62473886	*Pokec.sk/2.3.2 (Android 5.1.1; Sony D2303)*	234
62473886	*Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.103 Safari/537.36*	31
62473886	*Mozilla/5.0 (Windows NT 6.1; rv:46.0) Gecko/20100101 Firefox/46.0*	22

III. USED ANOMALY MODEL AND MOTIVATION

An anomaly model is developed for detection of compromised accounts with login data from the Pokec online social network. The login data of each account is recomputed into frequencies for given attributes and these frequencies are used to calculate anomalies of new logins from previous logins. Anomalies for particular attributes are combined into an anomaly score and if there is a login with the score higher than a threshold, the account is considered compromised. The model is inspired by [3] but login data is used instead of messages. At the same time, different attributes are employed and SQL tables are used. The models introduced for detection of compromised accounts in [3], [4], [9] cannot be employed directly as they need different data to what is available in the Pokec online social network. Collection of different data would require additional costs related to reimplementation of the network and additional long waiting time for the period of data collection. Any attempt to use traditional data mining techniques such as decision trees, Bayesian networks, nearest neighbors or neural networks lead to very bad accuracy for detection of compromised accounts. Let us define learning logins $\boldsymbol{L}_{\text{ts}}^{\text{L}}$ as the set of all logins $\boldsymbol{l} \in \boldsymbol{L}$ where $TS(\boldsymbol{l}) \leq$ ts, ts is a given timestamp. Testing logins $\boldsymbol{L}_{\text{ts}}^{\text{T}}$ is the set of all logins $\boldsymbol{l} \in \boldsymbol{L}$ where value $TS(\boldsymbol{l}) >$ ts. Suppose set $\boldsymbol{A} = \{a_1; a_2; \ldots; a_j; \ldots; a_{2806}\}$ is the set of unique accounts *IDUser*($\boldsymbol{l}$) from all $\boldsymbol{l} \in \boldsymbol{L}$. Suppose set $\boldsymbol{B} = \{B_3; B_4; B_5; B_6; B_7\}$ = {*IP*; *Status*; *HashPswd*; *Medium*; *UserAgent*} is the set of attributes whose anomalies are considered for the anomaly model. Attribute B_1 is not used as it only distinguishes different logins of one user B_2 and B_8 is not used because it only optimizes stored data for the database system of the network. The attributes in $\boldsymbol{B}$ were collected by the senior management of the Pokec online social network and the change of collected data and further selection was beyond our control. Suppose set $\boldsymbol{C}_{a_j,B_k}$ is the set of unique values $B_k(\boldsymbol{l})$ for all logins $\boldsymbol{l} \in \boldsymbol{L}_{\text{ts}}^{\text{L}}$ where $B_k \in \boldsymbol{B}$ and value *IDUser*($\boldsymbol{l}$)= a_j. Frequency f_{a_j,B_k,c_m} is the number of logins $\boldsymbol{l} \in \boldsymbol{L}_{\text{ts}}^{\text{L}}$ where *IDUser*($\boldsymbol{l}$)=a_j and $B_k(\boldsymbol{l}) = c_m$, $a_j \in \boldsymbol{A}$, $B_k \in \boldsymbol{B}$, $c_m \in \boldsymbol{C}_{a_j,\boldsymbol{B}_k}$. Average frequency $\bar{f}_{a_j,B_k}$ is formulated as

$$\bar{f}_{a_j,B_k} = \frac{\sum_{c_m \in \boldsymbol{C}_{a_j,B_k}} f_{a_j,B_k,c_m}}{\#(\boldsymbol{C}_{a_j,B_k})} \qquad (1)$$

where $a_j \in \boldsymbol{A}$ and $B_k \in \boldsymbol{B}$. The learning algorithm for creation of frequencies is as follows:

Step 1: Create frequencies f_{a_j,B_k,c_m} for all $a_j \in \boldsymbol{A}$, all $B_k \in \boldsymbol{B}$, and all $c_m \in \boldsymbol{C}_{a_j,\boldsymbol{B}_k}$. Go to Step 2.

Step 2: Compute average frequencies $\bar{f}_{a_j,B_k}$ according to equation (1) for all $a_j \in \boldsymbol{A}$ and all $B_k \in \boldsymbol{B}$. Go to Step 3.

Step 3: Store all f_{a_j,B_k,c_m} and all $\bar{f}_{a_j,B_k}$ for future use. End.

Frequencies f_{a_j,B_k,c_m} are stored as SQL tables and there is one SQL table for each $B_k \in \boldsymbol{B}$. A sample table for *UserAgent* is in Table III. In this table, $\boldsymbol{C}_{54473838,UserAgent}$ = {*Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36*; *Mozilla/5.0 (Linux; U; Android 4.1.1; sk-sk; SonyC1505 Build/11.3.A.2.33) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30*} and $\boldsymbol{C}_{62473886,UserAgent}$ = {*Pokec.sk/2.3.2 (Android 5.1.1; Sony D2303)*; *Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.103 Safari/537.36*; *Mozilla/5.0 (Windows NT 6.1; rv:46.0) Gecko/20100101 Firefox/46.0*}. If c = *Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36*, frequency $f_{54473838,UserAgent,c}$ = 96.

Stored frequencies and average frequencies are used for computation of anomaly scores $as(\boldsymbol{l})$, $\boldsymbol{l} \in \boldsymbol{L}_{\text{ts}}^{\text{T}}$, IDUser($\boldsymbol{l}$)= a_j, a_j is a specified account, $a_j \in \boldsymbol{A}$, in the classification part of the anomaly model. The anomaly scores themselves are calculated with given weights w_k for each $B_k \in \boldsymbol{B}$. The scores are compared with a given maximal threshold value $\text{as}_{\max} \in [0; 1]$. The classification algorithm for determination if a specified account $a_j \in \boldsymbol{A}$ is compromised or not is in the following steps:

Step 1: Take all logins $\boldsymbol{l} \in \boldsymbol{L}_{\text{ts}}^{\text{T}}$ where *IDUser*($\boldsymbol{l}$)= a_j and put them into a new set $\boldsymbol{M}$. Go to Step 2.

Step 2: Do the following for each $\boldsymbol{l} \in \boldsymbol{M}$ and $B_k \in \boldsymbol{B}$: If $B_k(\boldsymbol{l})$ is not in $\boldsymbol{C}_{a_j,\boldsymbol{B}_k}$, set anomaly for B_k $\text{an}_{B_k}(\boldsymbol{l}) = 1$. Otherwise, if $f_{a_j,B_k,B_k(\boldsymbol{l})} \geq \bar{f}_{a_j,B_k}$, set $\text{an}_{B_k}(\boldsymbol{l}) = 0$ and if $f_{a_j,B_k,B_k(\boldsymbol{l})} < \bar{f}_{a_j,B_k}$, set $\text{an}_{B_k}(\boldsymbol{l}) = 1 - \frac{f_{a_j,B_k,B_k(\boldsymbol{l})}}{\sum_{c_m \in \boldsymbol{C}_{a_j,B_k}} f_{a_j,B_k,c_m}}$. Go to Step 3.

Step 3: Compute anomaly score $as(\boldsymbol{l})$ for each $\boldsymbol{l} \in \boldsymbol{M}$ as $as(\boldsymbol{l}) = \sum_{B_k \in \boldsymbol{B}} w_k \cdot \text{an}_{B_k}(\boldsymbol{l})$. Go to Step 4.

Step 4: If there is any $as(\boldsymbol{l}) \geq \mathrm{as}_{\max}$ for all $\boldsymbol{l} \in \boldsymbol{M}$, then the specified account a_j is considered to be *compromised*. Otherwise, the account is considered *normal*. End.

The above mentioned algorithm allows all $a_j \in \boldsymbol{A}$ to be checked and as a result, compromised accounts can be processed. This helps their real holders to become the only holders again.

TABLE IV. CONFUSION MATRIX

		Detected	
		compromised	*normal*
Real	*compromised*	10	2
	normal	216	2578

IV. EXPERIMENTS

Achieved experimental results which were obtained with the anomaly model from Section III and the login data from Section II are described here. The anomaly model had been implemented in programming language Java 7 with open-source cluster-computing framework Apache Spark 1.6.0 and executed on a system with three nodes. Java is a general-purpose language that is object-oriented, class-based and designed to have very few implementation dependencies [6]. Computing framework Apache Spark allows distributed and in-memory processing of data and comes with Resilient Distributed Datasets and DataFrames [7]. Resilient Distributed Dataset is a resilient and distributed collection of records spread over one or many partitions. DataFrame uses capabilities of Resilient Distributed Dataset applies schema to the data so that a distributed collection of tabular data is organized into rows and named columns. This is conceptually equivalent to an SQL table and the Spark SQL execution engine allows to do SQL operations with the data stored in Apache Hive. Apache Hive facilitates querying and managing large datasets residing in distributed storage. In the experiments, the timestamp for division of logins into learning and testing was chosen so that the testing logins contained all logins from the last available thirty days and the learning logins had all the previous logins.

TABLE V. RESULTS

Measure/Method	**Anomaly Model**
Sensitivity (%)	83.33
Specificity (%)	92.27
Overall accuracy (%)	92.23

Frequencies and average frequencies were created for the learning logins involving 2806 accounts with 2794 normal and 12 compromised ones. After creation, the frequencies and average frequencies were used for detecting which accounts were compromised in the testing period of thirty days. The real situation and the detected outcomes were summarized into the confusion matrix in Table IV. The matrix has ten true positives (tp) formulated as the number of accounts that are compromised and detected as compromised. It has two false negatives (fn) defined as the number of accounts which are compromised and detected as normal. The confusion matrix also contains 216 false positives (fp) understood as the number of normal accounts identified as compromised ones and 2578 true negatives (tn) understood as the number of normal accounts which are correctly determined by the model as normal ones.

The confusion matrix in Table IV was used for computation of sensitivity, specificity and overall accuracy. Sensitivity is the ratio of compromised accounts detected as compromised ones correctly and its definition is $\text{sensitivity} = \frac{\text{tp}}{\text{tp+fn}}$. Specificity measures the ratio of normal accounts which are correctly detected as normal ones and it is defined as $\text{specificity} = \frac{\text{tn}}{\text{tn+fp}}$. $\text{Overall accuracy} = \frac{\text{tp+tn}}{\text{tp+tn+fp+fn}}$ measures the ratio of correct detections to all performed detections. Exact values for the measures are shown in Table V. The values are considered to be good if they are as high as possible. The overall accuracy of the anomaly model presented in Section III is 92.23 percent. Its sensitivity is 83.33 percent and its specificity is 92.27 percent, which is promising. High sensitivity means it does not happen that there are many compromised accounts considered normal with possible exposures of the accounts to non-holders. High specificity signalizes that high maintenance with unnecessary checking of normal accounts is avoided. This maintenance might include blocking of the account, an SMS authorization of the account or some investigations involving staff members of the network. Computations of the measures were performed for many possible weights of attributes and the weights giving the best results were chosen. The weight for the ip address was chosen as 0.3, status as 0.0, hashed password as 0.1, medium as 0.5, and the user-agent string as 0.1. Other combinations of weights also gave relatively good results and hence we decided to keep status. It might be useful for optimization in different situations or social networks. Many thresholds for anomaly scores were tested and value 0.9 gave the best results.

V. CONCLUSIONS

A system for automatic detection of compromised accounts based on measuring sudden changes in behavior with an anomaly model was presented. The model used historical logins into accounts for establishment of their typical behavior with frequencies and average frequencies. The frequencies considered the ip address of the user trying to log in the account, the success status of the login, hashed password given by the user, medium representing the device of the user and the user-agent string sent by the device of the user. Anomaly scores were computed for a set of new logins with the frequencies and average frequencies and an account was considered compromised if any of the scores exceeded a given threshold. The performance of the system was validated with data containing logins from 2806 accounts and with sensitivity, specificity and overall accuracy. The fact if compromised accounts were recognized was measured with sensitivity. The fact if normal accounts were left without additional maintenance was measured with specificity. With found weights for attributes and a threshold for anomaly scores, sensitivity was 83.33 percent and specificity reached 92.27 percent. The accuracy was 92.23 percent. Overall, the results showed that the presented system is interesting for finding compromised accounts and future development of the Pokec online social network.

REFERENCES

[1] M. R. M. Aburrous, Design and Development of an Intelligent Association Classification Mining Fuzzy Based Scheme for Phishing Website Detection with an Emphasis on E-Banking. UK: University of Bradford eThesis, 2010.

[2] Q. Cao, M. Sirivianos, X. Yang, T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in USENIX Conf. on Networked Systems Design and Implementation, 2012, pp. 1-14.

[3] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, "COMPA: Detecting compromised accounts on social networks," in Network and Distributed System Security Symposium, 2013.

[4] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, "Towards detecting compromised accounts on social networks," IEEE Transactions on Dependable and Secure Computing, vol. PP, no. 99, 2015.

[5] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, B. Y. Zhao, "Detecting and characterizing social spam campaigns," in ACM SIGCOMM Conference on Internet Measurement, 2010, pp. 35-47.

[6] J. Gosling, B. Joy, G. Steele, G. Bracha, A. Buckley, The Java Language Specification. USA: Oracle America, 2015.

[7] R. C. Maheshwar, D. Haritha, "Survey on high performance analytics of bigdata with Apache Spark," in Int. Conf. on Advanced Communication Control and Computing Technologies, 2016, pp. 721-725.

[8] J. A. Obar, S. Wildman, "Social media definition and the governance challenge: An introduction to the special issue," Telecommunications Policy, vol. 39, no. 9, pp. 745-750, 2015.

[9] X. Ruan, Z. Wu, S. Jajodia, "Profiling online social behaviors for compromised account detection," IEEE Transactions on Information Forensics and Security, vol. 11, no. 1, pp. 176-187, 2015.

[10] Statistical Office of the Slovak Republic, How Many of Us Are There, What Households We Form. Slovakia: Statistical Offfice of the Slovak Reputlic, 14 p., 2015.

[11] The Stastistics Portal, "Most famous social network sites worldwide as of January 2017, rangked by number of active users (in millions)," available at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, 2017

[12] G. Stringhini, C. Kruegel, G. Vigna, "Detecting spammers on social networks," in Annual Computer Security Applications Conf., 2010, pp. 1-9.

[13] World Newsmedia Network, Clobal Social Media Trends. UK : European Publishers Council, 2015.

[14] C. Xiao, D. M. Freeman, T. Hwa, "Detecting clusters of fake accounts in online social networks," in ACM Workshop on Artificial Intelligence and Security, 2015, pp. 91-101.

Comparison of the Battery Energy Storage and Fuel Cell Energy Source for the Safety-Critical Drives Considering Reliability and Fault Tolerance

Igor Bolvashenkov, Jörg Kammermann, and Hans-Georg Herzog
Institute of Energy Conversion Technology
Technical University of Munich (TUM)
Arcistrasse 21 80333 Munich, Germany
Email: igor.bolvashenkov@tum.de

Ilia Frenkel
Center for Reliability and Risk Management
Shamoon College of Engineering (SCE)
56 Bialik Street, 84100 Beer Sheva, Israel
Email: iliaf@sce.ac.il

***Abstract* - This paper describes the fault tolerance comparative analysis of a battery electric energy storage consisting of multiple battery submodules and fuel cell electric energy source for the traction drive of an electrical helicopter. The entire propulsion system reliability assessment is based on the reliability evaluation of the system components including electric energy sources. Considering the high requirements on reliability and fault tolerance, including the strict limitations on the installation space and weight of the overall system, the accurate assessment and correct choice of the type and design features of such energy sources is extremely important. Regarding the task of minimizing the weight and size of a storage device, in terms of a system approach, optimal types of electric energy source at today's technological level have been discussed and recommended.**

Keywords: battery storage, fuel cell, reliability, fault tolerance, Markov model, electrical helicopter

I. INTRODUCTION

The electric energy source is one of the main and important parts of the full electric vehicle's traction drive. The safety of helicopter's flight mainly depends on the sustainability of onboard electric energy sources. Today there are really many firms and research institutes all-around the world, which are all dedicated to the development of the Battery Electric Energy Storage (BEES) performance.

At the same time, due to the fact that the batteries at the current level of technological development have insufficient values of energy and power density in comparison with hydrogen, as is evident from Figure 1, it is expedient to carry out a comparative analysis of electric drive options with other competitive electric energy sources. As such an alternative option for comparison the Fuel Cell Electric Energy Source (FCEES) has been chosen.

Considering the purpose of using the battery electric energy storage, the accurate assessment of indicators, such as energy density, power density, storage capacity, failure probability, and fault tolerance of BEES and FCEES is becoming extremely important for the optimal choice of topology and the parameters of such supply devices. In case of an electric helicopter, the battery cells or fuel cells failures can lead to a reduced functionality, deterioration of operational modes, and even a catastrophic situation in the electric propulsion system.

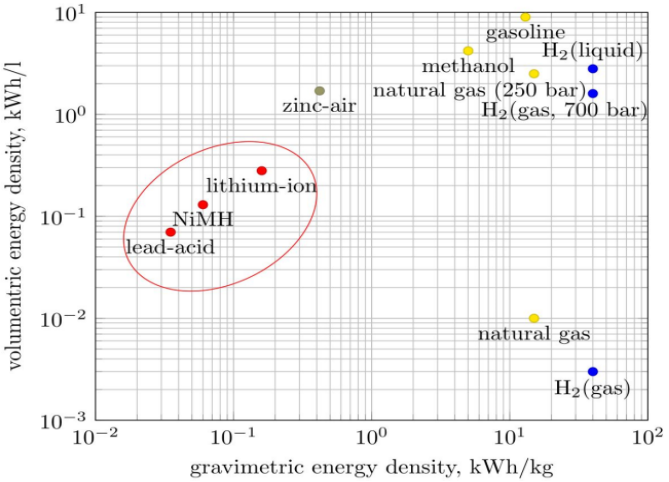


Fig.1: Energy density of various storage systems [4]

As is well known, the BEES as well as FCEES consist of multiple cells in parallel and/or serial connections in order to satisfy the high power and reliability design requirements to the traction drive of electrical helicopters. At the same time, it is necessary to consider the strict limitations on the installation space and weight of the whole electric propulsion system. The entire system reliability assessment is based on the reliability evaluation of system components including individual cells and stacks.

In this paper, one attempt of preliminary comparative assessment the reliability and fault tolerance indices of a BEES and FCEES consisting of multiple submodules for an electrical helicopter has been provided. As a comprehensive parameter for the assessment of fault tolerance of the components with a number of degradation states, it is proposed to use the criterion "degree of fault tolerance" (DOFT), of which the main theoretical principles are discussed in [1].

In order to evaluate the reliability of the whole battery module, the Multi-State System reliability Markov Model (MSS-MM) has been introduced. In this model the structural features and parameters of traction electric motor and electric inverter has also been considered to evaluate their reliability indices under different fail operational conditions. Comparative studies are carried out for the multilevel topology of BEES and FCEES in order to demonstrate the advantages of a reconfigurable system topology, hence, to improve the survivability and fault tolerance of the energy sources.

II. Objects of Investigation

The propulsion systems of an fully electrical "search and rescue" helicopter with a total take-off weight up to 3000 kg, schematically shown in Figure 2. One of the parts of such a helicopter is electric energy source.

As shown in paper [2]–[4], the best option of constructing the schemes of the traction electric drive for the helicopter in terms of the high level of the fault tolerance, is a 9-phase synchronous electric motor with permanent magnets with galvanically not connected stator windings and with multilevel cascaded H-bridge (CHB) inverter, shown in Figure 3. These options will be considered for the further investigations.

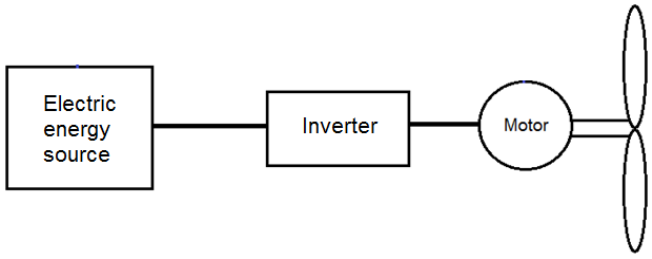


Fig.2: Structure of propulsion system of an electrical helicopter

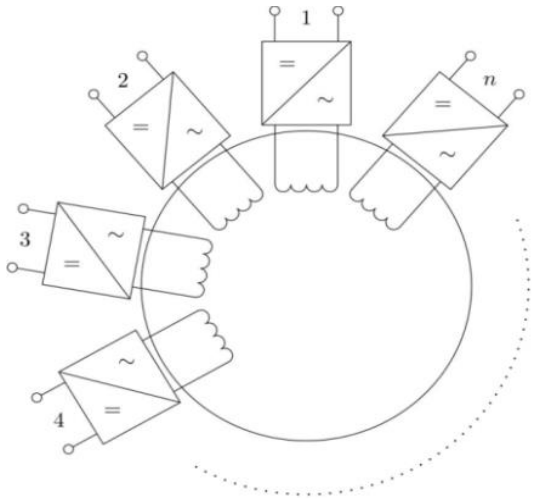


Fig.3: Schematic graphical system definition for a 9-phase traction motor, driven by an H-bridge inverter topology [2]

Considering the characteristics of the helicopter's traction drive and the parameters of the semiconductor devices, a 17-level CHB inverter has been recommended to use as electrical inverter [2]. One inverter submodule with the submodule of energy source is shown in Figure 4.

Thus, an open question requiring an electric traction drive topology optimization of the all-electric helicopter is a matter of choosing the optimal type of electric energy sources, considering reliability and fault tolerance at the minimum weight and dimensions.

While implementing this embodiment of the traction drive, BEES and FCEES are the only source of electric energy and, accordingly, it imposed more stringent requirements in terms of reliability and fault tolerance. The implementation of the helicopter's requirements regarding the safety flight can be achieved by fractional redundancy with a certain amount of reserve battery cells or fuel cells, considering the design limitations in weight and volume of the equipment. Today, the practical realization of an all-electric version of the helicopter is still quite difficult, due to the lack of sufficient electric energy and at the same time light and small enough batteries.

III. Battery Storage Topology

First of all, it is important to mention that that this study on the choice of battery cells type does not include the quality indicators (State of Charge, lifetime, etc.) of a particular type of cell, but only their declared technical data. In terms of the preliminary analysis and testing of the proposed method, this approach seems to be quite acceptable, since it allows in the future providing a similar assessment for the new promising options of battery cells based on the collected statistical data of BEES, which are already in use today.

Considering the weight and size characteristics of the various battery cells, presented in Figure 5, the best type of the batteries was chosen, which are practically realized today, i.e. lithium-ion batteries.

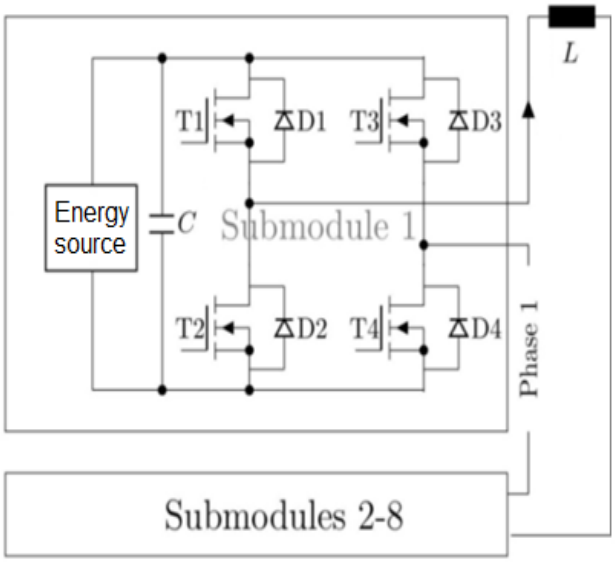


Fig.4: 17-level CHB inverter's one phase submodule [2]

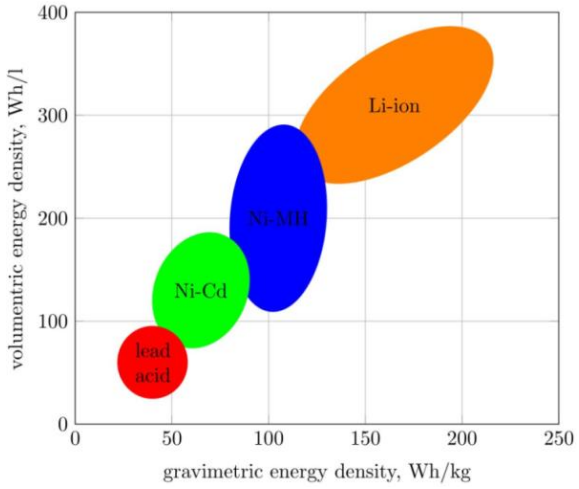


Fig.5: Modern battery types [5]

As the objects for further investigation two cells of different sizes were selected: Samsung INR18650-25R and GTF 26650. The results of the research presented in [6] show that the cells, which have a higher capacity, have also the best features of reliability and durability. However, in the present study at the carrying out of a comparative analysis by the authors, the same value of failure rate for the two considered battery cells, equal to 150 FIT (failure in time) has been accepted [7]–[14].

Taking into account the strict requirements to the level of fault tolerance of the helicopter's electric traction drive, it is advisable to use reconfigurable matrix topologies of BEES, shown in Figure 6, which have a higher reliability and fault tolerance indices than a conventional option, discussed in [15] and [16].

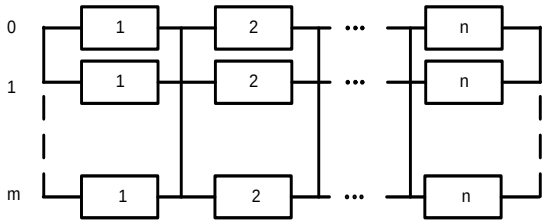
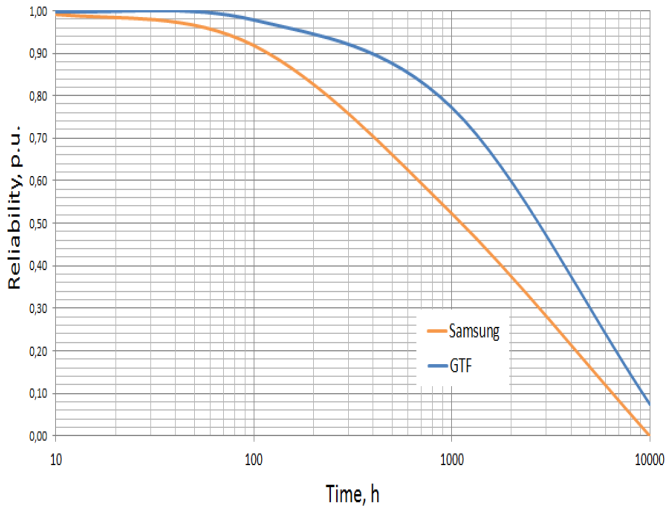


Fig.6: Reconfigurable BEES topology [16]

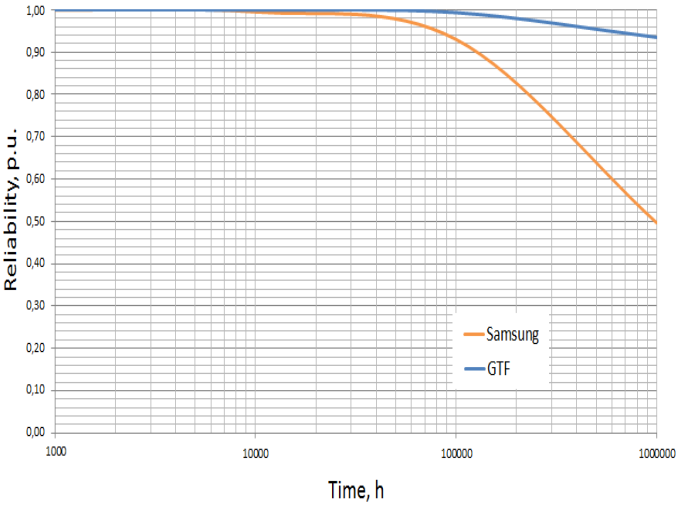
$$\lambda_{\text{BEES}}(t) = \frac{N(m+1)\cdot\lambda e^{-\lambda t}\cdot(1-e^{-\lambda t})^{m}}{1-(1-e^{-\lambda t})^{m+1}}$$

where λ is the failure rate of one element of the system and *m* the multiplicity of redundancy.

Considering the specified design parameters of the propulsion system of an electric helicopter, the driving power, and the voltage of 9-phase traction electric motor, the total failure rate of battery energy storage was computed. For the calculated value, the graphs of the BEES reliability functions, shown in Figure 7, till the first failure of the above two cell sizes, without, Figure 7(a), and with full redundancy, Figure 7(b), were plotted.



(a)



(b)

Fig.7: Reliability function for two BEES cells options without (a) and with (b) 100% redundancy

Figure 7 shows that in order to ensure the required value of the fault tolerance of the helicopter's traction drive, it is necessary to implement structural BEES redundancy. Since the full redundancy is not possible, because of the weight and dimension limits, the value of partial structural BEES redundancy – being sufficient to meet the requirements of the project – will be further evaluated.

IV. Fuel Cell Electric Energy Source

While analyzing the characteristics of the FCEES, it is necessary to consider the system from the position of the system approach, which means, together with hydrogen storage and hydrogen itself, respectively, with their weight and volume properties.

As an energy carrier, the hydrogen is used to provide the necessary amount of energy for the traction drive of electrical helicopter. Based on the analysis of modern technologies for hydrogen storage, two of the most suitable configuration for

hydrogen storage tanks were taken into account: a cylindrical tank [17] and a spherical tank [18].

If just recently, the fuel cells were losing significantly in almost all parameters an electrical traction drive with battery electric energy storage [19], at the present time, fuel cells at least not inferior to them in all main characteristics of [20]–[22]. Nevertheless, for the further assessment the real fuel cells data, installed on the Toyota Mirai were accepted [22].

Despite the fact that currently there are publications about the higher rates of reliability indices of the fuel cell stacks, for the reliability calculations the minimum MTTF value of 20,000 hours was accepted.

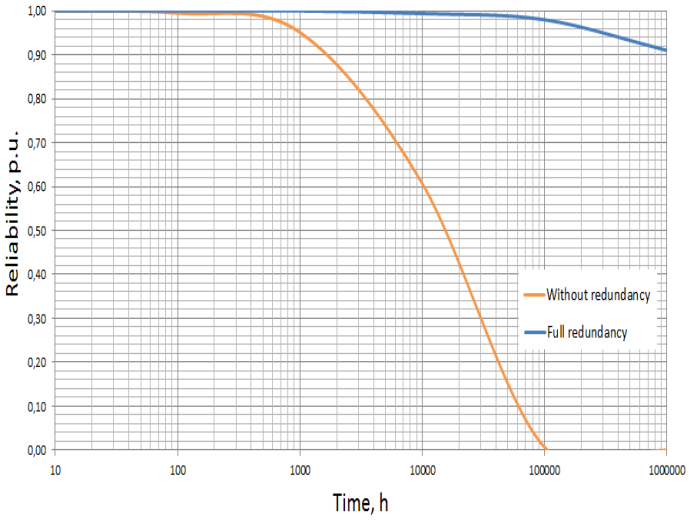


Fig.8: Reliability function of FCES

As can be seen from Figure 8 shows that in order to ensure the required value of the fault tolerance of the helicopter's traction drive, it is necessary to implement structural FCEES redundancy. Since the full redundancy is not possible, because of the weight and dimension limits, the value of partial structural FCEES redundancy – being sufficient to meet the requirements of the project – will be further evaluated

V. Markov Model and Results of Simulation

At the present time, a specific approach called the universal generating function (UGF) technique has been widely applied to MSS reliability analysis [23]–[26]. However, the main restriction of this powerful technique is that it may only be applied to random variables and therefore, concerning MSS reliability, it operates only with steady-states performance distributions. In order to extend the UGF technique application to dynamic MSS reliability analysis a special transform, called L_Z-transform, has been introduced [24] for discrete-state continuous-time Markov processes.

A. Battery Electric Energy Storage

Figure 9 presents states-transitions diagram of the Markov model of BEES consisting of NxM battery cells.

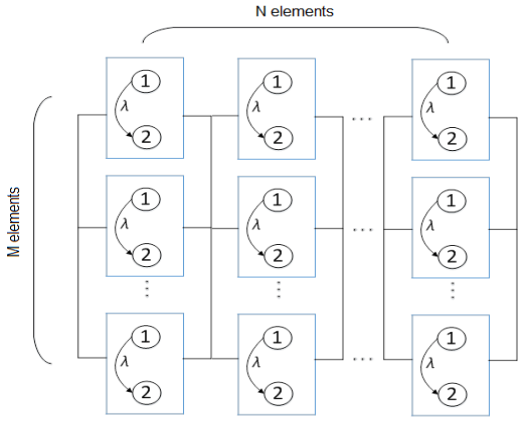


Fig.9: State-transitions diagram

As was mentioned, each battery cell is a device with two states of performance: a fully operational state with a nominal capacity G and a total failure corresponding to a capacity of 0.

According to the Markov method and the state-transitions diagram (Figure 9), the following system of differential equations has been constructed:

$$\begin{cases} \dfrac{dp_{i1}(t)}{dt} = -\lambda p_{i1}(t), \\ \dfrac{dp_{i2}(t)}{dt} = \lambda p_{i2}(t). \end{cases}$$

Initial conditions are: $p_{i1}(0)=1,\ p_{i2}(0)=0$

A numerical solution for probabilities can be obtained for each of this system of differential equations using MATLAB®. Therefore, one obtains for each accumulator the following output performance stochastic processes:

$$\mathbf{g}_i \in \{g_{i1}, g_{i2}\} = \{G, 0\};$$
$$\mathbf{p}_i(t) = \{p_{i1}(t), p_{i2}(t)\}.$$

Having the set $\mathbf{g}_i$, $\mathbf{p}_i(t)$ one can define L_Z-transform, associated with each unit output performance stochastic process:

$$L_z\{G_i(t)\} = p_{i1}(t)z^G + p_{i2}(t)z^0.$$

It was considered that all N-serial battery cells of the system are connected in M-parallel modules. Therefore, L_Z-transform, associated with the whole system is:

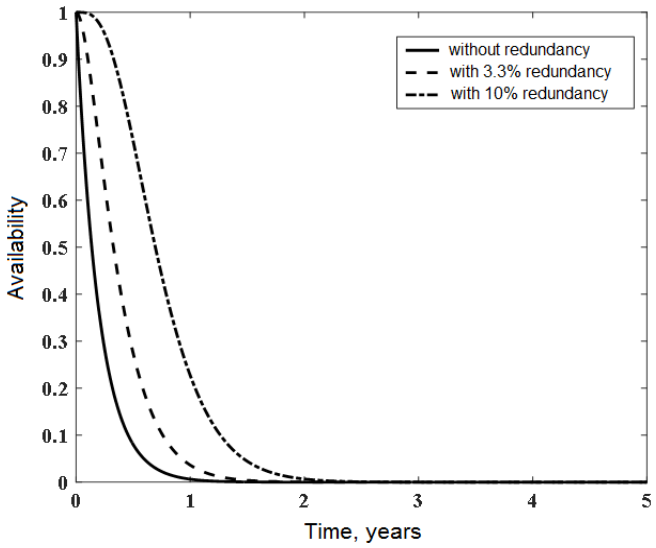
$$L_z\{G(t)\} = \Omega_{f_{par}}\left(L_z\{G_1(t)\}, L_z\{G_2(t)\}, \ldots, L_z\{G_M(t)\}\right) =$$
$$\Omega_{f_{par}}\left(p_{11}(t)z^G + p_{12}(t)z^0, p_{21}(t)z^G + p_{22}(t)z^0, \ldots, p_{M,1}(t)z^G + p_{M,2}(t)z^0\right)$$

All battery cells are similar, so, L_Z-transform of the whole system may be presented as M-times multiplication of the L_Z-transform of the N-serial module, where the performance of z

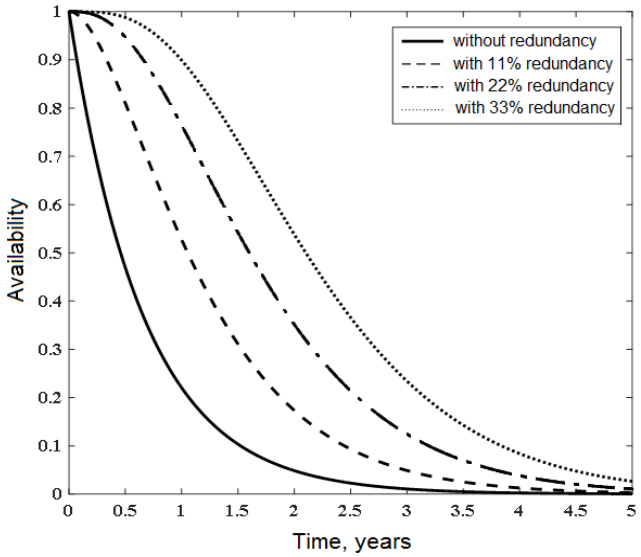
are found as sum values of performances of corresponding terms. The final expression of the whole system's L_Z-transform is of the following form:

$$L_z\{G(t)\} = \prod_{i=1}^{M}\{p_{i1}(t)z^{G} + p_{i2}(t)z^{0}\} = \sum_{k=0}^{M}\binom{M}{k}p_1(t)^{M-k}p_2(t)^{k}$$

Taking into account the requirements of the project, the reliability function has been calculated for the nominal the load level, shown in Figure 10.



(a)



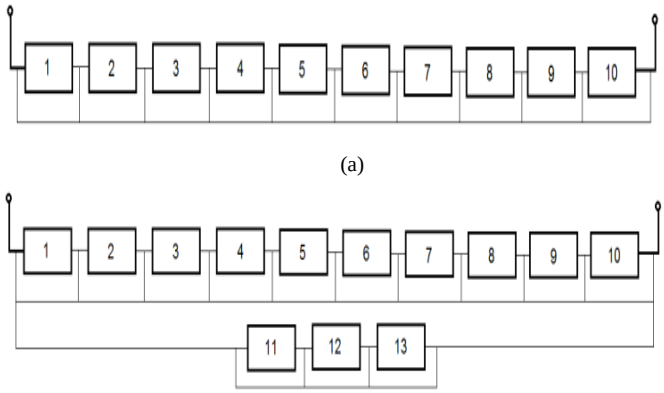
(b)

Fig.10: Mean instantaneous availability of BEES with Samsung cells (a) and GTF cells (b)

On the basis of simulation results, the level of redundancy of electric energy source, sufficient for the implementation of the required fault tolerance values has been defined. The results have shown that the fault tolerance indices of BEES could be improved to the required value for the GTF cells by implementing of more than 22% redundant battery cells. For BEES based on the Samsung battery cells, a much more backup cell is required. In this regard, further will be considered BEES based on the GTF cells.

B. Fuel Cell Electric Energy Source

Considering the requirements of the project, the following parameters for the electric traction drive have to be taken into account: the output nominal power is 500 kilowatts, the traction motor rotational speed is 400 rpm, and the voltage is 800 volts. In accordance with the above parameters, the minimally sufficient number of fuel cells and their connections schemes in terms of the maximum fault tolerance has been determined. The schemes are shown in Figure 11.



(a)

(b)

Fig.11: Fuel cell electric energy source, (a) – without redundancy, (b) – with 30% redundancy

According to the method described above, the each fuel cell stack can be represented as a device with 2 states of performance: a fully operational state with a stack power capacity G (50 kW) and a total failure corresponding to a capacity of 0, shown in Figure 12.

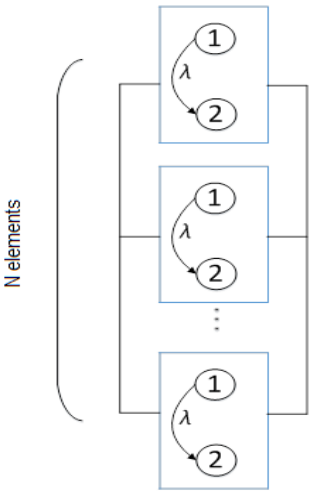


Fig.12: State-transitions diagram

Using the whole system's LZ-transform, as described above, it is possible to obtain the multi-state system (MSS) reliability function for the constant demand level. For the system without redundancy this function has the following form:

$$A_1(t) = \sum_{k=0}^{0} \binom{10}{k} p_1(t)^{10-k} \, p_2(t)^k = p_1(t)^{10}$$

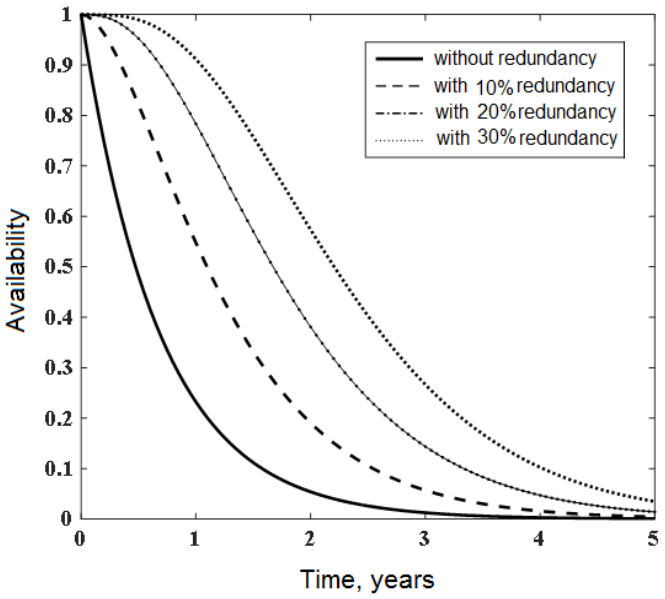


Fig.13: Mean instantaneous availability of FCEES

Figure 13 presents the simulation results of the MSS mean instantaneous availability for the fuel cells electric energy source with the different levels of redundancy: 0, 10%, 20%, and 30%, respectively. It should be noted that the reliability indices of the hydrogen storage system in the calculations were not considered.

Based on results of simulation, the level of redundancy of electric energy source, sufficient for the implementation of the required fault tolerance values, has been defined. The results have shown that the fault tolerance indices of FCEES could be improved to the required value by implementing of more than 20% redundant fuel cells.

VI. Weight Comparison

The assessment results discussed in [4] have shown that the optimum in terms of the weight and dimensions of the embodiment of the BEES submodules for both types of cells is a 9-phase version of a motor with a 17-level electric inverter.

On the basis of the given design parameters of the traction drive and the MOSFET's characteristics, for different option of multi-phase traction motors (from 6 to 11 phases), the number of battery cells, and the total weight of BEES for a one-hour flight were calculated. The calculation results are shown in Table I.

TABLE I. Total Weight of BEES in Kilogram

Redundancy \ Phase number		6	7	9	11
Without	Number	15600	15400	14400	15600
	Weight	1357	1340	1253	1340
With	Number	19032	18788	17568	19032
	Weight	1656	1635	1529	1656

For the correct calculation of the weight parameters of FCEES, it is necessary to take into account the inertia of the fuel cells, as shown in Figure 14. One possible way to improve the dynamic characteristics of FCEES is to use ultracapacitors [26]–[28]. Thus, the weight of the required number of ultracapacitors was taken into account in order to calculate the total weight of the FCEES.

Figure 15 shows the characteristics of ultracapacitors in comparison with other electrical energy storage devices.

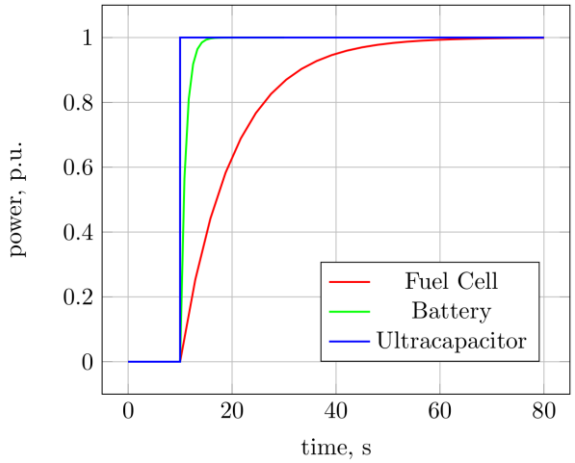


Fig.14: Power response of fuel cells, batteries, and ultracapacitor [27]

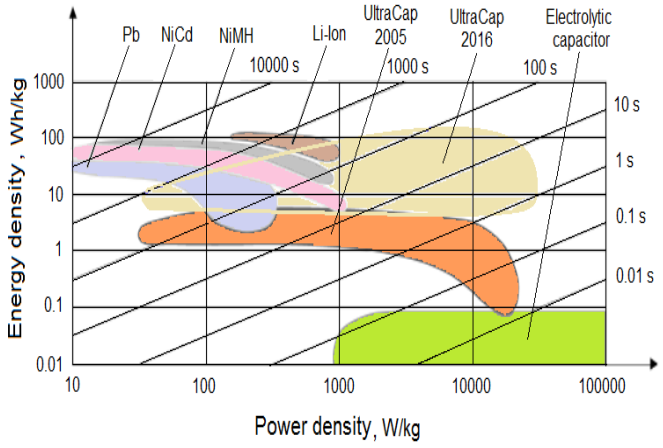


Fig.15: Energy and power density of electric energy storages

The value of 50 Wh/kg, achieved by the various researchers in practice on the basis of graphene's ultracapacitors, as the energy density has been accepted to calculate the weight characteristics of ultracapacitors. The ultracapacitors should provide sustainably the electric traction drive of the helicopter with energy for 60 seconds until the moment when it reaches the nominal mode of generating power.

Table II shows the power density values of the fuel cells, and Table III presents the energy density values of the hydrogen storage, used in further calculations.

TABLE II. TECHNICAL DATA OF FUEL CELLS

Fuel cell type	**Power density,** kW/kg	**Power density,** kW/l
Toyota Mirai	2.0	3.1

TABLE III. TECHNICAL DATA OF HYDROGEN STORAGE

Storage tank	**Energy density,** kWh/kg	**Energy density,** kWh/l
Cylindrical	3.0	1.2
Spherical	5.8	1.6

As hydrogen storage, the characteristic of the spherical variant of tank has been chosen for the calculation.

Table IV shows the weight characteristics of the particular components and the entire FCEES.

TABLE IV. TOTAL WEIGHT OF FCEES IN KILOGRAM

Component **Redundancy**	**Fuel cells**	**Tank**	**Ultracaps**	**TOTAL**
Without	338	195	167	700
With	406	195	201	802

Table V summarizes the results of the comparative analysis of the electric energy sources taking into account the required level of redundancy.

TABLE V. RESULTS OF COMPARISON IN KILOGRAM

Energy source	**BEES**	**FCEES**
Weight	1529	802

Considering the weight at the same level of fault tolerance, the BEES is significantly inferior to the FCEES, as can be seen from Table V.

VII. CONCLUSION

The results of analysis show that the reconfigurable topologies of BEES and FCEES have a higher reliability and fault tolerance than a conventional option. By using the structural fractional redundancy, the value of the total failure probability corresponds to the design requirements for the safety-critical system.

The amount of redundancy degree should be determined taking into account the restrictions on the weight and size of BEES and FCEES. Considering the task of minimizing the weight and size of electric energy source, in terms of a system approach, an optimal number of phases of the traction motor integrated with multilevel power inverter has been discussed and recommended.

It is advisable to carry out the choice of the optimal topology and parameters of BEES and FCEES, based on the sizes and parameters of the available types of battery cells and fuel cells, as well as the power, voltage, number of phases of the traction electric motor, and the number of submodules of the multilevel inverter.

Based on reliability simulation results, it has been concluded that the fault tolerance of BEES could be improved to the required value by implementing of 22% redundant battery cells and the fault tolerance indices of FCEES – by implementing of 20% redundant fuel cells.

Comparative analysis of the weight characteristics of various on-board electric energy sources for a one-hour flight allows drawing conclusions about the prospects of using the fuel cells together with the ultracapacitors and spherical hydrogen tank for the implementation of traction drive for an electric helicopter. In this case, the weight of the FCEES is two times lower than the weight of BEES.

Despite the fact that at the present level of technological development of battery cells, fuel cell, ultracapacitors and hydrogen tanks due to the relatively high weight and size characteristics, it is not quite easy to implement in practice an all-electric version of the helicopter, it may become real in the near future using new advanced types of the all such component of electric traction drive.

VIII. FURTHER WORK

A further extension of this work is a comprehensive study of the selected option of BEES and FCEES with other components of the electric traction drive and a comparative analysis of its parameters to other competitive types of electric energy storage, suitable for the use in a helicopter.

It is also advisable to carry out a comparative analysis of the sustainability of functioning fully electrical helicopter and hybrid-electrical helicopters with the different topology of the traction drive (serial, parallel, and combined) in the real operational conditions.

REFERENCES

[1] I. Bolvashenkov and H.-G. Herzog, "Degree of fault tolerance of the multi-phase traction electric motors: methodology and application", In Proc. of 16th IEEE Int. Conf. on Environment and Electrical Engineering (EEEIC'16), June 7-10, Florence, Italy, 2016, pp.1-6.

[2] I. Bolvashenkov, J. Kammermann, T. Lahlou and H.-G. Herzog, "Comparison and Choice of a Fault Tolerant Inverter Topology for the Traction Drive of an Electrical Helicopter", 3rd Int. Conf. on Electrical Systems for Aircraft, Railway, Ship Propulsion, and Road Vehicles (ESARS'16), 2-4 November 2016, Toulouse, France, 2016, pp.1-6

[3] I. Bolvashenkov, J. Kammermann, and H.-G. Herzog., "Reliability Assessment of a Fault Tolerant Propulsion System for an Electrical Helicopter", Proceedings of 12th International Conference and Exhibition on Ecological Vehicles and Renewable Energies (EVER), 11th – 13th April 2017, Monaco, 2017, pp. 1-6.

[4] I. Bolvashenkov, I. Frenkel, J. Kammermann, and H.-G. Herzog, "The Choice of an Optimal Structure and Parameters of Energy Storage for an Electrical Helicopter Traction Drive", Proceedings of 12th International Conference and Exhibition on Ecological Vehicles and Renewable Energies (EVER), 11th – 13th April 2017, Monaco, 2017, pp. 1-6.

[5] B. J. Landi, M. J. Ganter, C. D. Cress, R. A. Di Leoa and R. P. Raffaelle, "Carbon nanotubes for lithium ion batteries", Energy & Environmental Science, Vol.2, Iss.6, 2009, pp.638-654.

[6] N. Williard, W. He, M. Ostennan and M. Pecht, "Reliability and Failure Analysis of Lithium Ion Batteries for Electronic Systems", In Proc. of IEEE International Conference on Electronic Packaging Technology & High Density Packaging, 13-16 Aug. 2012, Guilin, China, 2012, pp.1051-1055.

[7] http://www.batteryspace.com, accessed: 08.03.2017.

[8] Y. Xing, N. Williard, K.-L. Tsui and M. Pecht, "A Comparative Review of Prognostics-based Reliability Methods for Lithium Batteries", In Proc. of Prognostics & System Health Management Conference (PHM), 24-25 May 2011, Shenzhen, China, 2011, pp.1-6.

[9] M. Liu, W. Li, C. Wang, M. P. Polis, L. Y. Wang and J. Li, "Reliability Evaluation of Large Scale Battery Energy Storage Systems", In IEEE Transactions on Smart Grid, 2016, Vol.PP, Iss.99, pp.1-11.

[10] M. Momayyezan, B. Hredzak and V. G. Agelidis, "A New Multiple Converter Topology for Battery/Ultracapacitor Hybrid Energy System", In Proc. of IECON, November 9-12 November 2015, Yokohama, Japan, 2015, pp.464-468,

[11] V. Musolino, L. Piegari and E. Tironi, "Storage systems for transportation, land handling and naval applications", In Proc. of IEEE Confer. on Electrical Systems for Aircraft, Railway and Ship Propulsion (ESARS), 16-18 Oct. 2012, Bologna, Italy, 2012, pp.1-9.

[12] C. Mikolajczak, M. Kahn, K. White and R. T. Long, "Lithium-Ion Batteries Hazard and Use Assessment", Fire Protection Research Foundation, Final report, MA, USA, 2011, 112 p.

[13] M. Abe, K. Nishimura, E. Seki, H. Haruna, T. Hirasawa, S. Ito and T. Yoshiura, "Lifetime Prediction for Heavy-duty Industrial Lithium-ion Batteries that Enables Highly Reliable System Design", Hitachi Review Vol.61, Iss.6, Japan, 2012, pp.259-263.

[14] B. Zhang and M. Kezunovic, "Impact of Available Electric Vehicle Battery Power Capacity on Power System Reliability", In Proc. of IEEE Power & Energy Society General Meeting, 21-25 July 2013, Vancouver, Canada, 2013, pp.1-5.

[15] Z. Liu, C. Tan, and F. Leng, "A reliability-based design concept for lithium-ion battery pack in electric vehicles", In: Reliability Engineering and System Safety, Elsevier, Iss.134, 2015, pp.169–177.

[16] S. M. Apollonsky and J. V. Kuklev, "The reliability and efficiency of electrical devices", Sankt-Petersburg, Russia, 2011, 448 p. (In Russian)

[17] K. Kunze and O. Kircher, "Cryo-compressed hydrogen storage", Oxford, September 28, 2012, 33 p.

[18] R.M. Sullivan et al., "Engineering analysis studies for preliminary design of light weight cryogenic hydrogen tanks in UAV applications", NASA, 2006, 27 p

[19] Eaves, S., Eaves, J., "A cost comparison of fuel-cell and battery electric vehicles", in Journal of Power Sources, Iss.130, 2004, pp. 208-212.

[20] Thomas, C.E., "Fuel cell and battery electric vehicles compared", in International Journal of Hydrogen Energy, Elsevier, Iss.34, 2009, pp.6005-6020.

[21] Eudy, L., Post, M., "American Fuel Cell Bus Project Evaluation: Second Report", National Renewable Energy Laboratory, Denver, Co, September 2015, 36 p.

[22] Yoshida, T. und Kojima, K., „Toyota MIRAI fuel cell vehicle and progress toward a future hydrogen society", The Electrochemical Society Interface, summer 2015, Chicago, IL, 2015, pp. 45-49.

[23] A. Lisnianski, I. Frenkel, Y. Ding, "Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers", Berlin, New York, Springer Verlag, 2010, 393 p.

[24] A. Lisnianski, "Lz-transform for a Discrete-state Continuous-time Markov Process and its Application to Multi-state System Reliability." In: Recent Advances in System Reliability. Signatures, Multi-state Systems and Statistical Inference. A. Lisnianski and I. Frenkel, Eds. London: Springer, 2012, pp. 79-95.

[25] I. Frenkel, I. Bolvashenkov, H.-G. Herzog, L. Khvatskin, "Performance Availability Assessment of Combined Multi Power Source Traction Drive Considering Real Operational Conditions", Transport and Telecommunication Journal, Vol.17, Iss.3, 2016, pp.179–191.

[26] T. Tsumura, K. Hirose, and M. Mino, "Application of a Markov process to evaluate the reliability of the power supply systems utilizing storage batteries", In Proc. of INTELEC, 13-17 October 2013, Hamburg, VDE Verlag GmbH, Berlin, Offenbach, 2013, pp.172-177.

[27] P. Thounthonga, S. Raël, and B. Davat, "Energy management of fuel cell/battery/supercapacitor hybrid power source for vehicle applications", in International Journal of Power Sources, Elsevier, Vol.193, 2009, pp. 376–385.

[28] D. Rotenberg, A. Vahidi, and I. Kolmanovsky, "Ultracapacitor Assisted Powertrains: Modeling, Control, Sizing, and the Impact on Fuel Economy", in IEEE Transactions on Control Systems Technology, May 2011, Vol.19, Iss.3, 201, pp.576-589.

[29] S. Schepmann and A. Vahidi "Heavy Vehicle Fuel Economy Improvement Using Ultracapacitor Power Assist and Preview-based MPC Energy Management", In Proc. of American Control Conference, 29 June -01 July 2011, San Francisco, CA, USA, 2011, pp.2707-2712.

Methodology for Quantitative Assessment of Fault Tolerance of the Multi-State Safety-Critical Systems with Functional Redundancy

Igor Bolvashenkov, Jörg Kammermann, and Hans-Georg Herzog
Institute of Energy Conversion Technology
Technical University of Munich (TUM)
Arcistrasse 21 80333 Munich, Germany
Email: igor.bolvashenkov@tum.de

***Abstract* – This paper describes a methodology of quantitative assessment of the fault tolerance of the multi-state safety-critical systems with functional redundancy. Such systems are the traction drives of electrical vehicles, consisting of the multiphase traction electric motor and multilevel electric inverter. It is suggested to consider such traction drive as a system with several degraded states. As a comprehensive parameter for the quantitative evaluating the fault tolerance, it is proposed to use the criterion of degree of the fault tolerance. For the approbation of proposed methodology, based on the multi-state system reliability Markov Model, the fault tolerance assessment of the traction train of an electrical helicopter has been carried out.**

Keywords: reliability, degree of fault tolerance, functional redundancy, multi-state system reliability Markov Model, transition probability, electric traction drive

I. Introduction

Regarding the constant growth of complexity of modern engineering systems it becomes more complicated to achieve the required level of its sustainable and safety operation. The task of implementing the specified requirements is closely related to the problem of the most accurate assessment of indicators of sustainable operation of the system, shown in Fig.1. It is particularly important to assess the required reliability and fault tolerance correctly. In the safety-critical applications, such as vehicle propulsion systems, the fault tolerance of all the equipment is obligatory. According to the plans for the electrification of various types of vehicles based on the electric energy generated by renewable sources, the tasks of a quantitative estimation of the global reliability of the safety-critical systems have now become very topical.

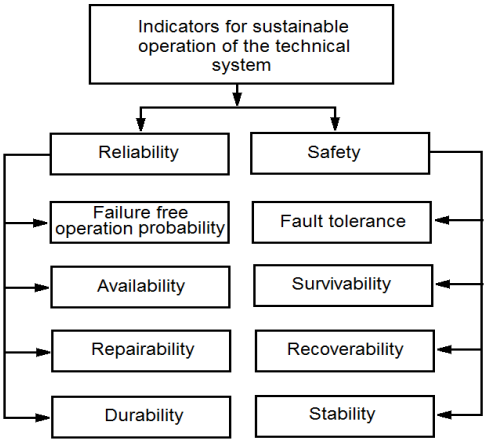


Fig.1. Indicators for sustainable operation

One of the most important requirements for the vehicle's propulsion system is the value of fault tolerance. In other words, the vehicle's propulsion system should operate and continue its sustainable functioning, even if one or more of its components have failed.

To implement this requirement, all components included in the system should be fault tolerant. As an example of practical use of the proposed method, the fault tolerance of two main parts of a vehicle's propulsion system was evaluated – the traction multiphase permanent magnet synchronous motor (PSM) and the electric inverter of electrical helicopter. In this case, according to the specified requirements of a safe flight of the designed electrical helicopter, the total failure rate for the entire traction drive of designed electrical helicopter should be less than 10^{-9}/h [1].

Today, a large number of publications are describing the comparison or analysis of different fault tolerant electric machines and electric inverter topologies for different vehicle applications [1]–[7] and [8]–[12]. Thus, it should be noted that the authors generally have studied various aspects of the fault tolerance, and in most cases only a qualitative assessment was performed. Modern methods for determining the degree of fault tolerance of electrical machines, power electronics, and the computer network topology are presented in [13]–[15] and [26]–[29]. The proposed methods have one typical common disadvantage – the lack of universality. Each of them allows solving a local specific problem for a particular object.

The authors have proposed a quite universal methodology of quantitative assessment the degree of fault tolerance (DOFT) of the multi-state safety-critical systems as a whole, as well as the DOFT of their components. As a mathematical apparatus for investigating the magnitude of fault tolerance, the Multi-State System Reliability Markov Models were used, which are well proven in the solution of reliability problems [16]–[18].

II. Approach and Methodology

A. MSS Markov Models and Transition Probabilities

Considering the above requirements on the probability of total failure of the electrical helicopter, as well as statistical data on the reliability of the traction electric motor and the electric inverter, it was determined that the optimal model for an assessment of the fault tolerance in such conditions is a Multi-State System Reliability Markov Model (MSS-MM), with K states, as shown in general form in Fig.2.

Theoretical base of this method is well known and described in [16]–[18] and examples of application in [23]–[25].

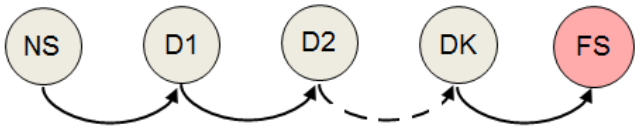


Fig.2. State diagram of Multi-State System Markov Model [16]

The first state NS corresponds to a completely failure-free operational normal state of the system. The states D1-DK are the states of degradation and correspond to failure cases – phase open-circuit failure, respectively, of the one, two, or more phases. The state FS of the model corresponds to the completely failed system, when the helicopter's drive completely loses the ability to operate. Thus, every following state of the degradation corresponds to one critical failure with a corresponding partial functionality loss of the traction drive.

At the same time, the number of the degraded states of the MSS-MM determined in accordance with the requirements of the project on the fault tolerance defines the technical capabilities of the system to continue functioning with reduced performance level as a result of the critical failure.

The most important and most difficult point for simulations by using the Markov Models is to determine the transition probabilities and the number of degraded states with a reduced level of functionality.

The values of the transition probabilities λ_1, λ_2,...λ_K are derived from the results of calculations of Degree of Fault Tolerance (*DOFT*) for the states 1, 2 and *K* respectively:

$$\lambda_{Ri} = 1 - DOFT_{Ri} \quad (1)$$

Here *R* is the value of reduced performance level according to the project requirements and *i*, the number of critical failures. The values of transition probabilities are affected by a large number of different factors, from design and environmental parameters to the using types of maintenance strategies, monitoring, and diagnostics.

As can be seen from equation (1), for the calculation of the transition probabilities the main importance is the correct calculation of DOFT for a given project required performance levels. Application cases of proposed methodology for the estimation reliability features of electric traction drive of the helicopter will be presented in the next section.

B. Degree of Fault Tolerance

Considering the definition of fault tolerance of a technical system as an ability to maintain the required functional level of the system, in case of one or more failures of its components, the *DOFT* can be defined as the amount of time, in which the system may remain in a degraded state without irreversible changes in its functionality. Mathematically, in general form this can be written by equation (2):

$$DOFT_i = \frac{W_R}{W_N} \cdot \frac{\Delta t_i}{\Delta t_N} \quad (2)$$

where W_R and W_N are the reduced and nominal values of the performance of the technical system; Δt_i and Δt_N are the duration of functioning after *i* failures and without failure, respectively.

As the value of the performance *W* can be considered the productivity, power, energy, quantity of information, etc. The value of Δt_i is defined by the overload capability of the system after *i* failures.

In the case when the required level of W_R has been predetermined by project requirements, it is useful to determine the *DOFT* for each corresponding level of performance in accordance with equation (3):

$$DOFT_{Ri}\left(W_R\right) = \frac{\Delta t_i}{\Delta t_N} \quad (3)$$

Δt_N is determined depending on the type of electric vehicles and largely depends on the specifications of its operation (aircraft, ships, trains, cars). For the helicopter, the flight operation "search and rescue" requires a sustainable and safe functioning for a duration of three hours. The value Δt_i is determined by the system's overload capability and its thermal stability. Δt_i indicates the duration of time, during which the system may operate in a critical failure mode without irreversible changes in the quality and functional inability concerning the further use.

Based on the considerations above, the procedure for the determination of the *DOFT* is shown as follows in Fig 3.

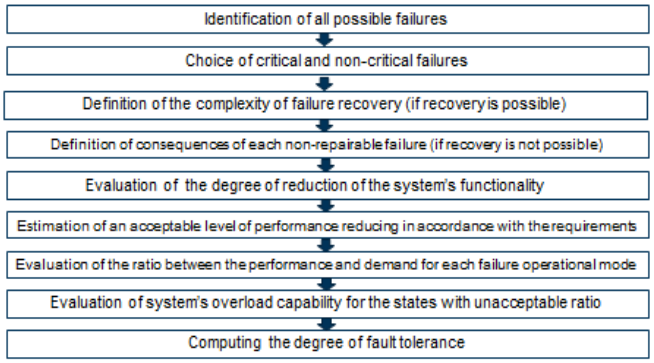


Fig.3. Algorithm for calculating the *DOFT*

As discussed in [1], the multiphase traction electric motor as well as multilevel electric inverter [30] can be considered as a multi-state system with the functional redundancy.

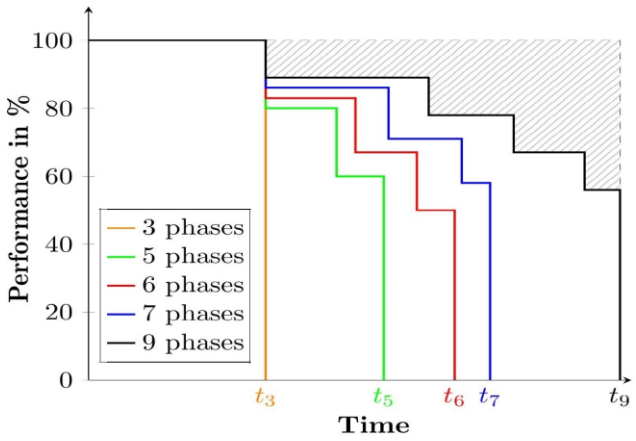


Fig.4. Graphical *DOFT* representation of the multiphase motor

In Fig.4 and 5 the value of *DOFT* is equal to the area below the degradation curves. The "gray" area above the curve of degradation of the 9-phase motor is equal to the probability of the transition of the multiphase motor in the state following on the critical failure.

The number of "steps" in Fig.4 corresponds to the number of degraded states of the electric machine after each critical failure, until the moment of time when the motor losses completely its functional ability. For a multiphase electric motor, this corresponds to the next critical phase failure.

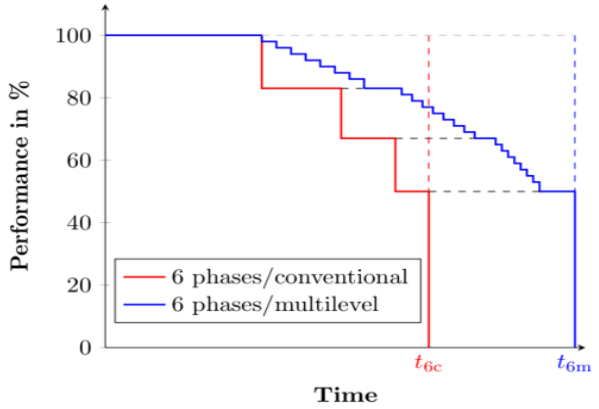


Fig.5. Graphical *DOFT* representation of the multilevel inverter

Additionally to the critical failures, the effect of small (non-critical) failures on the fault tolerance value of traction electric motor and electric power inverter, leading to partial or temporary loss of system functionality, can also be investigated using *DOFT*, as can be seen in Fig.6.

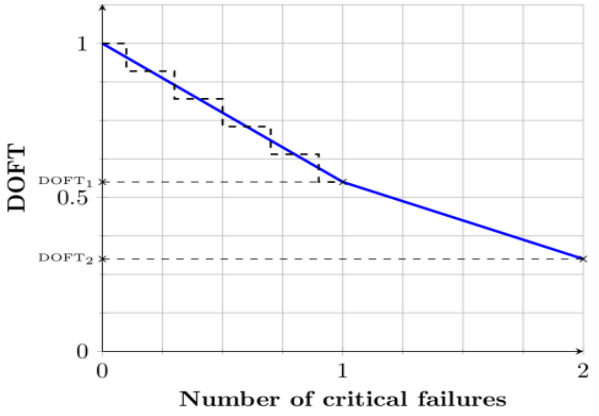


Fig.6. Graphical *DOFT* representation of different types of failure

A distinctive feature of the proposed methodology for a quantitative assessment of fault tolerance considering a technical system in comparison with existing techniques is its universality, i.e. the possibility to use it for different types of technical systems. Below, as an example of its practical use, the evaluation of the fault tolerance and transition probabilities of electric traction drive with multiphase motors and multilevel inverters was carried out.

III. Fault Tolerant Electric Traction Drive

The traction drive of electrical helicopter regarding the requirements on the reliability and fault tolerance is the safety-critical system. It means that all its components should be fault tolerant.

A. Electric Traction Drive Components

The power part of the traction drive of electric vehicles includes a source/storage of electric energy (the batteries, fuel cell, etc.), an electric energy converter/inverter, and a traction electric motor, as can be seen from Fig.7.

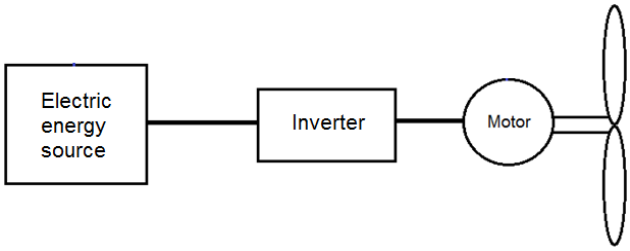


Fig.7. Traction drive components of electrical helicopter

In the present study, the electric energy source was not considered.

B. Critical Failures

In paper [6], it has been demonstrated that the vast majority of elements failures of the system "power inverter-traction electric motor" can be reduced to four basic types of failures of electric traction drive:

- open-circuit and short-circuit of the electric motor's phase;
- open-circuit and short-circuit of the inverter submodule.

In the development of an electric motor scheme is usually provided such a connection of the each inverter submodule with the protection system, which disconnects the electrical circuit of the failed submodule. This solution allows the failure "short-circuit" of the inverter's submodule to lead to the failure "open-circuit" of submodule.

The failure type "phase short-circuit" of electric motor can be reduced to the failure of "phase open-circuit" based on special design options of the stator winding performance, improved quality of insulation materials and advanced production technology. Excluding the possibility of failure type "short-circuit" on the basis of their reducing to the failures of the type "open-circuit" is one way to keep a functionality of multiphase electric motors in failure cases. Thus, in the simulation of functioning of electrical traction drive in failure conditions, as the critical failures will be considered the motor's "phase open-circuit" and inverter's "submodule open-circuit".

C. Multiphase Electric Motor

The results of a previous study by the authors [1] indicate that for the 3-phase PSM the specified requirements on the reliability for the entire power drive of a designed helicopter is not achievable without a functional and/or structural redundancy, and/or other activities that improve the indicators of reliability to the required value.

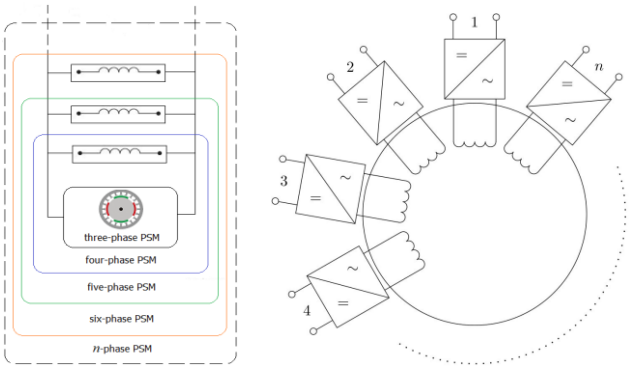


Fig.8. Multiphase traction motor as a multi-state system

One way to create the fault tolerant traction electric motors of high reliability is to increase the number of motor phases without changing the value of the motor's power. In Fig.8 is shown generally the schematic definition for a multiphase electric motor as the multi-state system.

The use of multiphase electric motors allows a reduction of the current value in each phase's open winding and to perform the power electronics unit integrally fabricated. Besides, it improves the efficiency of the electric motor and reduces the torque ripples, which is especially important in failure cases. At the same time, independent performance of each phase's switching channels provides increased reliability of the electric machine based on the principle of functional redundancy.

Unlike an electrical machine with a small number of phases in which the majority of critical elements failures leads to a complete failure of the machine, the multiphase electrical machine remains in operation up to a certain level of degradation and a corresponding change in the output characteristics. This allows realizing so a called functional redundancy.

Thus, on the one hand, increasing the number of phases of the electric motor reduces the impact of each failure in the power electronics control channel or in the phase of the motor on the characteristics of the whole traction drive.

On the other hand, increasing the number of phases leads to an increase of the failure probability in one of the phases. In this context, it is necessary to find an optimal compromise solution, based on the design requirements, the possibility of redundancy, and the reliability indices of the electrical machine and power electronics.

On the basis of the known values for the failure probability of each phase of the electric motor the optimal number of phases can be calculated, at which the required reliability and fault tolerance indicators of electric motor can be implemented, taking into account the possibility of one or more critical failures. The use of this redundancy method has its own features that must be considered in the study of physical processes and the design of the traction electric drive as a whole.

As shown in [1] the optimal electric machine for the safety-critical electric drives, considering system approach techniques, is a multiphase PSM with distributed stator windings and internal v-shaped arrangement of permanent magnets on the rotor. For a detailed study 5-, 6-, 7- and 9-phase PSM were selected.

For the traction electric motor of the helicopter considering high requirements on the drive's safety and fault tolerance, the overload capability in the fail operational modes is especially important. In such operating conditions, a stable operation for a specified time on the modes of reduced power without critical asymmetry of PSM's parameters is also extremely important.

Considering a normal, i.e. failure-free, operational mode, the electric motor can endure a short-term overload because the thermal capacity is sufficiently large, whereas for failure cases the situation is changing dramatically. As known from operating experience, the largest numbers of operational failures are caused by technological overloads [19]–[21].

Most of the total failures of traction electric motors occur because of stator winding faults and bearing failures, so that these components play a crucial role in the overall reliability value of the motor. Their lifetime and fault tolerance significantly affect the operating temperature, developed inside of the motor. Furthermore, the overheating causes quickly deterioration of the motor winding insulation and the bearing.

The causes for overheating of electrical machines can be technological overload of the motor as well as the occurrence of different failure modes. The most significant of them are the various types of short-circuits, unbalanced work at the loss of one or several phases, jamming of the rotor of the electrical machine.

Technological overload leads to an increase in temperature of the motor windings, a gradual deterioration and finally to its total failure. According to recent research [20], a long-term operation of the electric motor with an overcurrent by only 5% of the nominal reduces its lifetime by 10 times.

Most of the overcurrent failures of electric motors are associated primarily with damages inside the motor, leading to an asymmetric overcurrent and, as a consequence, to overheating. The main types of failures, which lead to dangerous overheating of the stator windings and to a total failure of the electric motor (without system of protection), are the short circuit faults:

- between turns;
- between coils;
- between phases;
- between wires and the housing of the motor.

Their effects are described in detail in the literature. These effects lead to dramatic increasing of the current in one or more phases of the motor and ultimately to the motor's overheating. At an effective system of protection against emergency situations, each of the above-discussed failures can be reduced to an embodiment of the loss of failed phases (or automatic shutdown).

When it is required to maintain the load at a given level, which is a common requirement for safety-critical systems, such as a helicopter traction drive, it is necessary to increase the phase currents in the remaining phases of the electric motor. This will result in a certain level of the motor's overheating and, in terms of reliability, a severely limited duration of operation in this mode of load.

For the traction drives of the electric helicopter the heaviest given load level for maintaining the performance in a failed operational mode is 113% of the nominal load (short time of operation).

In order to estimate the level of overheating of the windings of multi-phase traction motors and respective conclusions on its thermal stability in the case of failure in one or two phases, it is proposed to use the overheating factor K_T, which shows how many times the motor windings temperature exceeds the nominal value:

$$K_T = \frac{T_i}{T_N} \tag{4}$$

where T_i and T_N, respectively, describe the winding's temperature in failed operational mode in i-phases and in the nominal (N) failure-free operational mode. The overheating factor is graphically presented in Fig.9.

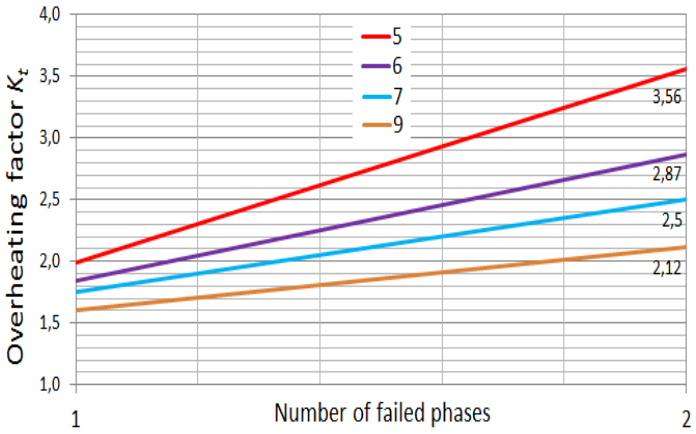


Fig.9. Overheating of the stator windings after the loss of phases

The main goal of the preliminary analysis of the possible overload modes of a traction electric motor at various critical failures is to find the critical points in terms of thermal stability and overload capacity.

The consequences of the overload in failure operational mode are overcurrent and overheating of PSM, which lead to a reduction of reliability indices and decrease lifetime of the motor, as can be seen in Fig.10.

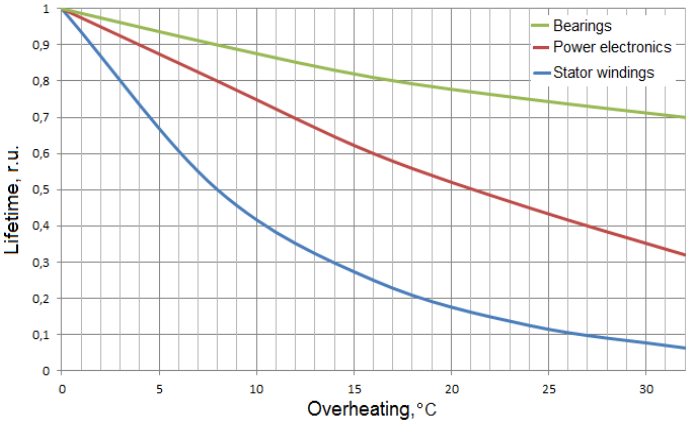


Fig.10. Lifetime of the parts of electric drive [22]

The main characteristic of the load modes of PSM for evaluation of *DOFT* is the thermal characteristics, estimated by formulas [20]:

$$t_N = \{\ln K^2 - \ln (K^2 - 1)\}/(A/C) \qquad (5)$$

where t_N describes the time of achievement of acceptable motor temperature value, K is the rate of exceeding the nominal value of the phase current, A stands for the heat irradiation of the motor, and C is the thermal capacity of the electrical machine.

Fig.11 shows graphically the results of calculating the thermal behavior of the electric motor in overload conditions, as a result of one or two critical failures considering thermal stability of the stator windings. The values of N_{deg} and N_{nom} in Fig.11 correspond to the values of traction drive performance (driving power) in degraded and nominal modes, respectively. On the basis of the analysis of the thermal behavior of the motor the values Δt_i for the different versions of the traction motor and the various failure modes can be determined. Based on the value Δt_i, the transition probabilities of Markov Models were calculated.

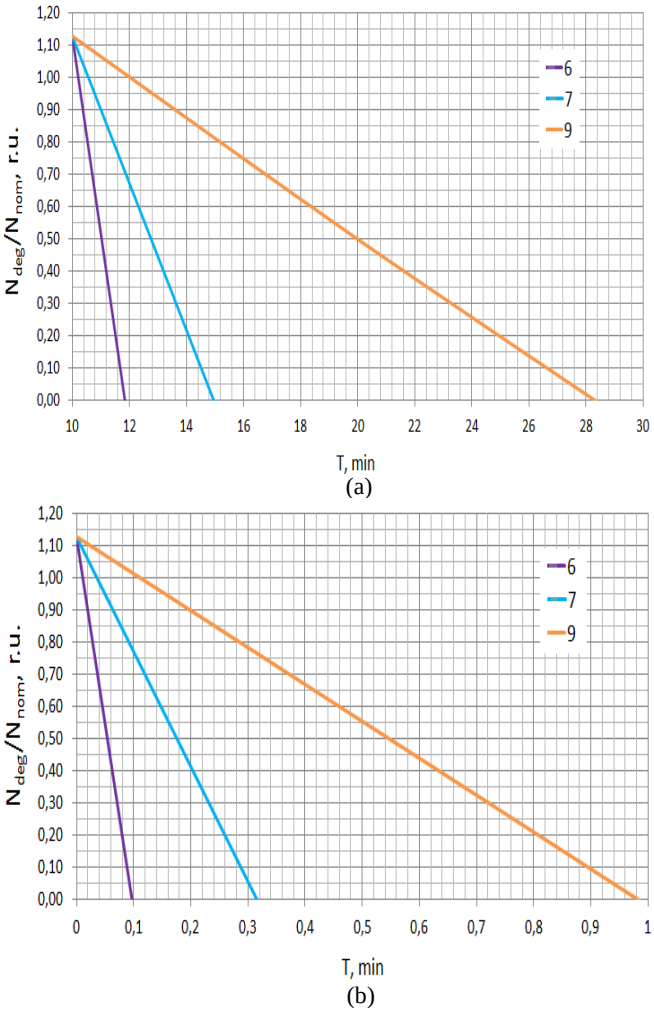


Fig.11. The duration of the safe motor operation at a one (a) and two (b) critical failures

The graphs of *DOFT* at 113% of nominal load, on the number of critical dangerous failures are shown in Fig.12.

In the general case, these graphical features allow carrying out a comparative analysis of different options of electric traction drives with a various number of safety critical failures. This is greatly helpful in the development of new types of traction electric drives and their components, considering the strict requirements on the fault tolerance.

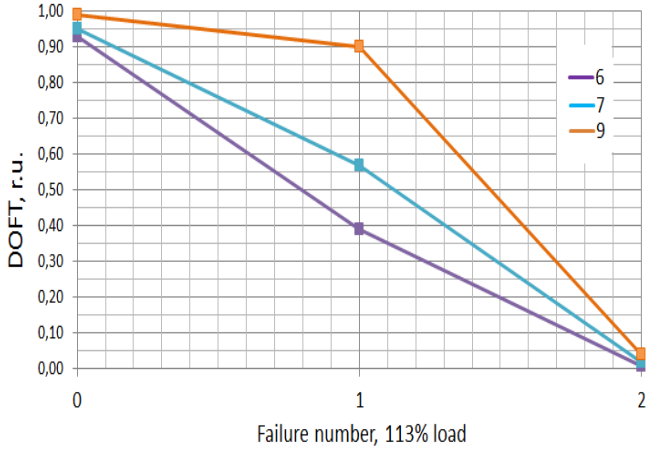
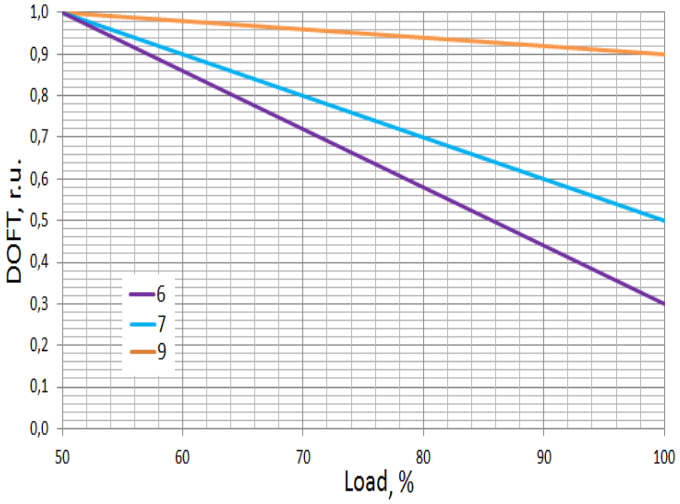


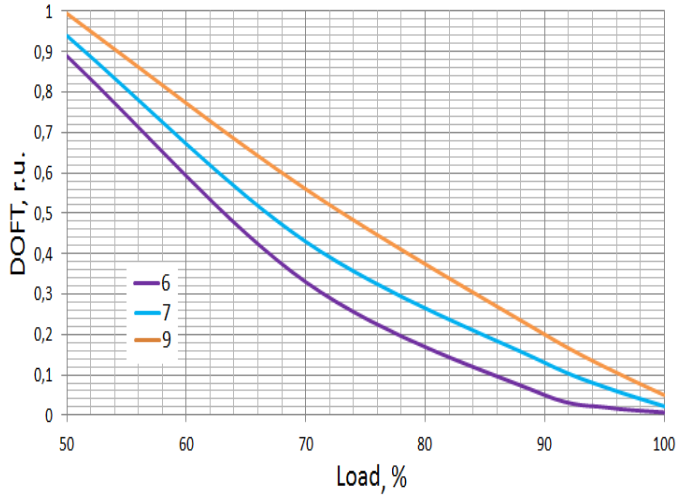
Fig.12. *DOFT* of multi-phase PSM in fail operations

Based on the constructed graphs a comparative analysis of *DOFT* for considered variants of the electric motors can be carried out quite easily. However, according to the

authors, more informative and convenient for practical use are the dependencies of *DOFT* on the value of a given load maintenance level, as shown in Fig.13. Thus, it is possible to assess the compliance of parameters of fault tolerance for each compared alternative to the design requirements.



(a)



(b)

Fig.13. DOFT of a given load level in fail operations: (a) – one failed phase, (b) – two failed phases

Generally, the graphical features shown in Fig.13 allow carrying out a comparative analysis of different options of electric traction drive for the different level of the required values of demand on the operational power.

Considering the project requirements it was determined that the optimal model for the analysis of fault tolerance in such conditions is a MSS-MM with a minimum of four states, as shown in Fig. 14, since it is not feasible to realize the required values of fault tolerance with only one degradative state of the motor.

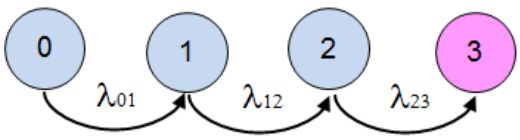


Fig.14. Multi-State System Markov Model of traction motor

For the traction electric motor of the helicopter considering high requirements on the drive safety and fault tolerance, the overload capability in the fail operational modes is especially important. In such operating conditions, it is also extremely crucial to be able to operate stably for a specified time on the modes of reduced power without critical asymmetry of PSM's parameters.

D. Multilevel Electric Inverter

As an inverter topology, the option of the multilevel cascaded H-bridge (CHB) inverter has been discussed. In this embodiment each phase of the multiphase traction motor has its own electric energy source and each phase is controlled by its own multilevel CHB inverter. It is well known that multilevel inverters offer several advantages compared to their two-level counterparts, discussed in [8] and [10]–[12]: smaller power filters, smaller voltage ratings for semiconductors, lower switching frequencies and less power losses. They offer also more modularity and they are more reliable.

Thus, CHB inverters are constructed on a series connection of single-phase inverters supplied by isolated DC electric energy sources. This kind of inverters gives a high modularity degree and consequently high reliability and fault tolerance. An approach based on the full inverter's power control could optimize the implementation and the reliability of the inverter while offering optimized operational behavior.

Based on the required design parameters of the traction drive of the electric helicopter, the preferred inverter option is a 17-Level CHB inverter [30]. So, in each phase there are 8 submodules. Fig.16 presents the basic topology of one phase of a CHB using battery electric energy storage on the DC side. Each battery module is connected to an H-bridge with 4 MOSFETs. The use of MOSFETs enhances the efficiency of the CHB inverter, because of the low conduction losses.

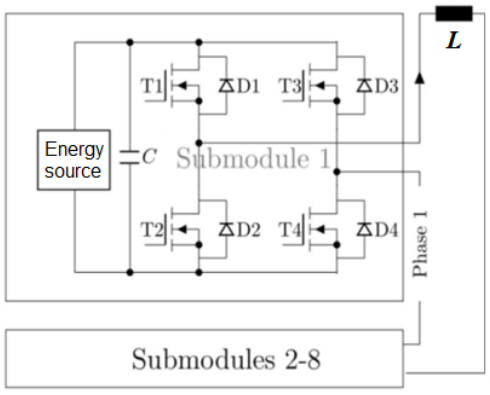


Fig.16. One submodule of the proposed 17-level CHB inverter

Considering the project requirements on the fault tolerance of safety-critical drives, as well as statistical data on the reliability of electric inverters, it was determined that the optimal model for the analysis of fault tolerance in such conditions is a MSS-MM with a minimum of five states, as shown in Fig.17.

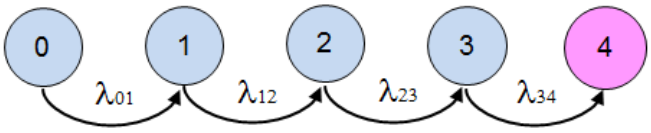


Fig.17. Multi-State System Markov model of multilevel inverter

On the basis of normative documents for the power electronics, the graph was plotted for electrical overload capability of the inverter, shown in Fig.18.

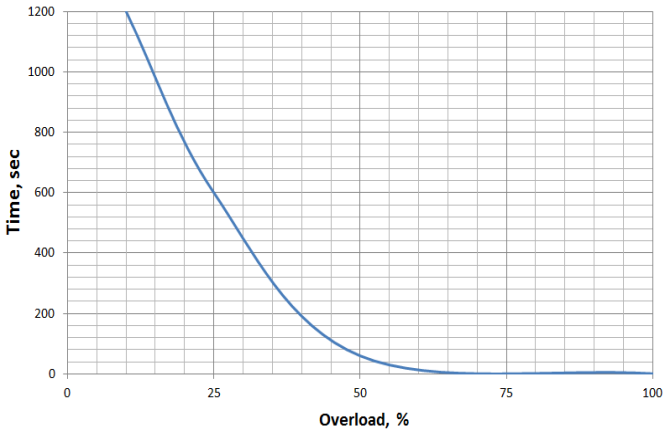


Fig.18. Overload capability of electric inverter

Considering the overload capability of multilevel inverter and real temperature modes during overload of an inverter in the case of successive failures of one, two, or three inverter submodules, the transition probabilities has been calculated for the MSS-MM, similar to the calculations for electric motors.

Based on the obtained data, the probabilities of a complete failure of the multiphase electric motor and a multilevel inverter were simulated and the results are presented in the next section.

IV. Simulation Results

In order to construct such models, the multiphase PSM as well as the multilevel inverter can be considered as a system with a loaded functional redundancy and consequently, with an appropriate reserve of fault tolerance. The transition probabilities for MSS-MM were calculated using the above mentioned *DOFT* method.

In this way the optimal choice the phases number and the windings type depends strongly on the requirements and parameters of a project and thus from the application.

Corresponding graphs for 6-, 7-, and 9-phase PSM at the 113% load level are presented in Fig.18. The horizontal axis indicates the operational time in hours and the vertical axis shows the probability of the total failure of the traction motor.

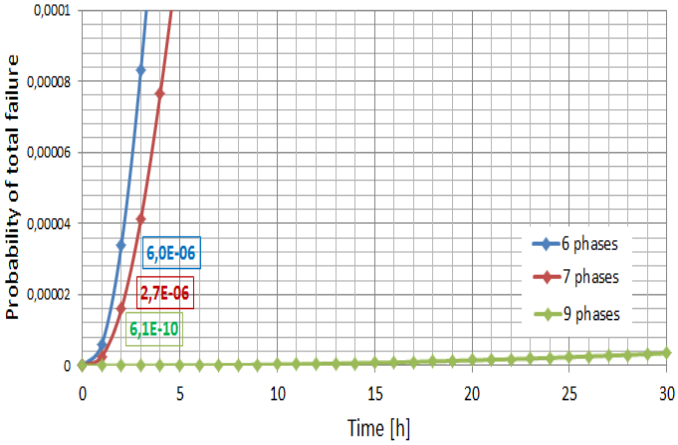


Fig.18. Probability function of total failure of PSM at the 113% load

For the simulation the worst case and critically dangerous variant of failure has been considered, i.e. the submodule failures occur in the same phase. In case of a simulation of a non-safety-critical failure, as well as the possibility of partial recovery of the power drive's operating ability in the degraded state, the value of the fault tolerance will be significantly higher.

Regarding the design requirements on fault tolerance, the reliability of the multilevel inverter was analyzed using the above MSS-MM, in case of emergency reducing the power to 113% of the nominal value. The corresponding graph for a different number of phases is presented in Fig.19.

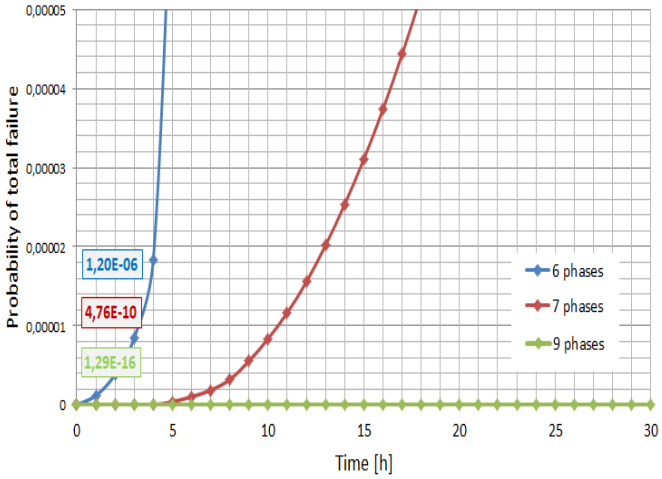


Fig.19. Probability function of a total failure of one motor phase with a multilevel inverter at the 113% load

The results of simulation of three consecutive critically dangerous failures allow quantifying the degree of fault tolerance of a 17-level inverter, which is one of the important parts of the traction drive of helicopters. The 7- and 9-phase options of the multiphase electric motor have shown the maximum compliance with the requirements relating to the safety-critical drives.

V. Conclusion

The paper presents a methodology for the quantitative assessment the fault tolerance of a multi-state safety-critical technical system, such as a vehicle's electric traction drive. The suggested evaluation of the operational reliability indices of fault tolerant topologies of electric traction drives is well formalized and suitable for practical application in reliability engineering to estimate fault tolerance indices of the multiphase traction motors as well as an electric power inverter, considering the aging process under the influence of operating conditions.

The evaluation of the fault tolerance of two important parts of a vehicle's electric propulsion system, the traction motor and the electric inverter, serve as an example of practical application of the proposed methodology. The results of the comparative analysis allow to conclude that for given project requirements on the level of reliability and fault tolerance of helicopter's electric traction drive in the real flying conditions only 9-phase motors in combination with 17-level CHB inverters completely fulfil the design requirements without any restriction.

Furthermore, the proposed method can be used as a universal tool for evaluation and optimization of the degree of fault tolerance in multi-state safety-critical technical systems, considering all the possibilities to increase functional and structural redundancy, monitoring, predictive control, and to choose the type of maintenance strategy.

REFERENCES

[1] I. Bolvashenkov, J. Kammermann, S. Willerich and H.-G. Herzog, "Comparative Study of Reliability and Fault Tolerance of Multi-Phase Permanent Magnet Synchronous Motors for Safety-Critical Drive Trains", in Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ'16), 4 th to 6th May, Madrid, Spain, 2016, pp.1-6.

[2] E. Levi, "Multiphase Electric Machines for Variable-Speed Applications", IEEE Transactions on Industrial Electronics, Vol.55, No 5, 2008, pp.1893-1909.

[3] M. Villani, M. Tursini, G. Fabri and L. Castellini, "Multi-Phase Permanent Magnet Motor Drives for Fault-Tolerant Applications", In Proc. of IEEE International Electric Machines & Drives Conference (IEMDC), 5-18 May 2011, Niagara Falls, Canada, 2011, pp.1351-1356.

[4] F. Scuiller, J.-F. Charpentier and E. Semail, "Multi-Star Multi-Phase Winding for a High Power Naval Propulsion Machine with Low Ripple Torques and High Fault Tolerant Ability", In Proc. of the IEEE Vehicle Power and Propulsion Conference (VPPC), 1-3 Sept. 2010, Lille, France, 2010, pp.1-5.

[5] E. Semail, X. Kestelyn and F. Locment, "Fault Tolerant Multiphase Electrical Drives: the Impact of Design", European Physical Journal - Applied Physics (EPJAP), Vol.43, Iss.2, 2008, pp.159-162.

[6] P.G. Vigrianov, "Assessment the impact of different failures on the power characteristics of the low power 7-phase permanent magnet synchronous motors", Moscow, Journal "Questions to Electromechanics", Moscow, Vol.128, Iss.3, 2012, pp.3-7. (in Russian)

[7] D. Fodorean, M. Ruba, L. Szabo and A. Miraoui, "Comparison of the main types of fault-tolerant electrical drives used in vehicle applications", In Proc. of International Symposium on Power Electronics, Electrical Drives, Automation and Motion, (SPEEDAM), June 11-13, Ischia, Italy, 2008, pp. 895-900.

[8] O. Josefsson, T. Thiringer, S. Lundmark and H. Zelaya, "Evaluation and comparison of a two-level and a multilevel inverter for an EV using a modulized battery topology", In Proc.of IEEE 38th Annual Conference on Industrial Electronics Society (IECON), Oct. 25-28, Montreal, Canada, 2012, pp. 2949-2956.

[9] A. V. Brazhnikov and I. R. Belozyorov,"Prospects for Use of Multiphase Phase-Pole-Controlled AC Inverter Drives in Traction Systems", European Journal of Natural History, Russia, Vol.2, 2011, pp.47-49.

[10] B. Sarrazin, N. Rouger, J. P. Ferrieux and J. C. Crebier, "Cascaded Inverters for electric vehicles: Towards a better management of traction chain from the battery to the motor?", In Proc. of IEEE International Symposium on Industrial Electronics, June 27-30, Gdansk, Poland, 2011, pp. 153-158.

[11] S. Fazel, S. Bernet, D. Krug and K. Jalili, "Design and Comparison of 4-kV Neutral-Point-Clamped, Flying-Capacitor, and Series-Connected H-Bridge Multilevel Converters", IEEE Transactions on Industry Applications, Vol.43, No.4, Jul.-Aug. 2007, pp. 1032-1040.

[12] M. Malinowski, K. Gopakumar, J. Rodriguez and M. Perez, "A Survey on Cascaded Multilevel Inverters", IEEE Transaction on Industrial Electronics, Vol. 57, No. 7, July 2010, pp. 2197-2206.

[13] B. A. Welchko, T. A. Lipo, T. M. Jahns and S. E. Schulz, "Fault Tolerant Three-Phase AC Motor Drive Topologies: A Comparison of Features, Cost, and Limitations", In: IEEE Transactions on Power Electronics, Vol.19, No 4, 2004, pp.1108-1116.

[14] U. De Pra, D. Baert and H. Kuyken, "Analysis of the Degree of Reliability of a Redundant Modular Inverter Structure", In Proc. of IEEE 12th International Telecommunications Energy Conference, 04-08 Oct. 1998, San Francisco, CA, 1998, pp.543- 548.

[15] S. Krivoi, M. Hajder, P. Dymora and M. Mazurek, "The Matrix Method of Determining the Fault Tolerance Degree of a Computer Network Topology", Sofia, Bulgaria, Publisher: ITHEA, Vol.13, No 3, 2006, pp.221-227.

[16] A. Lisnianski, I. Frenkel and Y. Ding, "Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers", Berlin, New York, Springer, 2010, 393 p.

[17] B. Natvig, "Multi-state systems reliability theory with applications", John Wiley & Sons, New York, 2011, 232 p.

[18] I. Bolvashenkov and H.-G. Herzog, "Use of Stochastic Models for Operational Efficiency Analysis of Multi Power Source Traction Drives", In Proc. of IEEE of International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO'16), 15-18 Feb. 2016, Beer Sheva, Israel, 2016, pp.124-130.

[19] D. Hann,"A combined electromagnetic and thermal optimisation of an aerospace electric motor", Int. Conference on Electrical Machines, ICEM, 6-8 Sept. 2010, Rome, Italy, 2010, pp.1-6.

[20] M. M. Katzman, "Electrical machines", Akademia, Moscow, Russia, 2001, 463 p. (in Russian)

[21] S. Mahdavi, T. Herold and K. Hameyer, "Thermal modeling as a tool to determine the overload capability of electrical machines", International Conference on Electrical Machines and Systems (ICEMS), 26-29 Oct. 2013, Busan, Korea, 2013, pp.454–458.

[22] I. Bolvashenkov and H.-G. Herzog, "Approach to Predictive Evaluation of the Reliability of Electric Drive Train Based on a Stochastic Model", In Proc. of IEEE 5th International Conference on Clean Electric Power (ICCEP'15), 16-18 June 2015, Taormina, Italy, 2015, pp.1-7.

[23] A. H. Ranjbar, M. Kiani and B. Fahimi, "Dynamic Markov Model for Reliability Evaluation of Power Electronic Systems", In Proc. of IEEE International Conference on Power Engineering, Energy and Electrical Drives (POWERENG), Malaga, Spain, May 2011, pp.1-6.

[24] M. Molaei, H. Oraee and M. Fotuhi-Firuzabad, "Markov Model of Drive-Motor Systems for Reliability Calculation", In Proc. of IEEE International Symposium on Industrial Electronics, 9-13 July 2006, Montreal, Canada, pp.2286-2291.

[25] T. Geyer and S. Schroder, "Reliability Considerations and Fault-Handling Strategies for Multi-MW Modular Drive Systems", In: IEEE Transactions on Industry Applications, Vol.46, No.6, Nov.-Dec. 2010, pp. 2442-2451.

[26] K. S. Trivedi, "Probability and Statistics with Reliability, Queuing, and Computer Science Applications", Second edition, Wiley, 2002, 848 p.

[27] S. J. Bavuso, J. B. Dugan, K. S. Trivedi, E. M. Rothmann and W. E. Smith, "Analysis of Typical Fault-Tolerant Architectures using HARP", In: IEEE Transactions on Reliability, Vol.R-36, Iss.2, June 1987, pp.176-185.

[28] N. Muellner and O. Thee, "The Degree of Masking Fault Tolerance vs. Temporal Redundancy", In: IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA), 22-25 March 2011, Biopolis, Singapore, 2011, pp.21-28.

[29] R. Ubar, S. Devadze, M. Jenihhin, J. Raik, G. Jervan and P. Ellervee, "Hierarchical Calculation of Malicious Faults for Evaluating the Fault-Tolerance", In Proc. of 4th IEEE International Symposium on Electronic Design, Test & Applications (DELTA), 23-25 Jan. 2008, Hong Kong, 2008, pp.222-227.

[30] I. Bolvashenkov, J. Kammermann, T. Lahlou and H.-G. Herzog, "Comparison and Choice of a Fault Tolerant Inverter Topology for the Traction Drive of an Electrical Helicopter", 3rd Int. Conf. on Electrical Systems for Aircraft, Railway, Ship Propulsion, and Road Vehicles (ESARS'16), Toulouse, France, 02-04 November, 2016, pp.1-6.

Stochastic Ageing and Maintenance Models for Unavailability Quantification of Complex Multi-Component Systems

Radim Briš
Department of Applied Mathematics
VŠB - Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Czech Republic
e-mail: radim.bris@vsb.cz

***Abstract*—Main objective of this paper is research and development of the previously developed original methodology so that to be used for unavailability calculation of a complex maintained system with maintained (both preventively and correctively) and ageing input components with optional distributions of the time to failure. Original computing methodology results from the stochastic alternating renewal process. Models I-III (for non-repairable component, repairable with both apparent and hidden failures) are in this paper completed by new Model IV with calendar based preventive maintenance policy in which the component is renewed as soon as it reaches a prescribed deterministic age. These models represent all frequently used component models with both preventive and corrective maintenance. Models were solved both analytically and numerically. Found component unavailability functions are used to quantify unavailability of a complex maintained system, here is used real power distribution network. System is represented by the use of directed acyclic graph, which proved to be very effective system representation for computing highly reliable systems.**

Keywords—unavailability quantification; alternating renewal process; maintenance and ageing models; directed acyclic graph

I. INTRODUCTION

To analyze unavailability of general and complex multi-component systems new or innovative stochastic ageing and maintenance models must be permanently developed. A complex multi-component system consists of finite number of non-identical components that can be realized as maintained and generally ageing components with different maintenance modes. Systems with non-identical components or components that are not operated in the same mode are analyzed in the paper. Such systems are widespread in practice. We can observe them for example in the electrical engineering industry (here we analyze a power electrical distribution network), nuclear industry, transport etc. Recently we can observe an increasing interest in publications to the development and especially maintenance optimization of multi-component systems. Obviously the selection of optimal maintenance policies for multi-component systems is more complex than that for single-component systems. The reason for this complexity is that there often exists one or more types of dependence (e.g. stochastic, structural or economic) between the components in a multi-component system [1]. Neglecting these dependencies while analyzing and optimizing these multi-component systems, may lead to wrong solutions of optimizing problems. Other more extended research works on maintenance models of multi-component systems are in [1-4].

Special methodology for high-performance computing which enables exact unavailability quantification of a real maintained and highly reliable system containing components with both preventive and corrective maintenance was demonstrated in our previous research [5]. The most important advantage of the methodology is that it enables the analyst to calculate arbitrary small values of the unavailability function during a mission time exactly, i.e. in full machine accuracy. Anyhow effective the methodology is, it had big restriction. An exponential distribution for the time to a failure is supposed, possibly for the time to restoration. Under this assumption, all frequently used models with both preventive and corrective maintenance were developed.

New generalization of the original methodology to be used for unavailability quantification of a system with ageing input components with optional distributions of the time to failure brings the paper [6]. For this purpose new findings in renewal theory were demonstrated and derived. For computing implementation an innovative recurrent linear integral equation was derived which helps us to compute unavailability function without knowledge the renewal density. The new equation is more effective than the corresponding equation resulting from the classic alternating renewal theory in which renewal density is included.

This paper brings further extension of the previous research. While paper [6] derives unavailability formulas for

basic three component models: i.e. Model I with components that cannot be repaired, Model II with repairable components for apparent failures and Model III with repairable components with hidden failures (model when a failure is identified only at special deterministically assigned times, appearing with a given period and in the case of its occurrence an restoration process starts), this paper develops new and realistic Model IV with calendar based preventive maintenance policy in which the component is renewed as soon as it reaches a prescribed constant age T_p.

Found component unavailability functions are used to quantify unavailability of a complex maintained system, as is power distribution network. In addition Model III and IV are confronted to demonstrate how the unavailability of the power network may be affected. System is represented by the use of directed acyclic graph, which proved to be very effective system representation [7] particularly for computing highly reliable systems, as demonstrated in [5].

II. System Representation and Unavailability Models

A. The directed acyclic graph

Example of a real system to be analyzed is shown in Figure 1. The system is demonstrated by means of a directed acyclic graph [8] (AG). A graph is composed of nodes and edges. The highest node (here u1) represents functionality of the whole system (success, failure), internal and terminal nodes represent subsystems and components. All of the nodes are bounded by edges. Direction of the graph is not explicitly marked in Figure 1 it is given by itself - by projection to vertical direction. The graph is acyclic which means that it cannot contain feedback loops.

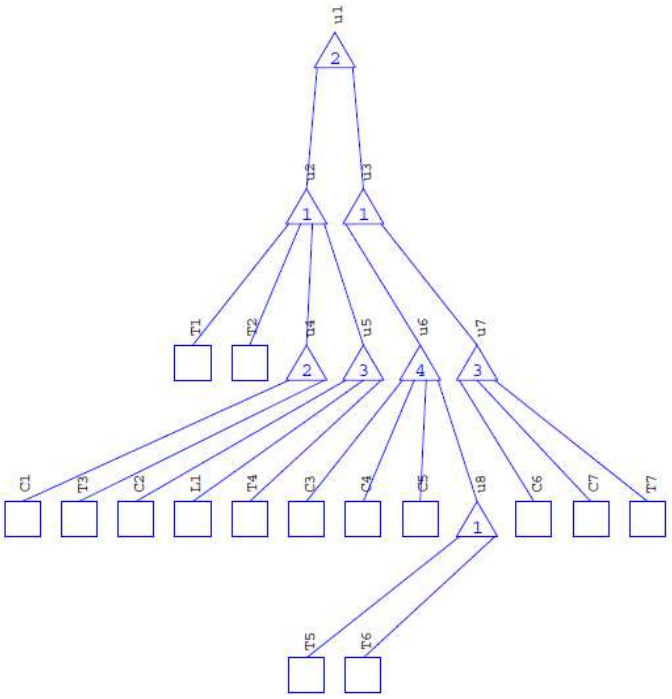


Figure 1. Graph structure of a real system

Terminal nodes, as for example T1 or T2, of the AG are marked by blue squares. They represent stochastic functionality of input system components given by a probability distribution of their time to failure and a maintenance model. From them we can compute a time course of the unavailability function of input components, using methodology of basic renewal theory, as for example in [9].

Internal nodes (non-terminal) are marked by blue triangles. They represent functionality of subsystems. A subsystem is well functioning in a given time point (success) just in the case when the number of well-functioning inferior edges reaches at least the number that is inside of the triangle, see Figure 1. Otherwise it is not-functioning (failure). For example internal node u3 is well-functioning when the number of well-functioning inferior edges is at least 1.

The key problem is to estimate the point (or instantaneous) unavailability at any time *t*, which is the probability that the system is unavailable at time *t* due to a failure or due to an ongoing repair after the detection of a failure. System may be composed of highly reliable components. In [5] we developed new procedure for exact reliability quantification of a highly reliable system. The procedure eliminates all errors made by a computing hardware system when calculations close to error limit are executed. But the procedure works only under assumption of exponential distribution of the time to failure of terminal nodes, which results in big restriction for practical use of the algorithm.

B. Unavailability models of terminal nodes

Most of component models (i.e. models of terminal nodes) including maintenance, both preventive and corrective, can be described by the following four models:

Model I with components that cannot be repaired. Final time dependent unavailability coefficient $U(t)$ for this simplest model is given by the distribution function of the time to failure of the component:

$$U(t) = F(t) = \int_0^t f(x)dx, \qquad (1)$$

where $F(t)$, $f(t)$ is distribution function and probability density function (*pdf*) of the time to failure.

Model II with repairable components (CM – Corrective Maintenance) for apparent failures, i.e. a model when a possible failure is identified at its occurrence and immediately afterwards it starts a process leading to its restoration. In Model II, two random variables are immediately connected, i.e. the time to failure X, characterized by distribution function $F(t)$ and density $f(t)$, and the repair (or restoration) time Y, characterized by distribution function $G(t)$ and density $g(t)$. In this model we can apply well known relations from renewal theory and alternating renewal processes. In [6] we derived and proofed the following Theorem for time dependent unavailability coefficient $U(t)$:

$$U(t) = \int_0^t f(x)\cdot[1-G(t-x)]dx + \int_0^t (f*g)(x)\cdot U(t-x)dx \quad (2)$$

where * means convolution. This Theorem can be considered as a recurrent linear integral equation which helps us to compute $U(t)$.

Model III with repairable components with hidden failures, i.e. a model when a failure is identified only at special deterministically assigned times, appearing with a given period (moments of periodical inspections). In the case of its occurrence at these times an analogical restoration process starts, as in the previous case. Inspections are carried out periodically. Let us denote inspection time points as k_1, k_2, k_3…, then $k_{i+1} - k_i = T_p$ is period of inspections. In [6] we derived numerical formula to calculate the time dependent unavailability coefficient $U(t)$:

$$U(t) = \int_{k_n}^t f(x)dx + \sum_{i=1}^n S_i \cdot \left\{ \int_{k_n-k_i}^{t-k_i} (g*f)(x)dx + [1-G(t-k_i)] \right\} \quad (3)$$

where S_i is the probability that in the inspection time k_i a renew was realized. Numerical formulas to find probabilities S_i are as well derived in [6].

C. Unavailability coefficient for Model IV

Model IV with preventive maintenance (PM) policy in which the component is restored (either repaired or renewed) as soon as it reaches the age which is equivalent to the length of period T_P. The model is a generalization of the previous Model III with periodically inspected and repaired (in case of failure) components. Difference is in the possibility of renew the component (as good as new) in each inspection time, even if it is in a functional state. Component is restored (either repaired or renewed) in each inspection time what results in prolongation of the time to failure. This model well respects in practice commonly used maintenance process applied to such components that are hard to access. In addition the model is useful to study ageing processes being underway in components with increasing failure rate.

Maintenance policies can be classified into two main categories: calendar-based and time-based policies. For maintained components of Model III and IV we suppose calendar-based policy. The calendar-based maintenance policy performs testing and maintenance at fixed times $k_1=T_F$ (T_F is first inspection time, which is considered as optional parameter in this paper), $k_2=T_F+T_P$, $k_3=T_F+2T_P$…, such as regular schedule of plant maintenance outages, for example. Consequently the asymptotic unavailability, which can be defined as the point-wise limit of unavailability function for infinite time, is a periodical function with period T_P.

Our calendar-based maintenance policy model is close to the traditional block-based maintenance policy in [10], however further distinction can be made between both models. The block-based maintenance policy performs preventive maintenance at fixed times k_1, k_2, k_3, k_4..., as well as calendar-based policy. This schedule is not updated when a failure occurs. The disadvantage of block-based maintenance is that preventive maintenance is sometimes performed shortly after a failure. However, it has the advantage that it allows for easier planning as it is exactly known in advance when preventive maintenance will be performed.

Furthermore, when multi-unit systems are considered, it allows for clustering preventive maintenance actions by choosing the same maintenance interval for each unit (component). Advantages of the block-based maintenance policy are fully-used in our calendar-based policy, however distinction between block-based policy and Models III and IV is made in that restoration process of Models III and IV is allowed, possibly launched, only at fixed times k_1, k_2, k_3, k_4,.... These models describe latent failures whereas traditional block-based policy is based on monitored failures.

Even though the algorithm for Model IV works without limitations, for better reading we take into account a realistic assumption for component of Model IV:

$$\int_0^{T_P} g(x)\cdot dx = 1 \quad (4)$$

i.e. restoration process must be finished within one inspection period. To find the course of $U(t)$, two following steps must be executed:

Step 1: to calculate probabilities of functioning states in all inspection times: PF_1, PF_2, …, as well as probabilities of non-functioning states (component under restoration) PR_1, PR_2, …

Apparently for the first inspection time k_1, we obtain:

$$\begin{aligned} PF_1 &= 1-F(k_1) \\ PR_1 &= F(k_1) \end{aligned} \quad (5)$$

Probabilities in other inspection times can be obtained, if probabilities of all possible state evolutions are summarized. For this reason we denote these evolutions and its probabilities A, B, C and D as follows, see Figure 2. In all time evolutions we differentiate between two functioning levels: upper green line means functioning state, lower red line means non-functioning state, where we further differentiate between solid (component under restoration) and dash (component waiting for restoration) line.

Consequently for PF_{i+1} and PR_{i+1} the following recurrent formulas can be obtained, see Figure 2:

$$\begin{aligned} PF_{i+1} &= A\cdot PF_i + C\cdot PR_i \\ PR_{i+1} &= B\cdot PF_i + D\cdot PR_i \qquad i=1,2,... \end{aligned} \quad (20)$$

$$A = 1 - F(T_P)$$
$$B = F(T_P)$$
$$C = \int_0^{T_P} g(x) \cdot [1 - F(T_P - x)] dx \qquad (21)$$
$$D = \int_0^{T_P} g(x) \cdot F(T_P - x) dx$$

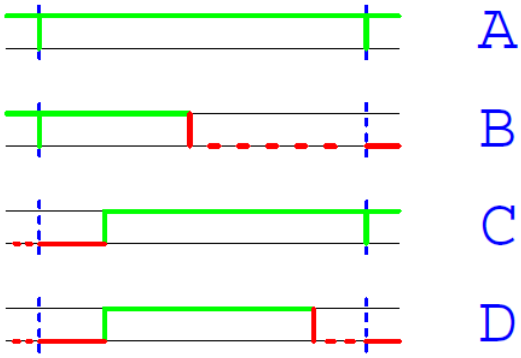


Figure 2. All possible state transitions between two consecutive inspection times

Step 2: Probabilities PF_1, PF_2, PF_3…, PR_1, PR_2, PR_3…, are thereafter used to calculate required point unavailability function $U(t)$:

For $t \in (0, k_1\rangle$

$$\Rightarrow U(t) = F(t) \qquad (22)$$

For $t \in (k_j, k_{j+1}\rangle$ $\quad j = 1,2,3...$, see Figure 3:

$$\Rightarrow U(t) =$$
$$PF_j \cdot F(t - k_j) + PR_j \cdot \int_0^{t-k_j} g(x) \cdot F(t - k_j - x) dx + \qquad (23)$$
$$PR_j \cdot [1 - G(t - k_j)]$$

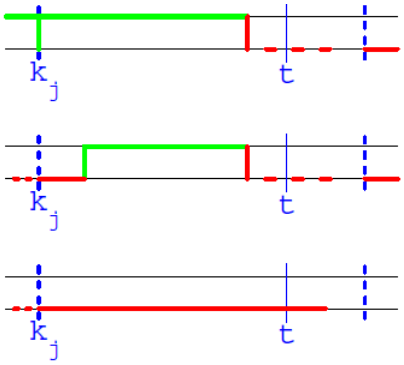


Figure 3. Three possible evolutions to calculate U(t) within (kj, kj+1>.

So that unavailability $U(t)$ is calculated as a sum of probabilities of three possible and exclusive evolutions.

III. Results with Tested Complex System

Different unavailability models of terminal nodes are demonstrated on real topical complex power distribution network, see Figure 4. It is a part of a real local distribution system (LDS) situated within the premises of the company focused on production and processing of pig iron as well as the metallurgical and machine engineering production. Technological diagram of selected network supplying very important manufacturing plant DS3 6 kV B (6 kV distribution substation) is in Figure 4. Such system is frequently used for ensuring of high reliable power supply in the Czech Republic.

This LDS is supplied from a 110 kV substation (DS1 110 kV) that is taken as a point with ideal reliability. In the selected network there are 22 kV substation (DS1 22 kV) fed from that 110 kV through transformer 110/22 kV (T1) and 6 kV substation (DS1 6 kV) fed from that 22 kV through transformer 22/6 kV (T5). To improve supply reliability of the manufacturing plant there are three parallel cables (C3 - C5) from the 6 kV substation. LDS operator must guarantee the quality of continuity of supply. That is why there are several possibilities of cold backups. Switch-over-to backup is simplified by switching time of circuit breakers (t_s 1 - t_s 5). As the switching time is very short, we consider switch-over as uninterrupted power supply and do not use it in further calculations.

All network consists of seven transformers (T), 7.35 km of cables (C), and 0.8 km of overhead power lines (L). Their reliability parameters are shown in Table 1 that contains additional data including periods of preventive maintenance T_P, as well as type of component model. We consider that lifetimes of all components follow Weibull distributions. Mean component lifetimes (denoted as MTBF) are further presented in Table 1 and the same shape factor β= 2 is adopted for all components. All these values are based on our experience coming from own database. The time to end of repair of each component is supposed to be exponential, characterized by mean value (denoted here as MTTR). However any applicable distribution is used for repair time we presume, that results will be identical, because system is very reliable, thus its failure followed by repair is a rare event.

Thus correct function of the system is defined so that the consumption point has to be supplied by the power. In highly reliable applications, when energy supply is crucial for safety reason, the cables C3 - C5 can be intended as one-out-of-three active parallel system, including three identical branches, as demonstrated in Figure 4. However in this specific network all of 3 cables have to be in functional operation mode. That is why operator uses so many backups.

Correct function of the considered distribution network from Figure 4 can be demonstrated by means of a directed acyclic graph, see Figure 1. The highest node u1 represents functionality of the whole system (success, failure), internal and terminal nodes represent subsystems and components. Main objective now is to estimate the instantaneous unavailability at any time *t*, which is the probability that the

system is unavailable at time *t* due to a failure or due to an ongoing repair after the detection of a failure.

Basic unavailability computation of the power distribution network was executed for original data from Table 1. The instantaneous unavailability is demonstrated in Figure 5 within the mission time of 5 years. In first view we see that unavailability is significantly influenced by PM, which ordinarily starts at the first inspection time T_F, which is identical for all components of the same type. For example, first inspection times $T_F(i)$ of all transformers T1-T4 are placed at the time of 7,885 hours. If these PM starting times are consecutively uniformly shifted within the length of T_P, i.e. PM of transformer T1 (T2; T3; T4) starts at the time T_F = 1,971 (3,942; 5,914; 7,885) hours, subsequently we obtain much more optimistic results at the same PM cost, compare unavailability courses in Figure 6, where maximal unavailability U_{MAX} = 6.5e-4 (compare with 5.1e-3 for original data). We will call this PM configuration with shifted times $T_F(i)$ as Base. It can be recommended as a first and basic improvement of the network before starting some further improvements. In addition it corresponds to operational practices. All maintenance activities are distributed uniformly during yearly time for the reason of optimal utilization of resources (workforce and technical equipment).

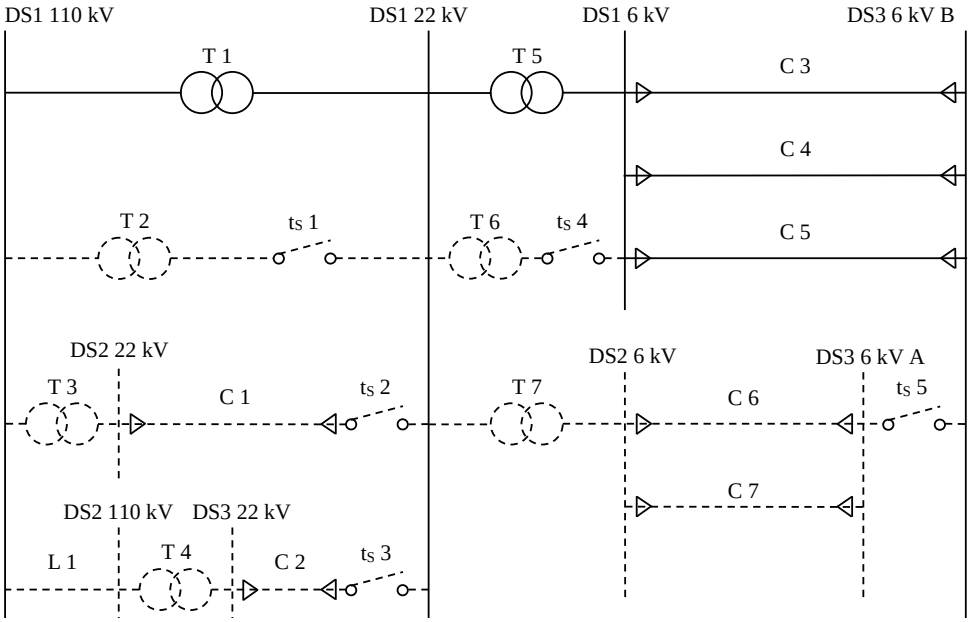


Figure 4. Real topical power distribution network – technological diagram

TABLE I. RELIABILITY PARAMETERS AND PM PERIODS OF THE SELECTED NETWORK

Component	MTBF (year)	β	MTTR (h)	Tp (h)	Component Model
Transformer T1	3.00	2	4.4	7,885	Model III
Transformer T2	3.00	2	4.4	7,885	Model III
Transformer T3	3.00	2	4.4	7,885	Model III
Transformer T4	3.00	2	4.4	7,885	Model III
Transformer T5	7.69	2	0.48	9,031	Model III
Transformer T6	7.69	2	0.48	9,031	Model III
Transformer T7	7.69	2	0.48	9,031	Model III
Cable C1	31.35	2	11.42	41,914	Model II
Cable C2	78.37	2	11.42	41,914	Model II
Cable C3	30.05	2	8.1	30,845	Model II
Cable C4	30.05	2	8.1	30,845	Model II
Cable C5	30.05	2	8.1	30,845	Model II
Cable C6	28.41	2	8.1	30,845	Model II
Cable C7	28.41	2	8.1	30,845	Model II
Power line L1	33.78	2	0.7	47,097	Model III

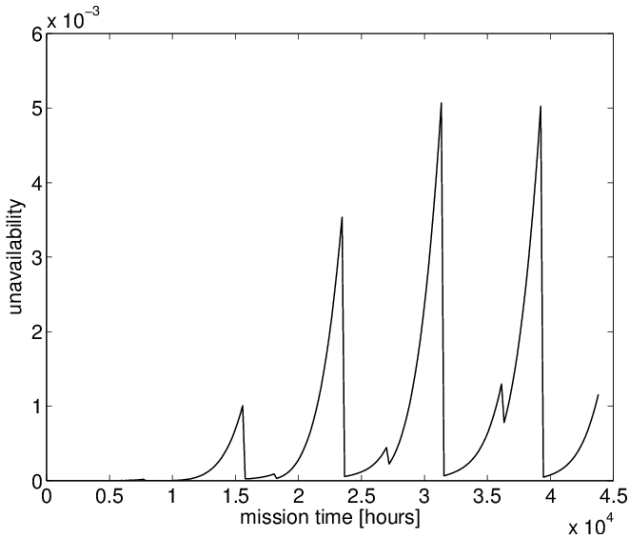


Figure 5. Instantaneous unavailability U(t) of the power distribution network. Original data.

Recently we frequently face the problem of penalization when system reliability indices are overstepped. One of the most important indices is SAIDI (System Average Interruption Duration Index), the average duration of interruptions for customers who experience an interruption during the year. Distribution companies make intensive effort to decrease the index even if the maintenance cost is increased. As a result of the effort next system configuration was computed where maintained components (all transformers) are modelled as Model IV, i.e. with PM policy in which the component is restored. However maintenance cost of this in-depth PM process is much more expensive.

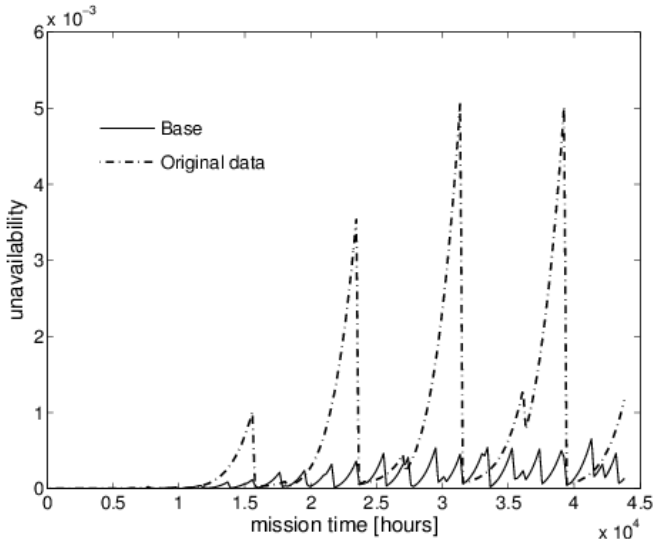


Figure 6. Unavailability U(t) of the power distribution network. Original data versus Base configuration with shifted first inspection times.

Comparing the configuration (marked here as Model IV) with Base (i.e. transformers as Model III) in Figure 7 one can conclude that maximal unavailability significantly decreased, in fact below the value 1e-6 (compare with Base U_{MAX} = 6.5e-4). However SAIDI of planned outages may be affected significantly so that final decision is up to responsible management team of the system.

Responsible managers are sensitive to the fact that maintenance is processed unnecessarily frequently without any significant effect. Consequently the maintenance cost is very expensive. That is why new maintenance strategy, saving the cost, could be found by doubling of PM periods. In Figure 8 we see that although the maximal unavailability U_{MAX} = 6.1e-3 is of about one order greater than Base, it is comparable with original data. In addition, cost of the configuration with doubling periods significantly dropped.

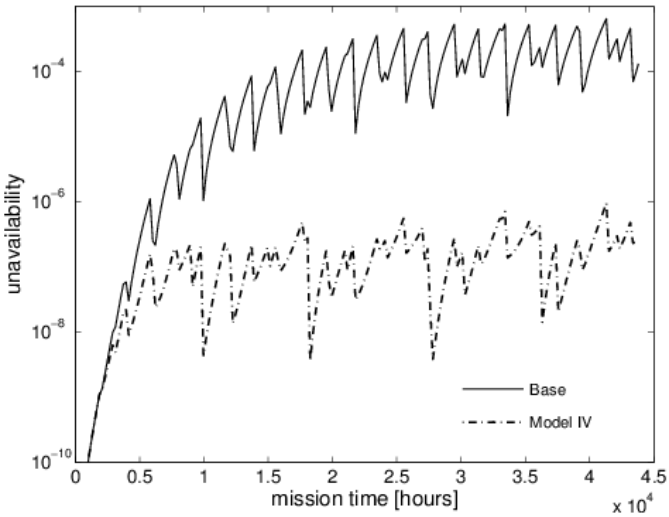


Figure 7. Instantaneous unavailability U(t) of the network. Transformers as Model IV versus Base (Model III).

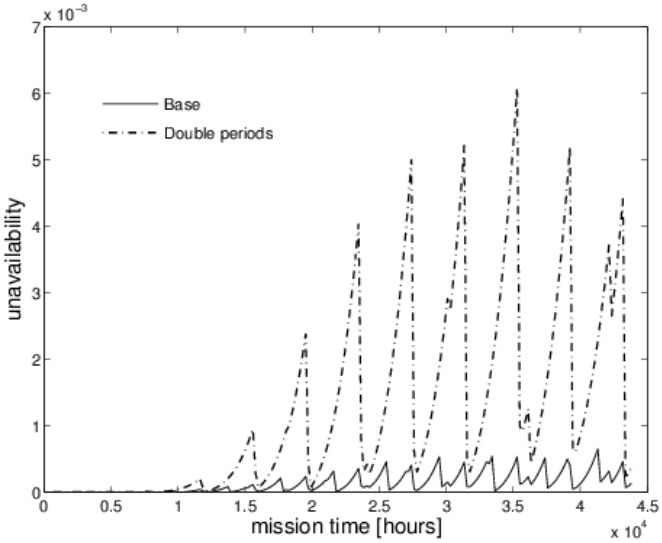


Figure 8. Unavailability *U*(t) of the power distribution network. Double periods versus Base.

IV. Conclusion

The innovative computing methodology enables exact unavailability quantification of multi-component systems with highly reliable and diverse components (Model I-IV) without any restrictions on the form of the probability distribution assigned to the time to failure or repair duration. It assumes that the system structure is mathematically represented by means of directed acyclic graph.

Models II-IV represent all frequently used component models with both preventive and corrective maintenance. These models were solved both analytically and numerically, by means of stochastic alternating renewal process without any restrictions on the form of the probability distribution assigned to time to failure or repair duration. Innovative formula, here called as a recurrent linear integral equation, for unavailability quantification of a repairable component has been derived and proved in previous research. It does not contain renewal density, which results in much easier numerical computation of unavailability $U(t)$ in comparison with traditional approach, where renewal equation must be used. In this paper close attention was paid to Model IV, with PM policy in which the component is restored as soon as it reaches the age which is equivalent to the length of period T_P. The model is a generalization of the previous Model III with periodically inspected and repaired components.

The computing process was demonstrated on the real complex highly reliable power distribution network. We clearly illustrated on the system, that different decision making situations from practice concerning maintenance or ageing can be solved by means of the innovative computing process. The computing methodology has been numerically realized within the high-performance language MATLAB. All computations above run on Intel (R) Core™ i7-3770 CPU @ 3.4GHz 3.9 GHz, 8.00 GB RAM.

Because the computing methodology admits to take into account also other lifetime distributions, in following research we plan to use other ageing models to be compared with the Weibull-one. In addition, innovative algorithms can be used for maintenance optimization what is intended by author of the paper as a natural continuation of this research. Basic intention is to create a new algorithm for realistic maintenance optimization which incorporates the innovative computing methodology as a main tool for unavailability calculations.

Acknowledgment

This work was supported by the internal grant agency of VŠB - Technical University of Ostrava, Czech Republic, under the project no. SP2017/56 ”Reliability and Risk Modeling”.

References

[1] R. Nicolai, R. Dekker, “Optimal maintenance of multi-component systems: a review,“ Springer Ser Reliab Eng Part D 2008:263–86.

[2] H. Wang, “A survey of maintenance policies of deteriorating systems,” Eur J Oper Res 2002;139:469–89.

[3] R. Dekker, R. Wildeman, Duyn Schouten Fvd, “A review of multi-component maintenance models with economic dependence,” Math Methods Oper Res 1997;45(3):411–35.

[4] D. Cho, M. Parlar, “A survey of maintenance models for multi-unit systems, “Eur J Oper Res 1991;51:1–23.

[5] R. Bris, “Exact reliability quantification of highly reliable systems with maintenance,” Reliability Engineering and System Safety 2010; 95:1286–1292.

[6] R. Briš and P. Byczanski, “Advanced computing methodology for general highly reliable systems,” Risk, Reliability and Safety: Innovating Theory and Practice – Walls, Revie & Bedford (Eds), © 2017 Taylor & Francis Group, London, p. 1466-1473, ISBN 978-1-138-02997-2.

[7] C. Simon, P. Weber and A. Evsukoff, “Bayesian networks inference algorithm to implement Dempster Shafer theory in reliability analysis,” Reliability Engineering and System Safety 2008; 93:950–963.

[8] R. Briš, “Parallel simulation algorithm for maintenance optimization based on directed Acyclic Graph,” Reliab Eng Syst Saf 2008;93:852-62.

[9] R. Briš, “Stochastic Ageing Models – Extensions of the Classic Renewal Theory,” Reliability: Theory & Applications, ISSN 1932-2321, 2007;2(3-4):19-27.

[10] J.A.M. Weide and M.D. Pandey, “A stochastic alternating renewal process model for unavailability analysis of standby safety equipment”, Reliability Engineering and System Safety 2015; 139:97–104.

The Effect of Current Delay Angle on Tripping of Residual Current Devices

Stanislaw Czapp, Jacek Horiszny
Faculty of Electrical and Control Engineering
Gdansk University of Technology
Gdansk, Poland
stanislaw.czapp@pg.gda.pl

***Abstract*—Power electronics converters applied in domestic or similar installations may utilize current delay (phase) angle control to change the level of transferred power. Due to application of such types of converters, earth fault current in the installation may be strongly distorted. The current distortion level depends on a value of current delay angle. This delay angle also influences the tripping threshold of residual current devices. In this paper, computer modelling of the impact of current delay angle on behaviour of residual current devices is conducted. Results of laboratory tests of the tripping of residual current devices are discussed. Conclusions regarding electrical safety are presented as well.**

Keywords—computer modelling; power electronics converters; residual current devices; tripping threshold

I. INTRODUCTION

Application of high energy efficiency equipment in electrical installations is one of the most important issues which are considered during selection of electric devices to low voltage systems. In order to save energy, electronic controllers/converters are widely used in industrial and domestic installations (e.g. lighting controllers, adjustable speed drives, etc.). One type of electronic converters is a controller which regulates power consumption by a delay (phase) angle control of a load current. Fig. 1 presents two structures of such type of controller. This figure also presents load current waveforms and earth fault (residual) current waveforms which may occur in the controlled circuit.

The shape of earth fault (residual) current influences operation of residual current protective devices. In low voltage systems, the most popular and widely used are AC-type and A-type residual current devices. The first type is provided for detection of sinusoidal currents of the power network frequency. The second one may also detect pulsating direct currents [1]. In practice, operational characteristics of both above mentioned types of residual current devices are not verified for residual currents with symmetrical phase control (i_Δ waveform in Fig. 1a). In the published works, the effect of the transients [2]–[5], high frequency [6], harmonics [7]–[10] and DC waveforms [11], [12] of residual current on electrical safety, especially tripping of residual current devices is considered. Special aspects of safety and operation of residual current devices in systems with generators are developed in [13], [14]. Some of these papers indicate that waveforms of residual current other than sinusoidal 50 Hz may have negative effect on tripping of residual current devices and the effectiveness of protection against electric shock may not be achieved.

This paper presents results of computer modelling and experimental tests of residual current devices behaviour when symmetrical residual current with delay (phase) angle control is forced.

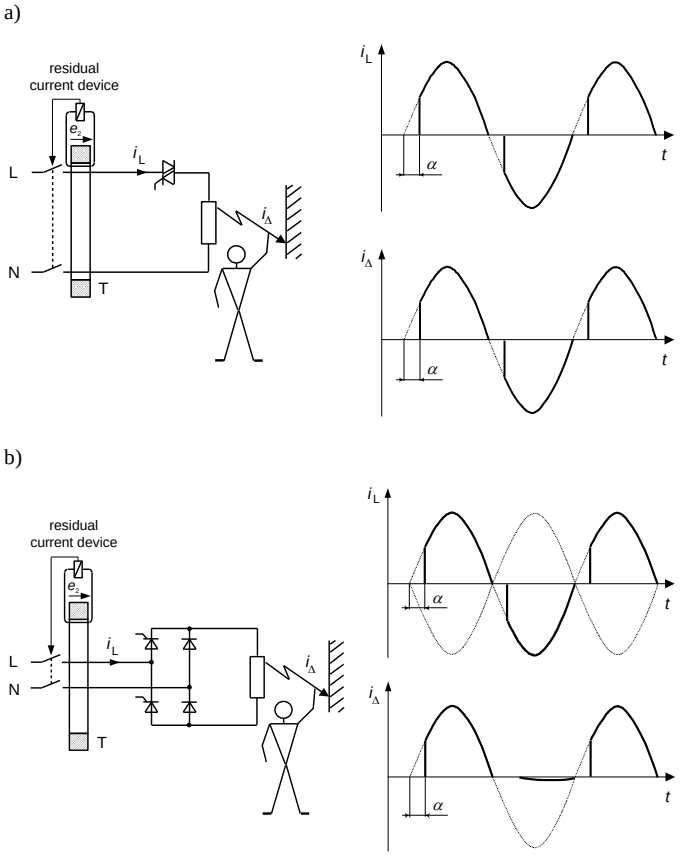


Figure 1. Simplified structure of converters for delay (phase) angle α control and characteristic waveforms: T – current transformer of residual current device, i_L – load current, i_Δ – earth fault (residual) current, e_2 – induced secondary voltage of transformer T. Converter producing earth fault current: a) bidirectional, b) practically unidirectional.

II. MODEL OF A RESIDUAL CURRENT DEVICE

A power electronic controller, which is able to control delay angle of a load current (residual current as well), and an equivalent circuit of a residual current device have been simulated with the use of ATP software (Fig. 2). Earth fault current i_Δ is a primary current i_1 of the current transformer T. The primary current i_1 is transformed to the secondary side of transformer T, and a secondary current i_2 then flows through the electromechanical relay K. If the value of the secondary current i_2 is relatively high (especially peak value), tripping of the electromechanical relay K occurs, and in consequence, the residual current device opens the main circuit.

The current transformer T has been simulated by means of the ATP library model of a transformer. It is formed by elements: L_1, L_2 – leakage inductances of windings: primary, secondary, respectively; R_1, R_2 – resistances of windings: primary, secondary, respectively; R_{Fe} – resistance representing excitation losses; T_i – ideal transformer for voltage and current transformation. Magnetizing branch has been removed from the model because it gives no possibility to determine the non-linear magnetizing characteristics with hysteresis. In return, the model is supplemented by an additional external element L_m, which is the inductance with the magnetic hysteresis. A simplified magnetic characteristic has been adopted for this element, as shown in Fig. 2c. This characteristic has been prepared on the base of laboratory test of an example current transformer core installed in a residual current device. The simplification of the measured characteristic has been done in order to achieve numerical stability of the solution. The elements R_{Th}, R_d and C_d connected in parallel with the switch S are introduced to the model for the same reason. Switch S simulates triac Th. The triggering signal for this element is generated in the generator G, which is represented by the source of square-wave signal with pulse position modulation. Parameters of the power supply (source voltage V_s and resistance R_s) have been chosen to force desired value of primary current i_1. The relay K is modelled by resistance R_K and inductance L_K.

A simulated response of the current transformer to the primary current with various current delay angles has been checked with the use of the above presented computer model of the residual current device. Fig. 3 presents waveforms of primary current i_1 and waveforms of induced secondary voltage e_2 for selected values of current delay angle α. Due to relatively wide hysteresis loop and reaching the saturation level, the induced secondary voltage is non-sinusoidal, even for sinusoidal primary current (Fig. 3a). The higher value of current delay angle α, the more narrow is the impulse in the secondary voltage waveform e_2. However, the peak value of the voltage e_2 rises then. A laboratory test of the selected current transformer of a residual current device indicates (Fig. 4) a similar dependency (as in the simulation).

Analysing the simulation and laboratory test results (Fig. 3 and Fig. 4), it may be concluded that the rising of peak value of the induced voltage e_2 with increasing the current delay angle α (for constant peak value of the primary current i_1), may cause lowering the tripping threshold of the residual current device. However, on the other hand, for relatively high value of the angle α, voltage impulses in the secondary voltage waveform e_2 (in consequence in the secondary current i_2 as well) may be too short/narrow to ensure effective reaction of the electromechanical relay K. In practice, these two mentioned factors may compensate each other.

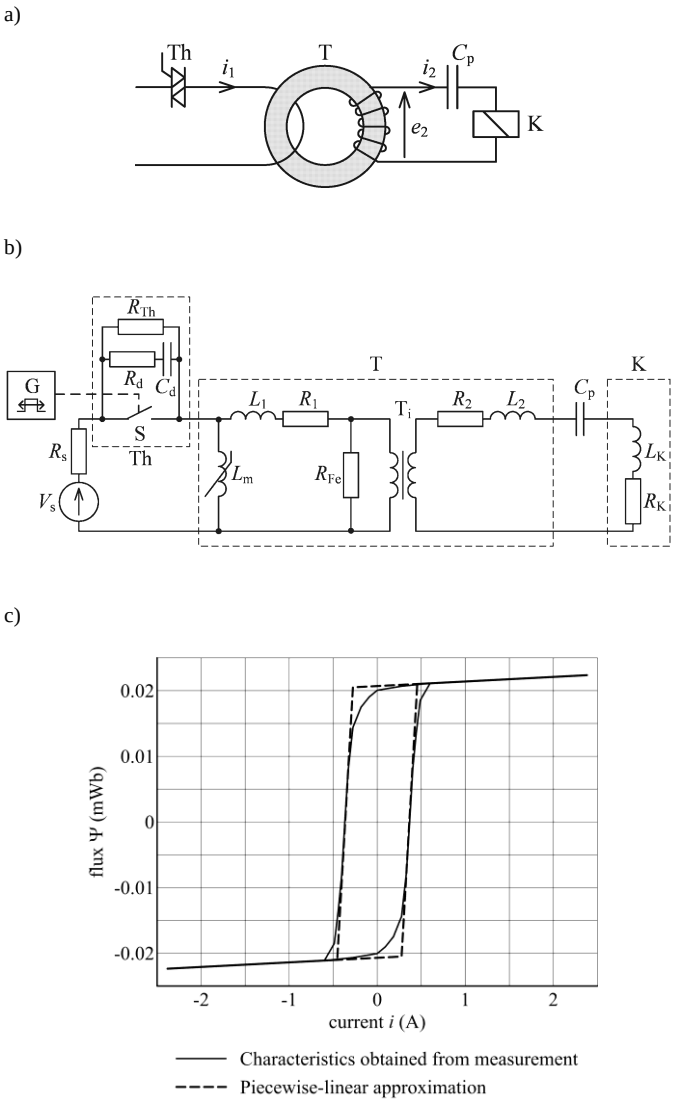


Figure 2. Model of the residual current device circuit (a), equivalent circuit performed in ATP software (b) and magnetizing characteritics of the iron core of the current transformer (c);
i_1 – primary current,
i_2 – secondary current,
e_2 – induced secondary voltage,
Th – triac for current delay/phase angle control,
T – current transformer,
Ti –ideal transformer model,
C_p – matching capacitor,
K – electromechanical relay,
C_d, R_s, R_{Th}, R_d, V_s, G, S – group od elements for ATP modelling of the controlled primary current,
L_m – inductance with magnetic histeresis,
L_1, R_1 – parameters of the transformer T primary winding,
L_2, R_2 – parametrers of the transformer T secondary winding,
R_{Fe} – resistance representing exitation losses,
L_K, R_K – parametrers of the relay K.

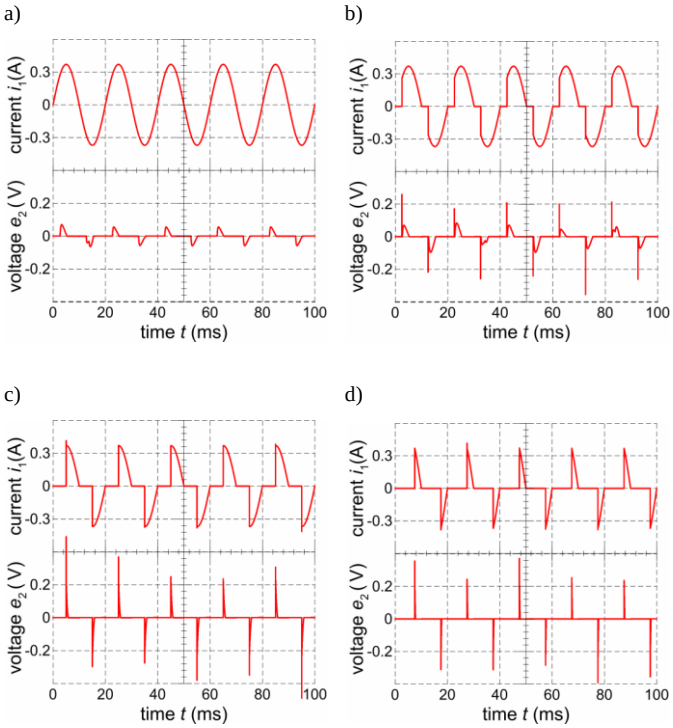


Figure 3. Computer analysis of the response of the modelled residual current transformer T (rated residual current: $I_{\Delta n}$ = 300 mA) to currents with controlled delay angle; i_1 – primary (residual) current, e_2 – induced secondary voltage. Delay angle α: a) 0°, b) 45°, c) 90°, d) 135°.

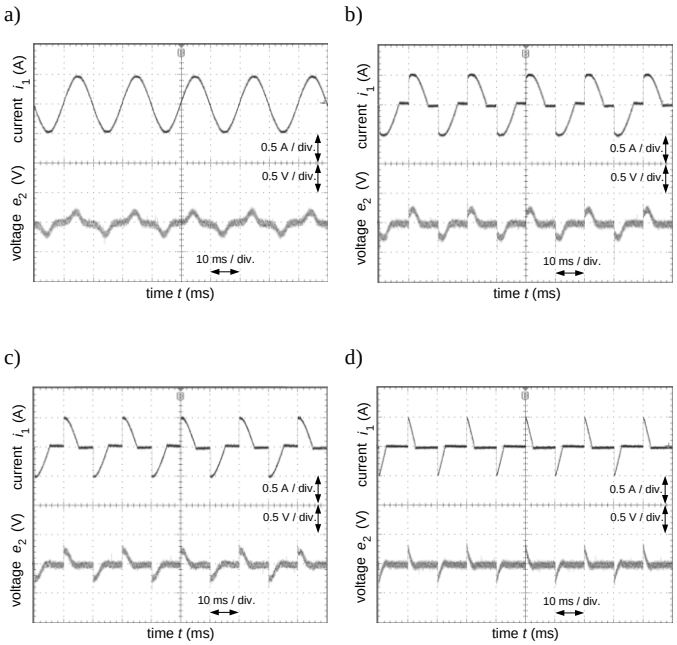


Figure 4. Laboratory test results of the response of the real residual current transformer (rated residual current: $I_{\Delta n}$ = 300 mA) to currents with controlled delay angle; i_1 – primary (residual) current, e_2 – induced secondary voltage. Delay angle α: a) 0°, b) 45°, c) 90°, d) 135°.

III. Experimental Verification of the Computer Modelling – Real Tripping Threshold

Sensitivity of residual current devices has been checked according to a diagram presented in Fig. 5. The test bench was supplied from laboratory 50 Hz AC supply. Only one pole of a residual current device was connected to the supply, in order to force primary (residual) current i_1 (i_Δ). Current delay angle α was adjusted in the electronic phase controller and value of current was increased by the variable resistance R. Similar to the theoretical analysis, the following four values of current delay angle were considered: 0º, 45º, 90º, 135º. Tripping threshold of residual current devices (both rms value "Irms" and peak value "Ipeak" of the testing current) was recorded by a digital oscilloscope. The residual current devices of the rated residual current equal to $I_{\Delta n}$ = 100, 300 and 500 mA, AC-type, A-type, no time delayed, short-time delayed (10 ms) and time-delayed (40 ms) [1] were tested. The tested devices were marked consecutively from RCD1 till RCD37.

For AC sinusoidal waveform, the tripping threshold of residual current devices should be within the range of $(0.5 \div 1.0)I_{\Delta n}$, where $0.5I_{\Delta n}$ is the non-operating current. In the test the same criterion range has been assumed. Tripping threshold (rms value) should not be $\leq 0.5I_{\Delta n}$ and $>1.0I_{\Delta n}$.

Fig. 6 presents results of the test of $I_{\Delta n}$ = 100 mA protective devices. For no time-delayed (RCD14) and short-time delayed (RCD13) residual current devices, rms value of tripping current "Irms" decreases when current delay angle increases. In case of RCD14, when current delay angle is equal to 135º, the tripping current is below the permissible limit – 50 mA ($0.5I_{\Delta n}$). What is important and advantageous is the behaviour of these residual current devices for strongly distorted residual current – due to high value of delay angle, does not mean too high tripping current and ineffective protection against electric shock like reported in [7]–[10]. On the contrary, residual current devices may be triggered by relatively low residual current (lower than $0.5I_{\Delta n}$) and it is called the nuisance tripping. A different behaviour is noticed for time-delayed RCD29 (Fig. 6c) – its rms tripping current rises with current delay angle increase.

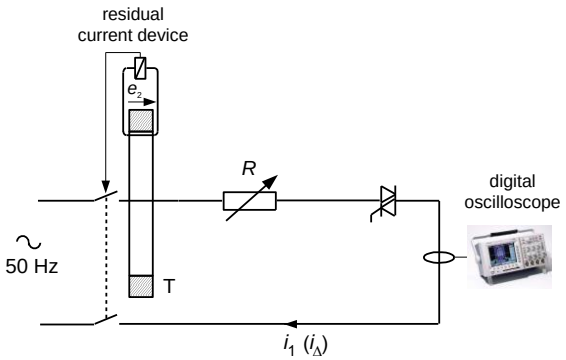


Figure 5. Diagram of the circuit for testing of sensitivity of residual current devices to the current delay angle. T – current transformer of residual current device, R – variable resistance, i_Δ – testing residual current.

A structure of the tripping circuit (secondary circuit of the transformer T) of time-delayed residual current devices is more complicated than in the other types of the devices (it contains additional electronics matching/resonant elements like diodes and capacitors in order to achieve delay in tripping), and this is the cause of the different behaviour mentioned.

It should be noted that for current delay angle α equal to 135º, the peak value of the tripping current "Ipeak" can be much higher than for $\alpha = 0$º. According to the information presented in the previous section, for triggering of the tripping mechanism of residual current devices, peak value of the current is responsible. A current impulse with high value of delay angle (90º, 135º) is very short/narrow, and to stimulate this mechanism, a higher value of peak current is necessary.

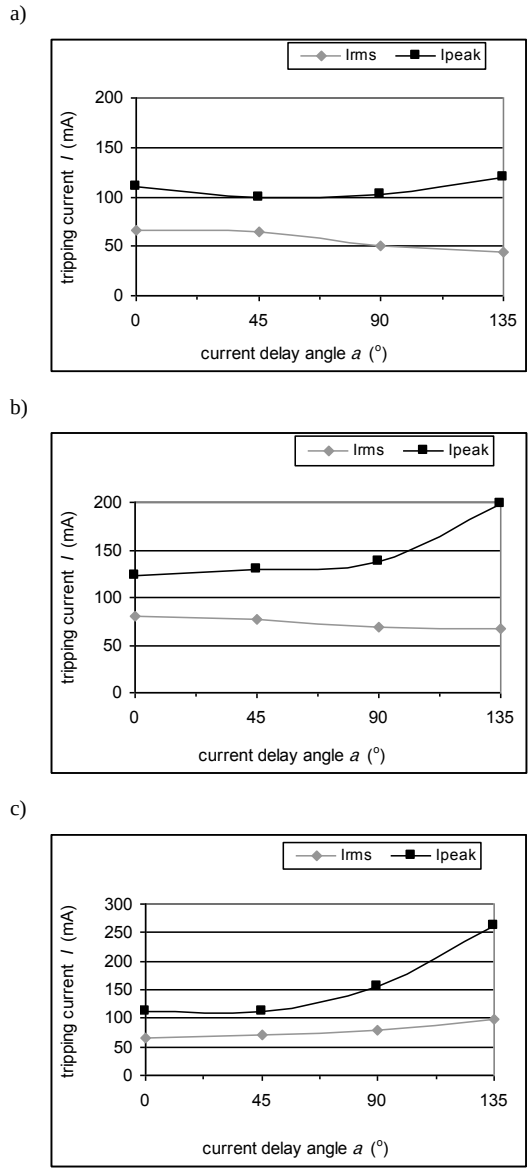


Figure 6. Tripping current I of the selected RCDs, $I_{\Delta n}$ = 100 mA: a) RCD14, AC-type, no time-delayed, b) RCD13, AC-type, short time-delayed, c) RCD29, A-type, time-delayed.

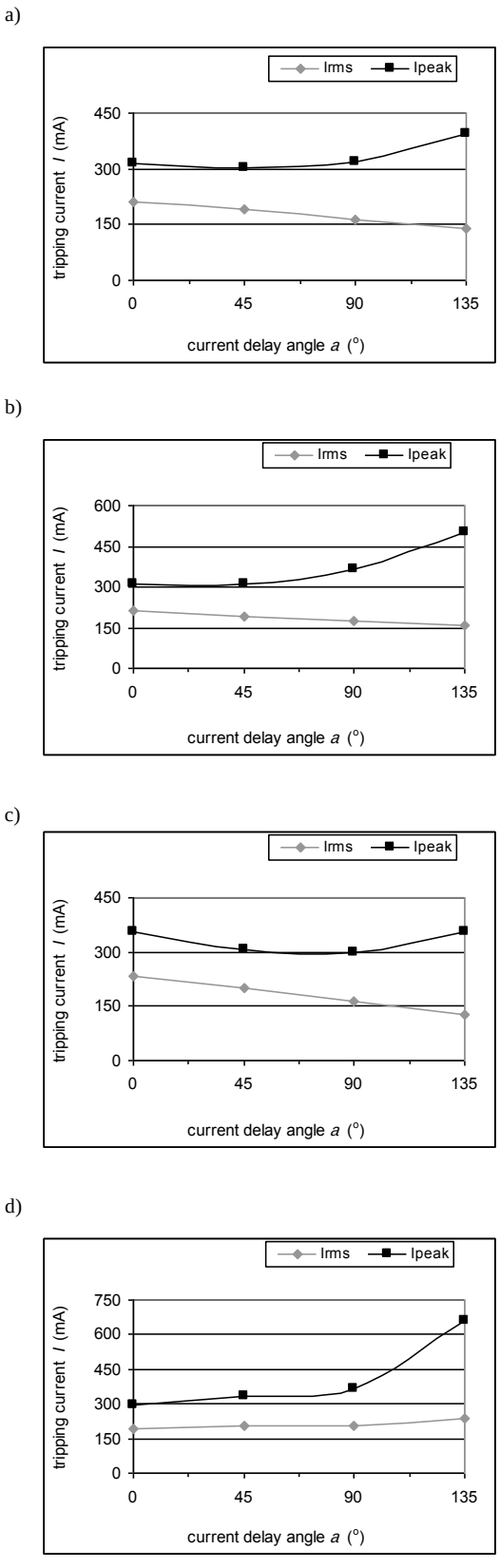
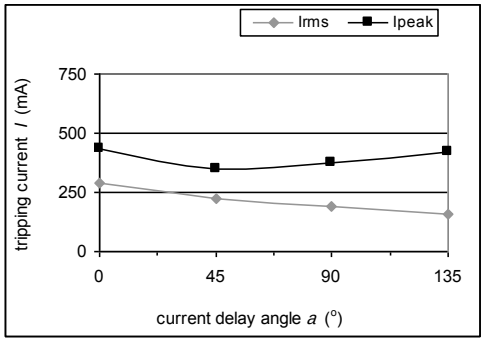


Figure 7. Tripping current I of the selected RCDs, $I_{\Delta n}$ = 300 mA: a) RCD18, AC-type, no time-delayed, b) RCD21, AC-type, time-delayed, c) RCD36, A-type, no time-delayed, d) RCD34, A-type, time-delayed.

a)



b)

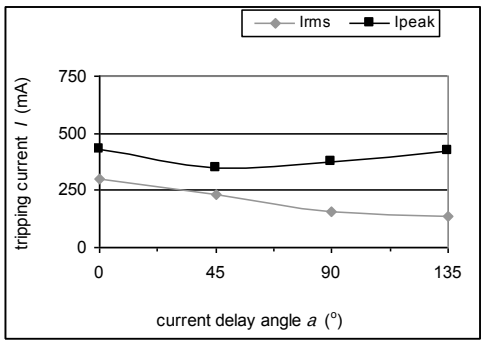


Figure 8. Tripping current I of the selected RCDs, $I_{\Delta n}$ = 500 mA: a) RCD22, AC-type, no time-delayed, b) RCD37, A-type, no time-delayed.

Similar conclusions flow from the analysis of tripping currents presented in Fig. 7 (RCDs of $I_{\Delta n}$ = 300 mA) and Fig. 8 (RCDs of $I_{\Delta n}$ = 500 mA). However, some differences in operation of time-delayed residual current devices are noticed comparing Fig. 7b (RCD21) to Fig. 7d (RCD34). The time-delayed A-type residual current device RCD34 (Fig. 7d) has similar tripping characteristic to the prior mentioned (Fig. 6c) – the rms value rises with current delay angle increasing. The rms value of tripping current of AC-type residual current device RCD21 (Fig. 7b), in turn, decreases with current delay angle increasing (like in no time-delayed and short time-delayed RCDs). The most probable cause of such difference in tripping characteristics lies in different structures of the tripping circuits of the considered devices.

IV. Conclusions

The current delay angle of residual current influences tripping threshold of residual current devices. Fortunately, high value of the current delay angle does not mean ineffective protection against electric shock but may result in nuisance tripping of residual current devices, because the tripping threshold (rms) can be lower than $0.5I_{\Delta n}$. In contrary to the other types of non-sinusoidal residual currents, a non-sinusoidal current characterized by high value of the delay angle is detected by the simplest residual current devices – AC-type residual current devices. Taking into account types of residual current devices in regard to intended undelayed/delayed operation, tripping characteristics as a function of delay angle of two or more residual current devices with the same rated residual current, can be different, but not dangerous in terms of electrical safety.

References

[1] IEC 60755:2008 General requirements for residual current operated protective devices.

[2] S. Czapp and K. Borowski, "Immunity of residual current devices to the impulse leakage current in circuits with variable speed drives," Elektronika ir Elektrotechnika, vol. 19, no. 8, pp. 15–18, 2013, DOI: http://dx.doi.org/10.5755/j01.eee.19.8.2883.

[3] G. Escriva-Escriva, C. Roldan-Porta, and E. C. W. de Jong, "Nuisance tripping of residual current circuit breakers in circuits supplying electronic loads," Electric Power Systems Research, vol. 131, pp. 139–146, 2016, DOI: 10.1016/j.epsr.2015.10.012.

[4] J. Lee, S. Chang, S. Myung, and Y. Cho, "Transient false tripping characteristic analysis of ground fault circuit interrupter," Int. Conference on Power System Technology (PowerCon 2004), 2004, DOI: 10.1109/ICPST.2004.1460050.

[5] A. Mohd Zaki and A. Rusnani, "Power quality analysis of Residual Current Device [RCD] nuisance tripping at commercial buildings," Symposium on Industrial Electronics and Applications (ISIEA), 2013, pp. 122–125, DOI: 10.1109/ISIEA.2013.6738980.

[6] F. Freschi, "High-frequency behavior of residual current devices," IEEE Trans. on Power Deliv., vol. 27, no. 3, pp. 1629–1635, 2012, DOI: 10.1109/TPWRD.2012.2191423.

[7] S. Czapp, "Comparison of residual current devices tripping characteristics for selected residual current waveforms," Elektronika ir Elektrotechnika, no. 4 (100), pp. 7–10, 2010, http://www.eejournal.ktu.lt/index.php/elt/article/view/9865/4853.

[8] S. Czapp, "The impact of higher-order harmonics on tripping of residual current devices," 13th Power Electronics and Motion Control Conference (EPE-PEMC 2008), Poznan, Poland, 2008, DOI: 10.1109/EPEPEMC.2008.4635569.

[9] S. Czapp, "The effect of earth fault current harmonics on tripping of residual current devices," Int. School on Nonsinusoidal Currents and Compensation (ISNCC 2008), Lagow, Poland, 2008, DOI: 10.1109/ISNCC.2008.4627489.

[10] X. Luo, Y. Du, X. H. Wang, et al.: "Tripping characteristics of residual current devices under nonsinusoidal currents," IEEE Trans. on Ind. Appl., vol. 47, no. 3, pp. 1515–1521, 2011, DOI: 10.1109/TIA.2011.2125939.

[11] A. Li, Y. Han, and J. Sun, "An innovative method to achieve minimum tripping current conformity for type A RCCBs," IEICE Electronics Express, vol. 7, no. 12, pp. 1–10, 2015, DOI: 10.1587/elex.12.20141220.

[12] X. Zeliang, M. Yingzong, D. Feng, Z. Yue, and M. Anheuser, "Type B RCD with a simplified magnetic modulation/demodulation method," Int. Power Electronics and Motion Control Conference (IPEMC '09), 2009, DOI: 10.1109/IPEMC.2009.5157488.

[13] BoTong Li, JianFei Jia, XiaoLong Chen, and ShiMin Xue, "Study on residual current protection in low-voltage network with distributed generators," 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2016, pp. 444–448, DOI: 10.1109/APPEEC.2016.7779543.

[14] K. Ludwinek, J. Szczepanik, and M. Sulowicz, „Experimental analysis of assessing of the tripping effectiveness of miniature circuit breakers in an electrical installation fed from a synchronous generator set," Electric Power Systems Research, vol. 142, pp. 341–350, Jan. 2017, DOI: 10.1016/j.epsr.2016.09.028.

Application of the Stewart Platform for studying in control theory*

Dmitrii Dobriborsci, Aleksandr Kapitonov, Nikolay Nikolaev
Department of Control systems and Computer Science
ITMO University
Saint Petersburg, Russia
Email: dmitrii.dobriborsci@corp.ifmo.ru

***Abstract*—In this paper a 2DOF Stewart platform for research and education in control systems and computer vision is described. We built parallel kinematic robotic platform for research nonprehensile manipulation tasks. It can be effectively used in the learning process using active educational methods. In this work we present four lab experiments of this setup at the course "Software of the mechatronics and robotics systems", which is being given at the Department of Control Systems and Computer Science, ITMO University. The main goal is to study students the basics of computer vision and control theory.**

I. Introduction

For the successful designing modern control systems for parallel kinematics manipulators there needs knowledge in trajectory planning and motion control algorithms, adaptive control, computer vision and sensory, modelling of complex mechatronic systems with respect to all forces which could act on the control object. To develop these competences the designing of proper laboratory setup is needed. We built a Stewart platform with two degrees of freedom "'Ball&Plate"' consisting the following components: webcam-based vision system, servodrives, square plate and connecting links. Local control system was implemented on the Arduino single-board computer, integration of the system was performed in development environment MATLAB/Simulink, including visual processing, signal filtering, trajectory planning and motion control algorithms.

Here students studying the basics of control theory and computer vision using active learning methods. Similar problem have been considered in a number of publications. In [1], [2] the role of software-hardware complexes of National Instruments in the formation of competencies, skills and abilities by the example of the course "Electrical Engineering and Electronics". In [3] the learning course "The integrated design and control systems" of "System analysis and control" master degree program at ITMO University with using of active learning methods is described. The results of this course implementation demonstrated high student motivation and improvement of their professional skills. In [4] the usage of modern teaching methods (e.g. active learning, learning in project implementation, etc.) was considered using Bioloid Premium robotic kit. Accomplishing course presented in [4] students know how to obtain the kinematic model and then know how to design a control system for walking robot. In [5] an upper-level undergraduate/graduate educational program at the Computer Science and Engineering Department of the University of Notre Dam in "Autonomous Mobile Robots" course is presented. The course emphasized active learning, where students engaged in the course by practicing concepts learned in lecture immediately on robots; cooperative learning, where students worked on projects in teams; and problem-based learning, where all lectures and assignments brought students back to the fundamental computational problems in mobile robotics.

In context to these works we present using of the 2DOF Stewart platform in education process at ITMO University for developing competencies in control theory, computer vision, modeling mechatronic systems and etc.

The paper is organized as follows. After the setup description in Section II, used active learning methods are presented in Section III. Section IV presents four lab works for students such as mathematical modelling of the whole system, designing computer vision for object detection, controller design for the ball stabilization problem for X-axis only. The next lab includes ball stabilization in both axes considering classical approaches of the control theory such as proportional-integral-differential (PID) controller. The last one consider adaptive output control design. We conclude the paper with discussion about the Ball and Plate setup application in learning process.

II. Setup description

We built a Stewart-like platform with two degrees of freedom consisting of the following components: webcam-based vision system, servodrives, square plate and connecting links. Local control system was implemented on the Arduino single-board computer, integration of the system was performed in MATLAB/Simulink development environment. Computer vision system (CVS) determines the coordinates of the object, the controller is implemented in MATLAB/Simulink calculates necessary rotational angles of the servomotors and through the communications log builds a 16-bit dataset, which transfers the data to Arduino-board through the usb interface. The data

*This work was supported by the Ministry of Education and Science of Russian Federation (Project 8.8885.2017/8.9) and by the Ministry of Education and Science of Russian Federation (Project 14.Z50.31.0031) and by the Russian Federation President Grant 14.Y31.16.9281-НШ.

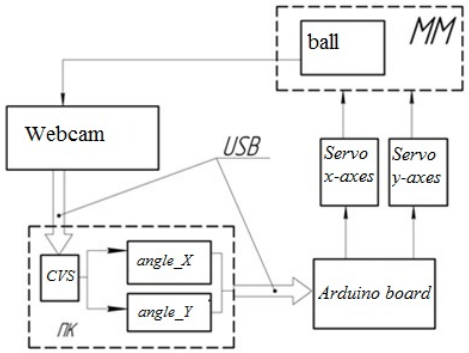


Figure 1. Block diagram.

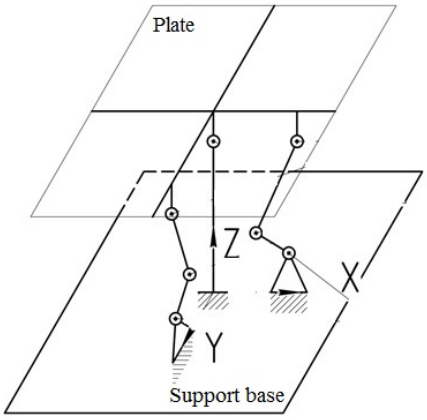


Figure 2. Kinematic scheme of the mechanism

undergoes a consistency test at Arduino-board and the data is sent straightforward to the servos. Each tilting axis is operated on by a servomotor. To detect object coordinates we use a webcam-based visual system. The functional configuration and kinematic scheme of the considered system are presented in the Fig. 1, 2, respectively.

III. Active learning

The main feature of this course is an application of active learning methods at lectures and seminars. This course correspond to new educational standard in Russian Federation. In this regard, systemically-active approach to educational process, which is also called competence-based approach, comes to replace the knowledge-based model of education. Independent work of students as one of the main components of the competence-based approach is aimed not only at achieving of educational goals, but mainly on the formation of personal qualities of the future specialist-self-knowledge, self-development and self-realization [6]. These qualities form a new competent person who meets the requirements of the labor market.

IV. Course overview

A. Lab 1. Computer Vision

Computer vision is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. The first we teach students how to design computer vision system in MATLAB/Simulink (Computer Vision System toolbox). Computer Vision System Toolbox provides algorithms, functions, and apps for designing and simulating computer vision and video processing systems. It provides an opportunity to perform feature detection, extraction, and matching; object detection and tracking; motion estimation; and video processing. The main task is to detect the object coordinates using webcam-based computer vision system. The camera is attached on the top of the laboratory setup above the platform, see fig. 4. The visualization of the CVS is presented in fig. 3.

B. Lab 2. Ball and Plate mathematical model

The second practical task is focused on simplified mathematical model of the 2DOF balancing system to illustrate performance of design a controller for ball stabilization in the specific point experiment. We consider the system model with the following simplifications

- There is no slipping for ball.
- The ball is completely symmetric and homogeneous.
- Friction forces are negleted.
- The ball and plate are in contact all time.

The angles of servo arms θ_x, θ_y are assumed to be the inputs, while the ball position on x, y axis are assumed to be the output. Here we derive dynamical equations of ball-on-plate system, by the help of Lagrangian. The following mathematical equations are based on [7], [8], [9]. So the nonlinear differential equations for the ball and plate system are presented below.

$$\left(m + \frac{I}{r^2}\right)\ddot{x} - m\left(x\dot{\alpha}^2 + y\dot{\alpha}\dot{\beta}\right) + mgsin\alpha = 0 \quad (1a)$$

$$\left(m + \frac{I}{r^2}\right)\ddot{y} - m\left(y\dot{\beta}^2 + x\dot{\alpha}\dot{\beta}\right) + mgsin\beta = 0 \quad (1b)$$

where m, r, I are mass of the ball, radius of the ball, mass moment of inertia x, y are position of axis, α, β, $\dot{\alpha}$,$\dot{\beta}$ are inclination angles of the plate, angular velocity of the plate, g, L, d are gravitation, plate side length, length between the joint and the center of the gear, respectively.
The relations between inclination angles of the plate α, β and θ_x, θ_y are the following

$$sin\alpha = \frac{2sin\theta_x d}{L} \quad (2a)$$

$$sin\beta = \frac{2sin\theta_y d}{L} \quad (2b)$$

In the case of a slow rate of change for the plate angles equations (1a), (1b) can be linearized

$$\left(m + \frac{I}{r^2}\right)\ddot{x} - \frac{2mgd}{L}\theta_x = 0 \quad (3a)$$

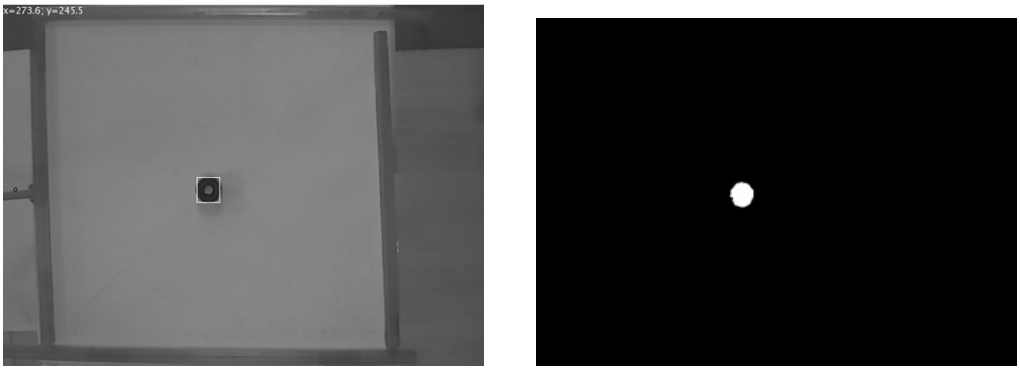


Figure 3. Visualization of the computer vision system.

Figure 4. Parallel kinematics robot Ball and Plate

$$\left(m + \frac{I}{r^2}\right)\ddot{y} - \frac{2mgd}{L}\theta_y = 0 \tag{3b}$$

The equations (3a), (3b) are equivalent because of the symmetry of the plate.
With the Laplace transformation, the following transfer functions can be obtained

$$P_x(s) = \frac{x}{\theta_x} = -\frac{2mgdr^2}{L(mr^2 + I)s^2} \tag{4a}$$

$$P_y(s) = \frac{x}{\theta_y} = -\frac{2mgdr^2}{L(mr^2 + I)s^2} \tag{4b}$$

where s is the Laplace operator.
Lab report should consist of mathematical model of the given system, calculated transfer functions, and model analysis in MATLAB/Simulink. Then we teach students how to design a PID controller using Ziegler- Nichols method [10].

C. Lab 3. Ball stabilization (X-axis)

The third practical exercise is devoted to control system design for a Ball and Plate system. Students should apply their knowledge they got during lecture course. Using the results obtained in previous labs students perform the next task: Design a PID- controller for ball stabilization in the specified coordinates in X-axis only (well-known "'ball and beam"' problem). Here students should implement Ziegler-Nichols tuning method for PID- controller for mathematical model they obtain in the second exercise and compare results with experiment data. The control system must be realized in MATLAB/Simulink. Except the control system students have to realize communication protocol for Arduino board and PC. Therefore, lab work consist of three tasks: modelling control system for Ball and Plate system for X-axis only tuned using Ziegler-Nichols method, CVS, experimental validation. Lab report should contain the following: modelling results, PID-controller performed in MATLAB/Simulink and protocol for Arduino-MATLAB/Simulink communication. In the lab report student must present transient-response plots (control gains, steady-state error, output position of the object), calculated coefficients for PID controller.

D. Ball and Plate control system for ball stabilization

Finally, students are ready for designing a full PID-controller of the 2DOF Ball And Plate system. The objective is to stable the ball in user-defined coordinates. As discussed the ball and plate system can well be approximated by two linear decoupled systems of the form (4a), (4b). Therefore, this system with two inputs and two outputs can be treated as two decoupled single input single output systems, this means that controllers designing can be realized independently. For the controller design we make no further assumptions about the model. We assume that the parameters of the system are unknown. The examples of the transient-response plots see in fig.5 - 10.

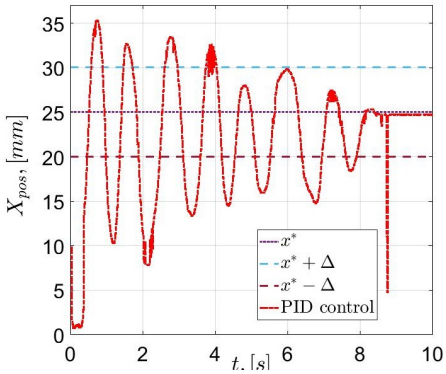


Figure 5. Position measured by CVS on X channel.

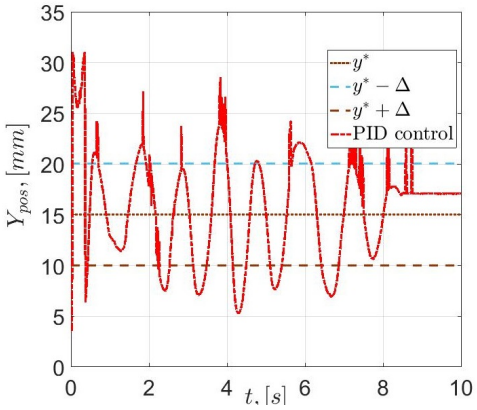


Figure 6. Position measured by CVS on Y channel.

V. Conclusion

The preparation of highly qualified specialists in robotics and control systems is impossible without evolvement of practical learning. It is clear that it should be realized combined with theoretical knowledge. Therefore, the lab activities on control for Ball and Plate system, which could be considered as a parallel kinematic robot manipulator, are aimed to improve the skills sets of students are proposed in this paper. They present a practical part of the course ”Software of the mechatronics and robotics systems” which is being taught at the Department of Control Systems and Computer Science of ITMO University.

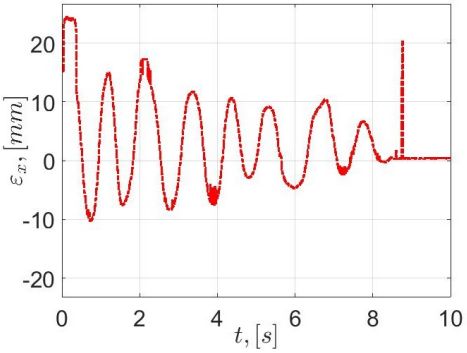


Figure 7. Error in X channel.

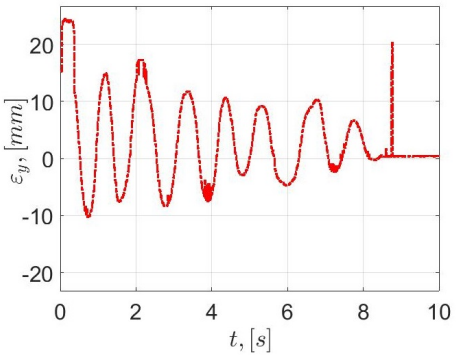


Figure 8. Error in Y channel.

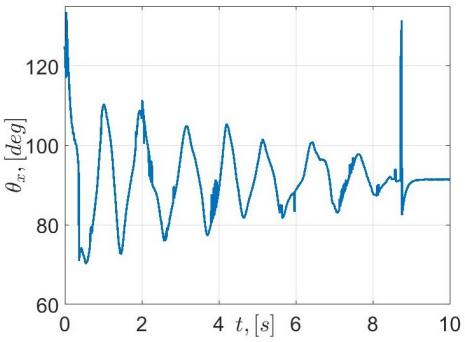


Figure 9. Control law for θ_x.

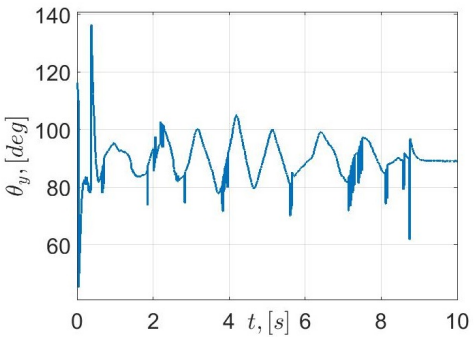


Figure 10. Control law for θ_y.

References

[1] M. Artemeva, N. Nikolaev, D. Dobriborsci, O. Nuyya, and O. Slita, “Ni elvis ii in the concept of cognitive and active learning technologies,” *IDT 2016 - Proceedings of the International Conference on Information and Digital Technologies 2016*, pp. 71–75, 2016.

[2] D. Dobriborsci, D. Bazylev, and A. Margun, “Teaching students the basics of control theory using ni elvis ii,” *International Conference on Smart Education and Smart E-Learning*, vol. 75, pp. 420–427, 2017.

[3] D. Bazylev, A. Shchukin, A. Margun, K. Zimenko, A. Kremlev, and A. Titov, “Applications of innovative “active learning” strategy in “con-

trol systems" curriculum," *Smart Innovation, Systems and Technologies*, vol. 59, pp. 485–494, 2016.

[4] K. Zimenko, D. Bazylev, A. Margun, and A. Kremlev, "Application of innovative mechatronic systems in automation and robotics learning," *Proceedings of the 16th International Conference on Mechatronics, Mechatronika 2014*, pp. 437–441, 2014.

[5] L. Riek, "Embodied computation: An active-learning approach to mobile robotics education," *IEEE Transactions on Education*, vol. 56, no. 1, pp. 67–72, 2013.

[6] A. Kremlev, D. Bazylev, A. Margun, and Z. K., "Transition of the russian federation to new educational standart: independent work of students as a factor in the quality of educational process," *ERPA International Congresses on Education 2015*, vol. 26, 2015.

[7] X. Fan, N. Zhang, and S. Teng, "Trajectory planning and tracking of ball and plate system using hierarchical fuzzy control scheme," *Fuzzy Sets and Systems*, vol. 144, no. 2, pp. 297–312, 2004.

[8] K.-K. Lee, G. Bätz, and D. Wollherr, "Basketball robot: Ball-on-plate with pure haptic information," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2410–2415, 2008.

[9] D. Yuan and Z. Zhang, "Modelling and control scheme of the ball-plate trajectory-tracking pneumatic system with a touch screen and a rotary cylinder," *IET Control Theory and Applications*, vol. 4, no. 4, pp. 573–589, 2010.

[10] N. Nichols and J. Ziegler, "Optimum settings for automatic controllers," *Journal of Dynamic Systems, Measurement and Control, Transactions of the ASME*, vol. 115, no. 2 B, pp. 220–222, 1993.

The basics of the identification, localization and navigation for mobile robots*

Dmitrii Dobriborsci, Aleksandr Kapitonov, Nikolay Nikolaev
Department of Control systems and Computer Science
ITMO University
Saint Petersburg, Russia
Email:kap2fox@gmail.com

***Abstract*—This article is an implementation of the results described in earlier papers for odometry modelling and navigation system synthesis. It'll be part of the introduction in Erasmus+** *Innovative Open Education on IoT*. **It's already used for MOOC course published on national education platform https://openedu.ru/, the e-learning control theory course used in ITMO University. The main goal of this course to give to the students an idea of the connection formulas, models and physical objects. The course consists from the DC-motor parameters identification, engine model description, linear and nonlinear controllers implementation, encoders and IMU-sensors odometry motion model.This experience was used for a formulation the navigation problem with the NXT differential drive mobile robot. Based on the method proposed in [1] the robot to goal movement with the obstacle avoidance task was solved. It presents a simple and demonstrative example for choosing a Lyapunov function candidate for controller design for a nonlinear system based on the robot-goal distance and the orientation error regarding the goal position.**

I. Introduction

Robots are becoming a part of our life. Pupils and students want to practice in robotics: robot design, robot control, participating in competitions. But few people know how to teach robotics today, like many years ago, computer science. Moreover, in general school teachers can't and don't know how to teach robotics. Also, the situation is complicated by the necessity of having the element base - robots. For solving of the big pat of a tasks in robotics control PID is enough. But if you want to reduce control costs or make you robot faster, for example in RoboCup competitions, the usage of a complex nonlinear methods is necessary. It's not so easy for a bachelor students to apply a well known result in practice. Robotics is complex scientific and engineering sphere. In one team we need mathematician, programmer and engineer. It's different specialization, and you need to organize work separately. But there many good examples with a good implementation of a difficult algorithms on practise that give a better understanding between team members. The article shows one of these algorithms. We introduce to students a new kind of the learning methodology based on Lego NXT platform using active learning methods [2]. Lego used as educational platform and had many different implementations, like in [3], [4] and [5]. We did it for first year students and high school students in preparation to the World Robot Olympiad. Positive feedback were received, since now students more interested in control theory applications, and will be interested in explanation of the difficult things in the future. We improved our course of the lab activities described in articles [6], [7], [8], and migrated course from NXT to EV3 with Python. In this paper the curricula of the control theory course is presented. The main goal of this course to give to the students an idea of the connection formulas, models and physical objects.

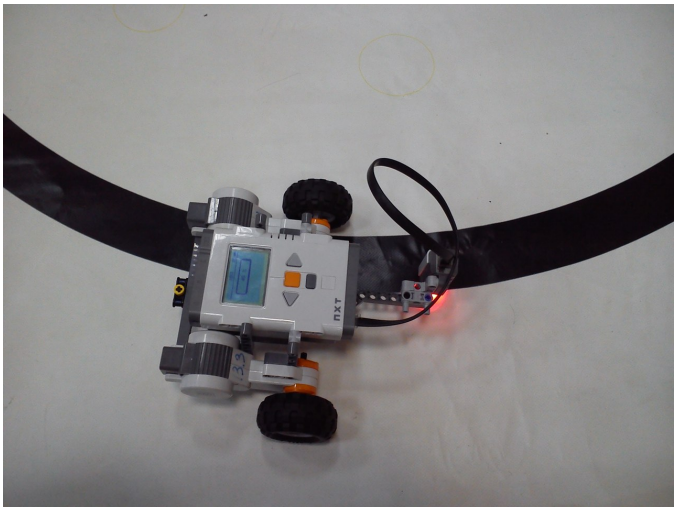


Figure 1. Line tracer robot.

The first sample includes Ziegler-Nichols' method application to the PID controller. Second sample is the identification NXT motor mechanical time constant and back EMF constant. Third sample is modelling and developing motor angle proportional controller. Fourth sample is measuring NXT motor torque constant. The last one is modelling and developing odometry motion and obstacle avoidance applied to the NXT differential drive robot.

II. Ziegler-Nichols method applied to mobile robot control

At first, students use a simple differential drive system and learn a methods, that don't requires profound knowledge in control theory. For this task the Ziegler-Nichols' ultimate gain method (closed-loop method) is used. During this work, students have time to learn programing software for NXT or EV3 (C or Python). They try Scilab for a visualisation of the

work results (Scilab an open source software for numerical computation). It's a good practice to work with open source software.

Ziegler-Nichols' method is used for tuning PID–controller, one of the most popular algorithm, and good example in control theory. For this task students construct simple line tracer robot 1 with one light sensor. The track line looks like number eight. We begin from the P–control law. After the robot is able to pass this line, we increase the proportional gain, until the robot starts to make stable oscillations, moving along the line. Maximum of the workable proportional coefficient, it is the ultimate gain. After we get measurements of the light sensor from robot, we commit the period of the signal oscillations.

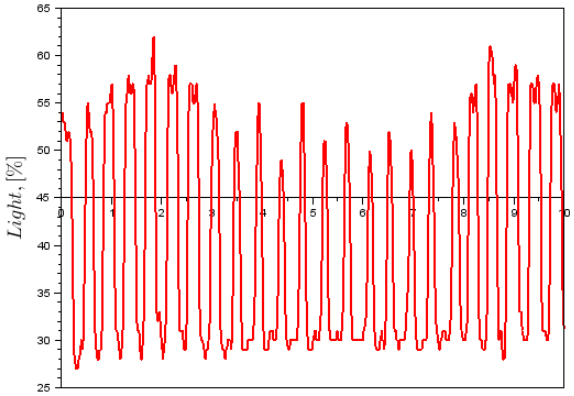


Figure 2. Measurements of light sensor in time.

In fig. 2 for our line tracer ultimate gain $K_u = 7$ and the period of oscillation $T_u = 0.43\ sec.$ After achieving the ultimate gain, the PID-controller gains can be calculated. Fig. 3 shows processes in the system with different control laws. The PID case is a good example for anti-Windup work demonstration.

III. The motor parameters identification

The second work starts series of works step-by-step description of the dynamic model for differential drive mobile robot. In the beginning students calculate a motor constant, which associate the values of control voltage and output torque. The first, a necessary value is a mechanical time constant. We can determine it with assignment constant control signal and a step response for a motor angle recording. Using a formula 1 for an angle-time dependence the mechanical time constant can be calculated:

$$\phi = \omega_{nls} t - T_m \omega_{nls} + T_m \omega_{nls} \exp(-t/T_m), \qquad (1)$$

where t – time, ϕ – turning angle, ω_{nls} – no load speed, T_m – mechanical time constant. Scilab has a special *datafit* function for model parameters estimation. The function result gives no load speed and mechanical time constant for NXT

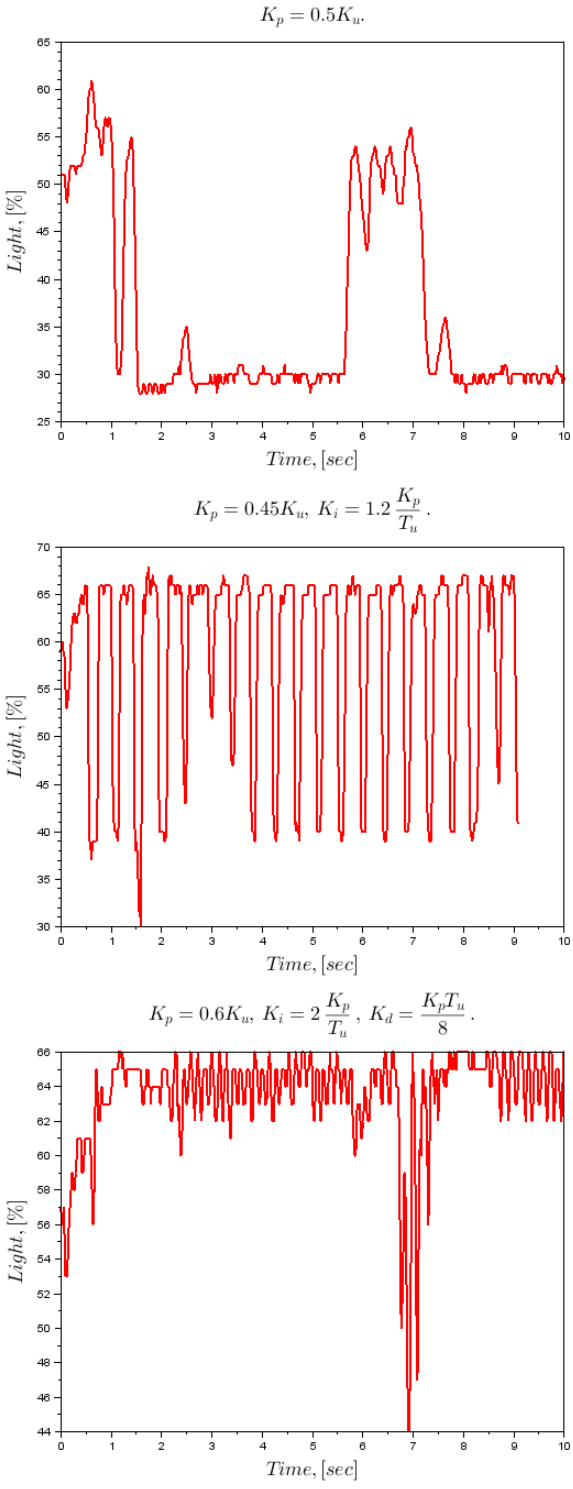


Figure 3. P,PI,PID control laws and coefficients calculation formulas.

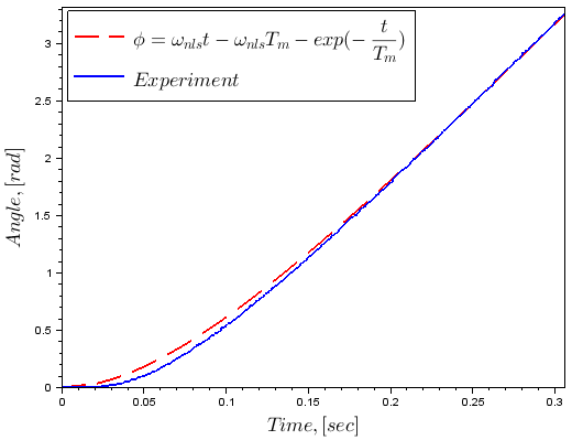


Figure 4. Identification of model parameters.

motor (fig. 4). In presented case $\omega_{nls} = 14.2\ \frac{rad}{sec}$ and $T_m = 0.079\ sec$. Control voltage value is $U = 7\ V$. Supposing that there is no viscous friction and other disturbances in motor, the ideal engine model is used. It means that a back EMF U_{emf} value equals to the control voltage in no load speed mode, and in the rotating motor current is $0\ A$ (the real friction gives a current value $I = 0,05\ A$).

$$K_e = \frac{U_{emf}}{\omega_{nls}}. \tag{2}$$

For the motor $K_e = 0.49V \cdot sec$. The next step is checking parameters in the open-loop model. Students synthesize control law $U = 7 * sin(\pi * t)$, where U – the control signal.

As you can see, these curves 5 too close to each other. It means that the motor parameters are identified correctly. But the initial point of oscillation is moving in the bottom part of graph, this is because the maximum voltage of PWM in different direction has difference $0.2\ V$. This value disperses in the bridge circuit.

IV. The proportional controller for NXT motor

In this work students check an obtained result in a closed-loop system. For this task synthesize control law for steering the angle of the motor by the proportional controller. This work includes programming and modelling a closed-loop system, it helps students to associate a mathematical formulas and text of the program.

Our control algorithm compares the current angle with a given one, in our case it is π, and based on the error generates a control signal to the motor [9].

$$U = Kp * (\pi - \phi). \tag{3}$$

From the closed-loop system modelling get the experimental curve has an error, not more than $2\ \%$. The calculations are made correctly, and students can proceed to the calculation of the torque constant.

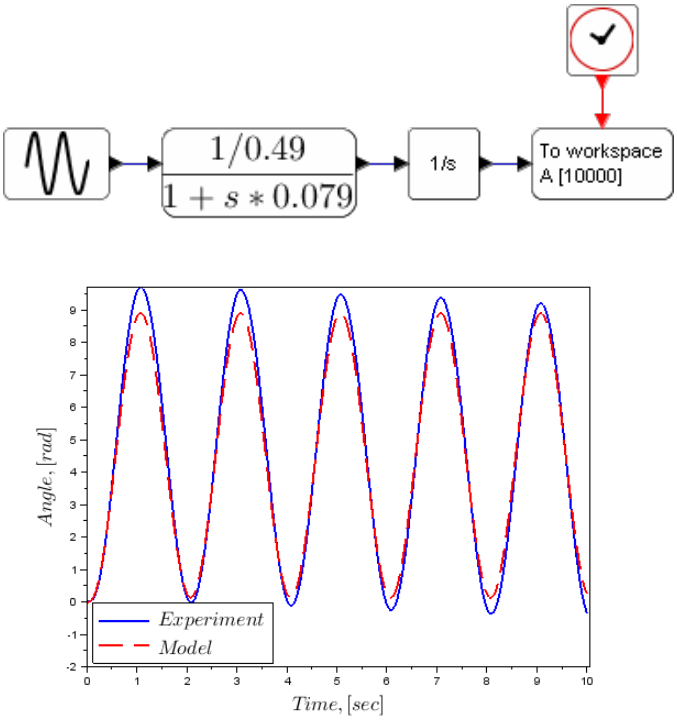


Figure 5. Checking model parameters.

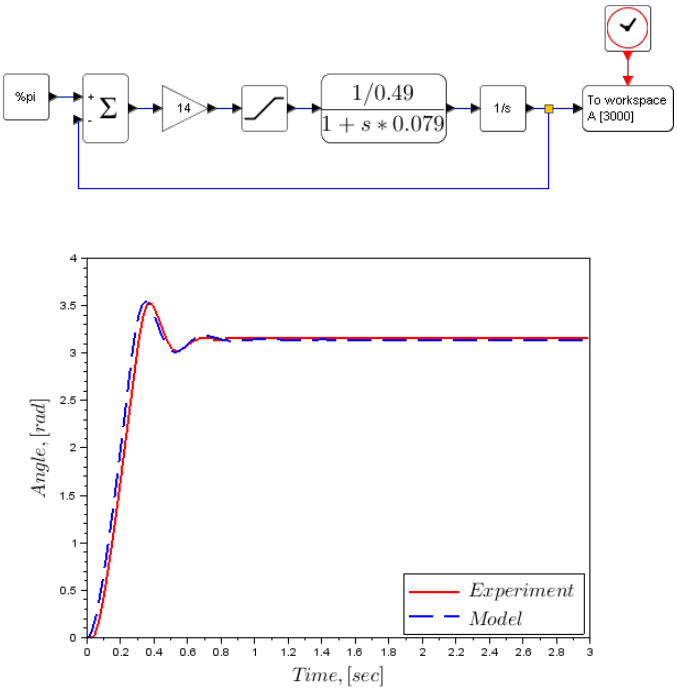


Figure 6. Model of the P-controller and graph of real motor.

V. The torque constant of the NXT motor

We need a mathematical description of the robotic system, and the torque constant K_t is needed for expanded model. Consider the equation of the motor without the viscous friction:

$$\begin{cases} \dot{\omega} = \frac{K_t}{J} I, \\ \dot{I} = \frac{1}{L} U - \frac{K_e}{L} \omega - \frac{R}{L} I. \end{cases} \quad (4)$$

,

where $J = 10^{-6}\ kg * m^2$ – moment of inertia of the rotor, $R = 6\,Ohm$ – electric resistance of the circuit, $L = 0.0047\,H$ – an inductance of the armature, ω – a rotor speed, I – a rotor current. Now, following formula 5, calculate the value of the torque constant.

$$K_t = \frac{M_{lrt} i^2}{I_{lrc}}, M_{lrt} = \frac{\omega_{nls} J}{T_m}, \quad (5)$$

where M_{lrt} – the locked rotor torque, I_{lrt} – the locked rotor current, $i = 48$ – the NXT motor's gear ratio. Substitute values of the parameters of the motor. Torque constant and back EMF constant values should be close, for NXT these constants are $K_t = 0.42\ \frac{Nm*m}{Ah}$ and $K_e = 0.49\ V * sec$. For input/output model we also need electrical time constant $T_e = \frac{L}{R} = 0.008\,sec$. Lego NXT motor controlled by PWM. But in our equation the continuous control signal is used. Students should check that using PWM and DC control give the same results. The graph 7 shows simulation results for NXT motor with a reduction gear.

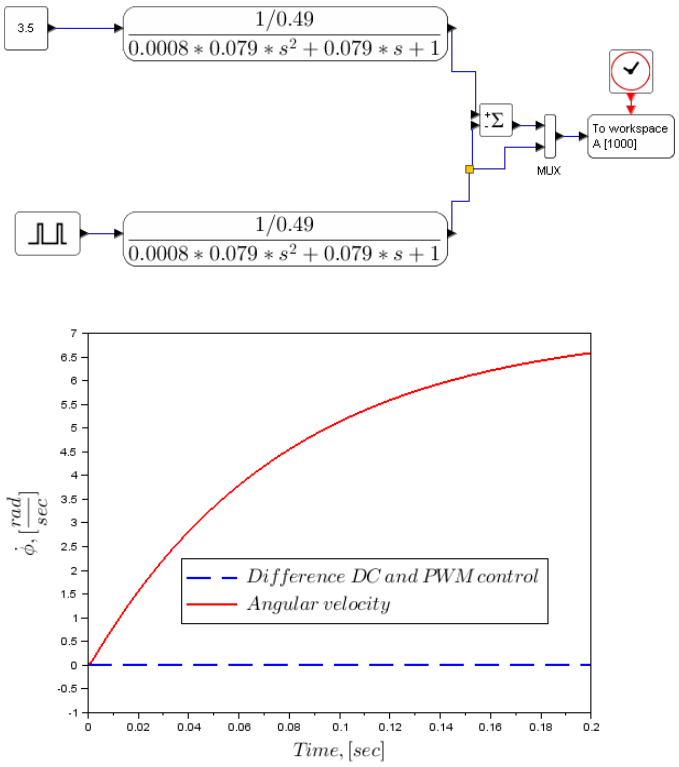


Figure 7. Comparison of DC and PWM control.

It should be noted, that measurements of the voltmeter show a correct constant value of the voltage. An oscilloscope should be used for demonstration of the PWM signal. After, for students should be clear, that the continuous control can be used for the modelling.

VI. The tangential escape for differential drive robot

In this work the mobile robot with a differential drive is used. In common case control signals are the linear and angular velocities [10], [11]. A sketch of the robot navigating towards its goal in a free space is given in 8.

$$\begin{aligned} \dot{\rho} &= \frac{\omega_1 + \omega_2}{2} r, \\ \dot{\psi} &= \frac{\omega_1 - \omega_2}{base} r, \end{aligned} \quad (6)$$

where r – the wheel radius, $Base$ – a distance between the wheels, ω_1, ω_2 – an appropriate wheel speed. The robot starts from the [0;0] point and $\psi = 0$. The [x;y] coordinates of the robot 7 and the full model of the mobile robot:

$$\begin{aligned} x &= \cos\psi \int \dot{\rho} dt, \\ y &= \sin\psi \int \dot{\rho} dt, \end{aligned} \quad (7)$$

The modelling diagram presented at fig. 10. The mathematical model describing the navigation, in polar coordinates. It's given by [12] implementing the tangential escape.

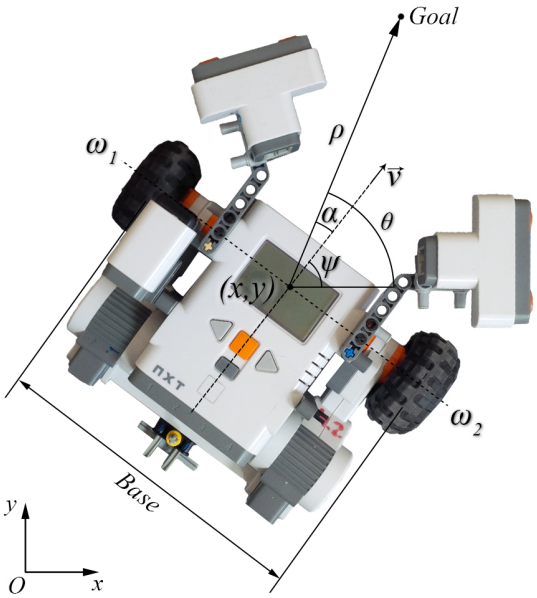


Figure 8. Differential drive mobile robot description.

$$\begin{aligned} \dot{\rho} &= -u \cdot \cos\alpha, \\ \dot{\alpha} &= -\omega + u \cdot \frac{\sin\alpha}{\rho}, \\ \dot{\theta} &= u \cdot \frac{\sin\alpha}{\rho}, \end{aligned} \quad (8)$$

The goal of the control system is during the time $t \to \infty$ $\lim_{t\to\infty} \rho(t) = 0$ and $\lim_{t\to\infty} \alpha(t) = 0$. Let consider the Lyapunov

function candidate:

$$V(\rho,\alpha)=\frac{1}{2}\rho^2+\frac{1}{2}\alpha^2, \tag{9}$$

whose derivative is:

$$\dot{V}(\rho,\alpha)=\rho\dot{\rho}+\alpha\dot{\alpha}<0. \tag{10}$$

Whereas the robot kinematic model 8, the Lyapunov function derivative 10 is:

$$\dot{V}(\rho,\alpha)=-\rho u\cdot\cos\alpha+\alpha(-\omega+u\cdot\frac{\sin\alpha}{\rho}), \tag{11}$$

and it's negative definite if the control variables u and ω are defined as:

$$\begin{aligned} u&=\rho\cdot cos\alpha,\\ \omega&=k_\omega\alpha+\tfrac{u}{\rho}\sin\alpha. \end{aligned} \tag{12}$$

where ρ – the distance robot-goal, u – the linear velocity of the robot (in the direction normal to the axis linking it's driven wheels), ω – the angular velocity of the robot, α – the orientation error (regarding the goal position), θ – the angle between the line linking the origins of the on-board and the inertial frames of coordinates and the horizontal axis, and ψ – the angle between the direction of movement and the horizontal axis.

For obstacle avoidance task used the next equation with $\hat{\alpha}$ angle correction:

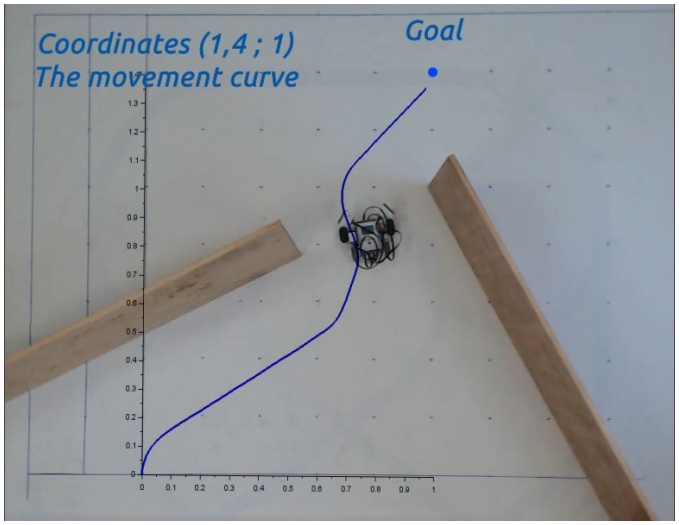


Figure 9. Point to point movement with obstacle avoidance.

$$\hat{\alpha}=\alpha-k_1\cdot err_1+k_2\cdot err_2, \tag{13}$$

where err_1, err_2 – distance between obstacle and mobile robot $0[m]<err_{1,2}<0.4[m]$, k_1, k_2 – proportional coefficients are selected empirically. If mobile robot has a sensor with point cloud, and the angle β between an obstacle and the robot orientation ψ can be measured, use a method proposed in [1]:

$$\hat{\alpha}=sign\beta\frac{\pi}{2}-(\beta-\alpha), \tag{14}$$

and the control law:

$$\begin{aligned} u&=\rho\cdot cos\hat{\alpha},\\ \omega&=k_\omega\hat{\alpha}+\tfrac{u}{\rho}\sin\hat{\alpha}. \end{aligned} \tag{15}$$

The work of the method presented at fig. 9.

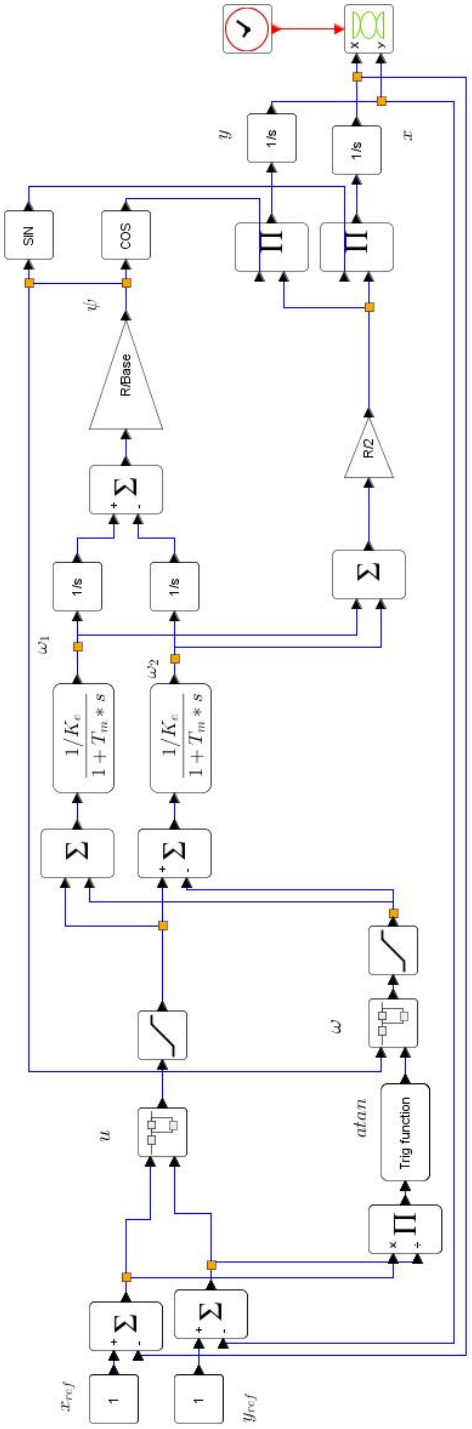


Figure 10. Differential drive mobile robot Xcos model.

VII. Course implementation

This course is already used in the ITMO University since 2015 year. Course passed several tens of the female and male students. It helps to work with more complex tasks and solve RoboCup competition missions. We start to train a competition teams for tracks with complex equipment and difficult goals, like trajectory following and formation control, simultaneous localization and mapping (SLAM).

VIII. Conclusion

Our department starts this course from the different underactuated systems [6], [13] like segway, Furuta pendulum, Kapitza pendulum, inverted pendulum on a cart and etc. But it was focused mainly on control theory, now we shifted to the world competition tasks. And our control theory background gives us good solutions in navigation and trajectory control. All projects of the ITMO University are available on our youtube channel http://www.youtube.com/itmo4robots. Video lectures of this course http://www.openedu.ru/.

IX. Aknowledgement

This work was partially supported by the Ministry of Education and Science of Russian Federation (Project 14.Z50.31.0031) and by the Russian Federation President Grant 14.Y31.16.9281-HIII. This material is based upon work supported by European Union under the Erasmus+ Key Action 2 (Strategic Partnership) projectIOT-OPEN.EU (Innovative Open Education on IoT: improving higher education for European digital global competitiveness), reference no. 2016-1-PL01-KA203-026471. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

[1] A. Ferreira, F. G. Pereira, R. F. Vassallo, T. F. Bastos Filho, and M. Sarcinelli Filho, "An approach to avoid obstacles in mobile robot navigation: the tangential escape," *Sba: Controle & Automação Sociedade Brasileira de Automatica*, vol. 19, no. 4, pp. 395–405, 2008.

[2] M. Artemeva, N. Nikolaev, D. Dobriborsci, O. Nuyya, and O. Slita, "Ni elvis ii in the concept of cognitive and active learning technologies," *IDT 2016 - Proceedings of the International Conference on Information and Digital Technologies 2016*, pp. 71–75, 2016.

[3] A. Valera, M. Valles, L. Marín, A. Soriano, A. Cervera, and A. Giret, "Application and evaluation of lego nxt tool for mobile robot control," 2011.

[4] S. Parsons and E. Sklar, "Teaching ai using lego mindstorms," 2004.

[5] P. Hurbain and M. Gasperi, "Extreme nxt: Extending the lego mindstorms nxt to the next level," 2007.

[6] A. A. Kapitonov, A. A. Bobtsov, Y. A. Kapitanyuk, D. S. Sysolyatin, E. S. Antonov, A. A. Pyrkin, and S. A. Chepinskiy, "Course of lab activities on control theory based on the lego nxt," vol. 19, no. 1, pp. 9063–9068, 2014.

[7] S. Kolyubin, A. Bobtsov, A. Pyrkin, A. Kapitonov, S. Shavetov, A. Khovanskiy, A. Krasnov, S. Vlasov, and A. Feskov, "Lego mindstorms nxt for students' research projects in control field," vol. 9, no. 1, pp. 102–106, 2012.

[8] A. Bobtsov, A. Pyrkin, S. Kolyubin, S. Chepinskiy, S. Shavetov, Y. Kapitanyuk, A. Kapitonov, A. Titov, M. Surov, and V. Bardov, "Using of lego mindstorms nxt technology for teaching of basics of adaptive control theory," pp. 9818–9823, 2011.

[9] A. L. Fradkov, I. V. Miroshnik, and V. O. Nikiforov, "Nonlinear and adaptive control of complex systems," vol. 491, 2013.

[10] P. Corke, "Robotics, vision and control: fundamental algorithms in matlab," vol. 73, 2011.

[11] C. Belta, A. Bicchi, M. Egerstedt, E. Frazzoli, E. Klavins, and G. J. Pappas, "Symbolic planning and control of robot motion [grand challenges of robotics]," *Robotics & Automation Magazine, IEEE*, vol. 14, no. 1, pp. 61–70, 2007.

[12] H. Secchi, R. Carelli, and V. Mut, "Discrete stable control of mobile robots with obstacles avoidance," vol. 1, pp. 405–411, 2001.

[13] D. Dobriborsci, D. Bazylev, and A. Margun, "Teaching students the basics of control theory using ni elvis ii," *International Conference on Smart Education and Smart E-Learning*, vol. 75, pp. 420–427, 2017.

Interference Comparison in Wi-Fi 2.4 GHz and 5 GHz Bands

Iwona Dolińska, Mariusz Jakubowski, Antoni Masiukiewicz
Vistula University (AFiBV)
Stokłosy 3, 02-787 Warsaw, Poland
i.dolinska@vistula.edu.pl; m.jakubowski@vistula.edu.pl; a.masiukiewicz@vistula.edu.pl

Abstract— **Adjacent and co-channel interference is one of the biggest problems in the IEEE 802.11 standard. It is believed that the issue mainly concerns the 2.4 GHz band due to insufficient number of non-overlapping channels. Co-channel interference occurs in both bands, however, and it is associated with the channel assignment method. The authors analyzed the level of interference in both bands for two AP distributions on the X-Y plane. The first distribution was symmetric, while the second one was a part of an actual multi-flat building. The Minimum-Spanning-Tree-Inspired (MISTI) algorithm, previously developed, was used for the channel assignment.**

Keywords— interference, MISTI, Wi-Fi, channel assignment

I. INTRODUCTION

Interference is one of the throughput limiting factors in the IEEE 802.11 standard. Adjacent channel interference occurs in the 2.4 GHz band and it is due to overlapping channels. In this band, we have only three so-called non-overlapping channels with numbers 1, 6, 11. There is also adjacent channel interference issue in the IEEE 802.11n standard, if 2.4 GHz channels are bonded to obtain a 40 MHz channel. The degree of overlapping can be described using correlation coefficients [10, 12]. There is no adjacent channels interference in the 5 GHz band, but still there are co-channel interference phenomena. They occur when we use too few channels in a given area due to negligence or a necessity, e.g. when 80 MHz and 160 MHz channels are taken into account. In this case, only two channels of 80 MHz width and one of 160 MHz width are available, so this interference type is inevitable. Table I and Table II show correlation coefficients for the 2.4 GHz band for 20 and 40 MHz channel widths. The channel interval means the difference between two channel numbers.

In the case of the CRC (Correlation Coefficient Ratio) of different radio channels equal to 100%, we are dealing with co-channel interference. This type of interference occurs both in the 2.4 GHz band and in the 5 GHz band. There are at least three separate bands in 5 GHz part of the spectrum: U-NII-1, U-NII-2 and U-NII-3. Hence, we can use more than twenty non-overlapping channels [2] in the following frequency ranges: 8 channels in U-NII-1, 11 channels in U-NII-2, 6 channels in U-NII-3. The worst-case interference value is the sum of adjacent and co-channel interference from all the stations within a transmission range. In practice, interference can also come from other sources. The interference value can be determined from the formula:

$$P_{N-INT} = \sum_{k=1}^{K} \sum_{x=1}^{m} P_{kx} \cdot CRC_z \qquad (1)$$

where $P_{N\text{-}INT}$ represents the interference power in a given channel. It is the sum of the power of all stations detected in all channels (P_{kx}) after we take into account the correlation coefficient (CRC_z), where k is the current channel, K is the number of available channels, z is the channel interval $z = |N - k|$, N is an interfering channel and x represents all other stations transmitting in each channel [10].

In this paper, the authors analyze the interference level in 2.4 GHz and 5 GHz networks for different channel assignment. The Minimum-Spanning-Tree-Inspired (MISTI) algorithm, previously developed [12], is used for the channel assignment. The interference level affects the real throughput, which is available for users, and even possibility of data transmission. The authors compare interference level in both bands and usefulness of those types of networks for two distributions of AP on 2-D plane: symmetrical and non-symmetrical. Non-symmetrical distribution is based on an actual multi-flat building blueprint, which allows to analyze a real-life example.

The rest of the paper is organized as follows. In Section 2, an overview of related works is presented. Section 3 gives the basic description of the MISTI algorithm [12], used for channel assignment. Calculation results are presented in Section 4 and the conclusions are given in Section 5.

TABLE I. CORRELATION RATIO COEFFICIENT (CRC) VALUES FOR 2.4 GHZ BAND AND 20MHZ CHANNEL WIDTH

Adjacent channel interval	CRC [%]
0	100
1	75
2	50
3	30
4	0
5	0
6	0
7	0
8	0
9	0
10	0

TABLE II. CORRELATION RATIO COEFFICIENT (CRC) VALUES FOR 2.4 GHZ BAND AND 40 MHZ CHANNEL WIDTH

Channel Sets		Bands Overlapping [MHz/MHz]	CRC [%]
1.5	1.5*	40/40	100
1.5	2.6	35/40	87.5
1.5	3.7	30/40	75
1.5	4.8	25/40	62.5
1.5	5.9	20/40	50
1.5	6.10	15/40	37.5
1.5	7.11	10/40	25

* 1.5 means that two 20 MHz channels with number 1 and 5 respectively, are bonded to obtain one 40 MHz channel.

II. RELATED WORKS

The IEEE 802.11 standard uses 5 GHz and 2.4 GHz bands. The literature points to a number of differences in the achievable parameters [1, 3, 4, 6] for these two bands. The main differences are: coverage, achieved throughput, level of interference from other systems and devices, number of available channels, adjacent- and co-channel interference, national regulations, bandwidth availability, security, and cost. The coverage in free space is greater for 2.4 GHz band; however, maximal range is not always necessary and with the support of repeaters, which are cheap and widely available nowadays, the coverage of 5 GHz band can be easily improved.

The actual bit rate depends on the signal level, channel width and Modulation and Coding Scheme (MCS). The situation may be different for both bands because for a given transmission, the 2.4 GHz band may be more advantageous in terms of the bit rate, e.g. due to higher power levels received and the possibility of using higher MCS schemes. For the 2.4 GHz band, there are a number of potential sources of interference from other systems [3, 4] such as microwave ovens, cordless phones, garage door controllers, IEEE 802.15.4, ZigBee, Wireless HART and Bluetooth devices. However, on the other hand, it should be bore in mind that in the U-NII-2 5GHz band, due to radar work, it is necessary to use DFS (Dynamic Frequency Selection) of channels [3, 5, 6] which in practice limits the use of this band. Few devices can actually use the U-NII-2 band.

The number of non-overlapping channels is significantly higher in the 5 GHz band. However, this is only a potential advantage that matters when we need more than three channels. The 5 GHz band from this point of view has great advantages in multi-residential buildings, where coverage does not need to be large (it may even be a disadvantage) but the high number of non-overlapping channels is very important. Adjacent-channel interference occurs only in the 2.4 GHz band. However, co-channel interference may occur in both bands. The extreme example is the transmission in the 160 MHz channel in the 802.11ac standard. You can build only one such channel by combining 20 MHz channels in the U-NII-1 band. In the case, when several users try to use this channel simultaneously, either it will fail or the interference level will increase considerably and all negative effects will occur. National regulations may exclude selected channels or entire bands. For example, in China, the use of the 5 GHz band is not permitted [7, 8].

The advantage of the 5 GHz band is the possibility to use very wide channels which are available in 802.11ac devices. Under certain circumstances this gives a really significant increase in bit rate difficult to achieve in the 2.4 GHz band. You may find that security in the 5 GHz band is better due to the smaller range [6]. It seems that this argument is not convincing. The cost of 5 GHz networks may be higher due to the coverage that may require more access points in a given area. However, if we consider other parameters such as channel width and maximum bit rate, costs may be lower for 5 GHz band. The intensity of the use of the band or its popularity is a fact in many countries. The 2.4 GHz band dominates in Poland and this fact limits the introduction of the 5 GHz band. One of the arguments is compatibility with older devices. It is impossible to tell authoritatively which band is better at a general level. It is always necessary to analyze and evaluate a given project.

III. MST-INSPIRED ALGORITHM DESCRIPTION

A Minimum-Spanning-Tree-Inspired Algorithm (MISTI) [12] is used in this paper to select channels for APs working on a limited area. The algorithm is inspired by the Prim's algorithm for finding minimum spanning trees (MSTs) in undirected graphs [13]. The algorithm is based on certain assumptions. In particular, each AP has its own unique number and the environment is static i.e., the number of APs is fixed and does not change. The coordinates (x, y, z) of each AP within the considered area are known. The new AP can be incorporated into the existing structure just by assigning a proper channel number.

In the MISTI algorithm, a set of APs is modeled as an undirected graph [12], as presented in Fig. 1. Between each AP pair (k, n), interference occurs with strength $F_{k,n} = F_{n,k}$ which depends both on the distance in the physical and channel space. Every pair of distinct vertices is connected by a unique edge, with weights. The values of weights $F_{k,n}$ are calculated according to formula:

$$F_{k,n} = \frac{\delta(C_k, C_n)}{L^2}, \qquad (2)$$

where $\delta(C_k, C_n)$ is the CRC value according to Table 1 [12], C_k is the channel number assigned to AP_k and C_n to AP_n, and L is Euclidean distance between APs in three-dimensional space.

Since the received RF power is inversely proportional in Friis loss model to the square of the distance from the source, L^2 was put in the denominator. It is noteworthy that for the 2.4 GHz band, when $|C_k - C_n| > 3$, $\delta(C_k, C_n) = 0$ and thus, $F_{k,n}$ is equal to zero regardless of the physical distance between the APs. As for the 5 GHz band, $F_{k,n}$ is equal to zero for any channel distance larger than 0 since channels in this band do not overlap. This measure does not include attenuating effects of the building structure and other obstacles, which can be considered by using a greater than 2 exponent.

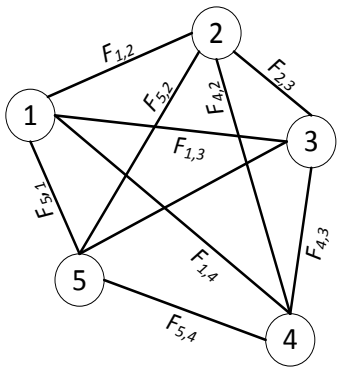


Figure 1 Set of APs as a complete weighted graph

Regarding the task of optimal channel assignment to APs, its primary goal is to minimize the total level of interference between APs, which can be expressed as the sum of interactions between every pair of APs:

$$F_{tot} = \sum_{k=1}^{N-1} \sum_{n=k+1}^{N} F_{k,n}, \quad (3)$$

where N is the number of APs in the set.

A straightforward approach to the channel assignment problem can be the exhaustive testing of all possible combinations of channels assigned to APs, starting from C_{1_min}, C_{2_min}, ..., C_{N_min} to C_{1_max}, C_{2_max}, ..., C_{N_max}, where C_{min} and C_{max} represent the lowest and the highest channel number available and N is the number of APs in the set. However, considering that the set of available channels and the set of APs contain M and N elements, respectively, all N-element variations with repetitions of the M-element set have to be tested. Thus, the number of possible assignments is M^N. It implies the exponential complexity of the exhaustive procedure, which makes it impractical in most applications in spite of its optimality. In contrast, the MISTI algorithm adopts a greedy approach in which the optimization process is carried out locally. As a result, the quadratic complexity is obtained, which is determined by the cost of generating the squared distance matrix.

First, to an arbitrarily chosen AP e.g., no. 1, the lowest channel number available C_{min} is assigned. Next, an AP is found which is the closest to the previously visited one (if there is more than one within the same distance, the one with the lowest number can be selected), and to this AP, the lowest channel number is assigned which minimizes the current total cost F_{tot_c}, which at this point is simply equal to the strength of interaction between the two APs. Next, the previous step is repeated i.e., the search for the closest neighbor of the previously visited AP which does not have a channel number assigned yet, and the lowest possible channel number is assigned according to the same rule – to minimize the up-to-date total cost. This is done by testing all available channel numbers and calculating F_{tot_c}. The process is continued up to the point when all the APs will get a channel number assigned (see Fig. 2). Using this cumulative approach, a final channel assignment is obtained.

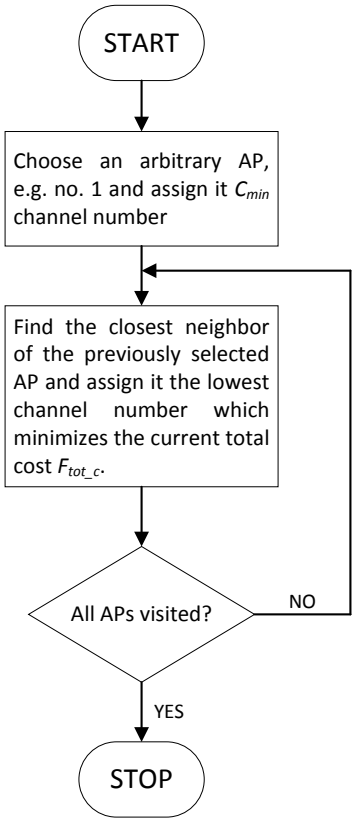


Figure 2. Flowchart of the greedy channel assignment procedure.

IV. Analysis of the Results

Calculations were made for two AP distributions on the 2-D plane. The first distribution, presented in Fig. 3, is a symmetric distribution in which 9 Access Points are located on a square grid.

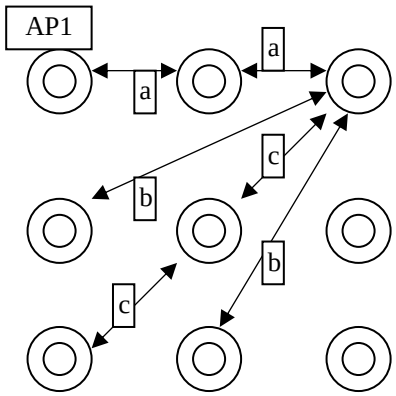


Figure 3 Access Points symmetrical distribution

TABLE III. Distances between APs

a	b	c	2a	2c
10 m	22.36 m	14.14 m	20 m	28.28 m

For the proposed distribution, the values of distances between APs are given in Table III.

It is assumed that the coverage in both bands is the same i.e. the power received at any point is the same. To meet this assumption, it is necessary to select the appropriate transmission power for each band. Since the 2.4 GHz and 5 GHz band gap loss is 6.6 dB, the transmission power in the 5 GHz band is assumed to be 13.3 dBm and in the 2.4 GHz band is 6.7 dBm. Table IV shows the results of the interference calculation as a function of the distance between APs.

The channel allocation for AP distribution presented in Fig. 3 was chosen for both frequency bands (2.4 GHz and 5 GHz) for single channel of 20 MHz width and for channel bonding. Therefore, in the 2.4 GHz band, channels of 20 and 40 MHz width are taken into account and channels of 20, 40, 80, and 160 MHz are taken into account in the 5 GHz band. The calculation was performed with use of both the exhaustive algorithm and the greedy MISTI algorithm [12]. Table V lists the assigned channel distributions for symmetrical AP distribution.

The interference power level of each AP from each AP was counted for channel assignments presented in Table V for the symmetrical distribution and in Table VIII for the non-symmetrical one. Interference power practically depends only on distance between each pair of AP because the transmitting power is chosen in the way which guarantees the same coverage. The results for MISTI-based assignments are shown in Fig. 4, but some characteristics were omitted due to identical shape. The characteristics have four attributes: frequency band, channel width, transmitting power and the list of channel numbers used for a given assignment. The lowest interference power level is obtained when three channels are used while the highest interference power occurs for the channel assignment with only one channel.

TABLE IV. INTERFERENCE POWER VALUES FOR TX(5 GHZ) = 13.3 DBM AND TX(2.4 GHZ) = 6.7 DBM

Distance [m]	Rx [dBm]	Rx [mW×10^{-5}]
10	-53.54	0.44
14.14	-56.55	0.22
20	-59.56	0.11
22.36	-60.54	0.09
28.28	-62.67	0.05

TABLE V. CHANNEL ASSIGNMENTS FOR SYMMETRICAL AP DISTRIBUTION

Configuration	Algorithm	A1	A2	A3	A4	A5	A6	A7	A8	A9
2.4 GHz, 3 channels, 20 MHz	Exhaustive	1	6	11	6	11	1	11	1	6
	MISTI	1	6	11	6	11	1	1	6	11
2.4 GHz, 2 channels, 40 MHz	Exhaustive	1.5	7.11	1.5	7.11	1.5	7.11	1.5	7.11	1.5
	MISTI	1.5	7.11	1.5	7.11	1.5	7.11	1.5	7.11	1.5
5 GHz, 3 channels, 20 MHz	Exhaustive	36	48	64	48	64	36	64	36	48
	MISTI	36	48	64	48	64	36	36	48	64
5 GHz, 4 channels, 20 MHz	Exhaustive	36	44	36	56	64	56	44	36	44
	MISTI	36	44	56	56	36	64	44	64	44
5 GHz, 8 channels, 20 MHz	Exhaustive	36	40	44	48	52	56	44	60	64
	MISTI	36	40	44	56	52	48	60	64	36
5 GHz, 2 channels, 40 MHz	Exhaustive	36.40	52.64	36.40	52.56	36.40	52.56	36.40	52.56	36.40
	MISTI	36.40	52.64	36.40	52.56	36.40	52.56	36.40	52.556	36.40
5 GHz, 3 channels, 40 MHz	Exhaustive	36.40	44.48	60.64	44.48	60.64	36.40	60.64	36.40	44.48
	MISTI	36.40	44.48	60.64	44.48	60.64	36.40	36.40	44.48	60.64
5 GHz, 4 channels, 40 MHz	Exhaustive	36.40	44.48	36.40	52.56	60.64	52.56	44.48	36.40	44.48
	MISTI	36.40	44.48	52.56	52.56	36.40	60.64	44.48	60.64	44.48

If we use three non-overlapping channels, the interference level will be the same in the 2.4 GHz band for 20 MHz and in the 5 GHz band for 20 and 40 MHz channels. In the case of 2.4 GHz band it is necessary to use channels 1, 6 and 11. Both algorithms give almost the same assignments with one difference for AP 7-9 where the channel 1 was replaced by the channel 6, 6 by 11 and 11 by 1 (see Table V). These changes however did not change the interference distribution. The highest interference level for both assignments is for the distance c = 14.14 m and such a situation occurs four times for channel distributions obtained with both algorithms. Interference power levels for both distributions are presented in Fig. 4. Interference power level varies from -48 to about -56 dBm.

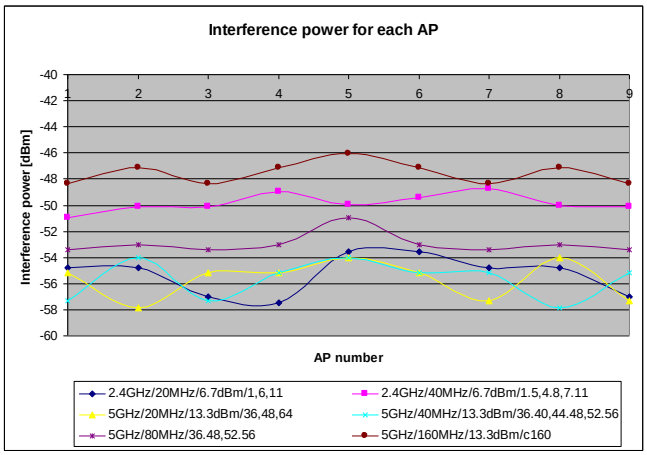


Figure 4 Interference power for each AP for symmetrical AP distribution for MISTI algorithm

In the 5 GHz band, there is more flexibility. We can use any three out of eight non-overlapping 20 MHz channels and any three out of four non-overlapping 40 MHz channels. It gives a number of 56 combinations for 20 MHz channels and 4 combinations for 40 MHz channels.

All assignments for three available channels for 20 and 40 MHZ width channels in the 5GHz band and for 20 MHz width channels in the 2.4 GHz band give the same interference distribution. We have to accept only co-channel interference and no adjacent–channel interference is present.

If we use more than three 20 MHz width channels or any number of combinations of 40 MHz channels in the 2.4 GHz band, the interference level will be higher than in the 5 GHz band for the same number of channels. This is the result of adjacent-channel interference which will be present for 4 and more 20 MHz channels in the 2.4 GHz band. In this situation, channels overlap and in dependence on their combination, the interference level can be different; however, always higher than in the case of the 5 GHz band. With 40 MHz channels in the 2.4 GHz band, the interference issue arises as no non-overlapping channels are available. Thus, the adjacent channels interference will be always present in the 2.4 GHz band for 40 MHz channels.

The interference fluctuations for given characteristics is within the range of 2 - 4 dBm what is the result of differences in distances between APs.

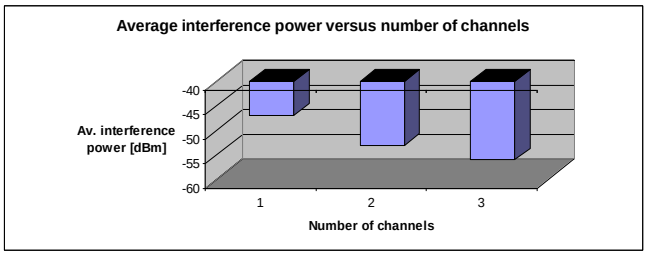


Figure 5 Average interference power versus the number of used channels

As the number of channel increases, the interference level decreases. The level of average interference for the symmetrical distribution is presented in Fig. 5. The difference is quite significant and could reach even 10 dBm between the scenario with one and three channels available.

A more realistic AP distribution was implemented on the plan of a selected floor of a multi-apartment building presented in Fig. 6.

For the calculation of AP location coordinates, the position of the left corner of the building has been taken as the origin of the coordinate system. The calculated coordinates are presented in Table VI.

The distances between all APs were calculated as well as the interference power. The results are presented in Table VII and VIII, respectively.

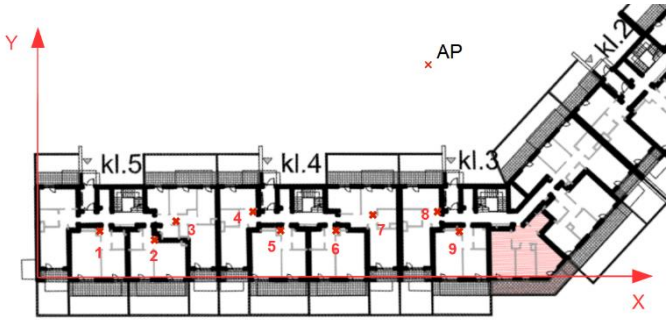


Figure 6 The AP distribution in an actual multi-flat building

TABLE VI. AP X, Y COORDINATES FOR NON-SYMMETRICAL DISTRIBUTION

AP number	X [m]	Y [m]
1	8.08	5.92
2	15.2	4.80
3	18.0	7.2
4	32.0	8.4
5	35.6	10.0
6	38.8	10.0
7	43.6	8.0
8	52	8.4
9	54.84	5.88

For the non-symmetrical AP distribution, presented in Fig. 6, the channel assignment was calculated. The results are presented in Table IX. The results of interference power distribution are presented in Fig. 7. The characteristics have four attributes: frequency band, channel width, list of used channels and the type of used algorithm. The lowest interference power level was achieved for eight available channels in the 5 GHz band for 20 MHz channel width. The interference power level is -78dBm for AP 1 and AP 9 and -100dBm for the rest. The highest interference power occurs in the situation when only two overlapping channels are used in the 2.4GHz band. The 20 MHz channels were bonded to achieve 40 MHz channel.

TABLE VII. DISTANCES BETWEEN APS IN THE NON-SYMMETRICAL DISTRIBUTION

	A2	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
A1	7.21	10.00	24.05	27.82	30.99	35.58	43.99	46.76
A2		3.69	17.18	21.05	24.17	28.58	36.98	39.65
A3			14.05	17.82	20.99	25.61	34.02	36.86
A4				3.94	6.99	11.61	20	22.98
A5					3.2	8.25	16.48	19.68
A6						5.2	13.30	16.56
A7							8.41	11.44
A8								3.80

TABLE VIII. INTERFERENCE POWER FOR EACH AP [DBM] FOR NON- SYMETRICAL DISTRIBUTION

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
A1		-51.0	-53.9	-61.5	-62.7	-63.7	-64.9	-66.7	-67.3
A2			-45.2	-58.5	-60.3	-61.5	-63.0	-65.2	-65.8
A3				-56.8	-58.9	-60.3	-62.0	-64.5	-65.2
A4					-45.8	-50.7	-55.1	-59.9	-61.1
A5						-44.0	-52.2	-58.2	-59.7
A6							-48.2	-56.3	-58.2
A7								-52.4	-55.0
A8									-45.4
A9									

TABLE IX. CHANNEL ASSIGNMENTS FOR THE NON-SYMMETRICAL AP DISTRIBUTION

Configuration	**Algorithm**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
2.4 GHz, 3 channels, 20 MHz	Exhaustive	1	6	11	1	6	11	1	6	11
	MISTI	1	11	6	1	11	6	1	11	6
2.4 GHz, 2 channels, 40 MHz	Exhaustive	1.5	7.11	1.5	7.11	1.5	7.11	1.5	7.11	1.5
	MISTI	1.5	7.11	1.5	7.11	1.5	7.11	1.5	7.11	1.5
5 GHz, 3 channels, 20 MHz	Exhaustive	36	48	64	36	48	64	36	48	64
	MISTI	36	64	48	36	64	48	36	64	48
5 GHz, 4 channels, 20 MHz	Exhaustive	36	44	56	64	36	44	56	64	36
	MISTI	36	44	56	64	36	44	56	64	36
5 GHz, 8 channels, 20 MHz	Exhaustive	36	40	44	48	52	56	60	64	36
	MISTI	36	64	60	56	52	48	44	40	36
5 GHz, 2 channels, 40 MHz	Exhaustive	36.40	52.56	36.40	52.56	36.40	52.56	36.40	52.56	36.40
	MISTI	36.40	52.56	36.40	52.56	36.40	52.56	36.40	52.56	36.40
5 GHz, 3 channels, 40 MHz	Exhaustive	36.40	44.48	60.64	36.40	44.48	60.64	36.40	44.48	60.64
	MISTI	36.40	60.64	40.48	36.40	60.64	40.48	36.40	60.64	40.48
5 GHz, 4 channels, 40 MHz	Exhaustive	36.40	44.48	52.56	60.64	36.40	44.48	52.56	60.64	36.40
	MISTI	36.40	60.64	52.56	44.48	36.40	60.64	52.56	44.48	36.40

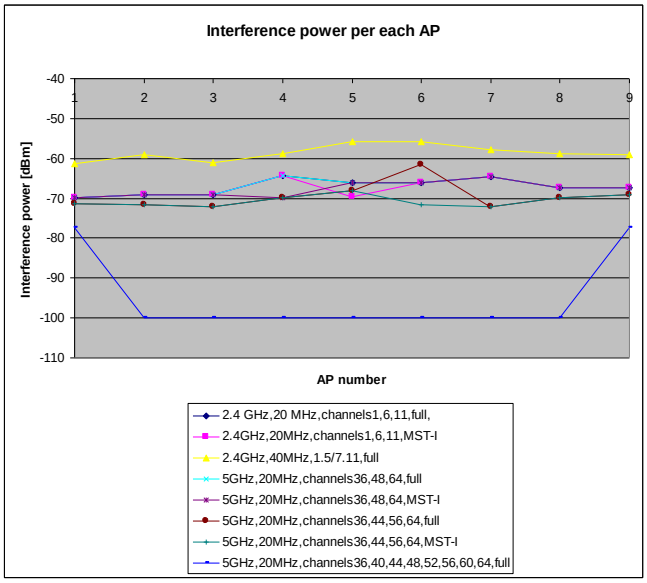


Figure 7 Interference power for each AP for non-symmetrical AP distribution for Ptx (5GHz)= 13 .3dBm, Ptx(2.4 GHz)=6.7 dBm

The highest difference in terms of average interference levels is between eight 20 MHz channel set in the 5 GHz band and two bonded 40 MHz channels with severe adjacent channel interference issue for the 2.4 GHz band. This difference reaches about 30 dBm value.

V. Conclusions

The 5 GHz band has a number of potential advantages from the point of view of interference. The large amount of non-overlapping channels and the availability of wider channels provide much greater network planning capabilities.

Obviously, it is not true that the interference power in the 5 GHz band is always lower than in the 2.4 GHz band. If we had three or less APs, the interference power distribution could be higher, lower or similar in any band. In such a case, the final result depends on the level of transmitted power and distances between APs. In environment with high station and AP density, the 5 GHz band seems to have more advantages than the 2.4 GHz band. However, the use of 5 GHz networks in high AP density locations must be coordinated; otherwise the benefits of this bandwidth may remain unused, e.g. when the same channel is used by many APs. The best solution in a multi-apartment building would be to assign a 2.4 GHz band channel number and a 5 GHz band channel number to the apartment number after channel assignment optimization. In Poland, the 2.4 GHz band continues to dominate due to the large number of 802.11b and g devices. In some situations, the 2.4 GHz band may be more advantageous, for example, when there is a low number of APs in a given area and larger range is required. It is advisable to perform environmental measurements to determine the amount of APs and to choose the best solution for a given location.

It will be advantageous to use any simulator to analyze the interference power distribution. Today the only one with the possibility of Phy layer analysis is network simulator NS-3. The newest version NS-3.26 and the latest samples enable some interference simulation but the scope of analysis is far from the need. The sample "wifi-spectrum-per-interference" available on www.nsnam.org_tab_examples/wireless [14] let us calculate the SINR but only for one pair AP and STA with the use of external interfering source which is not communication device. More complex sample has to be developed to support both channel assignment and interference power distribution estimation.

References

[1] What is the difference between 2.4 GHz and 5 GHz wireless frequencies? https://kb.netgear.com/29396/What-is-the-difference-between-2-4-GHz-and-5GHz?cid=wmt_netgear_organic access 03.2017

[2] Ł. Chruszczyk, A. Zając and D. Grzechca, Comparison of 2.4 and 5 GHz WLAN Network for Purpose of Indoor and Outdoor Location, International Journal of Electronics and Telecommunications, 2016, VOL. 62, NO. 1, PP. 71-79

[3] Nilsson Rolf, Compare 2.4 GHz and 5 GHz Wireless LAN in Industrial Applications, https://www.digikey.com/en/articles/techzone/2012/may/what-is-the-difference-between-24-ghz-and-5-ghz-wireless-lan-in-industrial-applications access 03.2017

[4] http://netgear-us.custhelp.com/app/answers/detail/a_id/29396/~/what-is-the-difference-between-2.4-ghz-and-5ghz%3F access 03.2017

[5] Zachary Hays, Grant Richter , Stephen Berger , Charles Baylis , Robert J. Marks,, Alleviating airport WiFi congestion: An comparison of 2.4 GHz and 5 GHz WiFi usage and capabilities, Wireless and Microwave Circuits and Systems (WMCS), 2014 Texas Symposium on

[6] Geier J., (2008), Whichn is better: 802.11n 2.4 or 5 GHz?, http://www.linuxplanet.com/linuxplanet/reports/6581/1 access 03.2017

[7] Angelakis, V.; Papadakis, S.; Siris, V.A.; Traganitis, A., Adjacent channel interference in 802.11a is harmful: Testbed validation of a simple quantification model, "Communications Magazine. IEEE. 49 (3), March 2011

[8] Garcia Villegas, E.; et al. , Effect of adjacent-channel interference in IEEE 802.11 WLANs . CrownCom 2007. ICST & IEEE.

[9] I.Dolińska, A.Masiukiewicz, G.Rządkowski, The mathematical model for interference simulation and optimization in 802.11n networks, Workshop, Concurrency Specification and Programming 2013, Warsaw University

[10] Dolińska I., Masiukiewicz A., Rządkowski G., Channel Selection in Home 802.11 Standard Networks, Digital Technologies Żylina 2014

[11] I. Dolińska, M. Jakubowski, A. Masiukiewicz, G. Rządkowski, Algorithms for Channels Assignment in 802.11 Networks, IDT Conference 2016, pp .83-89, IEEE Catalog Number CFP16CDT-USB

[12] Iwona Dolińska, Mariusz Jakubowski, Antoni Masiukiewicz, Grzegorz Rządkowski, Kamil Piórczyński A Minimum-Spanning-Tree-Inspired Algorithm for Channel Assignment in 802.11 Standard, International Journal of Electronics and Telecommunications , 4/2016 pp. 379-388

[13] D. A. Marcus – Graph Theory: A Problem Oriented Approach, Mathematical Association of America, 2008.

[14] https://www.nsnam.org/doxygen/wifi-spectrum-per-interference_8cc_source.html access 02.2017

Throughput Efficiency in IEEE 802.11n Networks

Iwona Dolińska, Mariusz Jakubowski, Antoni Masiukiewicz
Vistula University (AFiBV)
Stokłosy 3, 02-787 Warsaw, Poland
i.dolinska@vistula.edu.pl; m.jakubowski@vistula.edu.pl; a.masiukiewicz@vistula.edu.pl

***Abstract*— Throughput in IEEE 802.11 networks depends on a large number of factors related to the standard version, transmission conditions in the radio channel and a number of users of the network. In addition, the bit rate obtained for the actual conditions is significantly different from the theoretical maximum bit rate defined in the standard as it can be only the part of the maximum bit rate. The authors conducted an analysis of the efficiency of throughput in a Wi-Fi network compatible with the IEEE 802.11n, standard using the latest version of the NS-3 simulator. NS version 3.26 allows, among others, to analyze the impact of interference on the obtained bit rate, using a new model of the Wi-Fi physical layer called SpectrumWifiPhy.**

Keywords— NS-3 simulator, interference, throughput

I. INTRODUCTION

The value of the current throughput in any IEEE 802.11 network is variable in time as it depends on a very large number of factors. Theoretical throughput is reduced, among others, as a result of the need to transfer control, signaling, confirming data, and performing retransmission due to the phenomenon of collision. The last one is the effect of the method used for granting access to the radio channel in the MAC layer. DCF (Distributed Coordination Function) mechanism with the CW (Contention Window) value drawing may lead to a large number of collisions especially where there are many users connected to the network [1]. The amount of collision can even reach 50% for 10 users. The practical throughput is also very dependent on the conditions in the radio channel at a given point in time especially on both fast and slow power level changes in the received signal. Fig. 1 summarizes the factors affecting the practically achieved throughput.

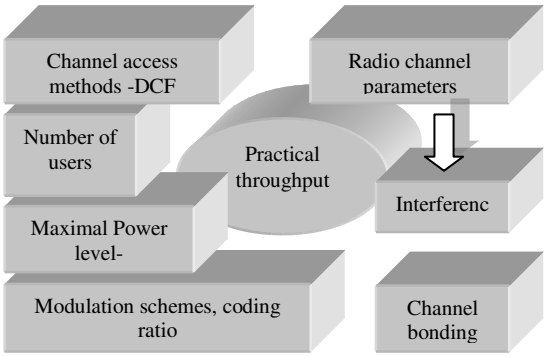


Figure 1. Factors of IEEE 802.11 theoretical throughput limitation

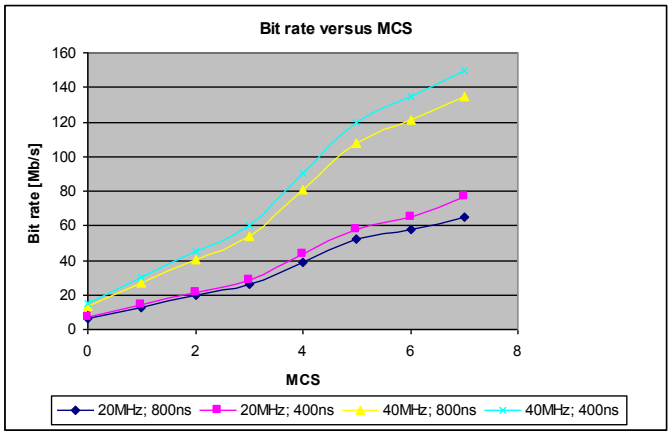


Figure 2. The maximum instantaneous theoretical bit rate in IEEE 802.11n standard for: guard interval 400/800 ns, channel width 20/40 MHz and modulation type BPSK (MCS=0), QPSK (MCS=1;2), 16QAM (MCS=3;4) and 64QAM (MCS=5;6;7)

Source: Juniper Networks, (2011), Coverage or Capacity-Making the Best Use of 802.11n, Juniper Networks Inc. 2011

The first parameter, the maximum theoretical bit rate of the device is usually defined by a standard version [2], but the value of the throughput obtained in practice can be defined rather by using statistical methods and the adoption of certain assumptions: eg. the time of day (important for links with long range), number of simultaneous users, the average load on the user. For 802.11n, the theoretical bit rate varies in the range of 6.5 Mb/s to 150 Mb/s for a SISO (Single Input Single Output) transmission type depending on MCS (Modulation and Coding Scheme). MSC define the permutation of code rate, channel width, the width of the guard interval and the type of modulation. Fig. 2 shows the maximum bit rate for 802.11n as a function of these parameters.

A number of factors affect the throughput achieved in practice, and the most important of them include: station distance from the access point (AP), the frequency of the radio channel, co-channel and adjacent channel interference, the number of stations connected to the AP, the amount of traffic, the number of collisions and the conditions in the radio channel, which in turn, are associated with numerous parameters. The basic theoretical Friis model of losses in the channel is generally an ideal model. To reflect the real conditions, dozens of models were built and a number of amendments to the Friis model were introduced [3,4]. Several

models have been designed within the framework of the simulator NS-3, among others, FriisPropagationLossModel and LogDistancePropagationLossModel [6]. The balance of power at the receiving point in free space according to the Friis model [4], using a logarithmic scale can be determined by the following formula:

$$P_{Rx}(r) = P_{Tx}[dBm] + G_{Tx}[dB] + G_{Rx}[dB] - L_{fspl}[dB] \quad (1)$$

where: $P_{Rx}(r)$ is the received power level, P_{Tx} transmitted power level, G_{Tx} i G_{Rx} transmitting and receiving antenna gains, L_{fspl} free space loss. L_{fspl} value can be calculated for r in kilometers and f in megahertz using the following formula [4]:

$$L_{fspl}[dB] = 32.44 + 20\log r[km] + 20\log f[MHz] \quad (2)$$

Friis loss model implies change of the signal level with the square of the distance. LogDistancePropagationLossModel is a more general model and allows the formation of the slope of decline in power as a function of the distance of the radio channel [5,6]. To calculate the loss of a radio channel, the following relationship is used:

$$L = L_0 + 10n\log(\frac{d}{d_0}) \quad (3)$$

where

n the path loss distance exponent,
d_0 reference distance (m),
L_0 path loss at reference distance (dB),
d distance (m),
L path loss (dB).

Both models have been used by the authors in the simulation. The authors performed simulations of IEEE 802.11n network using the latest version of the simulator NS-3.26. For analysis, the example Wifi-Spectrum-Per-Interference [8], made available in 2016 working with this version of the simulator, was used. Practical throughput to the theoretical throughput for the 802.11n standard was measured and compared. The calculations were performed for selected attributes such as signal propagation channel different models, the level of SINR (signal to interference and noise ratio), the distance between the station and the access point.

II. Related Works

The parameters of the radio channel are the first physical limitations of the throughput of the wireless system. In the first approximation, the theoretical maximum rate of a single channel can be estimated on the basis of several parameters, as follows: the transmit power level, the transmission frequency, the width of the radio channel, the distance between the transmitter receiver gain receiving antenna, thermal noise, and 1/f noise of the receiver. At the ambient temperature, throughput can be expressed with the following equation [4]:

$$C \approx Bw\lg_2 10^{\frac{P_{Tx} - 20\lg_{10} r - 20\lg_{10} f - 32{,}44 + G_{Rx}}{10(-173{,}98 - 10\lg_{10} Bw - NF)}} \quad (4)$$

where

P_{Tx} transmitted power,
C maximal theoretical throughput of the given channel (Mb/s),
Bw channel width (MHz),
NF Noise Figure - 1/f noise of the receiver (dB),
r distance between two stations (km),
f signal frequency value (MHz),
Go gain of the receiving antenna (dB).

Mechanisms for media access within the 802.11n standard have a significant impact on the throughput achieved in practice [9]. Standard is based on the assumption that the workstations are awaiting the release of the channel and after a specified period of inactivity in the channel they may attempt to gain access. Communication takes place within the standard limits the bandwidth in two ways. Firstly, the need to transfer control data, acknowledgments, retransmission, headers and preambles produces the additional data. Secondly, the standard used predefined periods of time separating the different activity within the standard. These periods result in dead time from the point of view of transmission possibility. Some of the factors are taken into account when determining the expected throughput [10] expressed by the formula:

$$E_{Th} = \frac{K \cdot L_{data} \cdot PRR}{DIFS + T_{BO}(PRR) + T_{Kdata} + SIFS + T_{ACK}} \quad (5)$$

where

E_{Th} expected throughput,
PRR Packet Reception Rate,
K aggregated packets number,
L_{data} data within the frame (bits),
$DIFS$ time interval (µs),
T_{BO} back off time (µs),
T_{Kdata} A-MPDU total transmission time with headers and other necessary information for given PHY rate (µs),
$SIFS$ time interval (µs),
T_{ACK} time for ACK or Block ACK (µs).

This formula includes three groups of factors. Some factors such as T_{DIFS}, T_{SIFS}, T_{BO} represents the time (wasted or dead time) when all stations are just waiting. T_{ACK} is the time for sending confirmation (acknowledgment frame). The second group of factors concerning the data structure are K, which is the number of aggregated frames and L_{data}, which is the payload carried per frame. The last group concerns selected parameters of the transmission e.g. PRR - the packet reception rate. The formula corresponds rather to the average value because it takes into account the average T_{BO}. One cannot calculate real transmission time, because T_{BO} value changes in every communication transaction. The formula (5) doesn't discuss the collisions, *RTS/CTS* (Request to Send/Clear to Send) messages time and the *SIFS* different share in successful transmissions versus collisions.

The *IFS* time intervals, separating every frame transmission, do not reduce the theoretical maximum throughput, but they reduce total throughput per time unit in the time domain. During some periods of time the radio channel is not working, waiting, although the radio channel is idle. In the case of

communication sessions, the dead time is the sum of T_{DIFS}, 3 T_{SIFS} and T_{BO} [16]. Some changes are introduced to improve the dead time problem. In 802.11ac maximum A-MPDU (aggregate MAC protocol data unit) length is increased from 65 535 octets to 1 048 575 octets. Also, the *RIFS* (reduced interframe space) time interval is not used any more. It is very difficult to estimate the loss of throughput resulting from both mechanisms mentioned above. Some authors suggest that the transmission of whole control data can reduce the throughput available for data by even about 50% [15]. The evaluation of a total throughput decrease is difficult, because it depends on the upper layer transmission type, e.g. UDP or TCP [14].

The NS-3 simulator is, in this regard, a tool that gives results much closer to real. You cannot avoid interference related to a Wi-Fi network. The IEEE 802.11 networks themselves produce harmful interference. Predominant are two mechanisms associated with interference in the channel used. Interference can come first from other stations operating on the same channel and from stations operating on adjacent channels. This type of interference is called co-channel and adjacent channel interference, respectively. The 2.4 GHz band allocated for 802.11a, b, g, n, and ax has width of 83 MHz only. This band is divided into channels of 20 MHz width, which carriers are shifted by 5 MHz. The spectra of individual channels overlap each other. As a result, there are only three non-interfering channels offset by 25 MHz in the frequency domain. At the x, y plane, the reduction of interference inside-channel is gained through proper distribution stations and appropriate design of channel allocation. In home networks using 802.11 standard, we have to deal with a difficult situation. Due to the lack of a license and limited protection from other users on a relatively small area (buildings, multi-family houses, office buildings) can be multiple networks using Wi-Fi technology, which projects were not in any way coordinated [11].

III. NS-3 SIMULATOR

Simulator NS-3 is an advanced network simulator. Performing the simulations with the NS-3 requires user knowledge on issues such as queuing, modeling and programming with the script. The result is that, in principle, the NS simulator can be used for scientific studies. Version NS3 is recognized as a basic tool for testing LAN, WAN and Wi-Fi [8]. One of the most important devices (netdevice) in NS-3 is a device of the Wi-Fi type. It represents the largest module in the NS-3. WifiNetDevice solution implements the IEEE 802.11 standard. It allows you to simulate different versions of MAC and PHY. Architecture of the IEEE 802.11 standard implemented in NS-3 was made on the basis of the solution used in the simulator Yans (Yet Another Network Simulator) by Mathieu Lacage and Tom Henderson [12]. The model developed in the NS-3 is very complex and contains 75 objects and many features. Work on the development of the simulator lasts continuously, and new versions are available almost every year. The current version of the simulator has the number 3.26. A number of new examples can be used with this version. Version 3.26 has implemented a new model of the physical layer, in addition to YansWifiPhy SpectrumWifiPhy can be currently used [7]. The new model of the physical layer allows for the analysis of parameters of the radio channel interference from other stations or resulting from the actions of other systems. To estimate the possibility of receiving packets, the SINR (Signal to Interference and Noise Ratio) factor is used. This factor is expressed in terms of:

$$SINR = \frac{P_{signal}}{P_{noise} + P_{\text{interference}}}[dB] \quad (6)$$

where

Pnoise white Gaussian noise (dBm),

Pinterference interference Power level (dBm).

Interference can come from other 802.11 stations, or other devices which also operate on frequencies used by this standard.

IV. WIFI-SPECTRUM-PER-INTERFERENCE EXAMPLE

Example "wifi-spectrum-per-interference" available on www.nsnam.org_tab_examples/wireless [13] was chosen for simulation with some small modifications introduced. Because the model allows the use of both available PHY models, it requires installation of the latest available version of the simulator with the number 3.26. The model contains three signal sources, two of them: the Access Point and the Workstation are Wi-Fi network devices. The third device is the source of interference. Wherein it is possible to change two parameters of the source: the power level and the position of this source on the XY plane. Distribution of the objects is shown in Fig. 3.

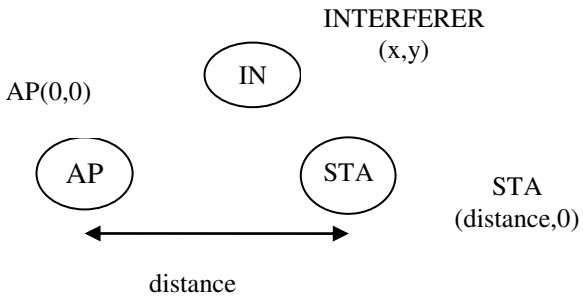


Figure 3. Objects distribution on X,Y plane

TABLE I. BASIC PARAMETERS OF THE SIMULATED NETWORK

Parametr	Value	Options
Network ssid	ns380211n	
Channel frequency [MHz]	5180	
Standard	802.11n	
Band [GHz]	5	
Channel width [MHz]	20	40
Guard interval [ns]	800	400
Number of streams	1	
Modulation	BPSK;QPSK;QAM	
MCS	0-7	
Network IP address	192.168.1.0	
Flow	tcp	udp
Phy model	SpectrumWifiPhy	YansWifiPhy
Loss model	FriisPropagationLossModel	
STA Tx [dBm]	16	

At the power level 0.0001 [W], the source of interference is off. The level of interference power at the receiving point, where the AP is located, can be adjusted through the level of interference power, which is waveformPower = [level of interference power] or by selecting a position of the source in the plane, or by changing the value of both parameters. Test network has a number of parameters which are listed in Table I.

During the simulation, it is possible to change the number of attributes. Some of them can be changed using commands available from the command line, part of them require changes making in source code of program.

V. THROUGHPUT EFFICIENCY ANALYSIS

A series of simulations were carried out for different values of the selected parameters. The practically achieved throughput as the function of the parameter changes was calculated. Table II summarizes the ranges of attribute variations used in analysis.

TABLE II. THE RANGE OF ANALYSIS ATTRIBUTES VARIATION

Atribut	Range
Guard interval [ns]	400; 800
Channel width [MHz]	20; 40
Distance [m]	2÷500
waveformPower [W]	0.0001÷1
x,y [m]	2-100
Loss model	FriisPropagationLossModel; LogDistancePropagationLossModel

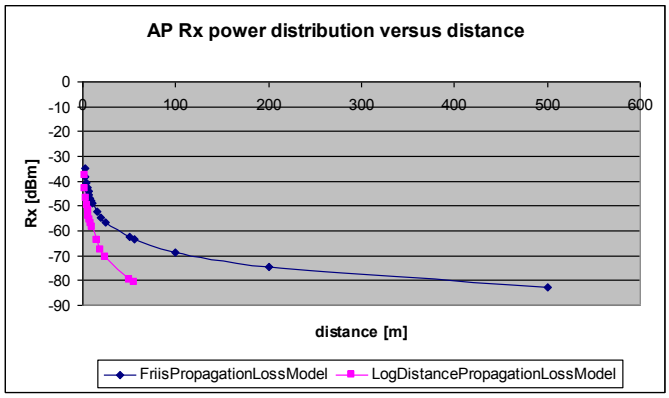


Figure 4. AP Rx power distribution for Tx=16 [dBm]

Power level of the interfering station was changed to obtain a certain level of interference power at the receiving point. Fig. 4 shows the of the power receiving by STA as a function of the distance between the AP and the STA for the two different models of signal propagation in the channel.

A sample of simulation results is presented in fig. 5. First, we have to set the values of selected parameters through the command line and then start the simulation. The following sequence of command have to be set:

access path to ns -3.26 filename cd/…… / ns-3.26$

./waf --run "wifi-spectrum-per-interference --simulationTime=2 --udp=0 --distance=10 --enablePcap=1 --waveformPower=0.1"

As a result, at first, we obtained confirmation of example building and than same parameters such as distance, number of sent packet and the value of TxPower are returned.

```
antoni@antoni-VirtualBox:~/workspace1/ns-allinone-3.26/ns-3.26$
 ./waf –run "wifi-spectrum-per-interference --simulationTime=2 –udp=0
 --distance=10 –enablePcap=1 --waveformPower=0.1"
Waf: Entering directory `/home/antoni/workspace1/ns-allinone-3.26/ns-
3.26/build'
Waf: Leaving directory `/home/antoni/workspace1/ns-allinone-3.26/ns-
3.26/build'
Build commands will be stored in build/compile_commands.json
'build' finished successfully (7.689s)
wifiType: ns3::SpectrumWifiPhy
distance: 10m; sent: 1000 TxPower: 16 dBm (40 mW)
```

index	MCS	Rate (Mb/s)	Tput (Mb/s)	Received packets	Signal (dBm)	Noi+Inf (dBm)	SNR (dB)
0	0	6.5	4.08336	705	-58.6571	-78.9711	20.314
1	1	13	9.37146	1618	-58.6571	-78.9376	20.2805
2	2	19.5	14.4742	2499	-58.6571	-78.9262	20.2691
3	3	26	11.862	2048	-58.6571	-78.9311	20.274
4	4	39	0	0	N/A	N/A	N/A
5	5	52	0	0	N/A	N/A	N/A
6	6	58.5	0	0	N/A	N/A	N/A
7	7	65	0	0	N/A	N/A	N/A
8	0	7.2	4.58726	792	-58.6571	-78.9648	20.3077
9	1	14.4	10.3329	1784	-58.6571	-78.9347	20.2776
10	2	21.7	15.6384	2700	-58.6571	-78.9247	20.2676
11	3	28.9	13.6286	2353	-58.6571	-78.9274	20.2703
12	4	43.3	0	0	N/A	N/A	N/A
13	5	57.8	0	0	N/A	N/A	N/A
14	6	65	0	0	N/A	N/A	N/A
15	7	72.2	0	0	N/A	N/A	N/A
16	0	13.5	9.7711	1687	-58.6571	-78.9363	20.2792
17	1	27	20.6543	3566	-58.6571	-78.9196	20.2625
18	2	40.5	31.0799	5366	-58.6571	-78.914	20.2569
19	3	54	32.3889	5592	-58.6571	-78.9134	20.2563
20	4	81	0	0	N/A	N/A	N/A
21	5	108	0	0	N/A	N/A	N/A
22	6	121.5	0	0	N/A	N/A	N/A
23	7	135	0	0	N/A	N/A	N/A
24	0	15	10.6341	1836	-58.6571	-78.9342	20.2771
25	1	30	22.5946	3901	-58.6571	-78.9181	20.261
26	2	45	34.6651	5985	-58.6571	-78.9128	20.2557
27	3	60	35.4297	6117	-58.6571	-78.9124	20.2553
28	4	90	0	0	N/A	N/A	N/A
29	5	120	0	0	N/A	N/A	N/A
30	6	135	0	0	N/A	N/A	N/A
31	7	150	0	0	N/A	N/A	N/A

Figure 5. Simulation results sample, for following simulation parametrs: simulationTime=2s, flow=tcp, distance=2m, waveFormPower=0.1W, wifitype=spectrumWifiPhy, lossModel=FriisLossPropagationModel, MCS=0-7, channel width and guard interval are included in index value

The simulation results are placed in sixth columns which are describe as: *index, MCS, Rate, Tput, Received, Signal, Noi+Inf, SNR*, where *index* shows the available simulation version from the point of view of MCS, *channel width* (20 or 40 MHz) and *guard interval* (400 or 800 ns), rate is a theoretical maximal rate of IEEE 802.11n standard for a given index, *Tput* is a practically obtained throughput, *Noi+Inf* represents the sum of white noise and interference while *SNR* is in practice SINR coefficient as it consider both noise and interference influence. If we set *enablePcap* value as true (=1)

we can also see all the communication within the simulation, separate file is produced for each index, using *tcpdump* pocket analyzer.

The sample simulation shows that Tput decrease significantly for MCS =3 and drops to 0 for higher MCS. This is the result of interference presence and SNR value which only slightly above 20 dB.

LogDistanceLossPropagationModel enable setting the value power coefficient of loss versus distance. Default value of power coefficient is 3. With this model, the receiving power decrease in much quicker as a function of distance than for Friis model. Fig. 6 shows the bit-rate efficiency in %, which was determined using the following formula:

$$ET[\%] = 100\frac{Th_{practical}}{Rate_{\max\ theoretical}} \quad (7)$$

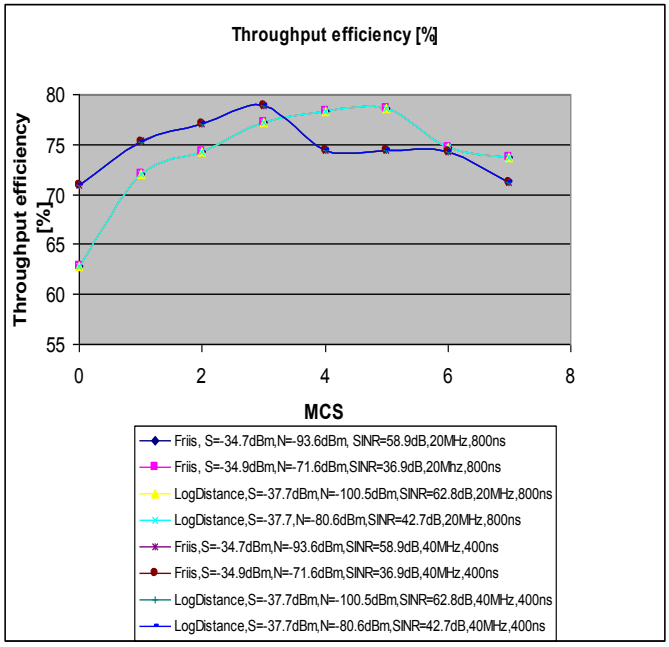


Figure 6. Throughput efficiency for distance=2 m

This efficiency rate when the distance of the station from the AP is 2m changes in the range of 62-78%. You can not see the impact of interference. Theoretically, higher rate for 40MHz channel width and 400ns guard interval does not result in greater efficiency, which is better for the channel 20MHz and 800ns guard interval for higher modulation schemas.

Fig. 7 shows the change in efficiency as a function of distance for the Friis model, the width of the channel 40MHz and 400ns guard interval. Effectiveness rate is maintained between 70-80%; however, when the signal level drops below a certain level, the modulation scheme becomes unavailable. Fig. 8 shows the decrease in the signal strength as a function of the distance and the theoretical power levels received conditioning the possibility of using the modulation level.

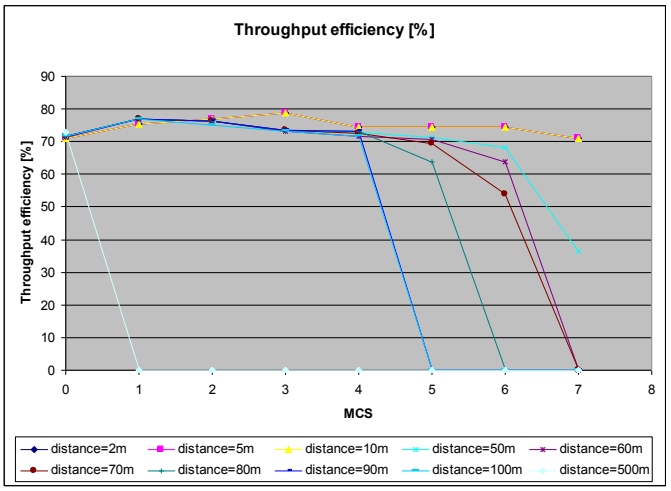


Figure 7. Throughput efficiency for Friis loss model, channel width 40MHz , guard interval = 400ns.

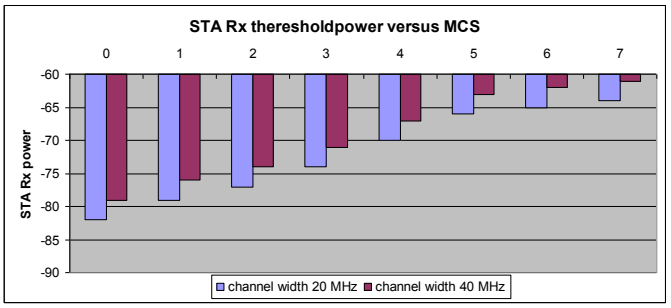


Figure 8. Theresholds minimal power required for given MCS

When the input voltage drops in the vicinity of the threshold value, one can no longer continue to use the given modulation scheme.

Throughput efficiency decreases with increasing interference level when a constant level of the signal received at the same time causes a decrease of SINR ratio. This results a large variation of the throughput efficiency which in consequence vary from 1 to 77%. For some number of modulation schemes, it is not possible to make the transmission. According to the literature, the SINR level necessary for the accurate transmission must take at least the values shown in Table III.

TABLE III. REQUIREMENTS FOR SINR VALUES

Modulation	Minimal SINR [dB]
BPSK	5
QPSK	8
16QAM	15
64QAM	25

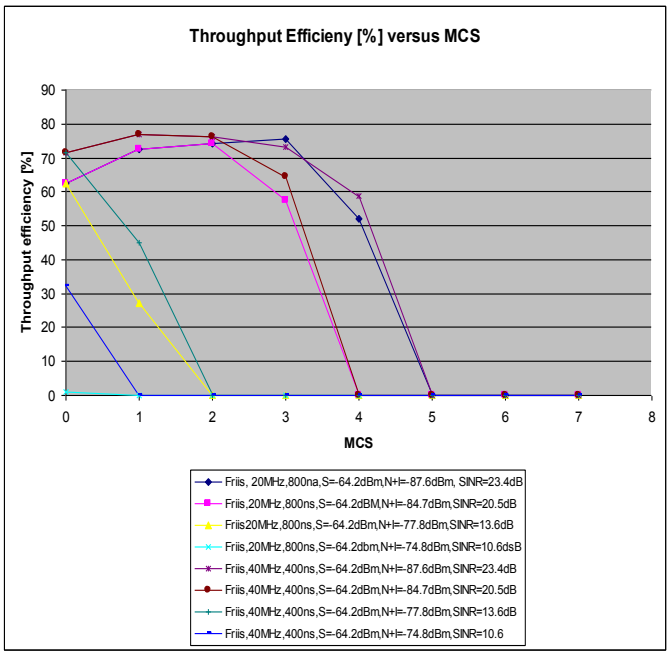


Figure 9. Throughput efficiency with interference presence, for Friis loss model, and distance = 60m

The simulation proves that there are some differences between values in Table III and simulation results. For MCS=3, 4 we have the 16QAM modulation, so the threshold in Table III is 15dB. However, the throughput decreases significantly for MCS = 3 and drops to 0 for MCS = 4 despite the fact that both the value of Signal level = -58dBm and SINR = 20dB are above requirements of successful transmission as it is shown in Fig. 5.

VI. Conclusions

The implementation of the PHY layer in the form of SpectrumWifiPhy significantly increases the possibility of interference analysis using the simulator NS-3. Until now it was not possible to take into account the interference within the channel, and inter-channel interference from devices operating in other systems. The analysis results show that when taken into account, the interference significantly alters the transmission conditions and even transmission possibility. We can observe the influence of interference increase. The practical obtained throughput efficiency is between 70 and 80% of maximum theoretical rate. If critical transmission parameters such as received signal power level and SINR are close to border values, the throughput significantly decreases and obtained throughput efficiency can drop to very small values eg. 1%. Some results show significant discrepancy between theoretical threshold of 802.11n standard and the simulation especially for minimum SINR requirements.

References

[1] Dolińska I., Masiukiewicz A., Rządkowski G., (2014), Collisions in DCF scheme used In 802.11 standard networks, Telecommunications Rewiev 2-3/2014

[2] Juniper Networks, (2011), Coverage or Capacity-Making the Best Use of 802.11n, Juniper Networks Inc. 2011

[3] Hodgkinson T.G, (2007), Wireless communications- the fundamentals, BT Technology Journal , Vol 25 No 2, April 2007

[4] Freeman Roger L., (2007), Radio system design for telecommunication, J.Willey&Sons, 2007

[5] Bingmann T., (2009), Accuracy Enhancements of the 802.11 Model and EDCA QoS Extensions in ns-3, Diploma Thesis at the Institute of Telematics Faculty of Computer Science University of Karlsruhe, 2009

[6] NS-3 Network Simulator, NS-3 Model Library, Release ns-3 dev, NS-3 Project, (2017), February 01.2017, www.nsnam.org access 02.2017

[7] Baldo N., Miozzo M., (2009), Spectrum-aware Channel and PHY layer modeling for ns3, NSTOOLS Italy 2009

[8] The ns-3 network simulator, (2016), http://www.nsnam.org access: 02.2017

[9] Akhavan M.R., (2006), Study the performance limits of IEEE 802.11 WLAN's, Master Thesis, Lulea University of Technology, Sweden 2006

[10] Deek L., Garcia-Villegas E., Belding† E.,Sung-Ju Lee, Almeroth K., (2011), The Impact of Channel Bonding on 802.11n Network Management, ACM CoNEXT 2011, December 6–9.2011, Tokyo, Japan.

[11] Dolińska I., Masiukiewicz A., Rządkowski G., (2014), Channel Selection in Home 802.11 Standard Networks, Proccedings of IEEE sponsored Digital Technologies Conference Żylina 2014

[12] Lacage M., Henderson T. R., (2006), Yet another network simulator. In WNS2 '06: Proceeding from the 2006 Workshop on ns-2: the IP network simulator, New York, NY, USA, October 2006. ACM.

[13] https://www.nsnam.org/doxygen/wifi-spectrum-per-interference_8cc_source.html access 02.2017

[14] R. Bruno, M. Conti and E. Gregori, 2008.,Throughput nalysis and Measurements in IEEE 802.11 WLANs with TCP and UDP Traffic Flows,IEEE Transactions On Mobile Computing, vol. 7, no. 3, pp. 1-16, http://dx.doi.org/10.1109/TMC.2007.70718.

[15] P. Gajewski and S. Wszelak, 2008., Wireless technology on teleinformatics networks , WKŁ, Warsaw, Poland. (in Polish)

[16] Dolińska I., Masiukiewicz A., Rzadkowski G., The Monte Carlo analysis of the media access time distribution in 802.11n MAC layer, Position Papers pf the 2014 FEDCCSIS Conference, Annals of Computer Science and Information Systems Volume 3

On the (mixed) integrated fractional Brownian motion

Charles El-Nouty
LAGA
Université Paris XIII, Sorbonne Paris Cité
99 avenue J-B Clément
93430 Villetaneuse, France
Email: elnouty@math.univ-paris13.fr

***Abstract*—We consider the integrated fractional Brownian motion with Hurst index $0 < H < 1$. The main aim consists in studying its links with the quasi-helix with approximately stationary increments class of centered Gaussian processes. More precisely, we compute some well-chosen quantities which depend on the moments of the integrated fractional Brownian motion. Firstly, we check if one can obtain appropriate upper or lower bounds of the above quantities. Secondly, focusing our attention on the small increments of the integrated fractional Brownian motion we derive an equivalent. Some surprising facts are observed. Next, we consider the mixed integrated fractional Brownian motion. Its main properties are studied. We also determine the values of H for which this process is a semi-martingale. Finally we compare the results obtained for the mixed integrated fractional Brownian motion with those obtained for the mixed fractional Brownian motion and for the sub-mixed fractional Brownian motion.**

I. Introduction

Let $\{B_H(t), t \geq 0\}$ be a fractional Brownian motion (fBm) with Hurst index $0 < H < 1$, i.e. a centered Gaussian process with stationary increments satisfying $B_H(0) = 0$, with probability 1, and $\mathbb{E}(B_H(t))^2 = t^{2H}, t \geq 0$. We also will use the fact that B_H is self-similar with index H, i.e. $B_H(at)$ and $a^H B_H(t)$ have the same distribution for all $a > 0$, and that its covariance function is

$$\forall\ t \geq 0\ \forall\ s \geq 0$$

$$\mathbb{E}(B_H(t)B_H(s)) = \frac{1}{2}\ (t^{2H} + s^{2H} - \mid t - s \mid^{2H}). \quad (1)$$

When $H = 1/2, B_{1/2}$ is a standard Wiener process or a standard Brownian motion (Bm). The name of fBm was first introduced by [1], but their sample path properties were already studied by Kolmogorov in the 1940's. Finally we insist on the fact that the fBm has a wide domain of applications such that, among the more recent, telecommunications, network traffic simulations, image recognition, etc.

Consider the integrated fractional Brownian motion (ifBm) defined as follows :

$$X_H(t) = \int_0^t B_H(u)\, du,\ t \geq 0.$$

We can easily remark that the ifBm X_H is a centered Gaussian process such that $X_H(0) = 0$ with probability 1 and X_H is self-similar with index $1 + H$. The process X_H is an interesting one on its own. Indeed, in the special case $H = 1/2$, it appears in several fields such that mechanics, biology and quantum probability. We refer to [2], [3], [4] and [5] and the references therein for further information on this process.

Let $\{Z(t), t \geq 0\}$ be a general Gaussian process. We introduce the mixed Z process defined by :

$$Y(t) = W(t) + Z(t),\ t \geq 0,$$

where W is a Bm independent of Z.

This type of Gaussian processes was introduced by [6] where the process Z was the fBm in order to solve some problems in mathematical finance, such as modeling some arbitrage-free and complete markets. The mixed fBm (mfBm) was extended by [7] and by [8]. The sub-mixed fBm (smfBm) was investigated by [9], i.e. when the process Z is the sub-fractional Brownian motion (sfBm). In this paper, we will focus our attention on the mixed ifBm (mifBm), i.e.

$$Y(t) = W(t) + X_H(t),\ t \geq 0,$$

where the ifBm X_H is independent of the Bm W.

In section 2, several properties of the ifBm are studied, namely:

- the covariance function,
- the non quasi-helix property,
- the non approximately stationary increments property.

In section 3, several properties of the mifBm are studied, namely:

- the basic ones,
- the non Markovian property,
- the semi-martingale property.

II. The IFBM and the QHASI class

The (correct) covariance function of X_H is given in the following lemma.

Lemma 1: We have for $t \geq s$

$$\begin{aligned}\mathbb{E}(X_H(t)X_H(s)) &= \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &\quad -\frac{1}{2}\,\frac{t^{2H+2}+s^{2H+2}}{(2H+1)(2H+2)} \\ &\quad +\frac{1}{2}\,\frac{(t-s)^{2H+2}}{(2H+1)(2H+2)},\end{aligned}$$

and therefore

$$\mathbb{E}(X_H(t)^2) = \frac{t^{2H+2}}{2H+2}.$$

Proof: Since (1) holds, direct computations imply for $0 \leq u \leq t$ and $0 \leq v \leq s$

$$\begin{aligned}\mathbb{E}(X_H(t)X_H(s)) &= \int_0^t\int_0^s \mathbb{E}(B_H(u)B_H(v))\,du\,dv \\ &= \frac{1}{2}\int_0^t\int_0^s (u^{2H}+v^{2H})\,du\,dv \\ &\quad -\frac{1}{2}\int_0^t\int_0^s \mid u-v\mid^{2H}\,du\,dv \\ &= \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &\quad -\frac{1}{2}\int_s^t\int_0^s (u-v)^{2H}\,du\,dv \\ &\quad -\frac{1}{2}\int_0^s\int_0^s \mid u-v\mid^{2H}\,du\,dv \\ &= \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &\quad -\frac{1}{2}\,\frac{t^{2H+2}-s^{2H+2}-(t-s)^{2H+2}}{(2H+1)(2H+2)} \\ &\quad -\frac{1}{2}\,2\int_0^s\Big(\int_0^u (u-v)^{2H}\,dv\Big)\,du \\ &= \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &\quad -\frac{1}{2}\,\frac{t^{2H+2}-s^{2H+2}-(t-s)^{2H+2}}{(2H+1)(2H+2)} \\ &\quad -\frac{s^{2H+2}}{(2H+1)(2H+2)} \\ &= \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &\quad -\frac{1}{2}\,\frac{t^{2H+2}+s^{2H+2}}{(2H+1)(2H+2)} \\ &\quad +\frac{1}{2}\,\frac{(t-s)^{2H+2}}{(2H+1)(2H+2)}.\end{aligned}$$

When $t = s$, we get the expression of $\mathrm{Var}X_H$. ■

Remark 2: When $H = 1/2$, we can verify that lemma 1 gives the classical expressions

$$\mathbb{E}(X_{1/2}(t)X_{1/2}(s)) = \frac{s^2}{6}\,(3t-s)\ and\ \mathbb{E}X_{1/2}(t)^2 = \frac{t^3}{3}.$$

Let us recall some definitions introduced in [10]. A centered Gaussian process $\{X(t), t \in I \subset \mathbb{R}\}$ belongs to the quasi-helix with approximately stationary increments (QHASI) class if it fulfills the five following assumptions:

A1: $X(0) = 0$ with probability 1,

A2: there exists $\lambda > 0$ such that X is self-similar with index λ,

A3: there exists $0 < C_1 \leq C_2 < +\infty$ such that $\forall\,(s,t) \in I^2$

$$\begin{aligned}C_1 \mid t-s\mid^{2\lambda} &\leq \mathbb{E}(X(t)-X(s))^2 \\ &\leq C_2 \mid t-s\mid^{2\lambda},\end{aligned}$$

A4: there exists $C_3 \in [C_1, C_2]$ such that

$$\begin{aligned}&\forall\,(s,t) \in I^2,\ t \geq s,\ st \neq 0, \\ &\qquad\qquad \text{when } t-s \to 0, \\ &\mathbb{E}(X(t)-X(s))^2 \sim C_3\,(t-s)^{2\lambda},\end{aligned}$$

A5: there exists $C_4 \in [C_1, C_2]$ such that

$$\forall\, t \in I,\ \mathbb{E}X(t)^2 = C_4\ \mid t\mid^{2\lambda}.$$

The QHASI class was introduced mainly because it contains several known processes, such that the Bm, the fBm, the bi-fractional Brownian motion, the sfBm and the sub-bi-fractional Brownian motion, each of them having a lot of practical applications. We refer to [10] and the references therein for further information on the QHASI class and on the processes mentioned above. The aim of this section is to answer to the following question:

does the ifBm **belong or not** to the QHASI class ?

The response searching is the main motivation of the author. Note that when this work was initiated, the author had no opinion concerning the answer. Since the assumptions A1 and A2 are obviously fulfiled by the ifBm, it suffices to study the two main properties of any element of the QHASI class : the quasi-helix one (assumption A3) and the approximately stationary increments one (assumption A4). To this aim, set for $t \geq s \geq 0$

$$\begin{aligned}f(t,s) &= \frac{t^{2H+2}+s^{2H+2}}{2H+1} \\ &\quad -\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &= \frac{(t-s)(t^{2H+1}-s^{2H+1})}{2H+1}.\end{aligned}$$

Then, we have by lemma 1

$$\mathbb{E}(X_H(t)-X_H(s))^2 = f(t,s) - \frac{(t-s)^{2H+2}}{(2H+1)(2H+2)}. \quad (2)$$

The two following lemmas give the answer to the previous question.

Lemma 3: The ifBm X_H is not a quasi-helix in the sense of Kahane.

Proof: Let $\gamma > 0$. For $t \geq s > 0$, set $u = \frac{t}{s} \geq 1$. Consider the auxiliary function

$$\begin{aligned} g(t,s) &= f(t,s) - \frac{\gamma}{2H+1}\ (t-s)^{2H+2} \\ &= \frac{(t-s)}{2H+1}\ (t^{2H+1} - s^{2H+1} - \gamma\ (t-s)^{2H+1}) \\ &= \frac{(t-s)s^{2H+1}}{2H+1}\ (u^{2H+1} - 1 - \gamma\ (u-1)^{2H+1}) \\ &:= \frac{(t-s)s^{2H+1}}{2H+1}\ h(1,u). \end{aligned}$$

The next task consists in studying the sign of the function g when γ describes the real positive line. We can remark that $h(1,1) = 0$. The derivative of $h(1,u)$ is

$$h^{'}(1,u) = (2H+1)\ (u^{2H} - \gamma\ (u-1)^{2H}).$$

When $\gamma = 1$, the function $h^{'}(1,u)$ is strictly positive. Then the function $h(1,u)$ is increasing. Since $h(1,1) = 0$, the function $h(1,u)$ is positive, and therefore $g(t,s)$. Hence, by using equality (2), we establish for any $t \geq s > 0$

$$\begin{aligned} \mathbb{E}(X_H(t) - X_H(s))^2 &\geq \frac{1}{2H+1}\ (t-s)^{2H+2} \\ &\quad - \frac{(t-s)^{2H+2}}{(2H+1)(2H+2)} \\ &= \frac{1}{2H+2}\ (t-s)^{2H+2}. \end{aligned}$$

The previous inequality holds also when $s = 0$.

When $\frac{1}{2H+2} < \gamma < 1$, we obviously get a smaller lower bound of $\mathbb{E}(X_H(t) - X_H(s))^2$ than the one obtained when $\gamma = 1$. Note that, when $0 < \gamma \leq \frac{1}{2H+2}$, the lower bound of $\mathbb{E}(X_H(t) - X_H(s))^2$ becomes negative.

Let us turn to the case $\gamma > 1$. The equation $h^{'}(1,u) = 0$ has an unique solution

$$u^* = u^*(\gamma) = \frac{\gamma^{1/2H}}{\gamma^{1/2H} - 1} > 1.$$

If $1 \leq u \leq u^*$, then the function $h^{'}(1,u)$ is positive, else negative. Therefore, if $1 \leq u \leq u^*$, then the function $h(1,u)$ is increasing and positive, else decreasing from $h(1,u^*) > 0$ to $-\infty$. Denote by $u^{**} = u^{**}(\gamma) > u^*$ such that $h(1,u^{**}) = 0$. Hence if $1 \leq u \leq u^{**}$, then the function $h(1,u)$ is positive, else the function $h(1,u)$ is negative.

Recall that $u = t/s$. If $1 \leq t/s \leq u^{**}$, then

$$f(t,s) \geq \frac{\gamma}{2H+1}\ (t-s)^{2H+2},$$

else

$$f(t,s) \leq \frac{\gamma}{2H+1}\ (t-s)^{2H+2}.$$

Thus, if $1 \leq t/s \leq u^{**}$, then we have by using equality (2)

$$\mathbb{E}(X_H(t) - X_H(s))^2 \geq \frac{\gamma(2H+2)-1}{(2H+1)(2H+2)}\ (t-s)^{2H+2},$$

else

$$\begin{aligned} \mathbb{E}(X_H(t) - X_H(s))^2 &\leq \frac{\gamma(2H+2)-1}{(2H+1)(2H+2)}\ (t-s)^{2H+2} \\ &\leq \frac{\gamma}{2H+1}\ (t-s)^{2H+2}. \end{aligned}$$

It implies that there is no constant $C_2 > 0$ such that, for any $0 < s \leq t$, $\mathbb{E}(X_H(t) - X_H(s))^2 \leq C_2\ (t-s)^{2H+2}$. This completes the proof of lemma 3. ■

Lemma 4: The ifBm X_H does not have approximately stationary increments.

Proof: For any fixed $s > 0$, set $t = s + h$. We get by equality (2)

$$\begin{aligned} \mathbb{E}(X_H(s+h) - X_H(s))^2 &= \frac{h}{2H+1}\ ((s+h)^{2H+1} \\ &\quad - s^{2H+1}) \\ &\quad - \frac{h^{2H+2}}{(2H+1)(2H+2)} \\ &= \frac{hs^{2H+1}}{2H+1}\ ((1+\frac{h}{s})^{2H+1} - 1) \\ &\quad - \frac{h^{2H+2}}{(2H+1)(2H+2)}. \end{aligned}$$

Since $0 < H < 1$, we have $2 < 2H + 2 < 4$. Then, when $h \to 0$, a Taylor expansion of order 2 yields

$$\begin{aligned} \mathbb{E}(X_H(s+h) - X_H(s))^2 &= s^{2H}\ h^2 + H\ s^{2H-1}\ h^3 \\ &\quad + o(h^3) \\ &\quad - \frac{h^{2H+2}}{(2H+1)(2H+2)}, \end{aligned}$$

and therefore

$$\mathbb{E}(X_H(s+h) - X_H(s))^2 \sim s^{2H}\ h^2.$$

This completes the proof of lemma 4. ■

Remark 5: When $H = 1/2$, we can go further. Indeed,

$$\mathbb{E}(X_{1/2}(s+h) - X_{1/2}(s))^2 = s\ h^2 + \frac{h^3}{3}.$$

Note that, when s is a function of h, the limit of $\mathbb{E}(X_{1/2}(s+h) - X_{1/2}(s))^2$, as $h \to 0$ can take different values. For example, choose $s = h^\alpha, \alpha \in \mathbb{R}$. If $\alpha > -2$, then the limit of $\mathbb{E}(X_{1/2}(s+h) - X_{1/2}(s))^2$ equals 0. If $\alpha = -2$, then this limit equals 1. If $\alpha < -2$, then this limit equals $+\infty$.

Hence, we have proved that the ifBm does not belong to the QHASI class. It was a real surprise to observe the importance of the two following facts : on one hand the size of the ratio t/s in the study of the quasi-helix property, and on the other hand the location of s in the study of the approximately stationary increments one. Note that, as a straight consequence of the proof of lemma 4, we can state the following corollary.

Corollary 6: We have for $t \geq s$

$$\mathbb{E}(X_H(t) - X_H(s))^2 \geq \frac{1}{2H+2}\ (t-s)^{2H+2}.$$

To end this section, let us make some comments on assumption A5. Although the ifBm does not fulfill assumption A3 (the constant C_2 does not exist), we can observe by combining lemma 1 with corollary 6 that, in some sense,

$$C_1 = C_4 = \frac{1}{2H+2}.$$

III. The mixed ifBm

The following lemma describes the basic properties of the mifBm.

Lemma 7: The mifBm Y satisfies the following properties:

- Y is a centered Gaussian process.
- $\forall s \in \mathbb{R}_+,\ \forall t \in \mathbb{R}_+$, such that $t \geq s$

$$\begin{aligned}\mathbb{E}\Big(Y(t)Y(s)\Big) &= s + \frac{1}{2}\,\frac{ts(t^{2H}+s^{2H})}{2H+1} \\ &- \frac{1}{2}\,\frac{t^{2H+2}+s^{2H+2}}{(2H+1)(2H+2)} \\ &+ \frac{1}{2}\,\frac{(t-s)^{2H+2}}{(2H+1)(2H+2)}\end{aligned}$$

- $$\forall t \in \mathbb{R}_+, \mathbb{E}(Y(t)^2) = t + \frac{t^{2H+2}}{2H+2}.$$

Proof: It is a direct consequence of lemma 1. ■

We insist of the fact that the nature of Y obviously implies that there is no self-similarity property. Now, we will study the Markovian property.

Lemma 8: For any $H \in \Big]0,1\Big[$, the mifBm is not a Markovian process.

Proof: By lemma 7 , Y is a centered Gaussian process such that $\mathbb{E}Y(t)^2 > 0$ for all $t > 0$. Then, if Y was a Markovian process, according to [11], for all $0 < s < t < u$ we would have:

$$\mathbb{E}\Big(Y(s)Y(u)\Big)\mathbb{E}\Big(Y(t)^2\Big) = \mathbb{E}\Big(Y(s)Y(t)\Big)\mathbb{E}\Big(Y(t)Y(u)\Big) \tag{3}$$

Let s be fixed and set $u = e^t$. The proof will be divided into the two following cases $H \neq 1/2$ and $H = 1/2$.

Consider the case $H \neq 1/2$. When $t \to +\infty$, combining Taylor expansions with lemma 7, we get

$$\begin{aligned}\mathbb{E}\Big(Y(s)Y(t)\Big) &= s - \frac{1}{2}\,\frac{s^{2H+2}}{(2H+1)(2H+2)} \\ &+ \frac{1}{2}\,\frac{s^{2H+1}}{2H+1}\,t \\ &+ \frac{1}{4}\,s^2\,t^{2H} + o(t^{2H}), \\ \mathbb{E}\Big(Y(t)^2\Big) &= t + \frac{t^{2H+2}}{2H+2}, \\ \mathbb{E}\Big(Y(t)Y(e^t)\Big) &= t - \frac{1}{2}\,\frac{t^{2H+2}}{(2H+1)(2H+2)} \\ &+ \frac{1}{2}\,\frac{t^{2H+1}}{2H+1}\,e^t \\ &+ \frac{1}{4}\,t^2\,e^{2Ht} + o(t^2\,e^{2Ht}), \\ \mathbb{E}\Big(Y(s)Y(e^t)\Big) &= s - \frac{1}{2}\,\frac{s^{2H+2}}{(2H+1)(2H+2)} \\ &+ \frac{1}{2}\,\frac{s^{2H+1}}{2H+1}\,e^t \\ &+ \frac{1}{4}\,s^2\,e^{2Ht} + o(e^{2Ht}).\end{aligned}$$

Consider the sub-case $0 < H < 1/2$ first. To verify equality (3), a necessary condition is

$$\frac{1}{2}\,\frac{s^{2H+1}}{2H+1}\,e^t \times \frac{t^{2H+2}}{2H+2} = \frac{1}{2}\,\frac{s^{2H+1}}{2H+1}\,t \times \frac{1}{2}\,\frac{t^{2H+1}}{2H+1}\,e^t,$$

i.e.

$$2H+2 = 2(2H+1) \Leftrightarrow H = 0,$$

which is impossible.

Consider the sub-case $1/2 < H < 1$. To verify equality (3), a necessary condition is

$$\frac{1}{4}\,s^2\,e^{2Ht} \times \frac{t^{2H+2}}{2H+2} = \frac{1}{4}\,s^2\,t^{2H} \times \frac{1}{4}\,t^2\,e^{2Ht},$$

i.e.

$$2H+2 = 4 \Leftrightarrow H = 1.$$

which is impossible.

Consider the case $H = 1/2$ now. Straightforward computations give

$$\begin{aligned}\mathbb{E}\Big(Y(s)Y(t)\Big) &= s + \frac{s^2}{6}\,(3t-s), \\ \mathbb{E}\Big(Y(t)^2\Big) &= t + \frac{t^3}{3}, \\ \mathbb{E}\Big(Y(t)Y(e^t)\Big) &= t + \frac{t^2}{6}\,(3e^t - t), \\ \mathbb{E}\Big(Y(s)Y(e^t)\Big) &= s + \frac{s^2}{6}\,(3e^t - s).\end{aligned}$$

To verify equality (3), a necessary condition is

$$\frac{s^2}{6}\,3\,e^t \times \frac{t^3}{3} = \frac{s^2}{6}\,3\,t \times \frac{t^2}{6}\,3\,e^t,$$

i.e.

$$\frac{1}{3} = \frac{1}{2},$$

which is impossible.

The proof of lemma 8 is complete. ■

Remark 9: Recall that, when $H = 1/2$, the mfBm and the smfBm are Markovian processes. Indeed, when $H = 1/2$, the mfBm and the smfBm are Bm. This property is not satisfied in the ifBm case.

The next lemma consists in studying the semi-martingale property of the mifBm Y.

Lemma 10: For every $T > 0$ and $H \in]0, 1[$, the mifBm $\{Y(t), t \in [0, T]\}$ is, in its own filtration, a semi-martingale equivalent in law to a Brownian motion.

Proof: Recall that the covariance function of the fBm is positive. Then, we get that the covariance function of the ifbm is positive definite and twice continuously differentiable on $[0, T]^2 - \{(t, s) : t = s\}$. When $t > s$, we have

$$\begin{aligned} \frac{\partial^2 \mathbb{E}(X_H(t)X_H(s))}{\partial t \partial s} &= \mathbb{E}(B_H(t)B_H(s)) \\ &= \frac{1}{2}\,(t^{2H} + s^{2H} - (t-s)^{2H}). \end{aligned}$$

Since

$$\begin{aligned} (t^{2H} + s^{2H} - (t-s)^{2H})^2 &\leq 2\,(t^{2H} + s^{2H})^2 \\ &\quad +2\,(t-s)^{4H} \\ &\leq 4\,t^{4H} + 4\,s^{4H} \\ &\quad +2\,(t-s)^{4H}, \end{aligned}$$

the function $\frac{\partial^2 \mathbb{E}(X_H(t)X_H(s))}{\partial t \partial s}$ is an element of $L^2([0, T]^2 - \{(t, s) : t > s\})$. Then, to complete the proof of the lemma, it suffices to apply Baudoin and Nualart's theorem 20 (2003, p.348,[12]). ■

Remark 11: Recall that if $3/4 < H < 1$, then the mfBm and the smfBm are semi-martingales, else they are not. This is an important difference with the mifBm.

IV. Conclusions

We have proved that the ifBm is not an element of the QHASI class. A careful reading of the proofs of lemmas 3 and 4 suggests that the QHASI class has to be extended. We have also studied some properties of the mifBm and compared them with those of the mfBm and the smfBm. This is a contribution of enlarging the modelling tool kit.

References

[1] B. Mandelbrot and J. van Ness, "Fractional brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, pp. 422–437, 1968.

[2] C. El-Nouty, "The lower classes of the integrated fractional brownian motion," *Studia Sci. Math. Hungar.*, vol. 41, pp. 17–38, 2004.

[3] D. Khoshnevisan and Z. Shi, "Chung's law for integrated brownian motion," *Trans. Amer. Math. Soc.*, vol. 350, pp. 4253–4264, 1998.

[4] W. V. Li and W. Linde, "Approximation, metric entropy and small ball estimates for gaussian measures," *Ann. Probab.*, vol. 27, pp. 1556–1578, 1999.

[5] W. V. Li and Q. M. Shao, "Handbook of statistics: Gaussian processes: Inequalities, small ball probabilities and applications," *Stochastic Processes: Theory and Methods, Handbook of Statistics*, vol. 19, pp. 533–597, 2001.

[6] P. Cheridito, "Mixed fractional brownian motion," *Bernoulli*, pp. 913–934, 2001.

[7] C. El-Nouty, "The fractional mixed fractional brownian motion," *Statist. Probab. Letters*, vol. 65, pp. 111–120, 2003.

[8] M. Zili, "On the mixed fractional brownian motion," *Journal of Mathematical Analysis and Applications*, 2006.

[9] C. El-Nouty and M. Zili, "On the sub-mixed fractional brownian motion," *Appl. Math. J. Chinese Univ.*, vol. 28, pp. 179–196, 2015.

[10] C. El-Nouty, "On approximately stationary gaussian processes," *International Journal for Computational Civil and Structural Engineering*, vol. 11, pp. 15–26, 2015.

[11] D. Revusz and M. Yor, *Continuous martingales and Brownian motion*. Springer-Verlag, 1991.

[12] F. Baudoin and D. Nualart, "Equivalence of volterra processes," *Stochastic Processes and their Applications*, vol. 107, pp. 327–350, 2003.

The Euler schemes for numerical modeling of stochastic differential equations

Darya Filatova
Production Engineering Department
Faculty of Management and Computer Modeling
Kielce University of Technology
al. Tysiąclecia Państwa Polskiego 7, 25–314 Kielce, Polska
Email: dfilatova@tu.kielce.pl

***Abstract*—This work presents a review of the progress in the development of the numerical methods for sample paths simulation of stochastic differential equations. The error of approximation, the order of the convergence, the stability improvement for Euler methods (explicit, implicit and composite schemes) are studied theoretically as well as numerically. MathLab programs support all the illustrations.**

I. Introduction

Stochastic differential equations are the convenient way to describe the nonlinear dynamics of many real-world systems. Let $B_t = \{B(t, \mathbb{H}), t \in [t_0, T]\}$, $0 \leq t_0 \leq T \leq +\infty$, be a Brownian motion (Wiener process). Define $\left(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, \mathbf{P}\right)$ a complete probability space where $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration generated by B_t, augmented by all the $\mathbf{P}$-null sets in $\{\mathcal{F}_t\}_{t\geq 0}$. The one-dimensional, time-homogeneous stochastic differential equation (SDE) has the form

$$dX = a(t, X)dt + b(t, X)dB_t,\ X(t_0) = X_0, \qquad (1)$$

where $a(t, X)$ and $b(t, X)$ are drift and diffusion coefficients, $t \in [t_0, T]$ is the independent variable corresponding to the "time", dB_t is an increment of Wiener process B.

We assume that the SDE (1) has a unique solution [1].

Only a few stochastic differential equations (1) have analytic solutions. Most of them belong to the group of linear equations with additive or multiplicative noise or equations which reduce to linear one [1], [2]. Consequently, the necessity of numerical solution methods construction appears. Presently there are many different approaches to the numerical solution of SDEs. One group of these methods base on the adaptation of numerical methods for ODEs solution with respect to stochastic integral [3], and the second group contains specially developed methods for SDEs [4]. Most of the researchers use the first approach as far as the theory of the ODEs numerical solution is well-developed, and it is easy to find analogies between ODEs and SDEs. The Euler scheme for SDE numerical solution is the simplest one. It was presented by Murayama in 1955 and developed later [2], [5]. This scheme has many useful properties (it has the rate of the strong convergence $\gamma = 0.5$), but at the same time has some limitations (sometimes it is not stable and has significant approximation error).

As in the case of ODEs numerical solution [6], we suppose that order of convergence, approximation error and stability can be improved through the expansion in approximation point using multiple derivatives of drift and diffusion functions of SDEs. This approach is known as Taylor method [2], [7], [8]. However, the success of Taylor method depends on multiple stochastic integrals approximations, which appear because of mentioned derivatives. To overcome these drawbacks, we can use multiple division of the approximation step (Runge-Kutta methods [2], [9], [10], [11]) or approximation results of previous steps (multi-step methods [3], [7], [10]). It is necessary to remind that development of numerical schemes is connected not only with mathematical aspects but also with physical nature of investigation object and with possibilities of the computer. Both ordinary and stochastic differential equation systems of importance in scientific modeling often demonstrate an undesired behavior when solved by classical numerical methods and can be considered as ill-posed systems. In the most cases, this unwanted behavior has the connection with very high instability of numerical solution caused by so-called stiffness. There are several possible explanations for this phenomenon. First reason is linked with computer technical characteristics. To reach the desired accuracy of the numerical solution, one should take smaller time step size repeatedly dividing initial integration interval. However, it brings round-off errors accumulation and, as a consequence, computer overflow. On the other hand using small time step sizes may require too much computational time. Second reason is related to the physical nature of the system. That means that the system describes processes of different speeds or gradients. It usually happens in the problems of the boundary layer (hydrodynamics), skin-effect (electromagnetism), chemical kinetics reaction, etc. At last, stiffness can be explained as a combination of first two reasons. Therefore, looking for more suitable numerical methods, one has to take into account all these crucial points. Development of numerical schemes for stiff systems frequently bases on ideas of [9], [12], [13], [14], [15]. They postulated that it was impossible to use explicit numerical schemes for stiff systems and presented approaches based only on implicit schemes. However, and it is necessary to mention, direct application of these methods is usually connected with extremely complicated procedures of the scheme parameters definition based on stability properties [9], [16]. All this makes proposed approaches unsuitable in the most of the mentioned applications. Nevertheless, researchers pointed out two significant mathematical properties of stiffness. First of all stiff systems have very wide spectrum. That indicates a presence of very different Lyapunov exponents [17]. Secondly, according to the existence and uniqueness of the solution, stiff

problems typically have large Lipschitz constants. Therefore, the main principles of numerical schemes for SDE solution (especially in stiff case) show the necessity not only of the construction of the new scheme but also software development.

The goal of this paper is to present the review of the Euler schemes, to show their properties and examples of MathLab code used for the simulation of the sample paths of (1).

II. Euler scheme and its properties

A deterministic ordinary differential equation (we suppose that it has a unique solution)

$$\frac{dX}{dt} - a\left(t, X\right) = 0, \qquad t \in \left[t_0, T\right], \qquad X\left(t_0\right) = X_0, \quad (2)$$

can be presented by the following operator

$$LX = \varphi\left(t\right) \quad (3)$$

where the symbol L denotes not only given equation, but also some restrictions (boundary conditions or initial values), $\varphi\left(t\right)$ denotes the right hand sides of the equation and restrictions. Therefore the equation (2) can be written

$$LX = \left\{ \begin{array}{c} \frac{d}{dt}X - a\left(t, X\right) \\ X\left(t_0\right) \end{array} \right\}, \quad \varphi\left(t\right) = \left\{ \begin{array}{c} 0 \\ X_0 \end{array} \right\}. \quad (4)$$

One of the possible and commonly used approaches for the numerical solution of (2) is to use the difference methods and in consequence to come from continuous to discrete time and deal with functions of a discrete variable. Thus, it is possible to speak about the discrete time approximation.

Consider a discretization $t_0 < t_1 < ... < t_n < ... < t_{n_T} = T$ of the time interval $\left[t_0, T\right]$ with some integer $n_T \geq 2$. This discretization is used to define the approximated solution of (2) at the nodal points t_n for $n = 0, 1, 2, ..., n_T$. We denote these values as $\left(Y_n\right)_{n \in \{0,1,..,n_T\}}$ and rewrite the differential equation as the difference one

$$L_n Y_n = \varphi_n, \quad (5)$$

where L_n denotes the structure of the difference equation and restrictions, namely with initial value $Y_0 = X_0$, φ_n is the task input data.

Let

$$\Delta_n = \int_{t_n}^{t_{n+1}} dt = t_{n+1} - t_n$$

is the length of the time discretization subinterval $\left[t_n, t_{n+1}\right]$. For simplicity, we consider equidistant time discretization

$$\Delta = \max_{n \in \{0,1,...,n_T - 1\}} \Delta_n = \frac{T}{n_T}.$$

Denoting the approximation of the derivative

$$\frac{X_{n+1} - X_n}{t_{n+1} - t_n} \approx \frac{Y_{n+1} - Y_n}{\Delta} \quad (6)$$

and taking into account different possibilities for its approximation we introduce the difference schemes

$$\begin{array}{rcl} L_\Delta Y_\Delta & = & \left\{ \begin{array}{c} \frac{Y_{n+1} - Y_n}{\Delta} - a\left(t_n, Y_n\right) \\ Y_0 \end{array} \right\}, \\ \varphi_\Delta & = & \left\{ \begin{array}{c} 0 \\ X_0 \end{array} \right\} \end{array} \quad (7)$$

and

$$\begin{array}{rcl} L_\Delta Y_\Delta & = & \left\{ \begin{array}{c} \frac{Y_{n+1} - Y_n}{\Delta} - a\left(t_{n+1}, Y_{n+1}\right) \\ Y_0 \end{array} \right\}, \\ \varphi_\Delta & = & \left\{ \begin{array}{c} 0 \\ X_0 \end{array} \right\}. \end{array} \quad (8)$$

Thus, the difference scheme (7) is the explicit Euler scheme

$$Y_{n+1} = Y_n + a\left(t_n, Y_n\right)\Delta, \quad (9)$$

and the scheme (8) is the implicit Euler scheme

$$Y_{n+1} = Y_n + a\left(t_{n+1}, Y_{n+1}\right)\Delta. \quad (10)$$

By the same reasoning, Euler schemes can be deduced for SDEs, however, in this case, it gives three possibilities

- explicit scheme
$$Y_{n+1} = Y_n + a\left(t_n, Y_n\right)\Delta + b\left(t_n, Y_n\right)\Delta B_n \quad (11)$$
- implicit by drift parameter
$$Y_{n+1} = Y_n + a\left(t_{n+1}, Y_{n+1}\right)\Delta + b\left(t_n, Y_n\right)\Delta B_n \quad (12)$$
- implicit by drift and diffusion parameters
$$Y_{n+1} = Y_n + a\left(t_{n+1}, Y_{n+1}\right)\Delta + b\left(t_{n+1}, Y_{n+1}\right)\Delta B_n, \quad (13)$$
where $\Delta B_n = B\left(t_{n+1}\right) - B\left(t_n\right)$.

A. The error of approximation and convergence

To study the error it is useful to approximate the **strong** solution for all $t \in \left[t_0, T\right]$ and not just at the nodal points t_n for $n = 0, 1, 2, ..., n_T$. To accomplish this $X\left(t\right) \approx \widehat{Y}\left(t\right)$

$$\widehat{Y}\left(t\right) = Y_n + \int_{t_n}^{t} a\left(t_n, Y_n\right) ds + \int_{t_n}^{t} b\left(t_n, Y_n\right) dB\left(s\right) \quad (14)$$

for $t_n \leq t \leq t_{n+1}$ and $n = 0, 1, ..., n_T - 1$.

Define the error as $\epsilon\left(t\right) = X\left(t\right) - \widehat{Y}\left(t\right)$ and suppose that $\widetilde{a_t} = a\left(t, X_t\right) - a\left(t_n, Y_n\right)$ and $\widetilde{b_t} = b\left(t, X_t\right) - b\left(t_n, Y_n\right)$ are two stochastic processes for $t_n \leq t \leq t_{n+1}$ and $n = 0, 1, ..., n_T - 1$ with appropriated measurability and integrability properties. Then the stochastic process $\epsilon = \left\{\epsilon_t, t \in \left[t_0, T\right]\right\}$ is characterized by the Itô differential

$$d\epsilon_t = \widetilde{a_t} dt + \widetilde{b_t} dB\left(t\right)$$

with initial value $\epsilon(t_0) = 0$.

Applying the Itô formula ([1]) to a continuous function $f(t, \epsilon(t)) = \epsilon^2(t)$, $t \in [t_0, T]$, we obtain the Itô differential

$$d\left(\epsilon^2(t)\right) = 2\epsilon_t\left(\widetilde{a}_t dt + \widetilde{b}_t dB(t)\right) + \widetilde{b}_t^{\,2} dt$$

or

$$\begin{aligned} d\left(\epsilon^2(t)\right) &= 2\left(X(t) - \widehat{Y}(t)\right)(a(t, X(t)) \\ &\quad -a(t_n, Y_n))dt \\ &\quad + (b(t, X(t)) - b(t_n, Y_n))^2 dt \\ &\quad +2\left(X(t) - \widehat{Y}(t)\right)((b(t, X(t)) \\ &\quad -b(t_n, Y_n))dB(t) \end{aligned}$$

with initial value $\epsilon^2(t) = 0$ and for $t_n \le t \le t_{n+1}$, $n = 0, 1, ..., n_T - 1$.

Hence,

$$\begin{aligned} \mathbb{E}\Big(\epsilon^2(t_{n+1})\Big) &= \mathbb{E}\Big(\epsilon^2(t_n)\Big) \\ &\quad +\mathbb{E}\int_{t_n}^{t_{n+1}} (b(t, X(t)) - b(t_n, Y_n))^2 dt \\ &\quad +\mathbb{E}\int_{t_n}^{t_{n+1}} 2\Big(X(t) - \widehat{Y}(t)\Big) \\ &\quad \times\Big(a(t, X(t)) - a(t_n, Y_n)\Big)dt \\ &\quad +\mathbb{E}\int_{t_n}^{t_{n+1}} 2\Big(X(t) - \widehat{Y}(t)\Big) \\ &\quad \times\Big(b(t, X(t)) - b(t_n, Y_n)\Big)dB(t). \end{aligned}$$

Using the inequality $|2cd| \le c^2 + d^2$ and the properties of stochastic integral

$$\begin{aligned} \mathbb{E}\left[\epsilon^2(t_{n+1})\right] &\le \mathbb{E}\left[\epsilon^2(t_n)\right] \\ &\quad + \int_{t_n}^{t_{n+1}} \mathbb{E}\left(X(t) - \widehat{Y}(t)\right)^2 dt \\ &\quad + \int_{t_n}^{t_{n+1}} \mathbb{E}\left|a(t, X(t)) - a(t_n, Y_n)\right|^2 dt \\ &\quad + \int_{t_n}^{t_{n+1}} \mathbb{E}\left|b(t, X(t)) - b(t_n, Y_n)\right|^2 dt. \end{aligned}$$

Taking into account that $(c+d)^2 \le 2c^2 + 2d^2$ and the functions a and b are non-anticipating we deduce

$$\begin{aligned} &\int_{t_n}^{t_{n+1}} \mathbb{E}\left|a(t, X(t)) - a(t_n, Y_n)\right|^2 \\ &\le 2\int_{t_n}^{t_{n+1}} \mathbb{E}\left|a(t, X(t)) - a(t_n, X(t_n))\right|^2 \\ &\quad +2\int_{t_n}^{t_{n+1}} \mathbb{E}\left|a(t_n, X(t_n)) - a(t_n, Y_n)\right|^2 \\ &\le 2k\left|t - t_n\right| + 2k\left|X(t) - X(t_n)\right|^2 \\ &\quad +2k\left|X(t_n) - Y_n\right|^2 \end{aligned}$$

and the same for

$$\begin{aligned} &\int_{t_n}^{t_{n+1}} \mathbb{E}\left|b(t, X(t)) - b(t_n, Y_n)\right|^2 dt \\ &\le 2k\left|t - t_n\right| + 2k\left|X(t) - X(t_n)\right|^2 \\ &\quad +2k\left|X(t_n) - Y_n\right|^2. \end{aligned}$$

We suppose the continuity of solution on $[t_0, T]$, namely $\mathbb{E}\left|X(t) - X(t_n)\right|^2 \le c\left|t - t_n\right|$, therefore

$$\begin{aligned} \mathbb{E}\left[\epsilon^2(t_{n+1})\right] &\le \mathbb{E}\left[\epsilon^2(t_n)\right] \\ &\quad + \int_{t_n}^{t_{n+1}} \mathbb{E}\left(X(t) - \widehat{Y}(t)\right)^2 dt \\ &\quad +4k(1+c)\int_{t_n}^{t_{n+1}} (t - t_n)\, dt \\ &\quad +4k\int_{t_n}^{t_{n+1}} \mathbb{E}\left(\epsilon^2(t_n)\right) dt \\ &\le \mathbb{E}\left[\epsilon^2(t_n)\right] + \int_{t_n}^{t_{n+1}} \mathbb{E}\left[\epsilon^2(t)\right] dt \\ &\quad +2k(1+c)\Delta^2 + 4k\mathbb{E}\left[\epsilon^2(t_n)\right]\Delta \\ &\le \mathbb{E}\left[\epsilon^2(t_n)\right](1 + 4k\Delta) \\ &\quad +2k(1+c)\Delta^2 + \int_{t_n}^{t_{n+1}} \mathbb{E}\left[\epsilon^2(s)\right] ds. \end{aligned}$$

Taking into account the Bellman-Grönwall [1] inequality we have

$$\begin{aligned} \mathbb{E}\left[\epsilon^2(t_{n+1})\right] &\le \mathbb{E}\left[\epsilon^2(t_n)\right](1 + 4k\Delta) + 2k(1+c)\Delta^2 \\ &\quad + \int_{t_n}^{t_{n+1}} e^{(t_{n+1} - t)}\Big[\mathbb{E}\left[\epsilon^2(t_n)\right](1 + 4k\Delta) \\ &\quad +2k(1+c)\Delta^2\Big]dt \\ &= e^{\Delta}\Big[\mathbb{E}\left(\epsilon^2(t_n)\right)(1 + 4k\Delta) \\ &\quad +2k(1+c)\Delta^2\Big]. \end{aligned}$$

Letting $R = e^{\Delta}(1 + 4k\Delta)$ and $S = e^{\Delta} 2k(1+c)\Delta^2$, then taking into account that $\mathbb{E}\left(\epsilon^2(t_0)\right) = 0$ for $n = 0, 1, 2, ..., n_{T-1}$ the inequalities

$$\mathbb{E}\left[\epsilon^2(t_{n+1})\right] \le R\mathbb{E}\left[\epsilon^2(t_n)\right] + S$$

yield

$$\mathbb{E}\left[\epsilon^2(t_{n_T})\right] \le S\frac{R^{n_T} - 1}{R - 1}.$$

Hence,

$$\begin{aligned} \mathbb{E}\left[\epsilon^2(t_{n_T})\right] &\le \frac{e^{\Delta} 2k(1+c)\Delta^2 e^{n_T\Delta} e^{4kn_T\Delta}}{e^{\Delta} - 1 + e^{\Delta} 4k\Delta} \\ &\le \Delta\frac{(1+c)\exp\left[(1+4k)T\right]}{2} \end{aligned}$$

So that, the *mean square error* satisfies

$$\mathbb{E}\left[\left|X(t_n) - Y_n\right|^2\right] \le \widehat{c}\Delta \tag{15}$$

[1] Let $\alpha, \beta : [0, T] \to \mathbb{R}$ be integrable functions with property $\alpha(t) \le \beta(t) + c\int_0^t \alpha(s)\, ds$ for $t \in [0, T]$ then $\alpha(t) \le \beta(t) + c\int_0^t \exp\left[c(t-s)\right]\beta(s) ds$.

for $n = 0, 1, ..., n_T$ where $\widehat{c} = \frac{1}{2}(1 + c)\exp[(1 + 4k)T]$.

Applying the Lyapunov inequality [2] to (15) we get the *absolute error*

$$\mathbb{E}[|X(t_n) - Y_n|] \leq \widehat{c}^{\frac{1}{2}}\Delta^{\frac{1}{2}} \tag{16}$$

In above-mentioned arguments, we have presented the uniform absolute and mean square errors bounds over the whole time interval $[t_0, T]$. These results can be formulated as the following theorem, which summarizes a well know results concerning the strong convergence of Euler's methods [18].

Theorem 2.1: Suppose that the functions a and b satisfy Lipschitz and uniform growth conditions and are Hölder continuous of order $\frac{1}{2}$ in the first variable, precisely

- $|a(t,x) - a(t,y)| + |b(t,x) - b(t,y)| \leq K|x - y|$,
- $|a(t,x)|^2 + |b(t,x)|^2 \leq K^2\left(1 + |x|^2\right)$,
- $|a(s,x) - a(t,y)| + |b(s,x) - b(t,y)| \leq K|s - t|^{\frac{1}{2}}$,

for some constant $K > 0$ and for all $s, t \in [t_0, T]$, $x, y \in \mathbb{R}$. Then for the Euler approximation there exist some positive constants $C_1 = C_1(T)$ and $C_2 = C_2(T)$, which do not depend on Δ, such that

$$\mathbb{E}[|X(t_n) - Y_n|] \leq C_1\Delta^{\frac{1}{2}} \tag{17}$$

and

$$\mathbb{E}\left[|X(t_n) - Y_n|^2\right] \leq C_2\Delta. \tag{18}$$

On the basis of (17) we can conclude that the Euler schemes converge to strong solution of Itô's SDE and have a strong order $\gamma = \frac{1}{2}$.

B. The stability improvement

Application of explicit scheme to the solution of linear ODE

$$dX_t = aX_t dt, Re(a) < 0,$$

gives

$$Y_{n+1} = (1 + a\Delta)Y_n.$$

The explicit scheme is stable if $|1 + a\Delta| < 1$.

Application of (10) to the same test equation gives

$$Y_{n+1} = \frac{1}{1 - a\Delta}Y_n$$

scheme is said to be stable for any $\Delta > 0$ if $\left|\frac{1}{1-a\Delta}\right| < 1$.

The derivative (6) can be also estimated inside the interval $[t_n, t_{n+1}]$ given the semi-implicit Euler scheme

$$Y_{n+1} = Y_n + \{(1 - \lambda)a(t_n, Y_n) + \lambda a(t_{n+1}, Y_{n+1})\}\Delta, \tag{19}$$

where $\lambda \in \mathbb{R}$ is called the degree of implicitness.

Definition 2.1: The equilibrium point or steady state $X(t) \equiv 0$ of the SDE is said to be *stochastically stable* if for all X_0 and for any $\varepsilon > 0$ and $t_0 \geq 0$ the following equality holds

$$\lim_{X_0 \longrightarrow 0} \mathbf{P}\left(sup_{t \geq t_0}|X(t; t_0, X_0)| \geq \varepsilon\right) = 0. \tag{20}$$

In addition to the above it is called *stochastically asymptotically stable*, if

$$\lim_{X_0 \longrightarrow 0} \mathbf{P}\left(\lim_{t \to \infty}|X(t; t_0, X_0)| = 0\right) = 1 \tag{21}$$

and *stochastically asymptotically stable in the large*, if

$$\mathbf{P}\left(\lim_{t \to \infty}|X(t; t_0, X_0)| = 0\right) = 1 \tag{22}$$

Definition 2.2: The equilibrium position $X(t) \equiv 0$ of the SDE is said to be *stable in p^{th} mean* if for any $\varepsilon > 0$ and $t_0 \geq 0$ there exists $\delta(t_0, \varepsilon) > 0$ which satisfies

$$\mathbb{E}[|X(t; t_0, X_0)|^p] < \varepsilon, \tag{23}$$

for all $t \geq t_0$ and $|X_0| < \delta(t_0, \varepsilon)$, and is said to be *asymptotically stable in p^{th} mean* if in addition there also exists $\delta_0(t_0, \varepsilon) > 0$ and the following equality holds

$$\lim_{t \to \infty}\mathbb{E}[|X(t; t_0, X_0)|^p] = 0 \tag{24}$$

for all $|X_0| < \delta_0(t_0, \varepsilon)$.

The cases where $p = 1$ or $p = 2$ are widely studded and called stability in mean ans mean-square stability or *MS-stability,* respectively [19].

The implicit Euler scheme for linear SDE

$$dX_t = aX_t dt + bX_t dB(t), X_0 = X(t_0) \tag{25}$$

can be written as

$$Y_{n+1} = R(\Delta, \Delta B_n)Y_n,$$

where

$$R(\Delta, \Delta B_n) = \frac{1}{1 - a\Delta - b\Delta B_n}.$$

It is clear that ΔB_n can take any values on time interval $(-\infty, +\infty)$, therefore $|R|$ can be unrestricted for any values of a, b or Δ. Thus, the implicit scheme can not be used and requires some modifications. So, we suppose that $\Delta B_n < -\frac{a\Delta}{b}$ or $\Delta B_n > \frac{2-a\Delta}{b}$, then $|R| < 1$ and value Y_{n+1} can be found by the implicit scheme with desired precision if value Y_n is sufficiently accurate. However, even in this case we can face to the problem. Value of $R(\Delta, \Delta B_n)$ can reach the infinity, if generated value ΔB_n lies in the neighborhood of $\frac{(1-a\Delta)}{b}$. In this case it is better to use the scheme implicit only by drift. In this manner we got the composite scheme of the type

[2]The Lyapunov inequality states that if X is not concentrated on a single point and if $\mathbb{E}[|X|^s]$ exists for some $s > 0$, then for all $0 < r < s$ and $Y \in \mathbb{R}$ the following inequality $(\mathbb{E}[|X - Y|^r])^{\frac{1}{r}} \leq (\mathbb{E}[|X - Y|^s])^{\frac{1}{s}}$ takes place

$$Y_{n+1} = \begin{cases} \frac{1+b\Delta B_n}{1-a\Delta} Y_n, & |R| > 1, \\ \frac{1}{1-a\Delta-b\Delta B_n} Y_n, & |R| < 1. \end{cases} \quad (26)$$

This scheme can written as follows

$$Y_{n+1} = Y_n + aY_{n+1}\Delta + [\lambda_n bY_n + (1-\lambda_n) bY_{n+1}] \Delta B_n, \quad (27)$$

where $\lambda_n \in \mathbb{R}$ is called the degree of implicitness, $\lambda_n \in [0,1]$ and is calculated on each integration step. To find the value of λ_n, we rewrite (27) once again

$$Y_{n+1} = \frac{1+\lambda_n b\Delta B_n}{1-a\Delta-(1-\lambda_n) b\Delta B_n} Y_n. \quad (28)$$

Let

$$f(\lambda) = \frac{1+\lambda b\Delta B_n}{1-a\Delta-(1-\lambda) b\Delta B_n},$$

then it is easy to get

$$\min_{\lambda\in[0,1]} |f(\lambda)| = \min\{|f(0)|, |f(1)|\}.$$

We introduce

$$\begin{array}{ll} Q_{n0} = & f(0) = \frac{1}{1-a\Delta-b\Delta B_n}, \\ Q_{n1} = & f(1) = \frac{1+b\Delta B_n}{1-a\Delta}. \end{array}$$

Then the criteria for optimal value selection of λ_n can be formulated as

$$\lambda_n = \begin{cases} 0, & |Q_{n0}| < |Q_{n1}|, \\ 1, & |Q_{n0}| > |Q_{n1}|. \end{cases}$$

The composite scheme for nonlinear SDE. Suppose that Y_{n+1} and Y^*_{n+1} are two numerical solutions of with Y_n and Y^*_n, then

$$\begin{aligned} Y_{n+1} &= Y_n + a(t_{n+1}, Y_{n+1})\Delta + \Big[\lambda_n b(t_{n+1}, Y_{n+1}) \\ &\quad + (1-\lambda_n) b(t_{n+1}, Y_{n+1})\Big]\Delta B_n \\ Y^*_{n+1} &= Y^*_n + a\left(t_{n+1}, Y^*_{n+1}\right)\Delta + \Big[\lambda_n b\left(t_{n+1}, Y^*_{n+1}\right) \\ &\quad + (1-\lambda_n) b\left(t_{n+1}, Y^*_{n+1}\right)\Big]\Delta B_n \end{aligned}$$

gives the possibilities for the approximation

$$\begin{aligned} Y^*_{n+1} &- Y_{n+1} \\ &\approx \frac{\left(1+\lambda_n \frac{\partial b}{\partial X}\Delta B_n\right)\Big|_{X=Y_n}}{\left(1-\frac{\partial a}{\partial X}\Delta+(\lambda_n-1)\frac{\partial b}{\partial X}\Delta B_n\right)\Big|_{X=Y_{n+1}}} \\ &\quad \times (Y_n - Y^*_n) \\ &\approx \frac{\left(1+\lambda_n \frac{\partial b}{\partial X}\Delta B_n\right)}{\left(1-\frac{\partial a}{\partial X}\Delta+(\lambda_n-1)\frac{\partial b}{\partial X}\Delta B_n\right)}\Bigg|_{X=Y_n} \\ &\quad \times (Y_n - Y^*_n). \end{aligned}$$

Let

$$R_{n0} = \frac{1}{1-\frac{\partial a}{\partial X}\Delta-\frac{\partial b}{\partial X}\Delta B_n}\Bigg|_{X=Y_n}$$

and

$$R_{n0} = \frac{1+\frac{\partial b}{\partial X}\Delta B_n}{1-\frac{\partial a}{\partial X}\Delta}\Bigg|_{Y=Y_n},$$

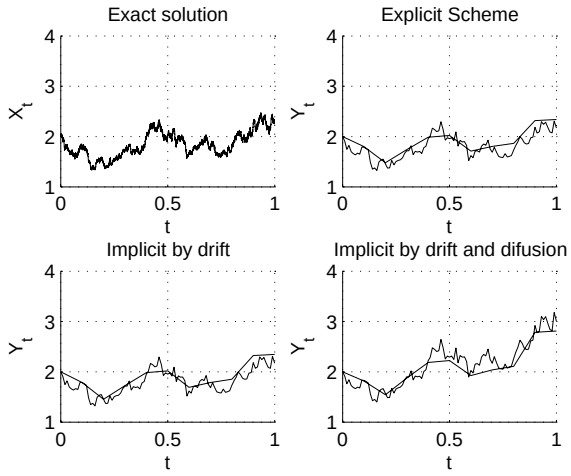


Fig. 1. Exact solution and Euler approximation for SDE (30)

then

$$\lambda_n = \begin{cases} 0, & |R_{n0}| \le |R_{n1}|, \\ 1, & |R_{n0}| > |R_{n1}|. \end{cases}$$

Solution can be obtain by iterations. Let

$$\begin{aligned} F(Y) &= Y - a(t_{n+1}, Y)\Delta - (1-\lambda_n) b(t_{n+1}, Y)\Delta B_n \\ &\quad - Y_n - \lambda_n b(t_{n+1}, X)\Delta B_n, \end{aligned}$$

then Newton-Raphson method is

$$Y_{n+1} = Y_n - \frac{F(Y_n)}{F'(Y_n)} \equiv \phi(Y_n). \quad (29)$$

It converges if

$$|\phi'(Y_n)| = \left|\frac{F(Y)\left(-\frac{\partial^2 a}{\partial Y^2}\Delta-(1-\lambda_n)\frac{\partial^2 b}{\partial Y^2}\Delta B_n\right)}{\left(1-\frac{\partial a}{\partial Y}\Delta-(1-\lambda_n)\frac{\partial b}{\partial Y}\Delta B_n\right)^2}\right| < 1.$$

III. Numerical illustrations

Example 3.1: we will illustrate the application of (11) - (13) and (26) for numerical solution for following SDE

$$dX_t = 0.5X_t dt + 0.5X_t dB(t) \quad (30)$$

with initial condition $X(t_0) = 2.0$ on time interval $t \in [0,1]$. The exact solution of (30)

$$X(t) = 2.0\exp[0.375t + 0.5B(t)].$$

is generated with step size $\Delta_{EX} = 10^{-4}$. The numeric approximation is completed twice for $\Delta = \{10^{-1}, 10^{-2}\}$ using MathLab codes listed in Appendex. The results are listed on figures 1 and 2. As it is possible to see the smaller stepsize gives the smaller error of approximation.

Example 3.2: approximation of non-linear SDE using Euler's numerical schemes (11) - (13) and (26). Consider the stochastic differential equation

$$dX_t = \frac{1}{2}a^2 mX_t^{2m-1}dt + aX_t^m dB_t, \qquad m \neq 1, \quad (31)$$

with analytical solution given as

$$X_t = \left(X_0^{1-m} - a(m-1)B_t\right)^{\frac{1}{(1-m)}}$$

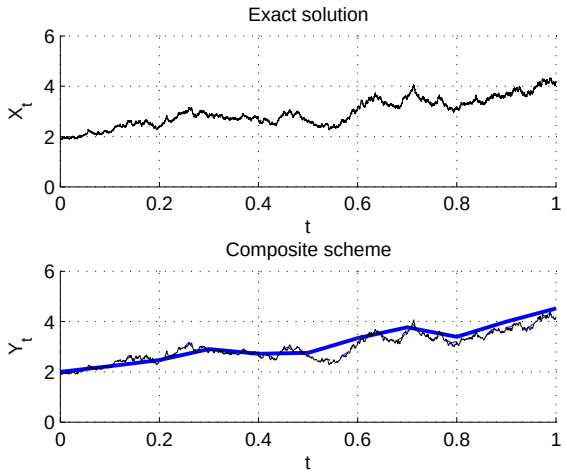


Fig. 2. The trajectories of exact and numerical solution by composite method for linear SDE (30)

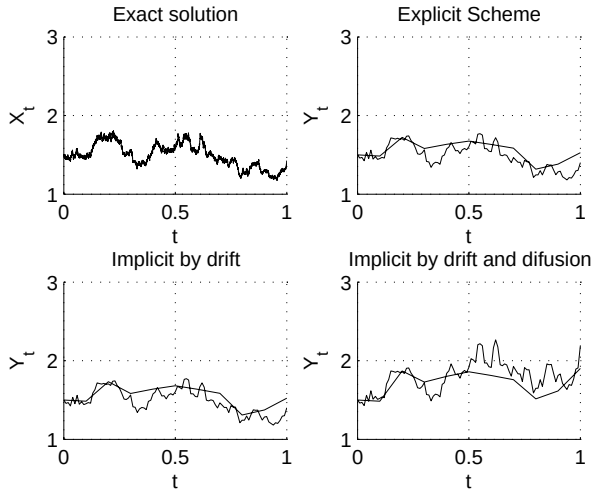


Fig. 3. The trajectories of exact and numerical solution for nonlinear SDE (31)

on time interval $t \in [0,1]$ and with initial value $X(t_0) = 1.5$ and parameters $a = 0.25$ and $m = 2$. The rest of experimental conditions are as in example 3.1. The calculation results are plotted in Fig. 3 and Fig.4.

Example 3.3: the strong convergence of the explicit Euler scheme. Consider two linear SDEs

$$dX_t = -1.5X_t dt + 0.5X_t dB(t), \quad (32)$$

$$dX_t = 10.0X_t dt + 2.0X_t dB(t) \quad (33)$$

with initial value $X(t_0) = 1$.

To examine (17) we repeat $M = 10000$ different simulations of sample paths of each test equations and their Euler approximations corresponding to the same sample paths of the Wiener process. Denote the values at time T of the kth exact solution and simulated trajectories by $X_{T,k}$ and $Y_{T,k}$, respectively. Therefore, the estimate of the absolute error (17) can be written as

$$\widehat{\epsilon} = \frac{1}{M}\sum_{k=1}^{M} |X_{T,k} - Y_{T,k}|. \quad (34)$$

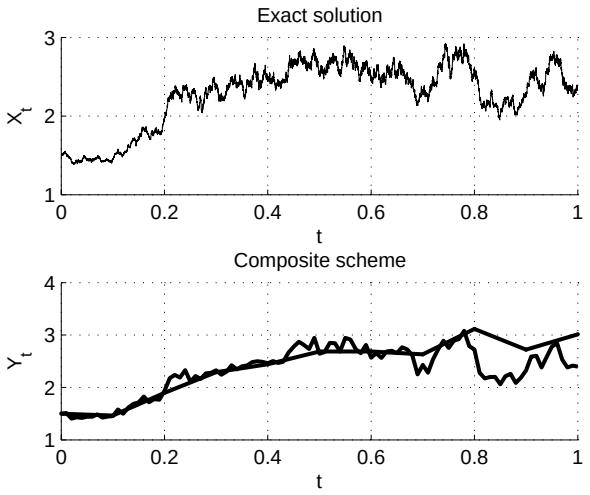


Fig. 4. The trajectories of exact and numerical solution by composite method for nonlinear SDE (31)

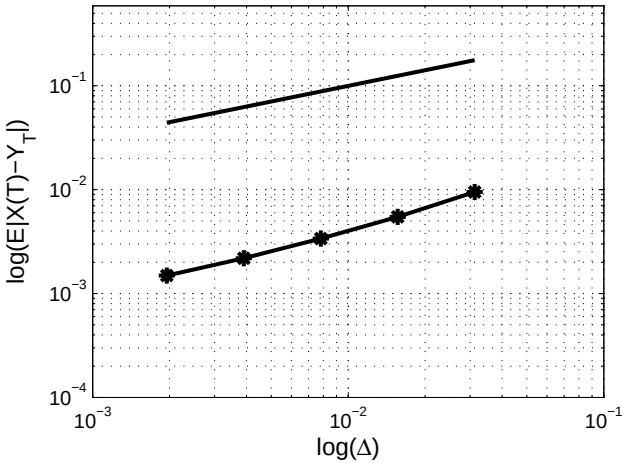


Fig. 5. The absolute error for SDE (32)

First we generated the trajectories of Wiener process with equidistant time steps of step size $\Delta_X = 2^{-9}$ and calculated the exact solution value at the point $T = 1$ for each of test equations. Trajectories of the Euler approximation (11) were simulated with different step size $\Delta_Y = \{2^{-9}, 2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}\}$ on the time interval $t\,[0,1]$. The statistics $\widehat{\epsilon}$ values defined by (34) are listed on figures 5 and 6.

APPENDIX: MATHLAB CODES

- Function for the independent random variables generation

```
%Independent random variables generation
function [G1,G2]=generate
v11=1; v21=1; w=v11*v11+v21*v21;
        while w > 1,
                v11=2*rand(1)-1;
                v21=2*rand(1)-1;
                w=v11*v11+v21*v21;
        end;
```

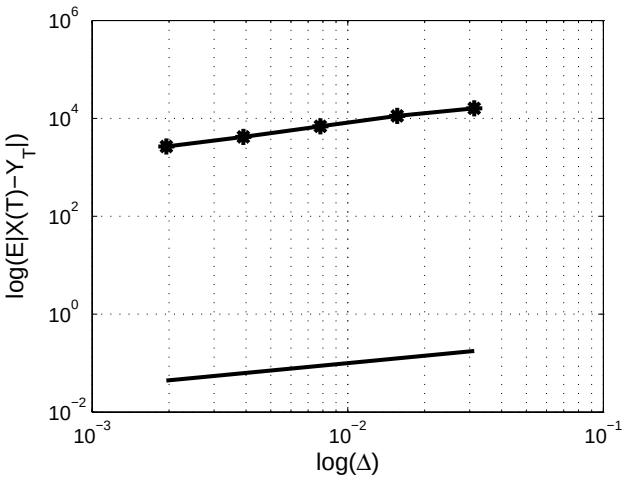


Fig. 6. The absolute error for SDE (33)

```
LW=log(w)/w; LW=sqrt(-LW-LW);
G1=v11*LW; G2=v21*LW;
```

- Function for Wiener process generation

```
% Wiener process generation
function [Wt]=Wiener(K,D);
Wt(1,1)=0;
for i=2:K+1
    [g1,g2]=generate;
    Wt(i,1)=Wt(i-1,1)+g1*sqrt(D);
end;
```

- The code for the numerical approximation by different Euler schemes

```
% SDE and schemes parameters
t0 = 0.0; tk = 1.0; x0 = 2.0; a = .50; b = .50;
delta = power(10, -4); numinv = (tk - t0)/delta;
eps = power(10, -6);
tx = (t0 : delta : tk)'; xt(1,1) = x0;
zt(1,1)=x0; yt(1,1) = x0; yyt(1,1) = x0;
[Bt] = Wiener(numinv,delta);

% Exact solution
for i = 2 : numinv + 1
    xt(i,1) = x0*exp((a - 0.5*b*b)*tx(i,1) + b*Bt(i,1));
end;

for h = 1 : 2
    deltay = power(10, -h); k = deltay/delta;
    numint=(tk-t0)/deltay;
    ty = (t0 : deltay : tk)';
    for j = 2 : numint + 1

% Explicit Euler Scheme
        zt(j,1) = zt(j-1,1) + zt(j-1,1)*a*deltay
            + zt(j-1,1)*b*(Bt((j-1)*k+1,1) - Bt((j-2)*k+1,1));

% Implicit by drift Euler Scheme
        yt(j,1) = yt(j-1,1);
        d = 1;
        yy = yt(j,1);
        ww = yt(j,1)*(Bt((j-1)*k+1,1) - Bt((j-2)*k+1,1));
        while d > eps,
            yn = yt(j,1) + a*yy*deltay + b*ww;
            d = abs(yn-yy);
            yy = yn;
        end;
        yt(j,1) = yn;

%Implicit by drift and diffusion Euler Scheme
        yyt(j,1) = yyt(j-1,1);
        d = 1;
        yy = yyt(j,1);
        ww = (Bt((j-1)*k+1,1) - Bt((j-2)*k+1,1));
        while d > eps,
            yn = yyt(j,1) + a*yy*deltay + yy*b*ww;
            d = abs(yn-yy);
            yy = yn;
        end;
        yyt(j,1)=yn;
    end;
end;


% Composite Euler Scheme
eps = 0.001;
for h = 1 : 3
    deltay = power(10, -h);
    k = deltay/delta;
    numint = (tk - t0)/deltay;
    ty = (t0 : deltay : tk)';
    yyt(1,1) = y0;
    for j = 2 : numint + 1
        yyt(j,1) = yyt(j-1,1);
        d = 1;
        yy = yyt(j,1);
        ww = (Bt((j-1)*k+1,1) - Bt((j-2)*k+1,1));
        Qn0 = 1/(1 - a*deltay - b*ww);
        Qn1 = (1 + b*ww)/(1 - a*deltay);
        if (abs(Qn0) <= abs(Qn1))
          lam = 0;
          else lam = 1;
        end;
        while d > eps,
        yn = yyt(j,1) + a*yy*deltay + (lam*yyt(j,1)*b
          + (1 - lam)*yy*b)*ww;
        d = abs(yn - yy);
        yy = yn;
        end;
        yyt(j,1) = yn;
    end;
end;
```

- The code for the absolute error estimation

```
% SDE and schemes parameters

randn('state',100)
a = -1.5; b = .5; X0 = 1; T = 1;
```

```
N = power(2, 9); dt = T/N; M = 10000;
Xerr = zeros(M,5);

% Estimate of absolute error

for s =1 : M
      dB = sqrt(dt)*randn(1,N);
      B = cumsum(dB);
      Xtrue = X0*exp((a - 0.5*b*b)*T + b*B(end));
      for p = 1 : 5
          R = power(2,p - 1); Dt = R*dt; L = N/R;
          Xtemp = X0;
          for j = 1 : L
              Binc = sum(dB(R*(j - 1) + 1 : R*j));
              Xtemp = Xtemp + Dt*a*Xtemp
              + b*Xtemp*Binc;
          end;
          Xerr(s,p) = abs(Xtemp - Xtrue);
      end;
end
```

ACKNOWLEDGMENT

This paper was supported by grant 04.0.10.00/2.01.01.01.0026 MNSP.ZKIP.17.005.

REFERENCES

[1] B. Oksendal, *Stochastic differential equations*. Springer, 2000.

[2] P. E. Kloden and E. Platen, *Numerical solution of Stochastic Differential Equations*. Springer, 1999.

[3] P. E. Kloden, E. Platen, and H. Schurz, *Numerical solution of SDE through computer experiments*. Springer, 1997.

[4] M. Carletti, K. Burrage, and P. M. Burrage, "Numerical solution of stochastic ordinary differential equations in biomathematical modeling," *Mathematics and computers in Simulation*, vol. 64, pp. 271–277, 2004.

[5] D. Higham, X. Mao, and A. Stuart, "Strong convergence of euler-type methods for nonlinear stochastic differential equations," *Journal of Numerical Analysis*, vol. 24, no. 3, pp. 1041–1063, 2002.

[6] J. C. Butcher, *Numerical methods for ordinary differential equations*. Wiley, 2003.

[7] D. F. Kuznetsov, "The three-step strong numerical methods of the orders of accuracy 1.0 and 1.5 for ito stochastic differential equations," *Journal of Automation and Information Sciences*, vol. 34, no. 12, pp. 22–35, 2002.

[8] G. N. Milstein and M. V. Tretyakov, "Evaluation of conditional wiener integrals by numerical integration of stochastic differential equations," *Journal of Computational Physics*, vol. 197, pp. 275–298, 2004.

[9] K. Burrage and T. Tian, "Stiffly accurate runge-kutta methods for stiff stochastic differential equations," *Computer Physics Communications*, vol. 142, pp. 186–190, 2001.

[10] T. H. Tian and K. Burrage, "Two-stage stochastic runge-kutta methods for stochastic differential equations," *BIT*, vol. 42, no. 3, pp. 625–643, 2002.

[11] A. Tocino and J. Vigo-Aguiar, "Weak second order conditions for stochastic runge-kutta methods," *Journal of Scientific Computations*, vol. 24, no. 2, pp. 507–523, 2002.

[12] A. B. Bakushinsky and M. Y. Kokurin, "On the construction of stable iterative methods for solving ill-posed nonlinear equations with nondifferentiable operators," *J. Inv. Ill-posed Problems*, vol. 11, no. 4, pp. 329 – 341, 2003.

[13] Y. G. Bulychev and A. V. Yeliseyev, "The stiffness problem for stochastic systems and a method of solving it," *J. Appl. Maths and Mechs.*, vol. 62, no. 6, pp. 877–882, 2017.

[14] J. R. Cach, "A comparison of some codes for the stiff oscillatory problem," *Computes Math. Applic*, vol. 36, no. 1, pp. 51–57, 1998.

[15] C. W. Gear and I. G. Kevrekidis, "Projective methods for stiff differential equations: problems with gaps in their eigenvalue spectrum," *SIAM J. Sci. Comput.*, vol. 24, no. 4, pp. 1091–1106, 2003.

[16] K. Burrage, P. Burrage, and T. Mitsui, "Numerical solutions of stochastic differential equations implementation and stability issues," *Journal of computational and applied mathematics*, vol. 125, no. 171–182, 2000.

[17] B. Schmalfuss, "Lyapunov function and non-trivial stationary solutions of stochastic differential equations," *Dynamical Systems*, vol. 16, no. 4, pp. 303–317, 2001.

[18] E. Platen and D. Heath, *A benchmark approach to quantitative finance*. Springer, 2006.

[19] I. Kolmanovsky and T. L. Maizenberg, "Mean-square stability of nonlinear systems with time-varying random delay," *Stochastic analysis and applications*, vol. 19, no. 2, pp. 279 – 293, 2001.

Necessary Optimality Conditions for Enterprises Production Programs

Dorota Bochnacka
Economics Department
Faculty of Management and Computer Modeling
Kielce University of Technology
al. Tysiąclecia Państwa Polskiego 7, 25–314 Kielce, Polska
Email: dbochnacka@tu.kielce.pl

Darya Filatova
Production Engineering Department
Faculty of Management and Computer Modeling
Kielce University of Technology
al. Tysiąclecia Państwa Polskiego 7, 25–314 Kielce, Polska
Email: dfilatova@tu.kielce.pl

***Abstract*—An idealized industrial group, consisting of two enterprises with vertical cooperation, is considered in this paper. We assume that the both enterprises, having necessary production capacities, act on the product markets once as the partners and once as the independent actors. The main problem of this kind of cooperation is to find the optimal conditions for the production programs for the both enterprises. The possible way of the mathematical problem formulation, as well as its relation to the optimal control problem, is the essence of this paper. The theoretical backgrounds for the solution of the optimal cooperation production program are developed using the Dubovitski-Milyutin method.**

I. Introduction

According to the World Bank national account data, the dominant part of GDP of high-income economics is manufactured by enterprises controlled by financial and industrial groups. Despite different expansion, various legal forms of production, the most important criteria that these groups meet for the efficient functioning are the possibilities for cash generation within a group, high degree of business diversification, optimal integration within the chosen business areas, high degree of property's concentration, corporate control minimized hierarchy (see [1]). Having in mind benefits coming from financial and industrial capital integration transition economies try to adopt these achievements in spite of political, legal, trade and investment problems, competition for entry into new markets and other barriers. These factors affect not only the efficiency of cooperation within the group but also the decision of cooperation's initiation(see [2], [3]). Taking into account the economic importance, the financial-industrial or industrial groups (FIG) determine the object and the maximization of economic efficiency of these groups makes the subject of paper's study. We insist, that "group" means any form of collaboration. Whereas the vertical cooperation in production chain motivates innovations, the object of study will be narrowed to FIG with vertical cooperation on the mesoeconomic level (see [4]).

The primary goal of the paper consists in the determination of those conditions of members' interaction, which could ensure the stable economic development of FIG and its maximum effectiveness on the long planning horizon using optimal control methods. We assume that there exist such information parameters (general for all FIG) allow on the statement of each problem as a stochastic/deterministic optimal control problem and its solution.

Research methodology is be based: on information approach to research (this allows the selection and analysis of the information characteristics of the research object, on the understanding the essence of the studied phenomenon and the principles of its development) and on system analysis (utilizing optimal control approach).

The structure of the rest of the paper is as follows. In Section II, we formulate the optimal cooperation model taking into account restrictions from the EU law and related documents. In Section III, firstly, we reduce the stochastic optimal control problem to the deterministic canonical form of the Pontryagin type; secondly, we derive the necessary optimality conditions for the production program. Finally, Section IV provides some conclusions.

II. Problem of optimal cooperation

To introduce the concept and mathematical models of the industrial group with vertical cooperation we use "Regulation No 2790/1999" and "Commission Regulation (EU) No 330/2010 of 20 April 2010 on the application of Article 101 (3) of the Treaty on the Functioning of the European Union to categories of vertical agreements and concerted practices" (see http://eur-lex.europa.eu)

A. The industrial group: basic concepts and notations

Some industrial group includes vertical cooperation of two business, which acquire the necessary resources for the manufacturing, i.e.: the enterprise-subcontractor (*Enterprise 1*) with *Product* A_1 and the enterprise-producer (*Enterprise 2*) with *Product* A_2. That is to say, in the vertical technological chain, *Enterprise 2* uses *Product* A_1 as a component to create its own product. Let $t \in [t_0, t_1] \subset \mathbb{R}_+$ be a time moment of cooperation and $j \in \{1, 2\}$. We assume that the manufactured volume X_j of the *Product* A_j depends only on the vector of production capacity parameters $\theta^{(j)} = \left[\theta_1^{(j)}, \theta_2^{(j)}, \theta_3^{(j)}\right]$, where the production growth $\theta_1^{(j)}$ is a function from $[t_0, t_1]$ to $\mathbb{R}$, the maximal production capacity $\theta_2^{(j)}$ is a function from $[t_0, t_1]$ to $\mathbb{R}_+$, and the rate of the defective unit production $\theta_3^{(j)}$ is a positive constant function of $t \in [t_0, t_1]$. The market price

is given by the function $p_j^{(0)}$ from $[t_0, t_1]$ to $\mathbb{R}_+$, the total volume of *Product* A_j available on the market is presented by function $X_j^{(0)}$ from $[t_0, t_1]$ to $\mathbb{R}_+$. The manufacturing activity of the group is considered only on product market, which is characterized by the demand function $d_j = \zeta\left(p_j^{(0)}, X_j^{(0)}\right)$, where ζ is a $\mathcal{C}^2$ (decreasing convex) function from $\mathbb{R}_+^2$ to $\mathbb{R}_+$. The selling price $p_j^{(1)}$ of the *Product* A_j is a function of the demand and the total volume of *Product* A_j available on the market, i.e. $p_j^{(1)} = h_j^{(1)}\left(d_j, X_j^{(0)}\right)$, where $h_j^{(1)}$ is a $\mathcal{C}^2$ (decreasing convex) function from $\mathbb{R}_+^2$ to $\mathbb{R}_+$. It is assumed that the overall production X_j does not exaggerate a demand d_j as well as the production of *Product* A_j does not bring the loss to *Enterprise j*, when its selling price covers the manufacturing costs.

The cooperation principle can be summarized as follows. *Enterprise 1* is interested to cooperate with *Enterprise 2* since it could guaranty the partial or even total distribution of the *Product* A_1. To make such cooperation successful *Enterprise 1* must offer to *Enterprise 2* the price - $p_1^{(*)}$ - below the selling $p_1^{(1)}$ price, in the contrary *Product* A_1 will be directly marketed, which could bring the decrease of income from its sale. *Enterprise 2* can buy *Product* A_1 from *Enterprise 1* or from the market. It is assumed that the demand on *Product* A_1 by *Enterprise 2* does not surpass the production capacity of *Enterprise 1*. However, since *Enterprise 1* can refuse to sell of the required volume of *Product* A_1 to *Enterprise 2* , *Enterprise 2* has to buy *Product* A_1 from the market. The income of *Enterprise 2* depends on the favorable purchase conditions of *Product* A_1, the demand d_2 and selling price $p_2^{(1)}$.

B. The production model

Let $B_t^{\mathbb{H}} = \{B(t, \mathbb{H}), t \in [t_0, t_1]\}$, $0 \le t_0 \le t_1 \le +\infty$, be a fractional Brownian motion (fBm) with Hurst parameter $\mathbb{H} \in (0, 1)$. For $\mathbb{H} = 0.5$ fractional Brownian motion is a Brownian motion. Define $\left(\Omega, \{\mathcal{F}_t\}_{t\ge 0}, \mathbf{P}\right)$ a complete probability space where $\{\mathcal{F}_t\}_{t\ge 0}$ is the natural filtration generated by $B_t^{\mathbb{H}}$, augmented by all the $\mathbf{P}$-null sets in $\{\mathcal{F}_t\}_{t\ge 0}$.

We assume that the overall production of *Product* A_1 and *Product* A_2 satisfies the following differential system:

$$dX_1(t) = f_1(X_1(t))\,dt, \quad (1)$$

$$dX_2(t) = f_2(X_2(t))\,dt + \theta_3^{(2)} X_2(t)\,dB_t^{\mathbb{H}}, \quad (2)$$

with $X_1(t_0) = x_{10}$, $(x_{10} \le \theta_2^{(1)}(t_0))$, $X_2(t_0) = x_{20}$, $(x_{20} \le \theta_2^{(2)}(t_0))$, where

$$f_1(X_1(t)) = \theta_1^{(1)}(t) X_1(t)\left(1 - \frac{X_1(t)}{\theta_2^{(1)}(t)}\right),$$

$$\begin{aligned} f_2(X_2(t)) &= u_{12}(t) X_1(t) \\ &\quad + \theta_1^{(2)}(t)\left(X_2(t) - u_{12}(t) X_1(t)\right) \\ &\quad \times \left(1 - \frac{X_2(t) - u_{12}(t) X_1(t)}{\theta_2^{(2)}(t) - u_{12}(t) X_1(t)}\right), \end{aligned}$$

$u_{12}(t)$ is the rate of the sale of *Product* A_1 to *Enterprise 2,* and $0 \le u_{12}(t) \le 1$, $dB_t^{\mathbb{H}}$ are uncorrelated fBm increments in the sense that for

$$\begin{aligned} X_2(t) - x_{20} &= \int_{t_0}^{t} f_2(X_2(\tau))\,d\tau \\ &\quad + \int_{t_0}^{t} \theta_3^{(2)} X_2(t)\,dB_\tau^{\mathbb{H}}, \end{aligned} \quad (3)$$

second integral can be understand as a stochastic Skorokhod integral with respect to the fBm. We also suppose that the conditions of *theorem 3.1* (see [5], p.338) are fulfilled, therefore Eq. (3) has a unique solution $X_2(t)$.

C. The profit model

We suppose that due to the defective units the overall production $X_j(t)$ depends on the coefficients $\gamma_j \in (0, 1)$, the profit π_j of *Enterprise j* is defined as a difference between the total revenue - $f_j^{(2)}$ (it can be the total income a firm receives) and the total costs - $f_j^{(1)}$ (it can be a sum of the variable and constant costs - VC_j and CC_j), namely [6]:

$$\pi_j\left(X_j^{\gamma_j}(t), \mathbf{u}(t)\right) = ff_j^{(2)}(t) - ff_j^{(1)}(t), \quad (4)$$

with

$$\begin{aligned} ff_1^{(1)}(t) &= VC_1(t) + CC_1(t), \\ ff_1^{(2)}(t) &= \left(u_{11}(t) p_1^{(0)}(t) + u_{12}(t) p_1^{(*)}(t)\right) X_1^{\gamma_1}(t) \end{aligned}$$

and

$$\begin{aligned} ff_2^{(1)}(t) &= u_{12}(t) p_1^{(*)}(t) X_1^{\gamma_1}(t) \\ &\quad + (1 - u_{12}(t)) p_1^{(0)}(t) X_1^0(t) + CC_2(t), \\ ff_2^{(2)}(t) &= u_2(t) p_2^{(0)}(t) X_2^{\gamma_2}(t), \end{aligned}$$

where $u_{11}(t) \ge 0$ and $u_{12}(t) \ge 0$ are the rates of *Product* A_1 sold by *Enterprise 1* on the market and *to Enterprise 2*, $0 \le u_{11}(t) + u_{12}(t) \le 1$, $u_2(t)$ is the rate of *Product* A_2 sold by *Enterprise 2*, $0 \le u_2(t) \le 1$, $\forall t \in [t_0, t_1]$, $\mathbf{u}(t) = [\theta_1^{(1)}(t), u_{11}(t), u_{12}(t), u_2(t), p_1^{(*)}(t)]$.

It is assumed that

- the total production of *Product* A_j does not exceed the total volume of *Product* A_j available on the market

$$X_j^{\gamma_j}(t) \le X_j^{(0)}, \forall t \in [t_0, t_1], \quad (5)$$

- the total sales of *Product* A_1 and *Product* A_2 do not exceed the demand, i.e.

$$\left(u_{11}(t) + u_{12}(t)\right) X_1^{\gamma_1}(t) \le d^{(1)}(t), \quad (6)$$

$$u_2(t) X_2^{\gamma_2}(t) \le d^{(2)}(t). \quad (7)$$

The integrated averaged profit of *Enterprise j* over the planning horizon $[t_0, t_1]$ can be calculated as

$$\Pi_j(t_1) = \int_{t_0}^{t_1} e^{-\kappa_j(t)t}\ \mathbb{E}\left[\pi_j\left(X_j^{\gamma_j}(t), \mathbf{u}(t)\right)\right] dt, \quad (8)$$

where $\kappa_j(t)$ is the discount rate used by *Enterprise j,* $\mathbb{E}[\cdot]$ stands for the mathematical expectation.

The goal of cooperation is to maximize the integrated averaged profit

$$\mathcal{J}\left(X_j^{\gamma_j}(t),\mathbf{u}(t)\right)$$

$$=\max_{\mathbf{u}(t)}\int_{t_0}^{t_1}\sum_{j=1}^{2}e^{-\kappa_j(t)t}\ \mathbb{E}\left[\pi_j\left(X_j^{\gamma_j}(t),\mathbf{u}(t)\right)\right]dt \quad (9)$$

taking into account dynamics of production, prices and market demand.

The problem (9) with related constraints is considered as a stochastic control problem due to the stochastic term in Eq.(2) (see [7], [8]). Its solution can be obtained by reduction to deterministic case. The next section discusses required transformations and the optimal task solution method.

III. Necessary optimality conditions for production program

A. Some required transformations

Let us transform the stochastic differential equation to the deterministic one. We denote $x_1=\mathbb{E}[X_1^{\gamma_1}]$, $x_2=\mathbb{E}[X_2^{\gamma_2}]$. Applying the fractional difference filter, we can find the approximations of the moments $\mathbb{E}[X_1^{\gamma_1}]$ and $\mathbb{E}[X_2^{\gamma_2}]$ and rewrite the dynamics of the overall production of *Product* A_1 and *Product* A_2 as follows ([9]):

$$x_1(t)-x_{10}^{\gamma_1}=\int_{t_0}^{t_1}f_1(x_1(\tau))\,d\tau \quad (10)$$

and

$$x_2(t)-x_{20}^{\gamma_2} = \gamma_2\int_{t_0}^{t_1}f_2(x_2(\tau))\,d\tau + \frac{\gamma_2(\gamma_2-1)}{2}\theta_3^2\int_{t_0}^{t_1}x_2(t)\,(d\tau)^{2\mathbb{H}}, \quad (11)$$

where for $0<\mathbb{H}<\frac{1}{2}$

$$\int_{t_0}^{t_1}x_2(t)\,(d\tau)^{2\mathbb{H}}=\mathbb{H}\int_{t_0}^{t_1}\frac{1}{(t-\tau)^{1-2\mathbb{H}}}x_2(t)\,d\tau \quad (12)$$

and for $\frac{1}{2}<\mathbb{H}<1$

$$\int_{t_0}^{t_1}x_2(t)\,(d\tau)^{2\mathbb{H}}=\mathbb{H}^2\left[\int_{t_0}^{t_1}\frac{\sqrt{x_2(t)}}{(t-\tau)^{1-\mathbb{H}}}d\tau\right]^2. \quad (13)$$

These transformations allow to treat the stochastic optimal control problem (the object equations (1)–(2) and related constraints (5) – (7)) as the deterministic one.

B. The Canonical optimal control problem of the Pontryagin type

Definition 3.1: Any $(\mathbf{x},\mathbf{u})$ is called an admissible point and and any $(\mathbf{x}(t),\mathbf{u}(t))$ is called an admissible process, if the conditions of the problem (14) - (18) are fulfilled.

Definition 3.2: A process $(\widehat{\mathbf{x}}(t),\widehat{\mathbf{u}}(t))$ is called an optimal one if there exists an $\varepsilon>0$ such that, for any admissible process $(\mathbf{x}(t),\mathbf{u}(t))$ satisfying the restriction

$$\|\mathbf{x}(\cdot)-\widehat{\mathbf{x}}(\cdot)\|_{\mathcal{C}(\Delta,\mathbb{R}^n)}<\varepsilon,$$

one has

$$\mathcal{J}(\mathbf{x}(\cdot),\mathbf{u}(\cdot))\geq\mathcal{J}(\widehat{\mathbf{x}}(\cdot),\widehat{\mathbf{u}}(\cdot)).$$

Let us introduce the canonical optimal problem of the Pontryagin type. Denote a time interval $\Delta=[t_0,t_1]$ and functions $\Phi:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}$, $\phi:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}$, $g_i:\mathbb{R}^n\to\mathbb{R}$ $(1\leq i\leq\ell_1)$, $\varphi_s:\mathbb{R}^r\to\mathbb{R}$ $(1\leq s\leq\ell_2)$, $\xi_k:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}$ $(1\leq k\leq\ell_3)$. In general case the task can be stated as follows:

$$\mathcal{J}(\mathbf{x}(\cdot),\mathbf{u}(\cdot))=\max_{\mathbf{u}(t)}\int_{t_0}^{t_1}\Phi(\mathbf{x}(t),\mathbf{u}(t))\,dt \quad (14)$$

subjected to

- the object equation:

$$\mathbf{x}(t)=\int_{t_0}^{t}\phi(t,\tau,\mathbf{x}(\tau),\mathbf{u}(\tau))\,d\tau \text{ a.e. on } \Delta \quad (15)$$

with $\mathbf{x}(t_0)=\mathbf{x}_0$, $\mathbf{u}\in U$,

- the phase constraints:

$$g(\mathbf{x}(t))\leq 0, \quad (16)$$

- the control constraints:

$$\varphi(\mathbf{u}(t))\leq 0, \quad (17)$$

- the mixed constraints:

$$\xi(\mathbf{x}(t),\mathbf{u}(t))\leq 0, \quad (18)$$

where $\mathbf{x}(\cdot)\in\mathcal{AC}(\Delta,\mathbb{R}^n)$, $\mathbf{u}(\cdot)\in\mathcal{L}^\infty(\Delta,\mathbb{R}^r)$, $\Phi\in\mathcal{C}^1$, $\phi\in\mathcal{C}^1$ in $\mathbf{x}$ and continuous in $\mathbf{u}$, the set $U\subset\mathbb{R}^r$ is arbitrary, $g\in\mathcal{C}^1$, $\varphi\in\mathcal{C}^1$, and $\xi\in\mathcal{C}^1$ are vector functions of dimension ℓ_1, ℓ_2, and ℓ_3 respectively. The admissible variables $(\mathbf{x},\mathbf{u})\in\mathcal{Q}$, where $\mathcal{Q}\subset\mathbb{R}^{n+r}$ is an open set.

The straight application of the above formulation faces some difficulties by applying the classical Dubovitski-Milyutin method [10] to the problem (9), (10) -(13) with related constraints. Indeed, this method requires the transformation of the goal function (9) to vector function. Since (9) presents the antagonist goals of the enterprises with respect to the price of the *Product A1* $p_1^{(*)}$ we propose a following method. Let, according to the agreement between the partners, *Enterprise 1* wants to have the guaranteed profit

$$\pi_1(x_1(t),\mathbf{u}(t))\in[\pi^*_{\min},\pi^*_{\max}],\ \forall t\in\Delta, \quad (19)$$

that corresponds to (18) and, thus, the goal function (9) takes the form (14)

$$\mathcal{J}(x_2(t),\mathbf{u}(t))=\max_{\mathbf{u}(t)}\int_{t_0}^{t_1}e^{-\kappa_2(t)t}\ \pi_2(x_2(t),\mathbf{u}(t))\,dt. \quad (20)$$

Besides the object equations (11) -(13) show singularity, so we have to extend the calculus of variation scheme.

C. Necessary optimality conditions

1) The case of $0 < \mathbb{H} < \frac{1}{2}$*:* Define a nonlinear operator

$$P : (x_1, x_2, \mathbf{u}) \in \mathcal{C} \times \mathcal{C} \times \mathcal{L}^\infty \to (z, \zeta) \in \mathcal{C} \times \mathcal{C}$$

where

$$z(t) = x_1(t) - \int_{t_0}^{t} f^{(1)}(x_1(\tau), \mathbf{u}(\tau))\, d\tau$$

and

$$\begin{aligned}\zeta(t) &= x_2(t) - \int_{t_0}^{t} f^{(2)}(x_2(\tau), \mathbf{u}(\tau))\, d\tau \\ &\quad -K \int_{t_0}^{t} \frac{x_2(\tau)}{(t-\tau)^{1-2\mathbb{H}}} d\tau,\end{aligned}$$

where $K = \frac{\gamma_2(\gamma_2-1)}{2}\theta_3^2\mathbb{H}$. The equation $P(x_1, x_2, \mathbf{u}) = 0$ is the equivalent of the system (15) in a form (11) -(13). Let $(x_1, x_2, \mathbf{u})$ be an admissible point of the problem. We assume that $g(\mathbf{x}(t_0)) < 0$ and $g(\mathbf{x}(t_1)) < 0$. The derivative of P at the point $(x_1, x_2, \mathbf{u})$ is a linear operator:

$$P'(x_1, x_2, \mathbf{u}) : (\overline{x}_1, \overline{x}_2, \mathbf{u}) \to \left(\overline{z}, \overline{\zeta}\right),$$

where

$$\begin{aligned}\overline{z}(t) &= \overline{x}_1(t) - \int_{t_0}^{t} f_{x_1}^{(1)}(x_1(\tau), \mathbf{u}(\tau))\, \overline{x}_1(\tau)\, d\tau \\ &\quad - \int_{t_0}^{t} f_{\mathbf{u}}^{(1)}(x_1(\tau), \mathbf{u}(\tau))\, \overline{\mathbf{u}}(\tau)\, d\tau\end{aligned}$$

and

$$\begin{aligned}\overline{\zeta}(t) &= \overline{x}_2(t) - \int_{t_0}^{t} f_{x_2}^{(2)}(x_2(\tau), \mathbf{u}(\tau))\, \overline{x}_2(\tau)\, d\tau \\ &\quad -K \int_{t_0}^{t} \frac{\overline{x}_2(\tau)}{(t-\tau)^{1-2\mathbb{H}}} d\tau \\ &\quad - \int_{t_0}^{t} f_{\mathbf{u}}^{(2)}(x_2(\tau), \mathbf{u}(\tau))\, \overline{\mathbf{u}}(\tau)\, d\tau.\end{aligned}$$

An arbitrary linear functional l , vanishing on the kernel of the operator $P'(x_1, x_2, \mathbf{u})$, has the form

$$\begin{aligned}l(\overline{x}_1, \overline{x}_2, \overline{\mathbf{u}}) &= \int_{t_0}^{t_1} \overline{x}_1(t)\, d\sigma_1(t) + \int_{t_0}^{t_1} \overline{x}_2(t)\, d\sigma_2(t) \\ &- \int_{t_0}^{t_1} \left[\int_{t_0}^{t} \left(f_{x_1}^{(1)}(\tau)\, \overline{x}_1(\tau) + f_{\mathbf{u}}^{(1)}(\tau)\, \overline{\mathbf{u}}(\tau)\right) d\tau\right] d\sigma_1(t) \\ &- \int_{t_0}^{t_1} \left[\int_{t_0}^{t} \left(f_{x_2}^{(2)}(\tau)\, \overline{x}_2(\tau) + f_{\mathbf{u}}^{(2)}(\tau)\, \overline{\mathbf{u}}(\tau)\right) d\tau\right] d\sigma_2(t) \\ &\quad -K \int_{t_0}^{t_1} \left[\int_{t_0}^{t} \frac{\overline{x}_2(\tau)}{(t-\tau)^{1-2\mathbb{H}}} d\tau\right] d\sigma_2(t),\end{aligned}$$

where $d\sigma_j(t)$, $j = \{1, 2\}$, is the Lebesgue-Stieltjes measure on $[t_0, t_1]$ with $\sigma_j(t)$ being the function of bounded variation on $[t_0, t_1]$. Changing the order of the integration

$$\begin{aligned}l(\overline{x}_1, \overline{x}_2, \overline{\mathbf{u}}) &= \int_{t_0}^{t_1} \overline{x}_1(t)\, d\sigma_1(t) + \int_{t_0}^{t_1} \overline{x}_2(t)\, d\sigma_2(t) \\ &- \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \left(f_{x_1}^{(1)}(\tau)\, \overline{x}_1(\tau) + f_{\mathbf{u}}^{(1)}(\tau)\, \overline{\mathbf{u}}(\tau)\right) d\sigma_1(t)\right] d\tau \\ &- \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \left(f_{x_2}^{(2)}(\tau)\, \overline{x}_2(\tau) + f_{\mathbf{u}}^{(2)}(\tau)\, \overline{\mathbf{u}}(\tau)\right) d\sigma_2(t)\right] d\tau \\ &\quad -K \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \frac{\overline{x}_2(\tau)}{(t-\tau)^{1-2\mathbb{H}}} d\sigma_2(t)\right] d\tau.\end{aligned}$$

we get

$$\begin{aligned}l(\overline{x}_1, \overline{x}_2, \overline{\mathbf{u}}) &= \int_{t_0}^{t_1} \overline{x}_1(t)\, d\sigma_1(t) + \int_{t_0}^{t_1} \overline{x}_2(t)\, d\sigma_2(t) \\ &- \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \left(f_{x_1}^{(1)}(t)\, \overline{x}_1(t) + f_{\mathbf{u}}^{(1)}(t)\, \overline{\mathbf{u}}(t)\right) d\sigma_1(\tau)\right] dt \\ &- \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \left(f_{x_2}^{(2)}(t)\, \overline{x}_2(t) + f_{\mathbf{u}}^{(2)}(t)\, \overline{\mathbf{u}}(t)\right) d\sigma_2(\tau)\right] dt \\ &\quad -K \int_{t_0}^{t_1} \left[\int_{\tau}^{t_1} \frac{\overline{x}_2(t)}{(\tau-t)^{1-2\mathbb{H}}} d\sigma_2(\tau)\right] dt.\end{aligned}$$

Now we write the Euler-Lagrange equation:

$$\begin{aligned}\alpha_0 \int_{t_0}^{t_1} &\left(\Phi_{x_1}(t)\, \overline{x}_1(t) + \Phi_{x_2}(t)\, \overline{x}_2(t) + \Phi_{\mathbf{u}}(t)\, \overline{\mathbf{u}}(t)\right) dt \\ &+ l(\overline{x}_1, \overline{x}_2, \overline{\mathbf{u}}) + \lambda\varphi'(\mathbf{u})\, \overline{\mathbf{u}} \\ &+ \int_{t_0}^{t_1} g'(x_1(t))\, \overline{x}_1(t)\, d\mu_1(t) \\ &+ \int_{t_0}^{t_1} g'(x_2(t))\, \overline{x}_2(t)\, d\mu_2(t) \\ &+ < h, (\xi'_{\mathbf{u}}(\mathbf{x}, \mathbf{u})\, \overline{\mathbf{u}}(\cdot) + \xi'_{\mathbf{x}}(\mathbf{x}, \mathbf{u})\, \overline{\mathbf{x}}) > \\ &= 0\end{aligned}$$

where $\lambda, h \in (\mathcal{L}^\infty)^*$, $\lambda \geq 0$, $h \geq 0$, $\lambda \langle \varphi(\mathbf{u}(\cdot))\, \overline{\mathbf{u}}(\cdot) \rangle = 0$, λ, h have the same dimension as φ and ξ; $d\mu_{ji} \in \mathcal{C}^*$, $d\mu_{ji} \geq 0$, $g_i(x_j(t))\, d\mu_{ji}(t) = 0$, $j = \{1, 2\}$, $1 \leq i \leq \ell_1^{(j)}$.

Let $\psi_j \in (\mathbb{R})^*$ be an adjoint variable of x_j ($j = \{1, 2\}$). Denote $\psi_j(t) = \int_{t_0}^{t} d\sigma_j(\tau)$. It is possible to admit that $\psi_j(t_1) = 0$. In addition, the positive independence of the gradients with respect to the mixed active constraints is fulfilled. Using the Euler-Lagrange equation we can get the adjoint equations and the local maximum principle. The theorem below, based on Dubovitski-Milyutin method ([10]), gives the possibility to find an optimal solution for the problem (14)-(17).

Theorem 3.1: Let $(\widehat{\mathbf{x}}(\cdot), \widehat{\mathbf{u}}(\cdot))$ be an optimal process ($\mathbf{x}(\cdot) \in \mathcal{AC}(\Delta, \mathbb{R}^2)$, $\mathbf{u}(\cdot) \in \mathcal{L}^\infty(\Delta, \mathbb{R}^r)$). Then there exist a number α_0 and functions $\mu_{ji}(t)$, $\{1, 2\}$, $1 \leq i \leq \ell_1^{(j)}$, $\xi_k(\mathbf{x}(t), \mathbf{u}(t))$, $1 \leq k \leq \ell_3$, a function of bounded variation $\psi(t)$ (which defines the measure $d\psi$), a function of bounded variation $\lambda_s(t)$, $1 \leq s \leq \ell_1$, (which defines the Radon measure $d\lambda_s$) such that the following conditions hold:

- nontriviality:

$$|\alpha_0| + \|\lambda\| + \|\mu\| + \|\xi\| > 0,$$

- nonnegativity:

$$\begin{aligned} \alpha_0 &\geq 0, \\ d\mu_{ji}(t) &\geq 0, \quad \forall i,j, \\ d\lambda_s(t) &\geq 0, \quad \forall s, \\ \xi_k\left(\widehat{\mathbf{x}}\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right) &\geq 0, \quad \forall k; \end{aligned}$$

- complementarity:

$$\begin{aligned} g_i(\widehat{\mathbf{x}}(t))d\mu_{ji}(t) &= 0, \quad \forall i,j, \\ \varphi_s\left(\widehat{\mathbf{u}}\left(t\right)\right)d\lambda_s(t) &= 0, \quad \forall s, \\ h_k\left(t\right)\xi_k\left(\widehat{\mathbf{x}}\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right) &= 0, \quad \forall k; \end{aligned}$$

- the adjoint equations:

$$\begin{aligned} -\psi_1'\left(t\right) &= \psi_1\left(t\right)f^{(1)}\left(\widehat{x}_1\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right) \\ &\quad +\alpha_0\Phi_{x_1}\left(\widehat{x}_1\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right), \end{aligned} \tag{21}$$

$$\begin{aligned} -\psi_2'\left(t\right) &= \psi_2\left(t\right)f^{(2)}\left(\widehat{x}_2\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right) \\ &\quad +K\left[\frac{\psi_2\left(t\right)}{\left(t_1-t\right)^{1-2\mathbb{H}}}-\int_t^{t_1}\frac{\psi_2'\left(\tau\right)}{\left(\tau-t\right)^{1-2\mathbb{H}}}d\tau\right] \\ &\quad +\alpha_0\Phi_{x_2}\left(\widehat{x}_2\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right), \end{aligned} \tag{22}$$

- the transversality conditions:

$$\begin{aligned} \psi_1\left(t_1\right) &= 0, \\ \psi_2\left(t_1\right) &= 0; \end{aligned}$$

- the local maximum condition:

$$\begin{aligned} &-\alpha_0\Phi_{\mathbf{u}}\left(\widehat{\mathbf{x}}\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right)-f_{\mathbf{u}}^{(1)}\psi_1\left(t\right)-f_{\mathbf{u}}^{(2)}\psi_2\left(t\right) \\ &\quad +\lambda^a\left(t\right)\varphi'\left(\widehat{\mathbf{u}}\left(t\right)\right)+h^b\left(t\right)\xi'\left(\widehat{\mathbf{x}}\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right) \\ &\quad =0, \end{aligned}$$

where $\lambda^a\left(t\right)\geq 0$ and $\lambda^a\left(t\right)\varphi\left(\widehat{\mathbf{u}}\left(t\right)\right)=0$, $h^b\left(t\right)\geq 0$ and $h^b\left(t\right)\xi\left(\widehat{\mathbf{x}}\left(t\right),\widehat{\mathbf{u}}\left(t\right)\right)=0$.

2) The case of $\mathbb{H}=\frac{1}{2}$*:* The necessary optimality conditions can be formulated in the classical manner, for more details see [11].

3) The case of $\frac{1}{2}<\mathbb{H}<1$*:* The reasoning is the same as in the case of $0<\mathbb{H}<\frac{1}{2}$ with the different object equation, namely

$$\begin{aligned} x_2\left(t\right)-x_{20}^{\gamma_2} &= \gamma_2\int_{t_0}^{t_1}f_2\left(x_2\left(\tau\right)\right)d\tau \\ &\quad +\frac{\gamma_2\left(\gamma_2-1\right)}{2}\theta_3^2\mathbb{H}^2\left[\int_{t_0}^{t_1}\frac{\sqrt{x_2\left(t\right)}}{\left(t-\tau\right)^{1-\mathbb{H}}}d\tau\right]^2 \end{aligned}$$

is presented as a system of two following equations

$$\begin{cases} x_2\left(t\right)=\Theta\left(y_2\left(t\right)\right)+\gamma_2\int_{t_0}^{t_1}f_2\left(x_2\left(\tau\right)\right)d\tau, \\ y_2\left(t\right)=y_2\left(t_0\right)+\int_{t_0}^{t_1}\frac{g(x_2(t))}{(t-\tau)^{1-\mathbb{H}}}d\tau, \end{cases}$$

where $t\in[t_0,t_1]$, $y_2\left(t_0\right)=\varrho$, $\Theta\left(y_2\left(t\right)\right)$ and $g\left(x_2\left(t\right)\right)$ are arbitrary smooth ($\mathcal{C}^1$) functions.

IV. Conclusions

Theoretical results of this paper can be used in decreasing of the level of uncertainty accompanying the decision-making processes related to issues of planning, designing, implementing, and managing the implementation of innovations and management of production systems not only of FIG but also of individual enterprise-producers looking for business partners and other opportunities to raise competitiveness and fighting for survival on the markets. Developing prototypes of information systems and indicating the optimal conditions for cooperation among members of FIGs can be motivating for the creation of such groups. The results of the paper allow deepening the knowledge of the research methods of complex economic objects on the mesolevel.

Acknowledgment

This paper was supported by grant 04.0.10.00/2.01.01.01.0026 MNSP.ZKIP.17.005. The authors would like to thank the anonymous referees for their valuable comments for improving the paper.

References

[1] M. A. Hitt, R. Duane, and R. E. Hoskisson, *Strategic Management: Concepts and Cases. Competitiveness and Globalization*. Cengage Learning, 2015.

[2] A. Singh, J. Glen, R. De-Hoyeos, B. Weisse, and A. Zammit, *Shareholder value maximization, stock market and new technology: should the US corporate model be the universal standard?* University of Cambridge, 2005.

[3] D. Mueller, "Corporate governance and economic performance," *International Journal of Applied Economics*, 2006.

[4] T. Santarius, "Investigating meso-economic rebound effects: production-side effects and feedback loops between the micro and macro level," *Journal of Cleaner Production*, vol. 134, pp. 406–413, 2016.

[5] T. Nguyen, "Fractional stochastic differential equations with applications to finance," *J. Math. Anal. Appl.*, vol. 397, pp. 334–348, 2013.

[6] J. M. Henderson and R. E. Quandt, *Microeconomic theory*. McGraw-Hill, 1971.

[7] L. Longjin, F. Y. Ren, and W. Y. Qiu, "The application of fractional derivatives in stochastic models driven by fractional brownian motion," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 21, pp. 4809–4818, 2010.

[8] N. T. Dung, "Fractional stochastic differential equations with applications to finance," *J. Math. Anal. Appl.*, vol. 397, pp. 334–348, 2013.

[9] D. Filatova, M. Grzywaczewski, and N. Osmolovskii, "Optimal control problem with an integral equation as the control object," *Nonlinear Analysis: Theory, Methods*, vol. 72, no. 3–4, pp. 1235–1246, 2010.

[10] A. Milyutin, A. Dmitruk, and N. Osmolovskii, *Maximum Principle in Optimal Control*. Moscow State University, 2004.

[11] A. Dmitruk, "Maximum principle for the general optimal control problem with phase and regular mixed constraints," *Optimality of Control Dynamic Systems*, vol. 14, pp. 26–42, 1990.

A Dynamically Reconfigurable VLSI Processor with Hierarchical Structure based on a Micropacket Transfer Scheme

Yoshichika Fujioka
Department of Engineering
Hachinohe Institute of Technology
Hachinohe, Japan
Email: fujioka@hi-tech.ac.jp

Michitaka Kameyama
Department of Information Technology and Electronics
Ishinomaki Senshu University
Ishinomaki, Japan
Email: michikameyama@isenshu-u.ac.jp

Martin Lukac
School of Science and Technology
Nazarbayev University
Astana, 01000, Kazakhstan
Email: martin.lukac@nu.edu.kz

***Abstract*—In this paper we propose improvements to the Micro-packet Transfer scheme in a multi-core device (also called reconfigurable VLSI or Network-on-chip). In particular we propose a new hierarchy based micro-packet control scheme that is especially effective for tasks require large number execution clock steps.**

I. Introduction

With the large amount of existing applications, the constant growth of user requirements, new technology and an constant need for knowledge transfer, hardware must adapt in order to permit such applications be run efficiently and problems being solved in reasonable time.

Additionally the advent of artificial intelligence for general problem solving and the extreme usage of machine learning for intelligent system leads to the requirement for intensive context switching in order to avoid a) the bias-variance effect and b) to allow problem specific instances of algorithms be used.

Consequently, a VLSI Computing platform oriented for real-world intelligent systems must be developed so that many kinds of adaptive algorithms can be autonomously replaced in the execution program memory. Only such hardware would lead to a real-time processing and the possibility to reduce the power consumption to a minimum.

To meet these requirements, standard general purpose processors suffer a cost overhead by implementing complex hardware and high power consumption. In order to avoid such downfalls a high-performance and low-power dynamic reconfigurable VLSI architecture for real-world intelligent systems is presented. This architecture uses a micro-packet routing scheme based on scheduling/allocation completed prior to execution [1], [2], [3], [4], [5].

The micro-packet routing scheme is effectively employed for making the reconfiguration cost effective by reducing the complexity of the data packet router to a minimum, because arbiters or buffer memories in a router are not required to avoid packet collisions. Therefore, great reduction of the configuration/control memory size is achieved at each hierarchy of the dynamic reconfigurable processor. For the application that requires many number of execution clock steps, the proposed micro-packet control scheme is very suitable for providing many cells on chip size constraint.

This paper is organized as follows. Section II introduces tha background to a reconfigurable VLSi processor, Section III describes an efficient design of a processing element. Section IV explains the micro-packet data transfer and Section V explains the design of a computational cell. Section VI describes the evaluation and experiments and Section VII concludes the paper.

II. Hierarchical structure of a dynamically reconfigurable VLSI processor

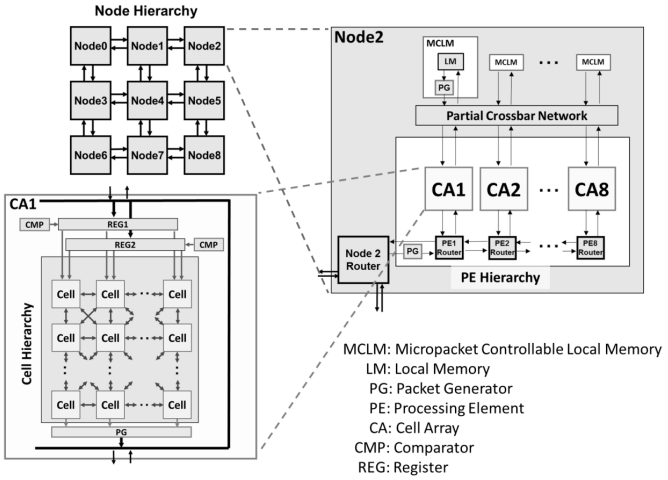


Fig. 1. Hierarchical structure of a dynamically reconfigurable VLSI processor

Figure 1 shows a dynamically reconfigurable VLSI processor architecture oriented for the algorithm selection scheme. This architecture is built in a VLSI platform for real-world intelligent applications requiring high speed real time recon-

figuration. For instance algorithm selection as proposed in [1], [2] is an example of ell suited application.

The architecture proposed has several levels of hierarchy. On the highest level, the architecture is a set of interconnected nodes and each node contains several cell arrays. Each cell array is a set of interconnected cells. Note that for each node a router is attached allowing to implemented a distributed packet communication scheme.

The architecture uses global memory for storing less used algorithms and data and local memories for very often used and required algorithms and data. Additionally the architecture has the following features:

1) From coarse-grain to fine-grain reconfigurable computing with hierarchical structure [6]
2) Direct allocation of Control Data Flow Graph (CDFG) at each hierarchy,
3) Configuration/control memory size reduction based on the micro-packet data transfer scheme,
4) Dynamic reconfiguration of local memories to solve memory-PE data transfer bottleneck,
5) Node-hierarchy power-gating with a processing completion signal.

Several types of nodes are defined such as a node composed of a Processing Element (PE) array and local memories, and nodes composed of a global memory.

As shown in Figure 2, almost all data transfer can be done between adjacent nodes by direct allocation of a Control Data Flow Graph(CDFG) at the node level. Similarly, the direct allocation of the CDFG at lower level is effectively employed for reducing complexity of data transfer between execution cores such as PEs. Figure 2 shows how a st of Processing stages from a CDFG is allocated onto the 2D node architecture.

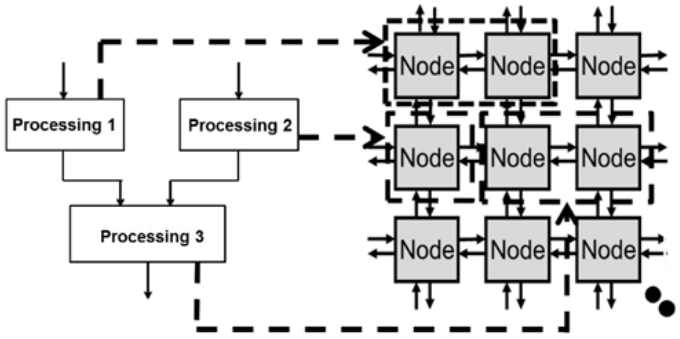


Fig. 2. Direct allocation at the node hierarchy

III. Design of a Processing Element

In general, for coarse-grain reconfigurable computing, the utilized ratio of functional units is not very high, because all kinds of functional units must be provided if they are used frequently or not. We propose the coarse-grain and fine-grain reconfigurable computing architecture. As shown in Figure 3, each PE contains eight linear cell arrays and the desired structure of functional units can be reconfigured using

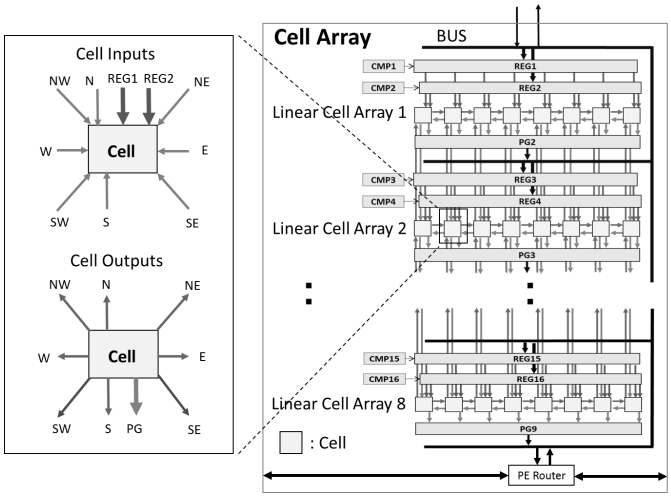


Fig. 3. PE structure with 8 linear cell arrays

these cells. Each cell has input/output connections to its eight neighborhood cells suitable for the direct allocation of CDFG. The direct allocation of CDFG in cell hierarchy makes the data transfer among the cells very simple. In addition, the linear cell array structure provides two inputs from registers and an output to PE are provided in each cell. Simple functional units can be dynamically reconfigured using the linear cell array. So maximally eight simple functional units can be reconfigured in each PE. More complex functional units can be reconfigured using multiple linear cell arrays together.

This fine-grain dynamic reconfiguration on the linear cell array level allows to increase efficiently the utilization ratio of cells and PEs.

IV. Micro-packet data transfer scheme

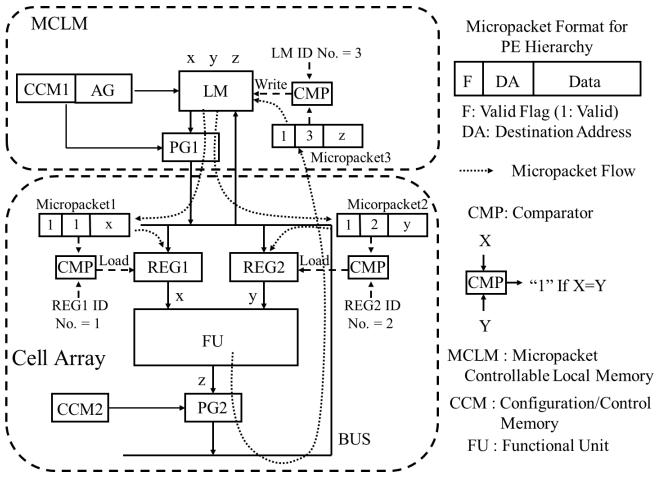


Fig. 4. PE structure based on a micro-packet data transfer scheme

The communication between the PE blocks is useing the micr-packet scheme [1], [2]. Figure 4 shows the PE structure based on the micro-packet data transfer scheme, which consists of the MCLM (Micro-Packet Controllable Local Memory) and

the cell. At the cycle when the active packet transfer to REG1, REG2 and LM is not necessary for the processor operation, the corresponding packet control information is not required to be stored in Control Configuration Memory (CCM). Because fundamental operations of a processor are attributed to data transfer between registers, automatic timing generation of data-receive can be done if a packet is appropriately sent to a routing network. This means that timing control for data-receive can be autonomously completed, which is different from conventional configuration specification. Therefore, the configuration memory size in the register-transfer-level packet data transfer scheme is greatly reduced in comparison with a conventional method.

V. Design of a Cell

As shown in Figure 5, the logic block, input multiplexers, an output crossbar switch and CCM are provided in each PE. Two Look-Up Table (LUT) and two D-type Flip Flop (D-FF) are provided in the logic block so that several kinds of 1 bit functional unit can be reconfigured such as a full adder. The desired structure of functional units and its data path are dynamically reconfigured by changing the LUT and the output crossbar switch control using CCM. The micro-

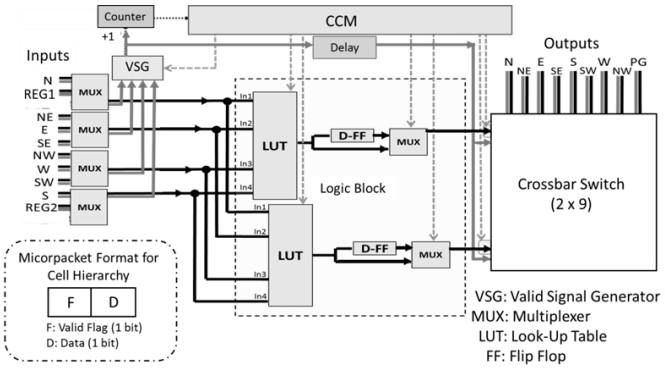


Fig. 5. Cell structure for the micro-packet transfer scheme

packet format in the cell hierarchy is shown in Figure 5. Each packet contains 1 bit valid flag and 1 bit data. The input multiplexers are controlled by the valid flag of the micro-packet instead of using CCM. So the control field for the input multiplexers becomes not necessary in CCM. In addition, the control timing of each cell is automatically generated by valid signal generator (VGG) shown in Figure 6.

The control timing is generated when all micro-packet necessary for the reconfigured functional unit reached the cell. The valid flags of the required micro-packets are used to generate the control timing. This signal is also used for micro-packet from the cell. To select the required valid flags, only 4 bit LUT input select field is provided in CCM. Therefore, the dynamically reconfigurable cell architecture based on micro-packet control scheme makes the reduction of the control fields for the input multiplexer control and the control timing possible.

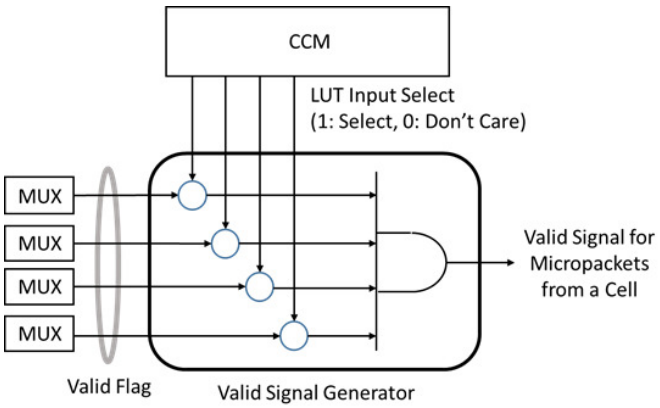


Fig. 6. Valid signal generator

VI. Evaluation of the Dynamically Reconfigurable VLSI Processor

A. CCM size

Let's consider the CCM size reduction in PE hierarchy and cell hierarchy based on the proposed micro-packet control scheme. For an example of the CCM size evaluation in PE hierarchy, we use the hardware model for execution of 4 input multiply-addition using 2 PE and 6 MCLM with a partial crossbar network as shown in Figure 8.

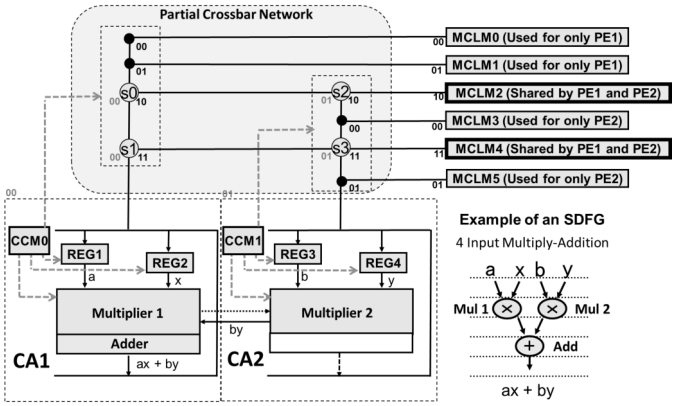


Fig. 7. PE structure for the execution of an SDFG

Table I shows the CCM size evaluation result. The CCM size of the conventional microprogram control scheme such that the control information is given at every clock cycles (Microprogram control scheme 1) is 1404 bits. By using the microprogram control scheme 1, many non active control information or don't care information should be contained in CCM even if the cell is not utilized, so that the CCM size becomes too large.

The CCM size of the other type of conventional microprogram control scheme such that the control signal and timing is given only when control information changes (Microprogram control scheme 2) is 154 bits. Since the non active control information or the don't care information can be reduced

TABLE I
CCM SIZE FOR THE EXECUTION OF AN SDFG ON PE HIERARCHY

	Microprogram Control Scheme 1	Microprogram Control Scheme 2	Micropacket Transfer Scheme
CCM Size (bit)	* 1404	** 154	103 (* 92% Reduced) (** 33% Reduced)

by using the microprogram control scheme 2, the CCM size becomes greatly reduced.

However, the CCM size of proposed micro-packet control scheme becomes 103 bits. In comparison with the microprogram control method 2, the CCM size can be more 33%

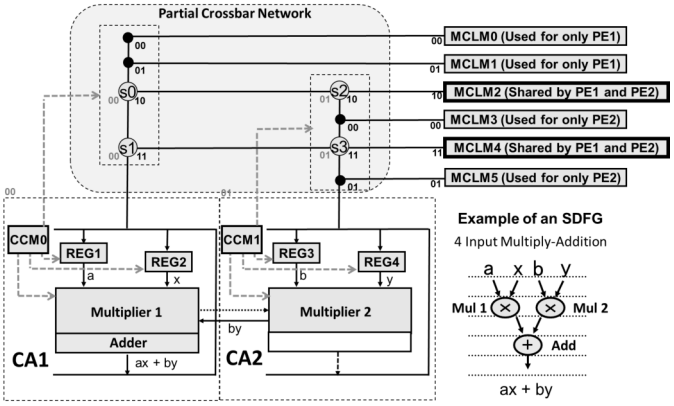


Fig. 8. PE structure for the execution of an SDFG

reduced. Since the automatic timing generation by destination address comparison or valid flag, in addition to the non active control information and the don't care information, the control timing information can be also reduced by using the proposed micro-packet control scheme.

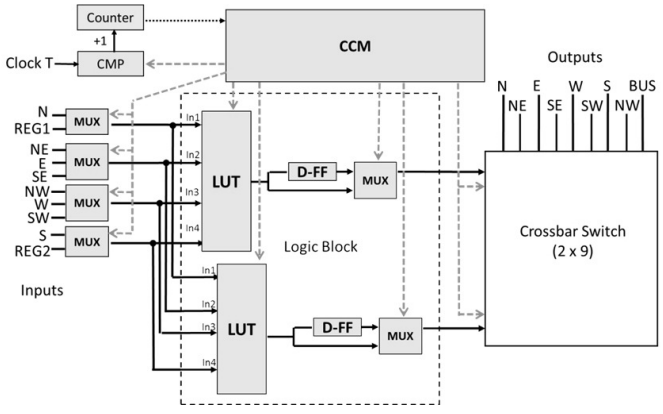


Fig. 9. Cell structure for microprogram control scheme 2

Let's consider the CCM size evaluation in the cell hierarchy. As shown in Figure 9, the input multiplexer control and control clock timing are required in the cell using the microprogram control scheme 2. On the other hand, instead of these control information, the input selection mask control becomes required in the micro-packet control scheme as shown in Figures 5 and Figure 6.

TABLE II
FIELD SIZE OF CCM IN A CELL

CCM Control Field	Clock Timing	Input MUX	Input Selection Mask	LUT	DFF Enable (MUX)	Output Crossbar Switch	Total (bit)
Microprogram Control Scheme 2	16	6	-	32	2	8	64
Micropacket Transfer Scheme	-	-	4	32	2	8	46 (28% Reduced)

Table II shows the field size of CCM in a cell. By using the micro-packet control scheme, the CCM size can be 28% reduced in comparison with the microprogram control scheme 2.

B. Advantage of the dynamically reconfigurable VLSI processor

From the above discussion, the proposed micro-packet control scheme is very useful for the CCM size reduction. However, additional hardware such as address comparators or packet generators are required. So let's consider the additional number of transistors for the micro-packet control scheme.

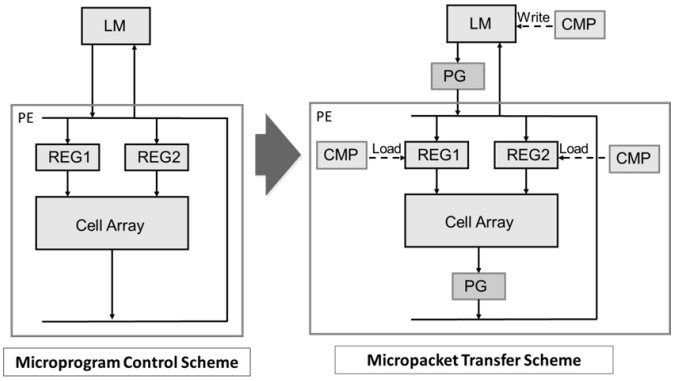


Fig. 10. Simplified PE structure model

Figure 10 shows the simplified PE structure model for this evaluation with no CCM. Table 3 shows the number of transistors for this model. In comparison with the conventional microprogram control scheme, the number of transistors becomes 2.4% increased.

TABLE III
NUMBER OF TRANSISTORS FOR THE SIMPLIFIED PE STRUCTURE MODELS WITHOUT CCM

	Microprogram Control Scheme 1 and 2	Micropacket Transfer Scheme
Number of Transistors	37776	38664 (2.4% Increased)

However, as shown in Figure 11, the total number of transistors of PE with CCM becomes reduced when the number of

execution clock steps is more than 27. Therefore, for the application that requires many number of execution clock steps, the CCM size reduction becomes very important for providing many cells on chip size constraint. The proposed micro-packet control scheme is very suitable for this requirement.

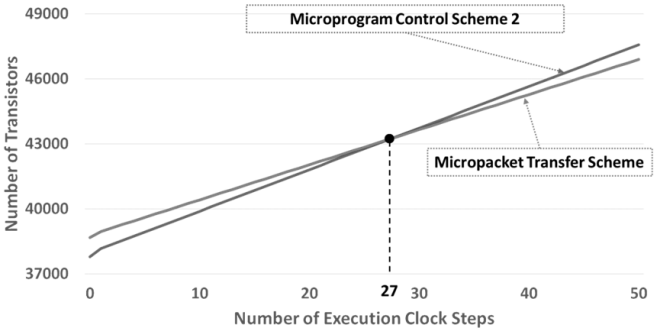


Fig. 11. Number of transistors of PE

VII. Conclusion

The proposed dynamically reconfigurable VLSI processor is suitable for a computing platform of real-world intelligent systems. The configuration/control memory size can be greatly reduced by the micro-packet routing, because efficient micro-packet formats are defined at each hierarchy. Moreover, concept of direct allocation at each hierarchy makes the interconnection complexity between execution cores very simple. The automatic mapping of the CDFG corresponding to each processing algorithm is left as a future important problem.

References

[1] Y. Honma, M. Kameyama, Y. Fujioka, and N. Tomabechi, "Vlsi architecture based on packet data transfer scheme and its application." in *In Proc. of 2005 IEEE Int. Symp. on Circuits and Systems*, 2005., pp. 1786–1789.

[2] Y. Fujioka, N. Tomabechi, and M. Kameyama, "Register-transfer-level packet data transfer scheme for a highly parallel vlsi processor," in *In Proc. of Int. Conf. on Computers and Devices for Communication*, 2006, pp. 9–13.

[3] Y. Fujioka and M. Kameyama, "Configuration memory size reduction of a dynamically reconfigurable processor based on a register-transfer-level packet data transfer scheme," in *SoC Design Conference (ISOCC), 2012 International*, nov. 2012, pp. 235 –238.

[4] M. Lukac, M. Kameyama, and Y. Fujioka, "Vlsi platform for real-world intelligent integrated systems based on algorithm selection," in *IADIS TPMC*, 2013.

[5] M. Lukac, K. Abdiyeva, Y. Fujioka, and K. Kameyama, "Algorithm selection platform in real-world intelligent systems," in *Proceedings of the 28th International Conference on Computer Applications in Industry and Engineering, CAINE 2015*, 2015.

[6] B. Mei, A. Lambrechts, J. Mignolet, D. Verkest, and R. Lauwereins, "Architecture exploration for a reconfigurable architecture template," *IEEE Design Test of Computers*, vol. 22, no. 2, pp. 90–101, March 2005.

A Multi-layer Multi-sensor Wearable Device for Physical and Chemical Environmental Parameters Monitoring (CO & NO_2)

Mostafa Haghi, Kerstin Thurow
Center for Life Science Automation
Rostock University
Rostock, Germany
{Mostafa.Haghi, Kerstin.Thurow}@celisca.de

Norbert Stoll
Institute of Automation
Rostock University
Rostock, Germany
Norbert.Stoll@uni-rostock.de

***Abstract*— The monitoring of the ambient working conditions is a crucial part in occupational medicine. A wearable system for monitoring environmental conditions including physical and chemical parameters is introduced. The system operates based on an integrated and add on-board multi-layer sensors and fusion approach. It uniquely consists of only a small and light wrist-worn device which embeds a microcontroller board with integrated temperature, humidity and barometer sensors (Physical parameters), add on-board gas sensors (Chemical parameters) in second top layer for monitoring hazardous gases in air and a vibration motor placed under of the board. For an early notification of the user about ambient parameters a haptic feedback pattern is actuated using a micro embedded vibration motor. Data are also sent to a smart phone simultaneously. The collected data are used to monitor the ambient environment and establish a long-term individual profile for employees in real-time. With this paper we introduce the system architecture. Efficiency, performance and power consumption of such unique multi-layer multi-parameter monitoring device in size and weight are evaluated.**

Keywords*— *Wearables; Sensor node; Gas sensor; Environment monitoring; Indoor Air Quality (IAQ)

I. INTRODUCTION

The monitoring of ambient working conditions is a crucial part in occupational medicine. Different regulations define limit values for different physical and chemical parameter.

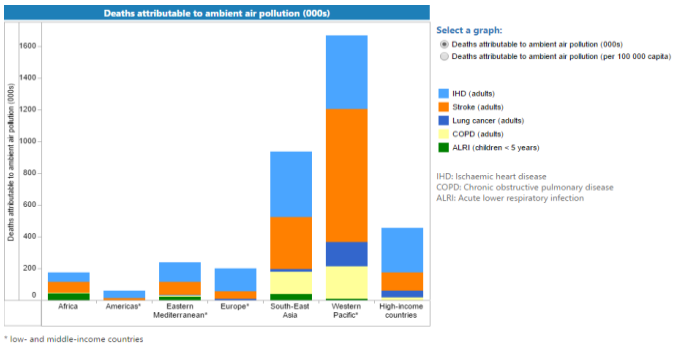


FIGURE 1. MORALITY FROM AMBIENT AIR POLUTION

Air quality and human health studies have proved that urban air pollution can affect human health (e.g. World Health Organization, 2000).

To ensure safe working conditions, the monitoring of irritating or toxic gases is required. Nitrogen dioxide (NO_2) and Nitrogen Oxide (NO) are identified as the most dangerous pollutants, those can affect quality of life and mortality rates (Fig.1) (e.g. World Health Organization, 2006). Both carbon monoxide (CO) and NO_2 are known to be respiratory sensitizers (e.g. McConnell et al., 2010) with greater effects on people with existing respiratory or cardiovascular conditions (e.g. HEI, 2010) [1].

In addition to hazardous gases, there are other parameters that are always useful to be aware of. Temperature, humidity and air pressure (Physical parameters) sensors are integrated in microcontrollers to fulfill a full range of environment observation. Wearable devices are becoming very popular and at the heart of just about every discussion related to the Internet of Things
(IoT), and the full range and advantage of new capabilities pervasive connectivity can bring [2], [3].

Body-worn wireless devices and the integrated sensors for long-term monitoring have to be reliable, small and light weight [4]. Thus the size of product, power consumption and safe data transition are of serious concerns, therefore deployment of a series of above mentioned body-worn sensors can lead to an effective health-monitoring mechanism [5]**.** There are sensor nodes in broad band of interested subject monitoring in harmful gases detection, movement, weight and vital signs which fusion of these parameters is resulting into an actual range of health monitoring. This study already is under progress in the Center for Life Science Automation (celisca, University of Rostock, Germany) [6]**.** These sensor nodes sometimes are fused with typical singular heartbeat or respiration sensors (often in smart phone), which have been commercially available in recent years**.** The combined information obtained from such systems can either be relayed directly to any health-monitoring personnel in the case of emergencies or can be logged and analyzed as a part of preventive health measures. However, the deployment of on-

body nodes must be designed carefully as many of wearer might be patients, and this device may interfere with their regular movements. On the hardware design aspect, **s**ome important challenges are raised up while the deployment of such mechanisms, where the nodes require additional on-body space and probably impose additional weight. Both of these undesirable factors are able potentially to make the device uncomfortable to the patient's body. The design and development of wearable health-monitoring systems (WHMS) has received lots of attention from the scientific community and industry during the last years [7]. Applications range from the environmental safety sensing, transportation systems and medical health to support diagnostics and even food safety and agriculture. Conventional gas detectors including the most widely used, relatively small Pallisters require large power consumption (hundreds of milli-Watts to Watts) with slow response time (tens of seconds). The technical specifications for gas sensors, indeed for all sensors, are challenging, and typically [8]: small and light, low power (e.g. less than 200 mW), long life (e.g. 5 years), selective (e.g. low cross sensitivities), range of operation -50°C to +125"C, range of operation between ***5*** and 95% relative humidity.

This paper is structured as follows: In section II some related research works are introduced. In part III, we concentrate on the proposed system architecture and the main components are described. In section IV, experimental results of a possible application scenario are presented to validate the efficiency of the proposed new multi-layers architecture design. Conclusion is summarized in section V.

II. RELATED RESEARCH WORKS

In the last decade, there have been numerous research efforts and products that can be classified as WHMS [8]. In fact, the presented work is not similar to any other work before. We suggest an innovative multi-layer device with add on-board sensors based on a round integrated platform, for monitoring several parameters. The device should be very small, light, wearable and power efficient. Related research which has been implemented in this area of interests includes e.g. the Advanced Care and Alert Portable tele medical Monitor (AMON) [9]. This system consists of a wrist-worn device, which is able to measure several vital signs such as blood pressure, skin temperature, blood oxygen saturation, and a one lead ECG. In addition, to observe user position during his daily activities a two-axis accelerometer is incorporated. All data collected by the wrist-worn are wirelessly sent to a station which enables physicians to analyze the received data remotely. Another relatively new approach consists of developing systems based on smart textiles. In the Wearable Health Care System (WEALTHY) project [10], several sensor elements targeted for various measurement, were integrated in fabric form (using conductive and piezo resistive materials) on a textile structure. The system has the capability of monitoring a three-lead ECG, electromyogram (EMG) placed on the arms, thoracic and abdominal respiration rate, body position and movement, skin temperature, and core temperature. The wearable system incorporates an analogue and digital signal-processing module with GPRS or Bluetooth wireless transmission capabilities. *Passow* and et al. have been working on wearable devices using an EFM32 microcontroller [11]. They designed and implemented some algorithms to compress the code and also secure data transmission in expose of electromagnetic interference to reduce power consumption and more efficient transmission either. *Sanfilippo* et al. investigated a system using new introduced platform by cooking hacks group [12]. In this preliminary platform all vital signs are measured by different wearable tools on the body and are sent to Arduino wire-based through e-health shield stock on the top of a microcontroller and from there to a smart phone by Bluetooth. However, health monitoring during daily ambulatory life is also possible. Despite of all these works, we deeply believe that, a health sensor monitoring system that features a multi-sensor fusion approach and meets a wearable device criteria (weight, wearability, power consumption and size) has not been investigated yet.

III. SYSTEM ARCHITECTURE

The framework is based on two main facts: first, multi-sensor fusion approach and second - which is the most important and innovative part, multi-layer add on-board sensors. In particular, a client-server pattern is adopted, in this architecture a smart phone is acting as a hub to communicate between client and server. The wrist-worn device is considered as a client and remotely communicates with a smart phone through Bluetooth. This is a two-way communication but the return way data communication from the smart phone to the wearable device is operated only when the data are sent from the smart phone to a server, stored there and thus are analyzed by a medical doctor. When appropriate decision is made according to the obtained data from the device, necessary recommendations will be sent to the applicant by his/her wearable device or some other defined algorithms before. Server is where the logic of this architecture is implemented. Now, the key elements of the system, their roles and device features are described. For more information, the reader is recommended to refer to Fig.2.

A. General concept

Many efforts have been performed to implement efficient wearable devices for ambient monitoring in medical care-

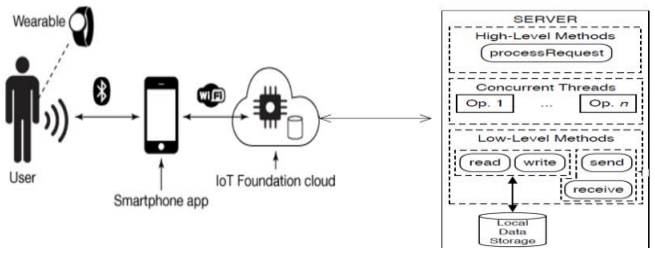


FIGURE 2. STRUCTURE OF THE SYSTEM

systems but for several reasons no such device has been introduced in academic level nor industry yet. In this part we concentrate on the general concepts of this ideal wearable and requirements in details.

We are going to measure both physical variants such as air pressure, temperature and humidity and chemicals, in particular hazardous gases. In our design the most critical parameters which are taken care of, are wearability and efficiency. To the best of our knowledge and our discussions with people whose are using such devices, long time monitoring, comfortness, weight and size are of serious concerns. These are important for general concepts, but what accomplish these factors, is a well-defined platform. This makes the platform flexible and compatible to adopt several number of nodes. A platform with light weight, high processing, high performance, small size and other features is required. The used platform in our work has been chosen due to performance, low power consumption, light, size and rounded shape which make it comfortable to be wrist worn. The platform as the backbone of our design is always kept fixed; external components, actuators and other requirements are all moved to the sensor node.

B. Sensor node concept

The sensor node is the most important part. In contrast to existing solutions, the unique design of our sensor node is compact, compatible with SMD and 2 and/or 3-lead gas sensors and low power consumption. Instead of designing and implementing a universal sensor node which makes it very big and inefficient, result in unnecessary high power consumption and is far away from wearability, we introduce the individual sensor node. These sensor nodes are replaced by the user on the top of the first layer (platform) with a simple board to board pin connection. Many of the sensors in the universal sensor nodes (previous efforts) due to the constant design, are continuously running to measure a parameter which does not really exist in that particular place and time, but changing the node configuration is not possible by individual, therefore this idle sensor operation consumes the battery, thus device length monitoring time is consequently reduced. Our sensor node provides this selectivity to the wearer to decide which sensor is required for a specific situation and environment. This could be a good beginning for the introduction of a new generation of sensor nodes. Many desired variables are measured with the least work for the wearer. Electrochemical SMD sensors with the minimum number of external components are the best candidates. To bring all theoretical concepts on operation, a BUS24 connector in very small size (7mm length) is adopted which accomplishes the multi-layer design. When the target gas is changed, user only has to replace another sensor target, all other requirements are on software part.

The electronics circuit of a sensor node for IAQ measurement is depicted in Fig.3. The circuit is divided into 3 parts, wherein the microcontroller is powered via a 3V Li-Ion battery, and the sensor is operating at 5V, a DC/DC converter with least leak out current has been carefully designed, the 5V output is fed to

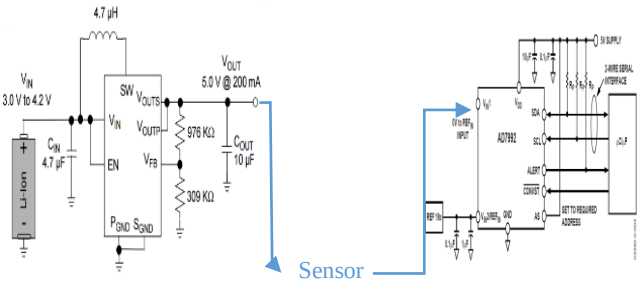


FIGURE 3. ELECTRONIC CIRCUIT FOR SENSOR NODE

the sensor. The output of the sensor is analogue and can be directly linked to the microcontroller, but to reserve the limited number of ADC pins on the microcontroller, we would prefer to reserve them for future developments. On the other hand to facilitate the number of sensors in the multi-layer design, I2C communication protocol is used.

The output signal of the sensor is converted through the second designed part circuit to digital which gives the I2C communication possibility.

C. Client (Wrist-Worn Device)

The proposed wearable device is working as a client. It gathers different bio-metric data from various integrated and add on-board sensors for ambient parameter and movement measurement. In particular, it is a light, small and easy to wear wrist-worn device (Fig.4-LEFT) that consists of a controller box and add on-board sensors which are powered by a LiPo / Li-Ion battery. The components housed in the controller box (watch-worn) are categorized in two groups, integrated sensors and add-on board sensors. A microcontroller is integrated as well. The microcontroller used in this work is suitable for wearable applications. The board is based on the Nordic Semiconductor's nRF51822 SoC, supporting Bluetooth 4.1 (Low Energy), and providing radio connectivity with integrated crystal antenna. A smart power control management block gives this capability of being used for long term monitoring with small battery, from days to several months. This small and round device is only 32 mm in diameter.

In such device, pin usability is a critical point due to number of sensors and other components that are used. There is a small integrated connector pin socket on board which provides the full range of pin usability. This gives a good opportunity to add PCBs on board. This makes the structure framework easy to maintain and enable the easy addition of new features in the future.

FIGURE4.-LEFT, PLATFORM, INTEGRATED AND ADD ON-BOARD SENSOR- RIGHT, WRIST-WORN WEARABLE DEVICE

Bluetooth 4.1 will be used for the communication of the wearable device. It is integrated on the board, thus no further external components are required [13]. Depending on the number of parameters to be measured, different sensors can be added. By using SMD gas sensors and components multi-layer nodes are able to stack on top of each other according to specific configuration for each and still keep the solution small and wearable.

In addition, a micro vibration motor is embedded at the bottom of proposed wrist-worn device. This micro vibrator motor is actuated according to the received data feedback from the micro controller or smart phone (depending in which steps it's actuating) about the ambient parameter. This feature is fundamental for improving the user's risk perception. This vibrator can highly improve the device efficiency in risk situations and make the applicant aware. Three different motor states are implemented. It does not vibrate when all measured parameters are below the threshold. On the other hand, if data values exceeds the threshold, the motor is vibrating in low frequency. The third state is defined for high risk, in this case the motor is frequently vibrating to gain wearers attention

D. Multi-Threading and Multi-Level Hierarchical Server

When data are collected by the wrist-worn watch from different sensors and sent to the smart phone, these data must be sent to a server to be stored permanently and also to a cloud computing which is communicating with the server. Recommendations and probably activities are performed and decisions are taken according to previously stored data and present the status of the applicant by medical doctors. Therefore, the server has to implement a multi-threading and multi-level control program, depending on the status and operations, different decisions are made in various levels. This allows to simultaneously and efficiently collect data from different users. To be able to perform a sequence of operations with lowest errors, strict real-time criteria are adopted. Moreover, the multi-threading and multi-level hierarchy, improve the performance and scalability.

IV. EXPERIMENTAL RESULTS

The described setup has been used to determine the concentration of NO2. In addition, investigations on the CO gas sensor have been done too. The blue line in Fig.5 represents the electrical output voltage of the sensor (CO) and the yellow line detects the level of the NO2 gas concentration.

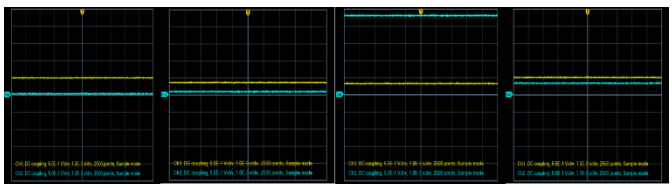


FIGURE 5. GAS SENSOR BEHAVIOUR TO CO AND NO2 CONCENTRATION

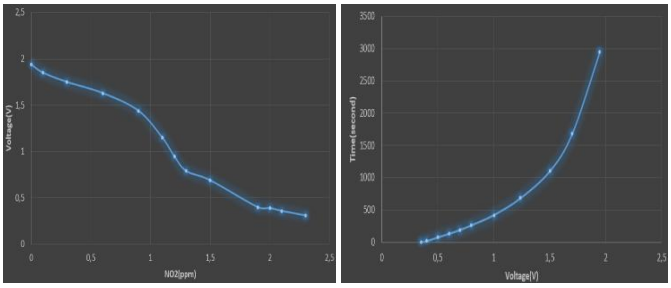


FIGURE 6. LEFT, OUTPUTVOLTAGE FOR CONCENRTATION, RIGHT-SENSOR RECOVERY TIME

to the nature of the gases and sensors' internal diagram, the reference level of both gases (even when no target gas is available) is 1.945V (NO2) and 2.6V (CO) (room temperature and 1atm. pressure), respectively. The CO gas is known as a reducing gas and NO2 is an oxidizing compound. Thus, when the sensor is exposed to the various levels of gas concentration the output voltage of the sensor is ascended for CO and is reduced for NO2 due to changes in load resistance, configured for each one.

Basically long exposure to NO2 can lead to serious health problem when it exceeds 2ppm. The NO2 concentration which normally is observed in urban area is not more than 0.7-0.8 ppm. In Fig.6-LEFT the results of sensor for the NO2 concentration between 0.03 to 2.3 ppm are given. The obtained results can be considered linear in the area of 0-1 ppm with a good estimation. The response time is very acceptable (< 30 seconds) and the stability time for different concentration varies but mostly is less than 60 seconds. The stability time is improved when the gas concentration is increased over 1.2 ppm.

A new concept is defined here as sensor recovery time. Sensor recovery time is a time that is required to make the sensor ready (come back to the reference level of voltage) to repeat the gas measurement for the next time. Although in this experiment relatively a long time (almost 50min.) is required for sensor recovery, but during the first 5 minutes (425 seconds) the sensor is recovered by 51%, in this segment the recovery time is quite linear (see Figmicro.6-RIGHT). In 18 minutes the sensor is recovered by 78%. Hereafter the recovery time rate is reduced, the remaining 22% of recovery is completed within 32 minutes which is long time, but sensor is prepared for reusability when it's recovered by 78% (calibration is required).

Data Transition Protocol and Power Mode When data from sensors are acquired by the microcontroller, a channel must be established between the microcontroller and the smartphone. To secure the transmission, data primary authentication is performed. A blue light indicates another user who is trying to connect to the device via Bluetooth but does not have the permission.

In this case, access is denied and a red LED is blinking. Data transfer is only possible for authorized users, after attempting

TABLE I. POWER CONSUMPTION IN DIFFERENT MODES

Sensors and connection mode	*Power consumption(mW)*
Motion + physical sensors (BLE not-connected)	111.29
Motion + Physical environmental sensors (BLE connected)	236.51/240.04
Chemical sensors only	43/76
All sensors (BLE not-connected)	279.51/312.51

to receive data, a green LED starts blinking which indicates the opening of a secure channel. The data from device are sampled in three modes. In the first layer of this device, integrated motion sensors are available which we don't pay attention in this paper but these sensors' data are sent every 20 ms for more accuracy tracking to the smart phone, the physical parameters (pressure, temperature and humidity) are sent every minute (the reason for providing two power consumption in Table 1, second column is because every one minute there is a coincidence). The chemical parameters are sampled every 20 minutes (due to stability, response time and recovery time), the NO2 and CO are sampled alternatively, that's the reason again there are two power consumption in column 3 in Table 1.

We allowed the battery to discharge to a medium level of 56%, and observed for continues working of the device for all active components, battery is discharged by less than 1% for less than 1 hour. This leads to a runtime of more than one week using appropriate configuration and power management block. In this work a polymer lithium ion battery - 250 mAh was used. It is placed under the device and is covered by the device properly (see FIG.4-RIGHT).

V. CONCLUSIONS

In this work, a preliminary innovative working prototype of a wrist-worn watch health sensor monitoring system for environment air quality (physical and chemical parameters) was introduced. A add- on board layer and sensor fusion approach was adopted. The system consists of several integrated and a stock layer on the top components embedded in a wrist-worn device. This small, light and ultra-low power consumption is distinguishing in monitoring of comprehensive parameters in physical environment as well as to observe harmful gases detection. This enables the biometric and medical monitoring in this device. It is possible to make the user aware of his health and environment through a micro vibrating motor located at the bottom of the watch according to the degree of risk. The presented structure enables a real-time monitoring of the studied parameters, which are sent via Bluetooth to a smart phone and to a health facility station for further subsequently analysis of the data and medical diagnosis. On the other hand, information can be sent wirelessly to a cloud computing system for permanent cloud-based data storage.

REFERENCES

[1] "Ambient (outdoor) air quality and health," World Health Organization. [Online].Available:http://www.who.int/mediacentre/factsheets/fs313/en.

[2] P. Soh, G. Vandenbosch, M. Mercuri and D. Schreurs, "Wearable Wireless Health Monitoring: Current Developments, Challenges, and Future Trends", IEEE Microwave Magazine, vol. 16, no. 4, pp. 55-70, 2015.

[3] D. Schreurs, M. Mercuri, P. J. Soh, and G. A. E. Vandenbosch, "Recent advances and design challenges in wireless health monitoring," Microwave Rev., vol. 19, no. 2, pp. 34–43, Dec. 2013.

[4] K. A. Townsend, J. W. Haslett, T. K. K. Tsang, M. N. El-Gamal, and K. Iniewski, "Recent advances and future trends in low power wireless systems for medical applications," in Proc. 5th Int. Workshop Systemon-Chip Real-Time Applications, July 20–24, 2005, pp. 476–481.

[5] B.-S. Lin, N.-K. Chou, F.-C. Chong, and S.-J. Chen, "RTWPMS: A real-time wireless physiological monitoring system," IEEE Trans. Inform. Technol. Biomed., vol. 10, no. 4, pp. 647–656, Oct. 2006.

[6] "Center for Life Science Automation," [Online]. Available at: http://celisca.de/

[7] M. Haghi, K. Thurow, and R. Stoll, "Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices," Healthcare Informatics Research, vol. 23, no. 1, p. 4, 2017.

[8] R.W. Cattrall (1997) Chemical Sensors, Oxford University Press, Oxford, 75pp.

[9] U. Anliker, J. A. Ward, P. Lukowicz, G. Troster, F. Dolveck, M. Baer, F. Keita, E. B. Schenker, F. Catarsi, L. Coluccini et al., "AMON: a wearable multiparameter medical monitoring and alert system," IEEE Transactions on Information Technology in Biomedicine, vol. 8, no. 4, pp. 415–427, 2004.

[10] A. Lymberis and R. Paradiso, "Smart fabrics and interactive textile enabling wearable personal applications: R&D state of the art and future challenges," in Proc. of the IEEE-EMBS 30th Annual International Conference of the Engineering in Medicine and Biology Society, 2008, pp. 5270–5273.

[11] W. Zhang, P. Passow, E. Jovanov, R. Stoll and K. Thurow „A Secure And Scalable Telemonitoring System Using Ultra-Low-Energy Wireless Sensor Interface For Long-Term Monitoring In Life Science Applications" 2013 IEEE International Conference on Automation Science and Engineering (CASE)

[12] F. Sanfilippo and K. Y. Pettersen „A Sensor Fusion Wearable Health-Monitoring System with Haptic Feedback"2015 11th International Conference on Innovations in Information Technology (IIT).

[13] "Home - Ultra Low Power Wireless Solutions from NORDIC ..." [Online].Available:https://www.nordicsemi.com/eng/Products/Bluetooth-low.

Command and Control Interface for a Navigation Lock at a Hydro Power Dam

Nicolae Luca Iacobici
POLITEHNICA University Timisoara, Romania
Power Systems Department
luca.iacobici@transelectrica.ro

Flaviu Mihai Frigura-Iliasa
POLITEHNICA University Timisoara, Romania
Power Systems Department
flaviu.frigura@upt.ro

Doru Vatau
POLITEHNICA University Timisoara, Romania
Power Systems Department
doru.vatau@upt.ro

Petru Andea
POLITEHNICA University Timisoara, Romania
Power Systems Department
petru.andea@upt.ro

***Abstract*— Increasing the safety of a navigation lock, by creating a modern and efficient communication between the human operator and the machine, for a prompt and accurate information offered to the operator, about the status of the equipment it serves, it is a requirement of today navigation worldwide. This paper aims to describe the implementation of a modern and efficient centralized tracking equipment for both operation and preparation of informative reports on operating activities, applied to a Danube Hydro Power Dam and its Locks. An up to date solution for all these issues consists in a SCADA system disposed on several hierarchical levels. The system thus gains greater reliability, increased modularity, and flexibility in operation. A simplified human-machine interface (interface between user and system) is one of the fundamental characteristics of these applications. Dedicated software development has led to an effective dialogue between the human user and the system implemented.**

Keywords— HM Interface, Navigation Lock, Control

I. GENERALITIES

Upgrading and improving all hydropower navigation and transportation systems is a necessity dictated by increasing their operational safety, but also a part of Romania's energy strategy. Upgrading these activities requires, in addition to large investments in infrastructure, establishing the correct algorithm for the command and control system.

Increasing efficiency throughout whole system is guaranteed primarily by the knowledge of all data and parameters involved in assembly operation. This can be achieved by implementing tele-computerized management systems at hydropower and navigation dam level. This modernization requires real-time monitoring of all system information of interest, control of system parameters and management of this information at all levels [1].

This application was developed by the authors for the Portile de Fier II Hydro Navigation dam and Power Plant, located on the Danube at the border between Romania and Serbia. It is a huge dam built in the ’80s both for hydro power and for improving navigation on the Danube. All applications are referring to the Romanian side lock.

Our team, from the Politehnica University of Timisoara, performed not only the necessary technical transfer but also, we look for specific customer needs for the Dam and Lock operator.

The main Romanian navigation lock located on Portile de Fier II (PdF II) is presented in Fig.1.

Fig. 1. Portile de Fier II main lock on Romanian side

PdF II main lock has a single room for balancing levels (SAS), equipped with the following main hydromechanical equipment:

- Service Plan Gate (PPS);
- Crash Plan Gate (PPA);
- Filling Valves Galleries (VU);
- Drain Valves Galleries (VG);
- Gates Protection Installation (IPPBS);
- Mitred-type Service Gate (PBS)

These devices are execution elements for sluicing system controls. In addition to all hydromechanical equipment, essential for a dam, several other parameters, signals, information are processed automatically in order to make vital decisions on the lock process deployment.

There are 3 options for each machine control:

- From the control tower;
- From the panels inside the electrical control rooms;
- Locally, from the control box for each individual device

One of the fundamental requirements of the sluicing process is the dynamic display of data about the sluicing process. The data which are collected and then stored in the database on industrial PCs located in the command structure and through data acquisition system, will be processed by the monitoring program.

Presentation of the sluicing process displayed on the monitor in graphic mode and simply operated by mouse and keyboard is a modern and efficient solution for communication and accurate information to the operator about the status of the equipment it serves.

II. System Description

The structure of the monitoring system is as follows:

- sensors and transducers implanted in the system for the collection of process parameters;
- an acquisition system which collects the signals from the transducers or actuators and automation system and transmits them to the PC, an industrial acquisition board type CP 1613, which is compatible with existing automation devices within the plant via the PLC
- software for data acquisition and data communication between the dispatcher and the industrial PCs (S7 1613)
- the monitoring Software Program itself.

The monitoring system must perform the following main functions:

a) Collecting data from the acquisition system to a central processing unit (Industrial PC);
b) Displaying the block diagram or dynamic images of the information collected;
c) Triggering alarms for limit exceeding of operating parameters;
d) Storing an alarm history;
e) Performing the function of calculation or centralization;
f) Making Trend Display function or record the evolution of some parameters;
g) Editing and reports trends that are made daily, monthly, etc;

Using databases as part of the monitoring systems is unavoidable, by consequent, in order to determine the instantaneous values and for monitoring their evolution in time, it is mandatory to have an efficient data base management. It might manage their history, the current events taking place, the alarms in case of special events.

Taking in consideration all these aspects, the monitoring system has to be very flexible, both as hardware and software.

We will briefly present each of them.

III. Data Base Management

To ensure the correct function of handling the database, it is necessary to have a strong DBMS that will call functions and procedures generated throughout higher level languages, such as the languages belonging to the C family (Visual C, C ++, Borland C etc.) [3].

Function using the database should be implemented based on the idea that human dispatcher supervising the system has extensive knowledge of programming or use databases, but only general knowledge of computer use [4].

Therefore, the application of the concept of user-friendly software involves a friendly DBMS too.

Figure 2 shows us the main visual interface for the DBMS developed by our team, for this particular application.

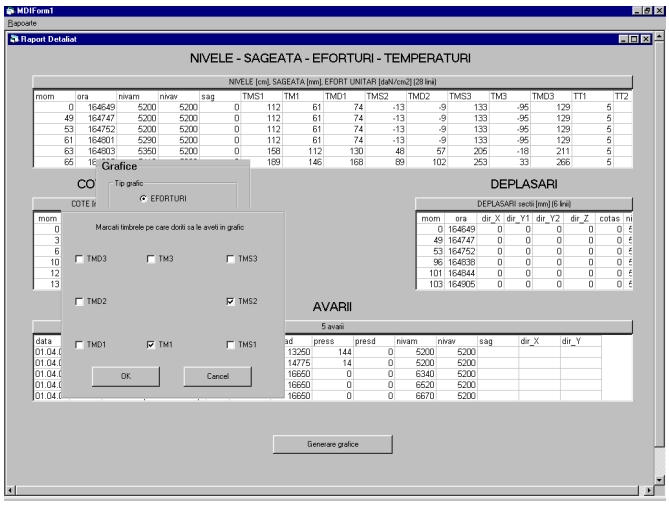


Fig. 2. Data stored and accessed by the DBMS

The schematic diagram of the controlled gate must also provide data elements related to the DBMS, like in Figure 3.

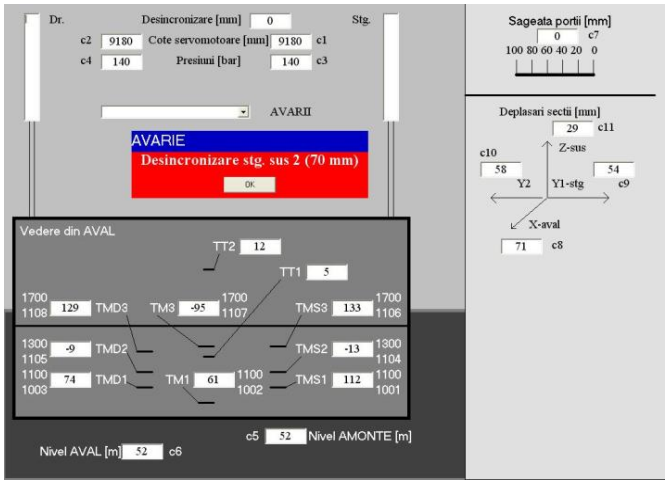


Fig. 3. Graphic representation of data

It also has to provide time evolution of the monitored data, as well as alarm windows in case of emergency. The resulting window for the time evolution (performed also under Visual C) is shown in Figure 4.

We will not insist on the software package we developed for this graphical application and we will present briefly the main windows obtained by using this dedicated software [5].

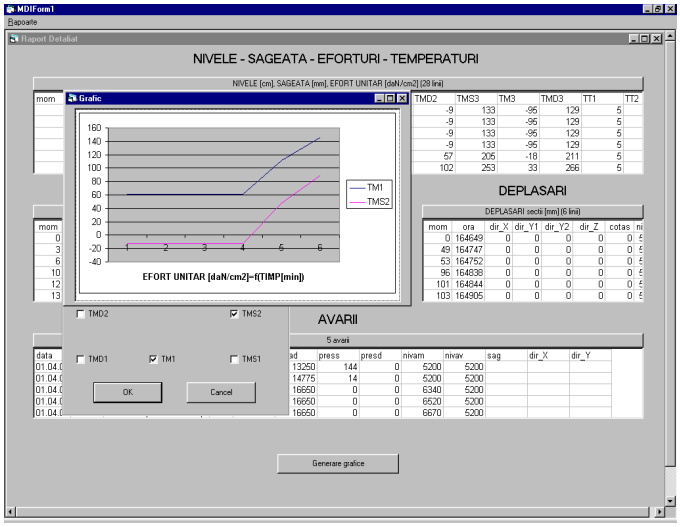


Fig. 4. Data time evolution

Over 800 data articles are collected from the whole system. Figure 5 shows us an alarm window for gate blocking.

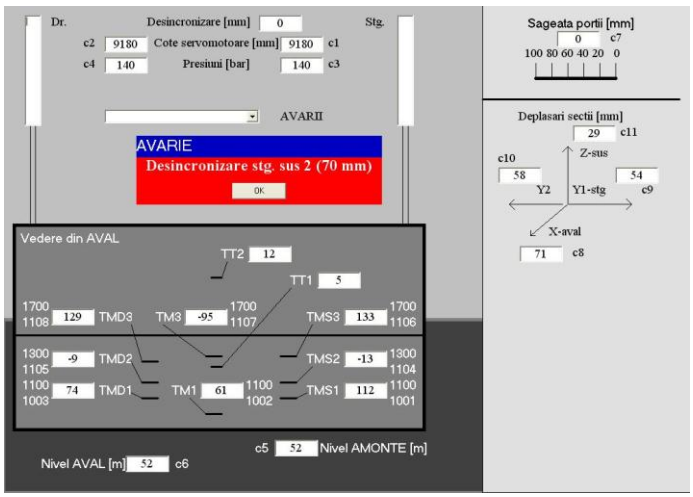


Fig. 5. Alarm window

The solution presented here depends on a database server in the recipient network. If the server exists (and the vast majority of applications running on the network use it for data storage) is probably useful that data stored on the server to be readily available to a third application that needs them (is it usually complex applications for data analysis). The existence of a database server also provides complete solutions for maintenance and back-up, or other similar tasks. In addition, recording and retrieving data will be the responsibility of the DBMS - site (Database Management System Site) and not of the visual program that does not know the internal structure of the database.

Another important part is about synoptic maps and command windows.

IV. Command Interface

A block diagram is a schematic representation of a technological installation. On this diagram are displayed as values , all the main parameters of the process. Representation in this way offers the advantage of a very good overview of the installation. The graphical display of values is dependent on their size. Like in previous alarm windows, exceeding the alarm limits leads to the display of other colors (eg yellow = warning alarms, red = general alarm) allowing the human operator to detect a single operation on the abnormal situations.

A particular case is the synoptic diagram showing the general scheme of the system which is depicted operating status of all devices and communication interfaces in the process [6].

Figure 6 shows an example of the synoptic diagram. There are 12 types of diagrams applied to the lock command.

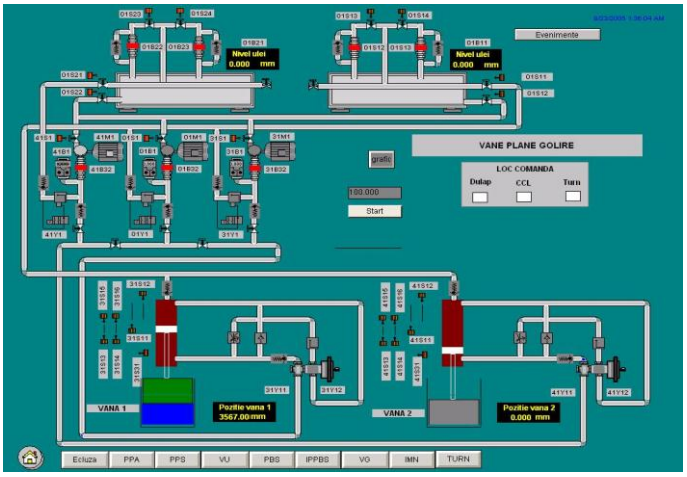


Fig. 6. Synoptic map example

Dedicated windows are used for command of the gates, too.

Figure 7 presents an example of command window used to directly command one of the gates.

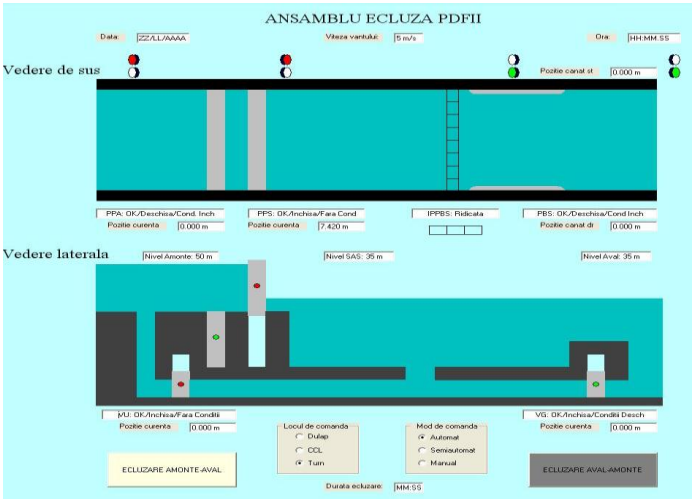


Fig. 7. Command window

Same windows are used for alarm events, too. Figure 8 shows an alarm event directly on the main command window.

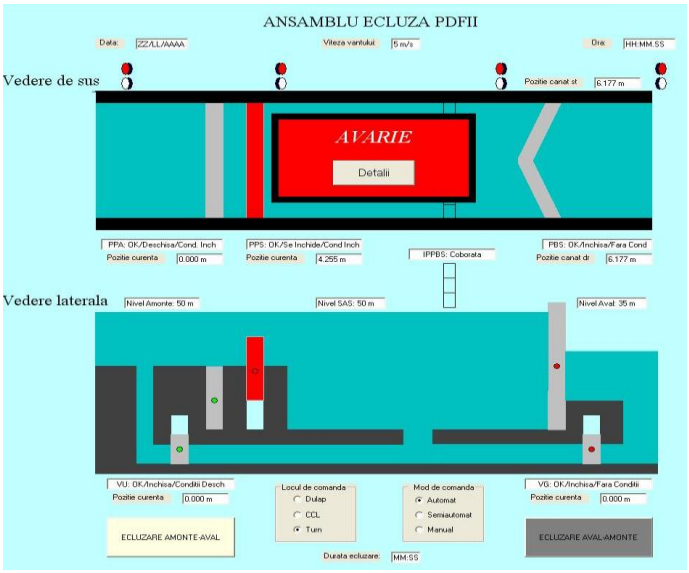


Fig. 8. Alarm on command window

We will not insist on the command and control sequence of each element. We only wanted to show the logic of this HMI and the results of the application.

Some of the applications (data base management and hardware architecture) briefly described here are still property of the manufacturing company and, we cannot give some more details about them.

V. CONCLUSIONS

The application made for PdF II demonstrates the possibility of designing a single monitoring and dispatching interface designed in this case on three hierarchical levels, having a subsystem for automation, one for measuring and monitoring and a subsystem for communication and dispatching. The efficiency of the assembly is guaranteed by introducing, together the most modern systems (transducers, actuators, motors from the latest generation) as well as equipment with programmed logic control facilities (programmable logic controllers -PLC) as well as industrial computers for the process.

This technique is the most advantageous solution from the point of view of operation, proving the efficiency and enhanced reliability in terms of a less demanding maintenance.

This paper is a brief description of the implemented solution, which was a more extended work, made by our team, as well as the team from the HIDROTIM S.A. (which is the manufacturer of all hard equipment existing there.

It is desirable that the existence of such automation solutions and control units applied to some hydraulic systems could prove that our solution could be taken in consideration by other dam operators from Romania and abroad. Our team from the POLITEHNICA University in Timisoara performed other kind of automation applications to other industrial segments, using existing working installations [7], mostly on district heating facilities. It is not something special, but it is an example of how things could be made in case of a less expensive refurbishment of the main gates and motors.

In practice, this application has demonstrated functionality since 2015, operating, making virtually thousands sluicing passes. Moreover, their implementation is accompanied by economic advantages arising from increased performance of the navigation system, increase operational safety and effective diagnosis of system status. The investment pays off shortly.

REFERENCES

[1] V. Iftime, F. M. Frigura – Iliasa, "A Few Measuring Solutions and Some Intelligent Management Systems Used by Heating Distribution Companies and Power Plants in Romania", *Proceedings of the Fourth International Power Systems Conference, Timişoara*, pp. 215 - 218 ISSN 1582 – 7194 (Romania) 8-9th of November 2001

[2] D. Durocher, "Langage: An Expert System for Alarm Processing", *Proceedings of the Eleventh IEEE Workshop on Power System Control Centers*, Montreal, Canada, 1990

[3] M. A.Laughton, "Expert Application in Power Systems", *Prentice Hall International*, 1990

[4] P. Millot, "Configurations homme- machine dans les procedées automatisées", *Edition Octares*, Marseille, Paris, 1990

[5] C. Pancu, A. Baraboi, M. Adam, T. Plesca, GSM Based Solution for Monitoring and Diagnostic of Electrical Equipment, *Proc. of the 13th International Conference on Circuits*, Rodos, Greece, July 22-24, 2009, pp. 58-63.

[6] Vătău Doru, Mușuroi Sorin, Bărbulescu Constantin, Babescu Marius, PV systems modelling and optimal control, Revue Energy Conversion and Management, Volume 84, August 2014, pp. 448-456, ISSN 0196-8904

[7] Vătău Doru, Surianu Flavius Dan, Monitoring of the Power Quality on the Wholesale Power Market in Romania, Proceedings of the 9th International Conference on Electric Power Systems, High Voltages, Electric Machines, Genova, Italy, October 17-19, 2009, pp.59-64, ISBN: 978-960-474-130-4, ISSN: 1790-5117.

Complexity Analysis of Business Processes

Martin Ibl
Institute of System Engineering and Informatics
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
Email: martin.ibl@upce.cz

Žaneta Boruchová
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic

Abstract— **Complexity is a term that is currently used not only in research articles but also in the methodologies and standards used to manage information and projects. The complexity can represent the size of a system, process, program or project, the number of functions and the cost of their acquisition, operation and maintenance. Within the context of this work, the complexity is a variable that represents system properties such as legibility, clarity, comprehensibility, usability, modifiability, easiness of implementation or predictability. With increasing complexity, these characteristics deteriorate and the system becomes more difficult and less effective, for example by increasing its cost, increasing the use of resources, increasing the time and expenses needed for training or maintenance, which may result in a loss of profit for businesses. Different systems or their parts can be compared by complexity measurement. If the system contains elements or bindings that are not necessary, simpler solutions can be created so that its complexity is minimal. This paper presents the process of quantification of complexity in Petri nets and then compares it with other existing approaches. The advantage of the presented complexity measure is the possibility to examine this variable at different levels of system load.**

Keywords—***complexity; process analysis; complexity measure; Petri nets; entropy***

I. INTRODUCTION

The concept of complexity becomes one of the most important concepts of contemporary science and implies the complicacy or size of a system. Such a system is difficult to define, comprehend, modify or use. Complexity refers to the intrinsic, implicit nature of the system, which affects both the properties of the interacting components and the nature of their interaction. The complexity of the system includes aspects such as uncertainty, fluctuation, singularity, internal dynamics, connectivity, and more. Descriptive complexity is how the system appears to an observer standing outside, usually static views. Natural complexity is the intrinsic, real matter of the system and forms the essence of the system. Both types of complexity are linked to information, both the information needed to describe the system and information to clarify uncertainty are embedded in it. These two complexities are in conflict. If we want to limit one, then the other is likely to grow or at best remain the same. This mutual exchange is one of the most important methodological bases of system science.

Business process is a workflow or activity that represents a dynamic component of a system. Each organization is an essentially organized set of processes and activities that interact with each other, run across organizational units and respond to various stimuli from the internal and external environment. In processes, inputs and resources are transformed into outputs that are valued by the process's customer. Processes exist within and between organizations. There is always a flow of work and activity from one person to another. The core of processes is the creation of value or benefit for the organization's customers. The most common division of processes is according to who is their customer and the added value they bring to them. The customer of the process may be a customer of a company, its employee or manager. Processes in an organization are divided into main, supportive, and managing. Major processes create value or benefit to an organization's customer, create a product or service. Supporting processes are all processes whose sole purpose is to ensure the functioning of the main processes and organization. Management processes and activities are all activities that coordinate, manage, organize and plan everything else. Business processes have their own complexity, which, if not controlled, can continue to increase over time, making processes prone to error, difficult to understand and maintain.

In recent years, several scientists have suggested several metrics that can be used to measure and thereby manage the complexity of business processes. Processes are not static. They are constantly undergoing revisions, adaptations, changes and adjustments to meet the needs of end-users. The complexity of these processes and their continuous development make it very difficult to ensure their stability and reliability. As the simplest measures of complexity can essentially be considered the size or length of the process.

Organizations are increasingly struggling with the issue of managing business processes, workflows, and more recently with web processes. One of the important aspects of business processes that has been overlooked is their complexity. A high complexity of processes can result in poor comprehensibility, errors, defects and exceptions, resulting in processes requiring more time to develop, test, and maintain. For this reason, it is necessary to avoid excessive complexity. The measurement of business processes is the task of empirical and objective assignment of numbers, due to their characteristics and in such a way as to describe them. Required attributes include complexity, cost, maintainability, and reliability. Metrics should be evaluated by theoretical (or empirical) validation principle, for example in terms of Weyuker's properties [1], to ensure that the metric is consistent and effective. Business process

management systems, referred to as BPMS, provide the basic infrastructure for defining and managing business processes. BPMS, such as workflow systems, has become a serious competitive factor for many organizations that are increasingly struggling with the issue of managing business applications, workflows, web services, and web processes. Business processes promise to mitigate several of the current challenges in infrastructure such as data, applications, and process integration. With the emergence of web services, the workflow process management system becomes crucial to support, manage and receive processes, both between businesses and within the enterprise. The measurement process deals with the derivation of a numeric value for process attributes. Measurements can be used to improve productivity and process quality. Designing and improving processes is a key aspect for businesses to remain competitive in today's market. Organizations are forced to improve their business processes because customers require better products and services. The business process consists of a series of activities, tasks or services that together lead to the goal.

The goal of this contribution is to compare existing measures of complexity with a previously defined measure [2] and evaluate its advantages and disadvantages. It is essentially an empirical validation of this measure. For the comparison process itself, the two most widely used measures were used, namely the McCabe's cyclomatic measure and Cardoso's CFC measure.

II. STATE OF THE ART

Analysing complexity at all stages of process lifecycle helps to avoid the disadvantages associated with high complexity of processes. At present, organizations have not accepted complexity metrics as part of their process management projects. As a result, simple processes can be designed unnecessarily complex. Using complexity analysis helps design and implement processes and workflows that are more simple, reliable, and robust. In-depth analysis is needed to correct defects in high complexity process parts. There are three questions that are often asked when measuring the complexity of a process [3]:

How difficult is the process to describe?

How difficult is the process to create?

What is the level of organization?

Complexity measurements can be grouped into the following categories depending on which question they are trying to answer:

- The difficulty of the description, typically measured in bits, such as information, entropy, algorithmic complexity, minimum length of description, Fisher information, Rényi entropy, length of code, Chernoff's information, Lempel-Ziv complexity, dimension and fractal dimension;
- The difficulty of creating, working with time, currency, or energy, such as computational complexity, time computational complexity, spatial computational complexity, information-based complexity, logical depth, thermodynamic depth and cost;
- A degree of organization that can be divided into the difficulty of describing the organizational structure and the amount of information divided between the parts of the system as a result of this organizational structure. This category includes, for example, stochastic entropy, sophistication, effective complexity, real complexity, ideal complexity, hierarchical complexity, schema length, homogeneous complexity, grammar complexity, information exchange algorithm, channel capacity or correlation.

Measurement has a long tradition and is a basic discipline in any type of engineering. Engineers have to be experienced in estimating and valuing, which means understanding the activities and risks associated with process development, forecasting and asset management, risk management, reliable delivery and proactive management to avoid a crisis. There is no single metric that would measure the complexity of the process. One of the most sophisticated methodologies to analyse complexity of processes has been created by Cardoso [4], which identifies four main views of complexity levels, namely complexity of activities, called AC, flow control complexity, also called control-flow, data stream complexity denoted as DFC and complexity of resources, labelled RC. The complexity of the AC simply counts the number of activities that the process has. While this metric is very simple, it is important to complement other forms of complexity. While control-flow complexity can be very low, the complexity of AC can be very high. For example, a sequencing process that has thousands of activities has control-flow complexity equal to zero, while its AC complexity is 100. The control-flow complexity is influenced by the design process. It is necessary to consider the existence of XOR, OR and AND operators. The complexity of DFC increases with the complexity of data structures, the number of formal activity parameters, and mapping between activity data. The metric may consist of several sub-metrics that include data complexity, complexity of the interface, and complexity of the integration interface. While the first two sub-metrics relate to static data aspects, the third metric is more dynamic in nature and is focused on data dependencies between different process activities. The RC complexity concerns process activities that need access to resources. Source is defined as any entity (e.g. human resources, IS resources, IT resources) that the activity requires during execution, such as document, database, printer, external application or role. Resources can be structured into the organization context. The structure that is used to form different types of resources can be analysed to determine its complexity. This analysis can help managers reduce administrative costs and optimize resource usage. The CFC metric can be used to analyse the complexity of business processes, as well as the workflows and processes associated with the website. The metric is validated using Weyuker's properties [1, 5], which provide an important basis for classifying complexity measures to determine whether they can be qualified as good, structured, and complex.

Other very popular complexity measure is the so-called cyclomatic complexity (MCC) defined by McCabe [6]. Since its development, MCC has been one of the most promising software metrics. The resulting empirical knowledge base has enabled software developers to calibrate their own software

measurements and gain some understanding of its complexity. Software metrics are often used to obtain a quantitative expression of program complexity. They cannot be confused with the complexity of algorithms that aim to compare the performance of the algorithm. It has been found that software metrics are useful in reducing software maintenance costs by assigning a numeric value that reflects the ease or difficulty with which the program module can be understood. MCC is a measure of the number of linearly independent paths in the program. The intention is independence of language and language format. The MCC bears an indication of the complexity of the program flow. From the module control representation graph, it was found that MCC is a reliable indicator of complexity in large software projects. This metric is based on the assumption that the complexity of the program relates to the number of control channels within the program. For example, a ten-line program with ten assignment commands is more comprehensible than a ten-line program with ten if-then commands. The MCC is defined for each module as e - n +2, where e and n is the number of edges and nodes in the control-flow graph. These graphs describe the logical structure of the software modules. Nodes represent computational commands or expressions, and the edges represent handover between nodes. Each possible realizable path of the software module has a corresponding path from the input to the output node of the control-flow graph of the module. An MCC value 10 indicates a simple program without a high risk, a value between 11-20 indicates a more complex program with a moderate risk, and a value between 21 to 50 indicates a complex high risk program.

Gruhn and Laue [7] have suggested a cognitive weight for business process models. This metric, referred to as CFS, is an adaptation of cognitive functional size. Cognitive degrees of complexity are based on cognitive informatics. Cognitive metrics suggest that there are three factors that lead to the complexity of software architecture, model input data, and model output data. This means that cognitive complexity is a function of these three factors. This metric is intended for use with enterprise-class business process models that emphasize visual communication with users but offer minimal formal semantics. The main limitation of this metric is that it ignores two of the three factors that involve cognitive complexity, namely inputs and outputs, and focuses only on flow control. They also suggested customizing the metric of the flow of information for business processes, and unlike Cardoso's IC, this metric does not include the length of the process.

Lassen and van der Aalst [8] have suggested three levels of complexity for the Petri net subclass, called a workflow network. Extension of Cardoso metric ECaM, extended ECyM cycling metrics, and structured SM metrics. ECaM extends CFCs by being tailored to support Petri nets. These metrics were implemented within Prom, a business process measure that focuses on monitoring BAM's business activities.

Vanderfeesten [9] proposed a metric called Cross-Connectivity, labelled CC, based on cognitive complexity. It is the predictive error rate that measures the strength of the bonds among the elements of the process model. It is based on the hypothesis that process models are more understandable and contain fewer errors if they have a high cross-linking CC. In addition to predicting errors, it can also measure the comprehensibility of the business process model. This metric has been empirically evaluated using Spearman's correlation coefficient and multidimensional logistic regression.

Mendling and Neumann [10] have suggested six metrics for errors that are closely related to complexity. These metrics are based on graph theory and include size, separability, context, structure, cyclicality and parallelism. Increasing the size increases the probability of an error. Increasing separability, context, and structure means decreasing the probability of error. Increasing cyclicality and parallelism also increases the likelihood of error.

III. MEASUREMENT OF COMPLEXITY IN PETRI NETS

The Petri net is a mathematical tool for modelling and simulating discreet dynamic event-driven systems and consists of places, transitions, and oriented edges connecting places and transitions. Places may contain tags that are called tokens. The number of tokens at the given places indicates the current state of the system. Transitions represent possible activities that can change the state of the system. Transitions triggers tokens from input to output. The Petri Net provide a visual method for examining the properties of the system.

The following types of Petri nets have been created successively:

- Condition / Event Petri nets, referred to as C / E;
- Place / Transition Petri nets, referred to as P / T;
- P / T Petri nets with inhibiting edges;
- P / T Petri nets with priorities;
- Timed Petri nets, referred to as TPNs;
- Colored Petri nets, referred to as CPN;
- The hierarchical Petri nets, referred to as HPN.

The Petri P/T network, which will be used in the context of this work, consists of places, transitions, oriented edges, capacities, weights, and initial markings. The places are graphically represented by a circle and transitions by a rectangle. Oriented edges point either from a place to a transition or from a transition to a place. Place capacity indicates the maximum number of tokens that may be present at one time. The complexity calculated by the Petri nets is expressed by entropy and represents the uncertainty of the system. The greater the entropy value, the more the model is complex.

Generalized P/T Petri net is a 5-tuple, $PN = (P, T, F, W, M_0)$ where:

- $P = \{p_1, p_2, p_3, \ldots, p_m\}$ – a finite set of places,
- $T = \{t_1, t_2, t_3, \ldots, t_n\}$ – a finite set of transitions,
- $P \cap T = \emptyset$ – places and transitions are mutually disjoint sets,
- $F \subseteq (P \times T) \cup (T \times P)$ – a set of edges (arcs), defined as a subset of the set of all possible connections,
- $W: F \rightarrow N_1$ – a weight function, defines the multiplicity of edges (arcs),
- $M_0: P \rightarrow N_0$ – an initial marking

Let $\boldsymbol{PN}$ be a Petri net and $\boldsymbol{A}$ its transition matrix, vector $\mathbf{u}$:$\boldsymbol{uA} = \boldsymbol{u}$ represents the stationary probabilities of all markings in $\boldsymbol{PN}$. Entropy of $\boldsymbol{PN}$ is then defined as:

$$H(PN) = -\sum_{i=1}^{|R(M_0)|} u_i \log_2 u_i \qquad (1)$$

where $|R(M_0)|$ is the number of all reachable markings for PN.

More details on the quantification of entropy in Petri nets can be found in [2, 11].

IV. Case Study - Granting a Loan Process

Large banks have realized that a new, modern infrastructure information system needs to be adopted in order to be competitive and efficient. Therefore, the first step in this direction was the adoption of the Workflow Management System (WfMS) to support its business processes. Given that the bank provides a number of services to its customers, the adoption of the WfMS has enabled the logic of the banking processes to be captured in the scheme. As a result, part of the services provided are stored and implemented through a workflow management system. One of the services offered by the bank is the process of providing a loan. The process of granting the loan to the client consists of 18 nodes representing activities marked A to R and twenty-four transitions. Four XOR operators and one AND operator are used.

The first activity is the client's entry into the bank's internet application. In order for the client to enter the application, he/she must fill in the password and enter the certificate. Then the client chooses to apply for a loan. The Bank offers three types of loans. Housing loan, education loan or car loan. The client may request only one loan within the process. The bank accepts the client's request and decides whether to approve or reject it. After the bank decides, the client is informed by e-mail, and then the credit application with the resulting decision is stored in the bank's database and the credit application process is completed.

The list of process activities is as follows:

- A - access to the bank's internet application;
- B – inserting passwords;
- C - inserting a certificate;
- D - selection of service;
- E - filling in the loan application and selecting the type of loan;
- F - housing loan;
- G - education loan;
- H - car loan;
- I - housing loan approval;
- J - rejection of housing loan;
- K - approval of credit for education;
- L - rejection of credit for education;
- M - approval of car credit;
- N - rejection of car loan;
- O - informing the client of the decision on the loan for housing;
- P - informing the client of the decision on the loan for education;
- Q - informing the client of the decision on the car loan;
- R - save the request to the bank database and end the process.

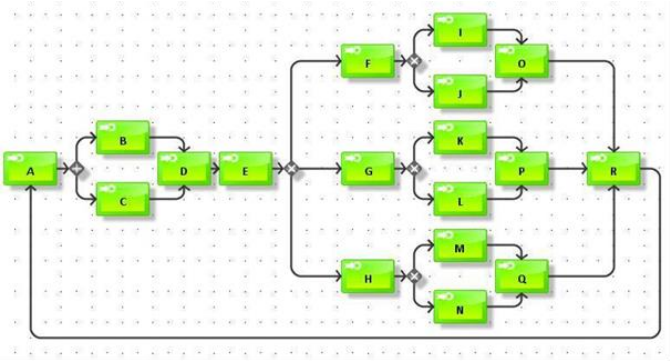


Figure 1. Process of obtaining loan - a case study.

Consequently, the complexity is computed for the process, firstly using a simple McCabe's MCC metric that ignores the used operators, then the Cardoso's CFC metric, which takes operators into account and ultimately entropy through the Petri nets.

MCC is computed using the formula e - n + 2. Where e is the number of edges and n is the number of nodes. The complexity calculated by this metric is equal to 8, which, according to the established limits, points to a simple process without great risk.

For each AND operator, the CFC complexity is equal to one. For the XOR operator, the CFC complexity of activity x is determined by the number of activities that follow from this activity, in other words by the number of outputs from activity x. For the presented process, individual CFC calculations are as follows:

- CFC_{AND} for A = 1;
- CFC_{XOR} for E = 3;
- CFC_{XOR} for F = 2;
- CFC_{XOR} for G = 2;
- CFC_{XOR} for H = 2.

By adding these individual complexities, an absolute CFC is obtain (equal to 10); the relative CFC is equal to 2.

The value of the process entropy modelled in the Petri nets is equal to 3.26.

Figure 2 and Figure 3 illustrate simplified versions of the original process.

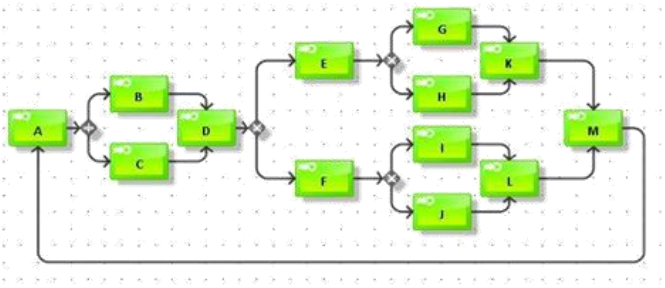


Figure 2. Simplified process A.

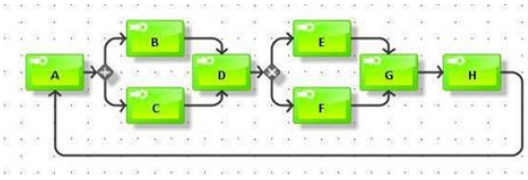


Figure 3. Simplified process B.

Figure *4* represents a more complicated version of the original process.

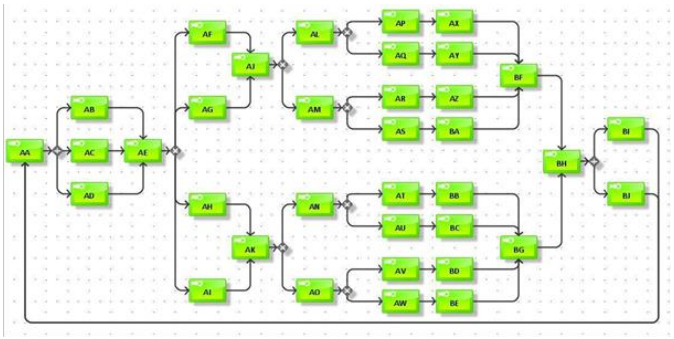


Figure 4. More complicated process.

The Figure 5 shows how the values of the individual complexities for different processes change. Interestingly, for the original process and its simplified variant A, the MCC value is lower than the CFCabs value. This is otherwise for simplified variant B, where the CFCabs value drops below the MCC, which is due to the fact that only two operators are used for this variant.

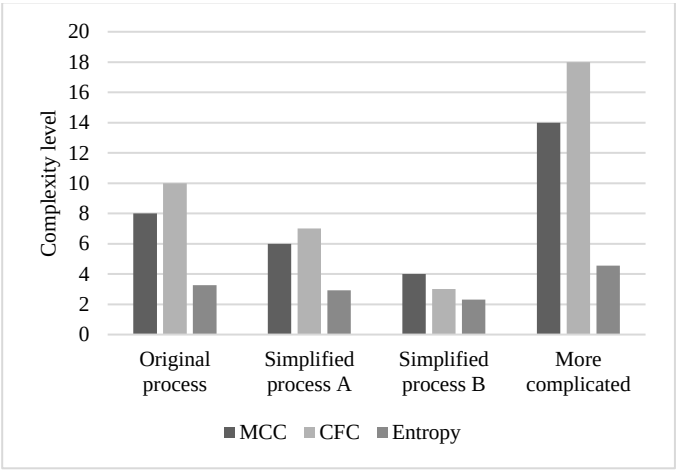


Figure 5. Comparison of different measures of complexity

The analysis shows that when modelling processes it is good to consider the number of activities and operators and try to minimize them to make the process as effective as possible.

A. Correlation Analysis

Based on the calculated values for the original process, its simplified variants A and B and for the more complicated process, a statistical correlation analysis of the individual complexity measures was performed in the SPSS Statistics program. Table 1 shows that there is a positive correlation between all metrics, the higher the value of one metric, the higher the value of the second metric. In all cases, this is a significant dependence, with the largest one being between CFC and entropy.

TABLE I. CORRELATION ANALYSIS

		MCC	CFC	Entropy
MCC	Pearson Correlation	1	,996**	,997**
	Sig. (2-tailed)		,004	,003
	N	4	4	4
CFC	Pearson Correlation	,996**	1	,999**
	Sig. (2-tailed)	,004		,001
	N	4	4	4
Entropy	Pearson Correlation	,997**	,999**	1
	Sig. (2-tailed)	,003	,001	
	N	4	4	4

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

B. Process Load Analysis

The analysis is carried out in Petri nets and examines the load of the original process and its simplified variants, namely how the complexity, in this case expressed by entropy, is changing, with the growing number of users who are applying for a loan at the same time. The increase in load of the processes is shown in Figure 6, which shows that with the increasing number of registered users the complexity of the processes is increasing but gradually the growth is slowing down.

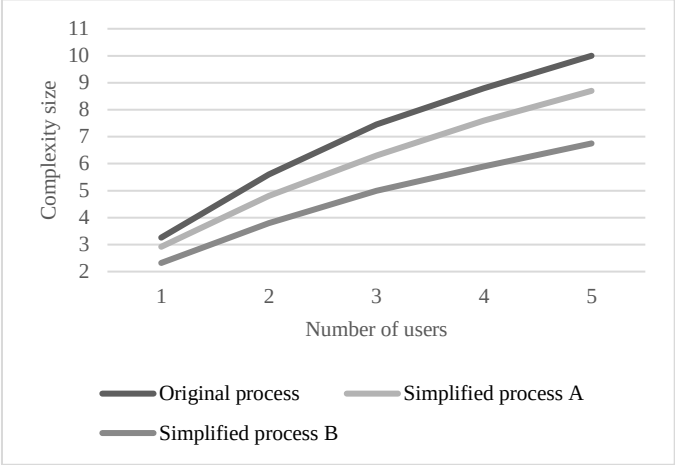


Figure 6. Comparison of different workload three processes.

V. DISCUSSION

From the empirical results, it is possible to see that the various complexity measures show a highly correlated dependence. It infers that the use of in this work specified measures is interchangeable. The main advantage of the complexity measure based on quantification of entropy in Petri nets is the possibility to simulate the increase/decrease of the load of individual states and monitor the response. The entropy measure of complexity in Petri nets therefore extends standard measures to the dynamic component. This makes it possible to achieve a more precise decision-making in general.

VI. Conclussion

Based on the comparative analysis, it can be stated that the individual complexity measures are comparable with statistical significance. In addition, it is possible to recommend the use of entropy in Petri nets as it extends the other measures with dynamic complexity analysis.

Acknowledgment

The paper was supported by the University of Pardubice, Faculty of Economics and Administration, Project SGS_2017_017.

References

[1] E. J. Weyuker, "Evaluating Software Complexity Measures," *IEEE Trans. Softw. Eng.*, vol. 14, pp. 1357-1365, 1988.

[2] M. Ibl, "Uncertainty Measure of Process Models using Entropy and Petri Nets," in *ICSOFT 2013 - Proceedings of the 8th International Joint Conference on Software Technologies*, 2013, pp. 542-547.

[3] S. Lloyd, "Measures of complexity: a nonexhaustive list," *IEEE Control Systems*, vol. 21, pp. 7-8, 2001.

[4] J. Cardoso, "Business Process Control-Flow Complexity: Metric, Evaluation, and Validation," *International Journal of Web Services Research (IJWSR)*, vol. 5, pp. 49-76, 2008.

[5] J. Cardoso, "Control-flow Complexity Measurement of Processes and Weyuker's Properties," *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, vol. Vol: 1, pp. 366-371, 2007.

[6] T. J. McCabe, "A Complexity Measure," *IEEE Trans. Softw. Eng.*, vol. 2, pp. 308-320, 1976.

[7] V. Gruhn and R. Laue, "On Experiments for Measuring Cognitive Weights for Software Control Structures," in *6th IEEE International Conference on Cognitive Informatics*, 2007, pp. 116-119.

[8] K. B. Lassen and W. M. P. van der Aalst, "Complexity metrics for Workflow nets," *Information and Software Technology*, vol. 51, pp. 610-626, Mar 2009.

[9] I. Vanderfeesten, H. A. Reijers, J. Mendling, W. M. P. van der Aalst, and J. Cardoso, "On a Quest for Good Process Models: The Cross-Connectivity Metric," in *Advanced Information Systems Engineering: 20th International Conference, CAiSE 2008 Montpellier, France, June 16-20, 2008 Proceedings*, Z. Bellahsène and M. Léonard, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 480-494.

[10] J. Mendling, M. Moser, G. Neumann, H. M. W. Verbeek, B. F. Dongen, and W. M. P. Aalst, "Faulty EPCs in the SAP Reference Model," in *Business Process Management*. vol. 4102, S. Dustdar, J. Fiadeiro, and A. Sheth, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 451-457.

[11] M. Ibl and J. Čapek, "Measure of Uncertainty in Process Models Using Stochastic Petri Nets and Shannon Entropy," *Entropy*, vol. 18, p. 33, 2016.

Grouping Genetic Algorithm for the Capacitated *p*-median Problem

Ľudmila Jánošíková, Patrik Vasilovský
Faculty of Management Science and Informatics
University of Žilina
Univerzitná 1, 010 26 Žilina, Slovak Republic
Ludmila.Janosikova@fri.uniza.sk

***Abstract*— The paper suggests an adaptation of a grouping genetic algorithm for solving the capacitated *p*-median problem. We propose a new encoding of the individual solutions that enables an efficient implementation of the crossover operation. A hybrid metaheuristic that combines the grouping genetic algorithm with the post-processing solver is proposed as well. Numerical experiments performed on benchmark instances proved the predominance of the grouping encoding over the standard encoding.**

Keywords—facility location; capacitated p-median problem; grouping genetic algorithm; linear programming

I. INTRODUCTION

The capacitated *p*-median problem (CPMP) is a classical discrete location problem. A fixed number of *p* facilities are to be located in the given area. Facilities serve customers with specified demands. In the most simple case, the demands do not depend on the distance or time travelled. The goal is not only to decide where facilities should be deployed but also which customers each facility will serve so that the total or average distance between facilities and customers could be as small as possible. To operate effectively, each facility can serve only a limited number of customers. The capacity limit may be alternatively defined as a maximum demand volume that can be assigned to the facility.

In the original mixed integer programming formulation of the problem customers' demands were allowed to be split among several facilities [1]. However, subsequently a pure integer formulation has occurred more frequently. Such a formulation assigns customers to exactly one facility, which is a common requirement in location problems in the public sector [2]. Examples of public facilities include public schools [3] or emergency medical stations [4, 5, 6]. The emergency medical service (EMS) is also behind our present research.

The *p*-median problem arises in the situation when an existing EMS system is to be improved. The aim is to find better locations of the stations where emergency vehicles are housed. Due to economic reasons the current number of stations will not increase. We focus on location of the stations in a large-scale area, e.g. the territory of a state. In that case a precise location of individual patients cannot input the model because i) the model is a plan for the future and we do not know in advance where the patients will occur; ii) if the patients were located in the specific address or street, the model would be intractable. That is why a macroscopic view must be applied. It means that patients are aggregated to municipalities. Thus the customers of the system are not individual patients but whole municipalities. In accordance with [7], the number of emergency calls arising in a municipality is supposed to be proportional to the number of its inhabitants, and so the volume of the customer's demand may be identified with the municipality population. At the same time, every municipality is regarded as a candidate location for a station. Every station can serve only a limited service area. The optimization criterion should reflect the accessibility of the service. A surrogate of the accessibility may be the average travel time of an ambulance from a station to a patient. An alternative formulation of the objective is the total travel time, where the time from a station to a municipality is multiplied by the municipality's population. The average travel time can be calculated by dividing the total travel time by the sum of all inhabitants. This way the problem of a station location becomes a weighted capacitated *p*-median problem.

The paper is organized as follows. Section II states the notation and formal definition of the weighted capacitated *p*-median problem. The relevant literature on solution methods is surveyed in Section III. Two versions of the genetic algorithm are described in Sections IV and V, respectively, and experimentally evaluated in Section VI. Section VII contains concluding remarks.

II. PROBLEM FORMULATION

Let us define the problem in a more formal way.

We are given a set I of n candidate locations, where emergency stations can be placed. Each station is equipped with one ambulance. The population of the region served by one ambulance cannot exceed Q people. Potential patients live in municipalities spatially spread in the given territory. We denote the set of the municipalities by J and the number of inhabitants of a municipality $j \in J$ by b_j. Further on, let t_{ij} be the shortest travel time of an ambulance from node $i \in I$ to node $j \in J$. The goal is to locate exactly p stations in nodes from the set I in order to minimize the total travel time needed to reach all potential patients.

The decision on opening a station must be done for each candidate location $i \in I$. To model this decision, we need a binary variable y_i, which takes the value 1 if a station is located in node i, otherwise it takes the value 0. The assignment of municipality j to the station located in node i is modelled by binary variables x_{ij}. Variable x_{ij} takes value 1, if municipality j will be served by an ambulance located in node i, otherwise $x_{ij} = 0$.

After these preliminaries, the model of the weighted capacitated p-median problem can be written as follows:

$$minimize \sum_{i \in I} \sum_{j \in J} t_{ij} b_j x_{ij} \quad (1)$$

$$subject\ to \sum_{i \in I} x_{ij} = 1 \quad for\ j \in J \quad (2)$$

$$x_{ij} \le y_i \quad for\ i \in I, j \in J \quad (3)$$

$$\sum_{j \in J} b_j x_{ij} \le Q \quad for\ i \in I \quad (4)$$

$$\sum_{i \in I} y_i = p \quad (5)$$

$$x_{ij}, y_i \in \{0,1\} \quad for\ i \in I, j \in J \quad (6)$$

The objective function (1) minimizes the travel time between stations and municipalities that is weighted by the population of the municipality. Constraints (2) assign every municipality j to exactly one station i. Constraints (3) ensure that if a municipality j is assigned to a node i, then a station will be open in the node i. Constraints (4) limit the total number of inhabitants in the region served by one ambulance. Constraint (5) limits the total number of stations that can be sited. The remaining obligatory constraints (6) specify the definition domains of the variables.

III. SOLUTION METHODS

The capacitated p-median problem is known to be an *NP*-hard problem. It means that only small to medium-sized instances can be solved to optimality. An efficient exact method was introduced in [8]. In principle, the method is a cut and branch procedure based on Fenchel cutting planes. The authors report that it outperforms another exact procedure – the branch-and-price algorithm by Ceselli and Righini [9] that exploits column generation, heuristics and branch-and-bound. The behaviour of both algorithms depends not only on the size of the instance but also on the p/n ratio. The branch-and-price algorithm was unable to find the optimal solution to most instances with $n \ge 100$ and $p \ge n/4$ in the time limit of 1 hour. However, in the case of hard instances the algorithms may be used as approximation procedures by limiting the calculation time. Among approximation methods, often cited is the Lagrangean/Surrogate Local Search Heuristic (LSLSH) by Lorena and Senne [10]. The heuristic exploits the Lagrangean/surrogate relaxation of the problem that is solved by a subgradient method and combined with an allocation heuristic to make the dual solution feasible. Two other interchange heuristics are used to improve the feasible solution. All popular metaheuristics were applied for solving the CPMP as well, e.g. a genetic algorithm [11] or a scatter search method [12].

Research in the past two decades has proved that the most successful solution methods combine concepts of various algorithmic approaches. A metaheuristic working in cooperation with another algorithm (either exact, heuristic or even another metaheuristic) or embedding other algorithmic components is called a hybrid metaheuristic. One of the first attempts to combine concepts of different metaheuristics was made by Osman and Christofides [13]. They proposed a simulated annealing method with some features adopted from the tabu search metaheuristic. They generated the first set of benchmark instances of size 50×5 and 100×10. Several hybrid approaches for the CPMP were devised by the research group around Masoud Yaghini from the Iran University of Science and Technology. In [14] they proposed a hybrid metaheuristic called GACO that combines two metaheuristics: a genetic algorithm (GA) and ant colony optimization metaheuristic (ACO). The GA guides the search. It acts with an incomplete representation of candidate solutions, specifically with sets of medians. The ACO is used to obtain corresponding actual solutions, it means to allocate demand nodes to the medians. The proposed metaheuristic was tested on small and moderate-sized instances in which the number of demand nodes ranged from 50 to 1030 and p from 5 to 70. Reference [15] introduces a hybrid metaheuristic based on a tabu search method that applies mathematical programming to explore the neighbourhood of the current solution. Numerical experiments were performed on instances with 50 to 200 nodes and 12 to 50 medians. Another hybrid metaheuristic that combines path relinking and scatter search principles was proposed by Díaz and Fernández [16]. Path relinking is used within the scatter search metaheuristic to combine two solutions from the reference set. The authors compare their hybrid metaheuristic with other published methods on small and medium-sized instances. Díaz and Fernández compare their results with the lower bound of the objective function value and report the largest gap of 2.36%.

In all cited papers the authors declare good performance of their proposed methods and their dominance to other methods in terms of the solution quality as well as computation time. However, none of the published metaheuristics was tested on large-scale instances with several thousands of demand nodes.

Finally we mention a state-of-the-art method introduced by Stefanello et al. [17] under the name IRMA (Iterated Reduction Matheuristic Algorithm). IRMA is an efficient heuristic based on a domain decomposition proposed by Taillard [18, 19]. It can be regarded as a hybrid heuristic as well since it utilizes the local optimization method as a master method and the integer programming solver as a slave solving a subproblem. In addition, the solution space is reduced by eliminating variables that probably will not be in the optimal solution. The authors evaluate the performance of IRMA on different instances including the 3038-node instances by Lorena and Senne [20]. They report very good results (the largest gap from the lower bound is between 0.57% and 3.15% for these large-scale instances).

The weakness of the decomposition methods in general is that they can hardly ensure global constraints. For example, if the capacitated *p*-median model was used to relocate current stations, one might require only a predefined percentage of the current stations to change their locations. Population-based metaheuristics seem to be more suitable in such situations, since they can easily control global constraints by a proper reproduction operator.

That is why we focus on a genetic algorithm in this study and more specifically on its grouping version. The motivation comes from successful applications of the grouping encoding scheme within population-based metaheuristics. The grouping encoding was proposed by Falkenauer [21]. It suits well on problems where objects are to be divided into groups. The CPMP falls into this problem category. It was already solved by two grouping evolutionary algorithms, namely a genetic algorithm and a harmony search algorithm [22]. In [22] both algorithms are combined with a specially tailored local search procedure and tested on randomly generated, medium-sized instances. In the presented study we propose a more efficient implementation of the grouping genetic algorithm and concentrate on large-scale instances since our practical problem is to deploy EMS stations around the whole state territory with thousands of municipalities (customers and candidate locations). For example, in the Slovak Republic there are 2928 municipalities and currently 273 EMS stations are operating.

IV. Genetic algorithm with the standard encoding and selective crossover

The genetic algorithm (GA) is a well-known population-based metaheuristic. It evolves the population of candidate solutions to a given problem by repeatedly applying operators based on a natural selection and genetic recombination to the current population until a stopping criterion is met.

In this section we describe an implementation of the GA for the CPMP using a standard encoding of individuals. The encoding was proposed for the uncapacitated *p*-median problem by Alp et al. [23] and used in further successful implementations for the problem with or without capacity limits [6, 11, 24].

The standard encoding of an individual means that the solution is recorded in the form of a *p*-dimensional vector. A vector element is a gene the allele of which represents the index of a candidate location selected as a median. The genome is interpreted as a set, which means that there are no duplicated indices and there is no ordering among the indices. This encoding ensures that constraint (5) is always satisfied.

The fitness function is identical to the objective function (1). To calculate the fitness value, the customers must be assigned to current medians encoded in the individual. The assignment itself is also an *NP*-hard problem since it must respect capacity limits of facilities. To obtain its optimal solution is a time-consuming task therefore a suboptimal solution is calculated instead, using a heuristic. An efficient heuristic for the assignment problem was proposed by Martello and Toth [25] and is known under the name MTHG.

The GA works on a population with fixed size. The size should be large enough to ensure that every candidate location occurs in the initial population at least once. In the case of the *p*-median problem it means that a minimum number of individuals is $\lceil n/p \rceil$ where $n = |I|$ is the number of candidate locations. However, this minimum number does not guarantee sufficient genetic diversity. Therefore we use a population of size $L = 100$ in our experiments as was suggested by Correa et al. [11].

The initial population is produced using random permutations of all candidate locations. The first permutation splits successively to $\lfloor n/p \rfloor$ individuals. The first individual consists of the first p members of the permutation, the second individual consists of members $p + 1$, $p + 2$, ..., $2p$ and so on. If n is not divisible by p, then n mod p members of the first permutation left and are available for the individual that is completed from the second permutation. The process continues until L individuals are created.

A crossover operator is applied in every generation to create two new individuals from a pair of selected individuals. The parents are selected according to the ranking-based method that was also used in [11]. The population is supposed to be ordered in ascending order by the fitness value. The first parent is the individual whose index r is given by the following formula:

$$r = L - \left\lfloor \frac{-1 + \sqrt{1 + 4rnd\left(L^2 + L\right)}}{2} \right\rfloor, \qquad (7)$$

where $rnd \in [0, 1)$ is a uniformly-distributed random number. The second parent is selected using the same formula.

Two implementations of the crossover operation with the standard encoding can be found in the literature. The implementation by Alp et al. [23] first merges the genes of the parents, obtaining an infeasible solutions with more than p medians. The fitness value of such an infeasible solution is calculated. Medians that are present in both parents are preserved in the offspring and the other medians are reduced by a greedy deletion heuristic. The heuristic deletes successively one median at a time until p medians left. The deleted median is the one that increases the fitness value by the least amount.

Correa et al. [11] use a one-point crossover. As a pre-processing step for the application of the crossover, two exchange vectors, one for each parent, are created. The exchange vector contains those parent's alleles that are not present in the genome of the second parent. For instance, let the two parents be the vectors $P_1 = [1, 2, 3, 4, 5]$ and $P_2 = [2, 5, 9, 10, 12]$. Then their respective exchange vectors are $E_1 = [1, 3, 4]$ and $E_2 = [9, 10, 12]$. The crossover is applied on the exchange vectors as follows: a random natural number k, varying between 1 and the number of elements in the exchange vectors minus 1, is generated. The parts of the exchange vectors behind this point are swapped. Let for example $k = 1$. Then the new exchange vectors are [1, 10, 12] and [9, 3, 4]. Two new individuals are produced. Each of them contains common alleles inherited from both parents ([2, 5] in our

example) and one of the new exchange vectors. So the offspring will be D_1 = [2, 5, 1, 10, 12] and D_2 = [2, 5, 9, 3, 4].

We propose a new crossover operator. The main idea is to preserve only positive traits of the parents in the offspring. The operator uses the problem knowledge to select high-quality alleles from the parents' genomes. First, a complete genetic information from both parents is collected. It means that a union operation of the parents' genomes is performed, producing the vector [1, 2, 3, 4, 5, 9, 10, 12] in our example. Then the fitness value of such a vector is evaluated, and priorities are assigned to the alleles, using some problem knowledge. Finally the alleles are sorted in a descending order according to their priorities, and one offspring is created, containing only p best alleles of the combined vector.

In general, the priority function should reflect the quality of the allele, i.e. the probability of being in the optimal solution. In the capacitated p-median problem, the priority value of a median may be defined as the number of customers that are assigned to it [24]. However, this may not work if there are customers with high demands close to the median's capacity because then the median may serve only one customer. The priority should rather be the total demand of the customers assigned to the median. If a median has (almost or entirely) exhausted capacity, it means that it is close to customers, and the assignment of customers to it increases the objective function value as little as possible. Such medians probably constitute a good solution.

In the following text we will denote this type of the crossover operation as selective crossover.

To preserve diversity in the population, the offspring goes through mutation with a pre-set probability that is fixed during the whole evolution. Mutation is performed as follows: we pick out an arbitrary median in the offspring and replace it by another randomly selected median that is not present in the solution.

The population is renewed in a steady-stated method: the offspring is included in the population if it has better fitness value than the worst individual in the old population. The process finishes after a pre-defined number of generations has evolved or a pre-defined amount of computation time has elapsed. The individual in the last population with the lowest fitness value represents the best solution.

V. GENETIC ALGORITHM WITH THE GROUPING ENCODING

The standard encoding described above has an important drawback: an individual stores only a part of the solution but not the whole solution. More precisely, only location of medians (variables **y**) are stored but the assignment of customers (variables **x**) is not encoded at all. Such encoding saves memory but has two negative impacts: i) the fitness value has to be calculated any time a new individual is generated, which is time consuming; ii) the approximation method used for fitness evaluation may destroy compact service areas, where customers are close to the service centre and the centre has no spare capacity. A grouping encoding scheme tries to avoid these drawbacks.

Reference [22] uses the grouping scheme in two metaheuristics for solving the CPMP. The grouping scheme divides the encoding of an individual into a group part and an object part. The group part records the location of medians, i.e. it is a p-dimensional vector that stores the indices of candidate locations selected as medians. So the group part corresponds to the simple encoding described in the previous section. The object part records the assignment of customers to the medians. For instance, let us suppose that $I = J$, $|I| = |J| = 12$ and $p = 5$. Then an individual can be [1, 2, 3, 4, 5 | 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 5, 5]. The first part of the vector is the group part meaning that the medians are located in nodes 1, 2, 3, 4, and 5. The second part (after the slash symbol) contains the indices of medians which individual customers are assigned to. One can see that the first and sixth customer are assigned to median 1, the second and seventh customer belong to median 2 and so on, finally median 5 serves customers 5, 10, 11, and 12. In [22] the crossover operation is implemented in a rather inefficient way since the crossover may produce an infeasible individual with exceeding number of medians or broken capacity limits. Infeasible individuals must be subsequently repaired by a heuristic procedure that still may not be able to produce a feasible solution and the generated individual must be discarded. This wastes time.

We propose an alternative encoding that enables the implementation of a more efficient crossover. The proposed crossover operation is an adaptation of a general scheme by Falkenauer [21]. In principle, we use the two-point crossover that works on the group part of the parents. It always produces two offspring maintaining the service areas of the inherited medians. The crossover operation proceeds through these steps:

1) Select at random two crossing points, delimiting the crossing section in each of the two parents.
2) Copy the crossing section of the first parent into the offspring. Copy the assignment of customers to the medians present in this crossing section.
3) Complete the offspring with the medians of the second parent that are outside the crossing points. If a median is already in the offspring, choose an arbitrary median from the crossing section of the second parent instead and insert it to the offspring. Copy the assignment of customers to the medians inherited from the second parent unless they have been assigned yet to the medians from the first parent.
4) If there is an unassigned customer left, assign it to the nearest median with lose capacity margin.
5) Find a better location of the median in each cluster. A candidate location with sufficient capacity that minimizes the total distance to the other nodes becomes the new median.
6) Apply steps 2 through 5 to the two parents with their role reversed in order to generate the second child.

For illustration let the two parents be P_1 = [1, 2, 3, 4, 5 | 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 5, 5] and P_2 = [2, 6, 7, 8, 9 | 6, 2, 7, 8, 9, 6, 7, 8, 9, 6, 6, 6]. Let the crossing points 2 and 4 be generated in step 1. Then the crossing sections will be positions 2, 3, and 4 in the group parts of the parents, it means medians [2, 3, 4] in the first parent and [6, 7, 8] in the second parent, respectively. The first offspring after step 2 will be [-, 2, 3, 4, - | -, 2, 3, 4, -, -, 2, 3, 4, -, -, -]. In step 3, the offspring is completed with medians from the second parent that are outside the crossing

section and are not in the first parent. Since median 2 is already in the offspring, we get [-, 2, 3, 4, 9 | -, 2, 3, 4, 9, -, 2, 3, 4, -, -, -]. To fill in the missing median we pick out an arbitrary median from the crossing section of the second parent that is not in the first parent, e.g. 7. After step 3, the offspring will be [7, 2, 3, 4, 9 | -, 2, 3, 4, 9, -, 2, 3, 4, -, -, -]. The unassigned customers are subsequently assigned to the closest median with a loose capacity margin. After step 4, the offspring may be [7, 2, 3, 4, 9 | 7, 2, 3, 4, 9, 4, 2, 3, 4, 7, 4, 9]. In step 5 the service areas are recalculated so that a node with the minimum distance to the nodes in the area could be a new median. The improvement operation is illustrated in Fig. 1.

The crossover must be supported by a proper data structure. In our representation, the object part is not saved as an array of customer indices but rather each median keeps track of assigned customers in a bit array of size n (Fig. 2). This enables to use logical operations in step 3 and so check easily whether a customer is already assigned to any median inherited from the first parent. Although this representation is more memory demanding, it is worth of it because of time savings and regarding the fact that memory is not a problem in present-day computers.

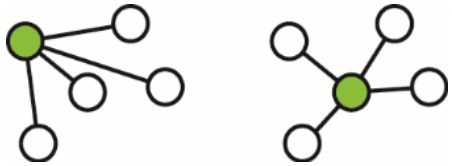
The crossover may produce an inefficient assignment of a customer to a distant median even if the customer is a median itself, e.g. node 7 is serviced from median 2 in the new individual. In such a case another interchange heuristic may improve the assignment. The interchange heuristic tries to assign each customer to another median supposing that the other median is closer and has a sufficient capacity. Since this is a time consuming operation, it is applied to an individual with a small probability.

The mutation operation is implemented in the way described in the previous section, which means that a randomly selected median is replaced by another medial that is not present in the individual. However, such severe change of the group part of the solution causes that current assignment of customers is not valid anymore, and the whole object part of the solution must be recalculated using the MTHG heuristic. As preliminary experiments revealed, the evolution process using the grouping encoding converges quickly to a local minimum. So mutation should be more aggressive than with the standard encoding. That is why we obey the recommendation by Augustín-Blas et al. [26] and increase the mutation probability during the evolution according to the formula

$$p_m(k) = p_{mi} + \frac{k}{TG}\left(p_{mf} - p_{mi}\right), \tag{8}$$

where $p_m(k)$ is the probability of mutation used in generation k, TG stands for the total number of generations, and p_{mi} and p_{mf} are the initial and final values of probability considered, respectively.

The remaining parameters of the GA, namely the size of the population, its initialization, the rules for the parents selection and the population renewal, and the stopping criterion are not affected by the grouping encoding and remain the same as with the standard encoding.



a) b)

Figure 1. Recalculation of a service area in step 5; a) before b) after.

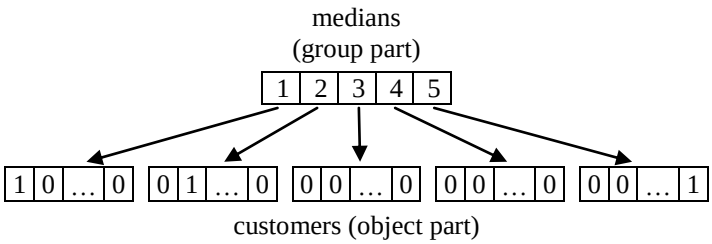


Figure 2. Grouping encoding of an individual.

VI. COMPUTATIONAL EXPERIMENTS

To evaluate the efficiency of the grouping GA we conducted a series of computational experiments using five benchmark instances that were created by Lorena and Senne [20] and are available on the site http://www-usr.inf.ufsm.br/~stefanello/instances/cpmp/. The instances are named *p*3038_600, *p*3038_700, *p*3038_800, *p*3038_900, and *p*3038_1000. Each of them contains 3038 nodes. The number of medians ranges from 600 to 1000. The sets of candidate locations and customers are identical, i.e. every customer can be a median. Since the optimal solutions have not been published so far, our results are compared to the best solutions produced by the IRMA algorithm reported in [17].

The experiments were performed on a personal computer equipped with the Intel Core i7 processor with 1.60 GHz and 8 GB RAM. The main application was implemented in Java language, and the solver Gurobi Optimizer 6.5.0 was used.

The goal of the first experiment was to examine the impact of the encoding on the quality of the final solution. Two versions of the GA were compared, one with the standard encoding and the other with the grouping encoding. Regarding the former version, preliminary experiments revealed that a combination of the one-point and selective crossover operators produces better results than using only one of them does. The selective crossover manages to improve the solution quickly and significantly although after the first 500 generations the decrease becomes negligible. That is why we decided to use the selective crossover during the first 500 generations and then switch to the one-point crossover. We experimented with the mutation rate using medium-sized instances. We found out that the mutation rate proposed by Correa et al. [11] is too low to ensure sufficient diversity. Finally we decided to set the mutation probability to 0.2 as was suggested also by Yaghini et al. [14]. We will denote this implementation as the standard GA in the following text.

TABLE I. PERFORMANCE OF THE STANDARD GA

Instance	IRMA	Standard GA				
		Min	*Avg*	*Max*	*St. dev.*	*Gap* (%)
*p*3038_600	122711.17	132260.56	132757.63	133269.02	296.12	7.78
*p*3038_700	109677.30	119343.34	119852.09	120308.23	318.97	8.81
*p*3038_800	100064.94	109754.79	110351.76	111028.56	373.09	9.68
*p*3038_900	92310.09	101687.85	102162.91	102757.95	347.98	10.16
*p*3038_1000	85854.05	94917.95	95935.70	96620.23	459.15	10.56
Average					359.06	9.40

TABLE II. PERFORMANCE OF THE GROUPING GA

Instance	IRMA	Grouping GA				
		Min	*Avg*	*Max*	*St. dev.*	*Gap* (%)
*p*3038_600	122711.17	128419.30	130373.75	132037.84	1170.46	4.65
*p*3038_700	109677.30	116860.80	118438.83	119579.61	1010.90	6.55
*p*3038_800	100064.94	108971.45	110124.19	112636.65	1074.49	8.90
*p*3038_900	92310.09	101379.05	102559.10	103923.47	783.39	9.82
*p*3038_1000	85854.05	94402.79	98217.97	105977.56	3241.75	9.96
Average					1456.20	7.98

TABLE III. PERFORMANCE OF THE STANDARD GA WITH POST-PROCESSING BY THE SOLVER

Instance	IRMA	Standard GA with post-processing				
		Min	*Avg*	*Max*	*St. dev.*	*Gap* (%)
*p*3038_600	122711.17	125319.03	125928.54	126527.85	395.06	2.13
*p*3038_700	109677.30	111944.25	113768.53	119538.69	2273.09	2.07
*p*3038_800	100064.94	102685.01	105633.34	110188.18	2892.77	2.62
*p*3038_900	92310.09	96331.23	99167.77	101676.07	1909.93	4.36
*p*3038_1000	85854.05	89602.33	92805.12	95585.54	2088.87	4.37
Average					1911.94	3.11

TABLE IV. PERFORMANCE OF THE GROUPING GA WITH POST-PROCESSING BY THE SOLVER

Instance	IRMA	Grouping GA with post-processing				
		Min	*Avg*	*Max*	*St. dev.*	*Gap* (%)
*p*3038_600	122711.17	124618.52	125391.96	125971.25	440.37	1.55
*p*3038_700	109677.30	111556.61	112632.70	113208.68	502.57	1.71
*p*3038_800	100064.94	101968.84	104207.52	110282.76	2134.22	1.90
*p*3038_900	92310.09	96790.38	99545.80	101149.04	1683.88	4.85
*p*3038_1000	85854.05	88800.12	92765.42	95774.98	2386.88	3.43
Average					1429.59	2.69

The latter version of the GA is based on the grouping encoding and the crossover operation described in Section V. The initial mutation probability of p_{mi} = 0.2 and final mutation probability of p_{mf} = 0.5 are used. This version will be referenced as the grouping GA.

The stopping criterion in both implementations is set to 3600 seconds. Ten replications of the GA for each problem instance are evaluated.

Table I presents the results for the standard GA. The first column gives the instance name followed by the best solution found by IRMA. The other columns give the results for the GA: the minimum, average and maximum values of the objective function out of 10 runs, the standard deviation from the average, and the relative gap between the best GA and IRMA solutions. The last row contains the average standard deviation and gap through all instances so that we could compare alternative methods mutually. Table II presents analogous results for the grouping GA.

As one can see, the grouping encoding improves the behaviour of the GA. The best solution found by the grouping GA is always better than the solution found by the standard GA although the difference decreases with increasing p. The average gap between the best GA solution and the published solution over all instances is 9.40% for the standard GA and 7.98% for the grouping GA. However, the grouping encoding causes great variability of the results as the standard deviation suggests. Therefore we was concerned with enhancing the grouping GA by its hybridization.

Recently we have experimented with a mathematical programming solver acting as a post-processing technique after the GA [6]. In the present study we conducted a similar experiment with the aim to examine whether the solution of the grouping GA is a good starting point for the solver and can be further improved.

The only difficulty with the solver is the dimension of the problem instances that have got three thousands of candidates and customers. In such a case the number of assignment variables $\mathbf{x}$ in the mathematical programming model is

enormous. In order to decrease memory and time complexity, we propose a reduction of the solution space. The model is reduced by heuristic elimination of those variables **x** which are less likely to belong to a good or optimal solution. The elimination is based on the assumption that customers will not be served by those centres that are too far away. That is why only those variables x_{ij} are included in the model whose coefficient t_{ij} is less than a predefined threshold. The threshold is defined by the value $\alpha \cdot t^{\max} / \sqrt{p}$ where α is a parameter (α = 1.5 in our experiments) and $t^{\max} = \max\{t_{ij}: i \in I, j \in J\}$.

As in the former tests, the total running time is 1 hour, when the GA runs 45 minutes, and then the solver starts from the best solution and runs 15 minutes. The results are summarised in Tables III and IV.

Both versions of the hybrid GA outperform the pure GA. The average gap between the best solution found by the hybrid GA and the published solution is 3.11% for the standard encoding and 2.69% for the grouping encoding. The hybrid grouping GA shows a more stable behaviour since the average standard deviation is 1429.59 versus 1911.94 of the standard GA. That is why we recommend the combination of the grouping GA with the post-processing solver as a promising solution technique.

VII. Conclusions

The paper deals with two improvements of the genetic algorithm for solving the capacitated *p*-median problem:

1) grouping encoding,
2) hybridization of the grouping GA with a mathematical programming solver.

The grouping encoding enables to preserve service areas of inherited medians while generating the offspring. The result is that only a part of customers need to be re-assigned in the offspring, which saves time. Therefore more solutions can be inspected in the defined running time. The computational experiments with the large-scale instances show that the grouping GA performs approximately twice as many crossover operations as the standard GA in the delimited time. The grouping encoding is better by 1.42% on average compared to the original GA with the traditional encoding. The post-processing solver is able to further improve the objective function value.

Although the grouping GA is inferior compared to the reference decomposition heuristic, we believe that it makes sense to develop population-based heuristics. They may be useful especially in more practical problems when global constraints are added to the basic formulation (1)-(6). One example of such a modification was mentioned in Section III, namely the requirement that the location of medians cannot be changed too much compared to the present location. A decomposition heuristic is not able to cope with such constraints but genetic operations can easily manage them.

In this study we did not make a serious effort in finding the best parameter setting but we have chosen a reasonable set of values. The parameter setting will be the objective of future research.

Acknowledgment

This research was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences under project VEGA 1/0463/16 „Economically efficient charging infrastructure deployment for electric vehicles in smart cities and communities" and by the Slovak Research and Development Agency under project APVV-15-0179 "Reliability of emergency systems on infrastructure with uncertain functionality of critical elements".

References

[1] J. Holmes, F. Williams, and L. Brown, "Facility location under maximum travel restriction: An example using day care facilities," Geographical Analysis, vol. 4, pp. 258–266, 1972.

[2] V. Marianov and D. Serra, "Location problems in the public sector," in Facility Location: Applications and Theory, Z. Drezner, H. W. Hamacher, Eds. 1st ed. Berlin: Springer, 2004, pp. 119–150.

[3] R. C. Menezes and N. D. Pizzolato, "Locating public schools in fast expanding areas: application of the capacitated p-median and maximal covering location models," Pesquisa Operacional, vol. 34, no. 2, pp. 301–317, 2014.

[4] L. Gábrišová and J. Janáček, "Design of capacitated emergency service system," Communications : scientific letters of the University of Žilina, vol. 17, no. 2, pp. 42–48, 2015.

[5] Ľ. Jánošíková and M. Žarnay, "Location of emergency stations as the capacitated p-median problem," in Proceedings of the International Scientific Conference Quantitative Methods in Economics – Multiple Criteria Decision Making XVII, Virt, Slovak Republic, 28 – 30 May 2014, Bratislava: University of Economics, 2014, pp. 116–122.

[6] Ľ. Jánošíková and M. Haviar, "Hybrid Genetic Algorithms for the Capacitated p-median Problem," in SOR ´15 : Proceedings of the 13th International Symposium on Operational Research in Slovenia, Bled, Slovenia, 23 – 25 September 2015, Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2015, pp. 176–181.

[7] S. Felder and H. Brinkmann, "Spatial allocation of emergency medical services: minimising the death rate or providing equal access?" Regional Science and Urban Economics, vol. 32, pp. 27–45, 2002.

[8] M. Boccia, A. Sforza, C. Sterle, and I. Vasilyev, "A cut and branch approach for the capacitated p-median problem based on Fenchel cutting planes," Journal of Mathematical Modelling and Algorithms, vol. 7, no. 1, pp. 43–58, 2008.

[9] A. Ceselli and G. Righini, "A branch and price algorithm for the capacitated p-median problem," Networks, vol. 45, no. 3, pp. 125–142, 2005.

[10] L. A. N. Lorena and E. L. F. Senne, "Local search heuristics for capacitated p-median problems," Networks and Spatial Economics, vol. 3, no. 4, pp. 407–419, 2003.

[11] E. A. Correa, M. T. A. Steiner, A. A. Freitas, and C. Carnieri, "A genetic algorithm for solving a capacitated p-median problem," Numerical Algorithms, vol. 35, pp. 373–388, 2004.

[12] S. Scheuerer and R. Wendolsky, "A scatter search heuristic for the capacitated clustering problem," European Journal of Operational Research, vol. 169, pp. 533–547, 2006.

[13] I. H. Osman and N. Christofides, "Capacitated Clustering Problems by Hybrid Simulated Annealing and Tabu Search," International Transactions in Operational Research, vol. 1, no. 3, pp. 317–336, 1994.

[14] M. Yaghini, J. Lessan, and H. Gholami Mazinan, "An efficient hybrid metaheuristic for capacitated p-median problem," International Journal of Industrial Engineering & Production Research, vol. 21, no. 1, pp. 11–15, 2010.

[15] M. Yaghini, M. Karimi, and M. Rahbar, "A hybrid metaheuristic approach for the capacitated p-median problem," Applied soft computing, vol. 13, pp. 3922–3930, 2013.

[16] J. A. Díaz and E. Fernández, "Hybrid scatter search and path relinking for the capacitated p-median problem," European Journal of Operational Research, vol. 169, pp. 570–585, 2006.

[17] F. Stefanello, O. C. B. de Araújo, and F. M. Müller, "Matheuristics for the capacitated p-median problem," International Transactions in Operational Research, vol. 22, pp. 149–167, 2015.

[18] É. D. Taillard and S. Voß, "POPMUSIC: Partial Optimization Metaheuristic Under Special Intensification Conditions," in Essays and Surveys in Metaheuristics, C. C. Ribeiro and P. Hansen, Eds. New York: Springer, 2002, pp. 613–629.

[19] É. D. Taillard, "Heuristic methods for large centroid clustering problems", Journal of Heuristics, vol. 9, pp. 51–73, 2003.

[20] L. A. N. Lorena and E. L. F. Senne, "A column generation approach to capacitated p-median problems," Computers & Operations Research, vol. 31, pp. 863–876, 2004.

[21] E. Falkenauer, "A new representation and operators for genetic algorithms applied to grouping problems," Evolutionary Computation vol. 2, no. 2, pp. 123–144, 1994.

[22] I. Landa-Torres, J. Del Ser, S. Salcedo-Sanz, S. Gil-Lopez, J. A. Portilla-Figueras, and O. Alonso-Garrido, "A comparative study of two hybrid grouping evolutionary techniques for the capacitated p-median problem," Computers & Operations Research, vol. 39, pp. 2214–2222, 2012.

[23] O. Alp, E. Erkut, and Z. Drezner, "An efficient genetic algorithm for the p-median problem," Annals of Operations Research, vol. 122, pp. 21–42, 2003.

[24] M. Herda, "Combined genetic algorithm for capacitated p-median problem," in CINTI 2015 : 16th IEEE international symposium on Computational intelligence and informatics, Óbuda University, Budapest, Hungary, 2015, pp. 151–154.

[25] S. Martello and P. Toth, Knapsack problems: Algorithms and computer implementations, John Wiley & Sons, 1990.

[26] L. E. Augustín-Blas, S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, and J. A. Portilla-Figueras, "A new grouping genetic algorithm for clustering problems," Expert Systems with Applications, vol. 39, pp. 9695–9703, 2012.

Cyber security assurance approaches for FPGA-based safety platform configuration tool

V. Kharchenko
Director
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

K. Leontiiev
Technical Director
Research and Production Corporation Radiy
Kirovograd, Ukraine

A. Kovalenko
Senior Researcher
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

E. Babeshko
Senior Researcher
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

***Abstract*—This paper represents possible approaches to cyber security assurance for implementation the configuration process of Field Programmable Gates Array (FPGA) based platform for safety critical applications. It also contains results of conducted analysis for secure configuration process in existing platforms, similar in terms of functionality, but based on different technologies. Protection concepts for RadICS Platform Configuration Tool (RPCT) and appropriate license key file are presented. Requirements to key management system are provided considering RPCT features.**

Keywords—FPGA; platform; configuration; security

I. INTRODUCTION

Nowadays, there are plenty of various Instrumentation and Control (I&C) systems under operation thorough the world, covering all the industry branches. Such I&C systems are based on different technologies and principles, performing the most critical functions in the required way. [1]

One of the modern trends in I&C systems design is in application of FPGA technology. Physically, FPGA technology is represented by a chip (complex programmable device), which can be configured to perform a required function by appropriate electronic design (a set of instructions in hardware description language).

Another resource-effective modern trend is in modular design of I&C systems on the basis of some more or less "generic" safety platform certified in required way. One of such platforms in digital I&C platform RadICS, which consists mostly of a set of general-purpose building modules that can be configured and used to implement application-specific functions and systems. The RadICS platform is composed of various standardized modules, each based on the use of FPGA chip(s) as logic solvers. Security assurance process, in turn, is a challenging modern problem. [2-7]

In turn, RPCT is designated to provide a use the opportunity to make changes to user's program which describes application logic (the way it performs specific functions) in RadICS Platform. RPCT also allows to do on-line monitoring and tuning parameters. RPCT has to be protected against its unauthorized use, which is one of the requirements of international standards on cyber security and functional safety. In addition, the RPCT protection is required for:

- Control of RPCT copies use;
- Protection against download of incompatible application logic to a certain version of RadICS Platform (compatibility issue);
- Protection against download of safety-critical application logic into modules that are not designated for this logic;
- Control on changes the accessible tuning parameters in user's application logic.

II. EXISTING SOLUTIONS AND DONGLE TYPE SELECTION CRITERIA

A. Existing solutions: dongles

At present, there are several main types and implementations of dongles. They consist, as a rule, the dongle, connected to the computer USB port, and software (drivers for different operational systems and a module built in the program that is protected). The hardware part of those dongles is developed on FLASH-memory microcircuits, PIC-controllers, and on the specific ASIC-chips

FLASH-memory based dongle is the most simple and can be easily cracked. Its main idea is to write some data or parts of the programming code to dongle before it has been sold, and on the stage of the software validity check to read those data from the dongle. The protection like that can be cracked by the following sequence of actions: understanding the algorithm of information exchange between the computer and the dongle, then information from the dongle FLASH-memory is read and

the corresponding emulator is developed (a driver which replaces authorized driver of the dongle, and – instead of exchange with a real device – it transfers already prepared data to the applied program).

Dongles developed based on PIC- or ASIC-chips are one level more crack resistant. Both those microcircuits are actually controllers, that have processor, some Random Access Memory, FLASH-memory for commands and memory for microprogram storage. Firmware and internal memory are usually protected against external read-in (scanning), therefore, it is quite difficult to develop the hard copy of the dongle.

The main difference between PIC-based and ASIC-based dongles is the following:

- PIC-chips are programmed by the dongle developer (that is he can relatively easy modify algorithm of its operation);
- ASIC-chips are industrial microcircuits (that is, their algorithms cannot be changed after the stage of manufacture).

That explains the fact that ASIC-based dongles are cheaper than those developed based on PIC-chips, the same reason makes them less reliable (if data processing algorithms has been identified in one of ASIC-chips, it is possible to develop emulator dongle for the whole lot).

Besides, some dongles have additional capacities: non-volatile timer (to limit the time of protective program operation), non-volatile memory (to limit the number of installations or program activations), possibility to use the same dongle to protect several application software packages (data coding algorithm selection depending on the protective program identifier).

B. Existing solutions: Flexnet

Flexnet allows controlling the software activation inside the network, using the license server. The system controls the number of software versions activated in the network, limit the product life time, and the number of used versions. The generic protection system structure has four components:

- License manager;
- Vendor daemon;
- License file;
- Software that is being protected.

The license manager is activated on the server and it reads configuration from the license file. According to configuration, software vendor demons then activated, they check the corresponding license lines. Software requests license server each time when it has to perform important function and also according to timer approximately once in 120 seconds – so called heartbeat. Server location is identified according to specific variable. Its value is a path to the license file or of "port@servername" type.

C. Security methods used by other vendors

Table 1 below summarizes results of conducted analysis.

TABLE I. SUMMARY ON PROTECTION OF EXISTING PLATFORMS

Vendor	Platform	Protection technologies in use	Description
Emerson	Ovation	Safenet Sentinel Proprietary	Hard USB-dongle plus its own protection implementation
Siemens	SPPA-T3000	Flexnet	Standard solution with reference to equipment wothout a separate dongle
Westron	Vulkan-M	WEK	Dongle (in-house implementation)

D. Protection concept for RPCT

The RPCT protection concept is based on using personal keys, which are actually text license files where every line is protected by hash. The following key types are supposed to be available; they determine RPCT functionality (its options):

- Developer key;
- Integrator key;
- Engineer key;
- Operator key.

The following roles are identified to describe RPCT user possibilities:

- Developer of a platform/systems;
- Integrator;
- Engineer;
- Operator.

All four types of users can activate RPCT (option "Start RPCT" in Fig. 1), in this case we have read and analysis of a license file (option "Check license file" in Fig. 1). Depending on the type of license, a user can be given the privileges of designer, integrator, engineer or operator.

It has to be noted that a user of "developer type" is the most privileged. An user of this type can only be assigned by an employer of the RadICS platform vendor and he can be used within the vendor's facility.

A designer (Platform vendor) has access to all options: developing of new application logic, changing existing application logic, loading of application logic, changing tuning parameters, loading of firmware (platform logic), and runtime monitoring (options "Design application logic", "Download application logic", "Create HW configuration", "Download HW configuration", "Perform runtime monitoring" and "Change tuning parameters" in Fig. 1).

Integrator has access to development and downloading application logic into the Platform (options "Design application logic", "Download application logic", "Create HW

configuration", "Download HW configuration", "Perform runtime monitoring" and "Change tuning parameters" in Fig. 1). Engineer has access to do runtime monitoring and changing tuning parameters (options "Perform runtime monitoring" and "Change tuning parameters" in Fig. 1).

Operator has access to runtime monitoring (see Fig. 1 where use case diagram is presented).

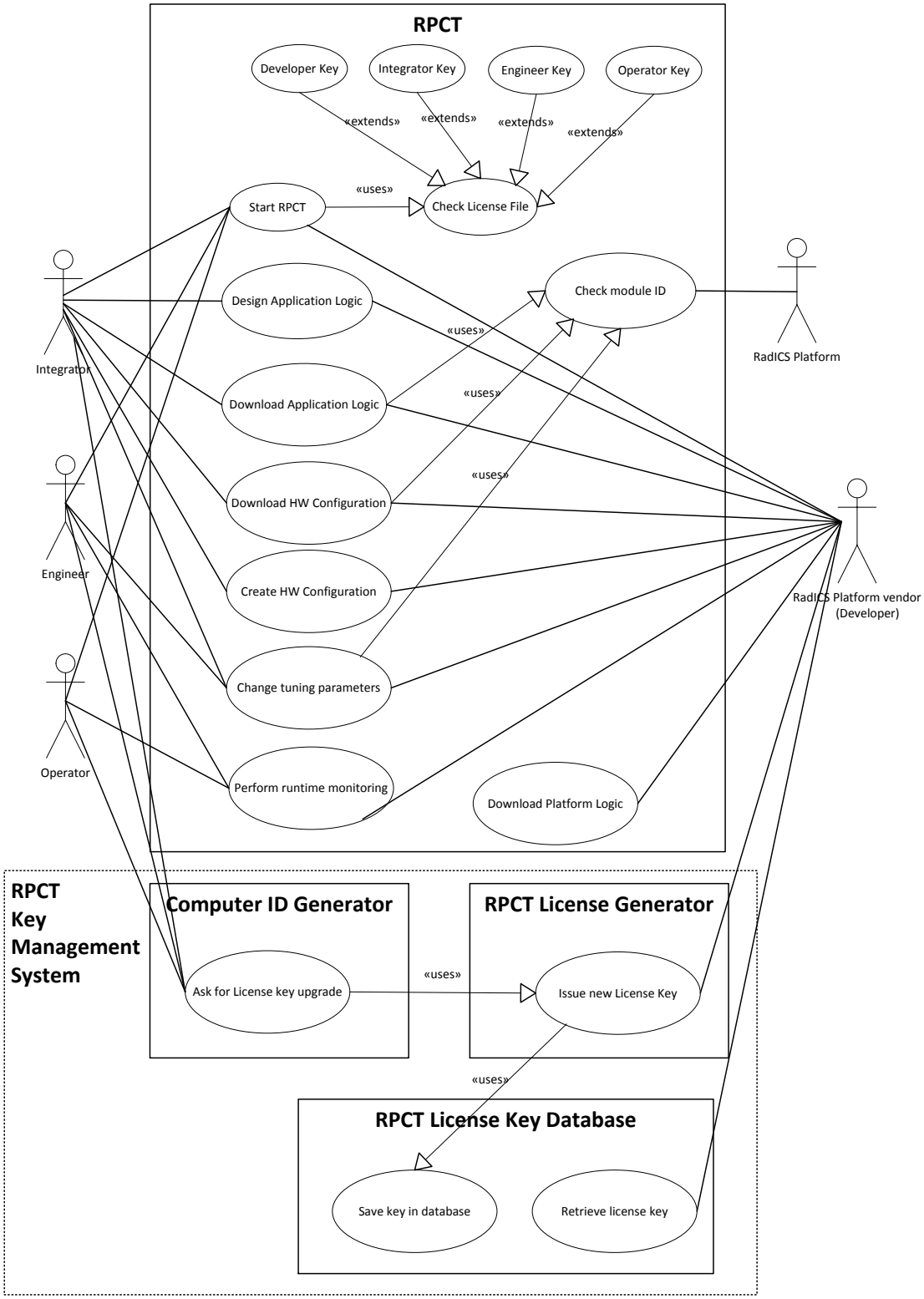


Figure 1. Use case diagram for RPCT.

Activation the RPCT on user's computer starts reading the license file, which contains information on modules, operation with those is supported by this RPCT version. The license file also has information about options allowed by the license.

When the Platform is transferred to the configuration mode, the module identifier and the version number of firmware being read from the service flash memory automatically. Such information then automatically transferred into all units, which perform its processing.

When a use tries to download configuration or application logic to the Platform, RPCT compares module identifier and the version of its firmware that are obtained as the result of information exchange with the Platform with identifies read from the license file. The downloading process is allowed only in a case if RPCT licensed functionality corresponds to platform identification information.

During the application logic download, module identifiers (actor "RadICS Platform" in Fig. 1) are compared with identifiers that are obtained from the license file. If comparison is successful, the application logic load is authorized. In other cases the application downloading is not permitted.

A third-party RPCT user (integrator, operator or engineer) can ask for a new license key and provide a vendor with a new computer identifier (option "Ask for License key upgrade" in Fig. 1). The vendor, using special software, can generate a new license file (option "Issue new License key" in Fig. 1), which is stored in RPCT License Key Database (option "Save key in database" in Fig. 1) and in case of necessity it can be retrieved from the database (option "Retrieve license key" in Fig. 1).

III. License Files Concept

The analysis of the existing types of protection that are discussed above enables to introduce the following protection concept: a user is provided with license key, which is a text file where every line is protected by hash.

The license file has the following information:

- Permitted version number of RPCT to start;
- License type;
- Identifier of the computer, which has permission for this RPCT copy start;
- License expiration date;
- List of modules serial numbers and their firmware versions, which are permitted to use with the given version of RPCT;
- List of permitted module configurations.

IV. Requirements to Key Management System

The key management system is required to provide license files management functions (keys) and their metadata at the vendor's site only in case of inquiry from authorized users and with all applicable constrains.

The following metadata can be associated with license keys:

- Key Label: a text line that has a set of key descriptors ;
- Key Identifier: the element that is used by the key management system to choose a specific key from the set (primary key in the database);
- Owner Identifier: the identifier of user's computer which is going to use RPCT;
- Schemes/Modes of Operation: the identifier of permitted schemes/modes of operation used for cryptography performed by the key;
- Key Type;
- Key Access Control List: has data on Platform's module identifiers and their hardware configurations.

In case the system user needs new or additional key, he has to ask a vendor to create additional key. The latter, then, has to control that the key type and other parameters (for example, metadata) correspond to the user type and to the systems he acquired.

The main information need to be obtained from the user to develop the key is identifier of the computer, which is going to be used to work with RPCT. To obtain identifier, a user launches at a target computer a special software, which performs identifier generation according to the built-in algorithm.

Based on the information obtained from the user a vendor can generate license, using special software (license generator). The license in this case after being generated (with corresponding metadata) is automatically transferred to the license database, which is inseparable part of key generation process.

While keys and metadata are being transferred, they have to be protected by providing confidentiality and access control. To provide confidentiality one can use either physical protection or cryptography. To transfer the key it is mandatory to provide physical protection by using a reliable courier or physically protected channel.

Conclusions

Cyber security assurance process for FPGA-based safety platform configuration tool is an open modern problem and it greatly depends on underlying technologies. The results of performed analysis revealed additional issues that have to be resolved in the scope of initial problem, including the following:

- Necessity in additional module identifier storage in service FLASH memory;
- Introduction and support of additional entity – RPCT license key file;

- Establishment and maintenance of dedicated and protected key management system.

REFERENCES

[1] V. Kharchenko, M. Yastrebenetsky (ed.) Nuclear Power Plant Instrumentation and Control Systems for Safety and Security. – IGI Global, 2014. - 450 p.

[2] CJ Clark, FPGA Security, FPGA Configuration, FPGA Bitstream, FPGA Authentication. Business Considerations for Systems with RAM-Based FPGA Configuration, Intellitech, 2009.

[3] S. Drimer, Security for volatile FPGAs. Technical Report UCAM-CL-TR-763, University of Cambridge, 2009.

[4] M. Majzoobi, F. Koushanfar, and M. Potkonjak. FPGA-oriented Security. Introduction to Hardware Security and Trust, Springer, 2011.

[5] V. Kharchenko, O. Illiashenko, A. Kovalenko, V. Sklyar, and A. Boyarchuk, "Security Informed Safety Assessment of NPP I&C Systems: GAP-IMECA Technique", Proceedings of 22nd International Conference on Nuclear Engineering, Volume 3: Next Generation Reactors and Advanced Reactors, Nuclear Safety and Security, 2014.

[6] V. Kharchenko, A. Kovalenko, O. Siora, and V. Sklyar, "Security assessment of FPGA-based safety-critical systems: US NRC requirements context", Proceedings of 2015 International Conference on Information and Digital Technologies, pp.132-138.

[7] V. Kharchenko, A. Kovalenko, and V. Sklyar, "Secure environment establishment for FPGA-based safety-critical systems", Proceedings of East-West Design & Test Symposium (EWDTS) 2015 IEEE, pp. 88-92.

Modeling of Industrial FPGA-based Controllers with ForSyDe

A. Panarin
Lead Designer
Research and Production Corporation Radiy
Kirovograd, Ukraine

V. Kharchenko
Director
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

V. Sklyar
Professor
National Aerospace University "KhAI"
Kharkiv, Ukraine

A. Kovalenko
Senior Researcher
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

E. Babeshko
Senior Researcher
Centre for Safety Infrastructure-Oriented Research and Analysis
Kharkiv, Ukraine

***Abstract*—The paper includes results of the theoretical research and practical application of soft-processors testing, in particular Nios compiler of Altera SoPC Builder core. The model is based on Model-Based Testing concept. ForSyDe programming formal language is used as instrument for the development of soft-processor reference program model. Stages of model development with the follow-up analysis and comparison of performance results are introduced.**

Keywords—modeling; Model-Based Testing; soft-processor; ForSyDe; Altera SoPC Builder; Nios

I. INTRODUCTION

When the soft-processors software is developed for Field Programmable Gates Array (FPGA) technology, the issue of the final product verification and validation comes forward, as by doing those tests one can see if equipment is robust and reliable. The issue was explored by many researchers [1, 2, 3], however, no result is unique and can be applied to each specific case. This paper presents the discussion of the issue in terms of its practical application.

When the program code development process is finished, we come to the phase of error detection [1]. Of course, the designer's gross errors will be reflected in the errors of a compiler, but the system stability is hidden not only in the proper syntax of the written code. The hidden errors are difficult to detect, and they can be revealed at different stages. One of the factors that can make an impact on the program development is a human factor.

Therefore, we need tools that could identify even the smallest deviation in the program behavior, and the identification process should be maximally automatic. Many compilers are known to have means to emulate the program operation, even with self-diagnostics, but those are particular cases that are not the focus of this paper. We are trying to find more specific approach to testing that could be applicable to many compilers or development tools for FPGA program development. One of the most relevant approaches that can cover mentioned above requirements is Model-Based Testing.

The purpose of the research is to analyze how a specific tool can be used for implementation of the concept on the development of alternative IP-cores, which allow to perform independent verification, and diverse implementation for industrial FPGA-based controllers.

II. ANALYSIS OF APPROACHES AND TOOLS APPLICABLE TO TESTING OF FPGA-BASED CONTROLLERS

A. Model-Based Testing

Model-Based Testing is one of the approaches to testing, which implies manual test development, automatic test development or model-based test generation, when tests are generated based on the testing system behavior model or situation model, which relates to system operation. [2, 3]

Behavior model formalizes requirements to the system, that is, it describes what external impacts in which situations are tolerable, and how system is expected to react and operate under those impacts.

Situation model formalizes the structure of possible testing situations, which consider external impacts, system state and system environment; the model identifies different situation

types and their importance in terms of quality control. Usually, the model like that has the final set of equivalent situation types, meaning that while testing it is sufficient to check system's operation at least in one situation of each type. Also, to specify a situation model, a final set of elementary events is often used, where every situation corresponds to a certain set of those events.

Situation model is used to solve two closely connected tasks: to identify the criterion of adequacy or testing coverage and to identify testing coverage number metrics. It should be noted that the testing coverage here means meeting certain criterion (let's say having 80% of code coverage), but not the exhaustive try of all practically possible situations. Testing coverage criterion defines features for the test set, which provides a full scope of the testing system's behaviors while solving certain number of tasks. Test coverage metrics are given in percent of the checked types of different situations. [4]

Development of tests based on models consists in the development of the testing system model of behavior, model of situations which show the main priorities and risks of the project; more often the model uses the structural elements of the behavior model, then a set of tests is built (generated automatically, developed manually, or developed using tools). Those tests check the difference between the real behavior of the testing system and its behavior model. The testing set is developed so that it meets the criterion of the testing coverage, which is specified by the situation model.

B. ForSyDe tool

The most obvious method is developing the model with Very high speed integrated circuits Hardware Description Language (VHDL). This method is practically perfect, and can be widely used, but if the project development time is taken into consideration, this method turns out ineffective. The number of scientific researches shows that the most optimal tool is ForSyDe, which provides application of functional programming paradigm. [5]

C. ForSyDe as a functional programming in Haskell

When the functional programming paradigm is used, the computing process is interpreted as the calculus of functions values in their mathematical understanding, which assumes constancy of values when the function is caused by the same arguments. The approach like this is obviously the most adequate way to describe calculations.

However, in practice, the most wide spread is the paradigm of imperative programming, which describes computing process as the sequence of state change, which is connected with the change of variables. In this case, the function is considered as the subprogram variation (procedure that brings back parameters).

The foundation of the functional programming theory is the works of American mathematicians Curry Haskell in combinatory logic and Alonzo Church in formal λ-calculations. Based on those research works John McCarthy in 1958 developed the first functional programming language LISP. Among the most noticeable achievements in the field of functional languages implementation, there are two systems of industrial level with a wide user environment and support – Ericsson Erlang and Harlequin ML Works. In addition, the Haskell language is well known in academic circles, and the last standard for the language is Haskell-98.

As the functional programming advantages, the following features are mentioned: smartness, effectiveness, determinacy, etc. The disadvantages are explained by the fact that functional languages are not popular enough, and, therefore, they are not provided by sufficient compatibility, transformation, tool support, etc.

An example of how to use functional programming for digital systems modeling is the tool called ForSyDe (Formal System Design), which was developed in the Royal Institute of Technology (KTH, Stockholm, Sweden). [6]

The main concept of ForSyDe is in the resolution of the "gap" problem by developing methodology of system design with high level of abstraction, which is based on the so called transformational design refinement approach. In this case, the system is modeled as a net of interactive processes, connected by signals. The processes perform calculations by transforming input signals into output signals. Thus, communication functions and calculations functions are isolated. Using abstract time makes it possible to implement several calculation models, the main is simultaneous model. [7]

Considering technological maturity of ForSyDe and fundamental nature of theoretical concepts that were used as its basis, it looks reasonable to use it for alternative development and verification of IP-Cores for System on Chip (SoPC), which implement functions of parallel digital systems. Because functional programming provides convenient view of digital parallel systems it is possible to see SoPC or its part (built-in IP-Core) in diverse execution with the help of ForSyDe. If the results of main and diverse SoPC functioning converge, the results of verification will be recognized as successful. SoPC diverse presentation can be used in working projects to provide protection against common cause failure and to strengthen so called "defence in depth". The approach like this would make it possible to correct mistakes in SoPC project at the initial stages of its development.

The purpose of the research is to analyze how ForSyDe can be used for implementation of the concept on the development of alternative IP-cores, which allow to perform independent verification, and diverse implementation for SoPC. In the research, FPGA-based industrial controllers are treated as SoPC.

D. The gap between specification and heterogeneous SoPC implementation

High level of abstraction at the development of digital SoPCs is the answer to the market challenge, the result of it is a maximum reduction of time to develop IT products.

State-of-the-art SoPC are multicore heterogeneous systems that also include microprocessors emulators. The example of those is Intellectual Property Core – IP-Core Nios made by company Altera and Cortex made by company Actel. Those cores are designed to be implemented in FPGA.

SoPC are developed and verified by imperative programming means, as the standard sequence program system. In this case the focus is not a physical implementation, which is actually a parallel digital system.

At the same time the requirements specification to SoPC, as a rule, has "parallel" form, including requirements to processing/generation of input/output signals, requirements to internal and external communication, to diagnostics, etc. Therefore, there is a 'gap" between the source specification requirements and development environment, and also between development environment and final implementation. Each of those gaps requires considerable information transformation, when SoPC goes from one form to the other, which creates a source for defects.

In principle, such approach could be justified by the requirements to the high level of abstraction, which is described at the rapid development of complex systems.

However, the imperfectness of integrated development environments (IDEs) of SoPC causes hardware in-circuit problems (synchronization, "needles", internal and external clocking). As a rule, those problems depend on the specific implementation of SoPC, however, manufacture's development guides do not provide description of approaches with mentioned problems. This leads to degradation of effectiveness and quality of development processes and products.

One of the possible approaches to solve the "gap" problem in SoPC with built-in processor cores is to use for those cores [6] functional programming as the way of analytic description and subsequent implementation.

III. An Operation Analysis of a FPGA-based Discrete Signals Input Module

As a model project, we choose discrete signals input module. Tasks to process discrete signals are typical for industrial control systems. In particular, the module under discussion is a part of digital I&C Platform RadICS that is used for NPP safety systems [8]. The main designation of the module is reading digital data from the input port and their transfer to the logic control module, it also transfers diagnostic information to the diagnostic module.

Altera FPGA chips are used as a programmable component. FPGA electronic designs are developed in IDE Altera Quartus II, using processor which is based on the Altera Nios core. Nios microprocessor belongs to the microprocessors class for built-in application, that is, its main task is to work in a real time mode. Using Nios in SoPC gives additional possibility – to implement commands and hardware co-processors that are picked up and developed by the user.

Fig. 1 shows brief algorithm of the program, which is written with the programming language C for a compiler from Nios core Altera Sopc Builder. The program is designed for a continuous work; therefore, it is performed as a non-stop cycle. Program operation completion or its discard can only be done if FPGA circuit is power off. The most important program operations are performed with 10 ms periodicity. The information transfer to the external units is done upon receiving data transfer request.

Algorithm has several steps (see Fig. 1):

- Begin – at the program initiation, on default all the variables are initialized in a certain state. All input/output devices are also initialized;
- Rx Request from LCM – request wait to send data to logic control module (Tx Message to LCM). Upon message completion the transfer starts initializing with the help of a protocol UART (Universal Asynchronous Receiver Transmitter), using DMA (Direct Memory Access), therefore, the program needs to follow only DMA ready flags;
- Rx Request from LCM_d – wait and execution request to send data to the diagnostics module is performed in the same way;
- 10 ms timer – overflow flag wait 10 ms timer to transfer the program to read updated data from external input ports to the global variables.

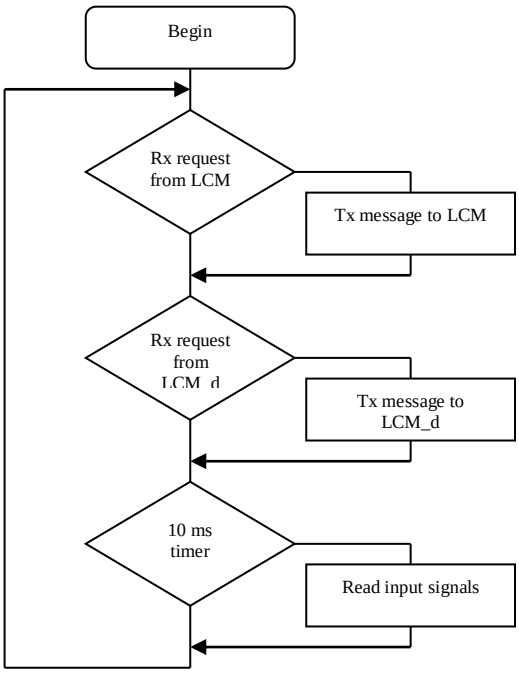


Figure 1. Flowchart for discrete input module.

IV. An Approach to Modelling of a FPGA-based Discrete Signals Processing Module in ForSyDe

The general methodology of modeling digital systems in the ForSyDe environment includes the following steps:

- Introducing system as modules that process input/output signals;
- Analyzing combinations of input/output signals and forming the functions of signal processing modules;

- Developing system block diagram, that includes processes and signals;
- Developing program with Haskell; program describes functions (processes) of system block diagram;
- Compiling program text;
- Forming input data array and executing modeling in ForSyDe environment;
- Code convertion from Haskell to VHDL;
- Analyzing received code in VHDL in the design environment for FPGA electronic designs and analyzing the compilation correctness;
- Studying digital device in the design environment of FPGA electronic designs and/or as a part of FPGA microcircuit with implemented electronic design.

Let us analyze this approach in its application to the discrete signals input module.

The first step to develop program model in ForSyDe environment is to introduce system block-diagram (describing system operation), (see Fig. 1) as input and output signals and modules to process them. The system has 4 main input signals (Input signals, 10 ms, Rx request LCM_d, Rx request LCM_d) and 2 outgoing signals (Tx message to LCM, Tx message to LCM_d), their processing is performed in System module.

All the input signals have a certain logic state, that makes impact on their processing and, therefore, on the output signals. The diagram of the model in ForSyDe environment is characterized by the following features:

- System input signals –make impact on the processing of data that comes from external sources or system modules: input – port to read external digital data of the system; req and req_d – request ports to transfer data to external units of the system;
- Emulation module for 10 ms timer – allows to develop the timer, that uses system frequency of clock pulse;
- Data processing module – with the pulse overflow 10 ms timer the data is read from the external input ports to local variables;
- Output, Output_d – modules that wait for the byte accepting request to external system modules. Its availability makes it possible to send data package to external system unit;
- Outgoing system signals – those are signals resulted out of the input data processing by the system, with two transceivers, each of those has 2 signals at the output (data port output, output_d and availability ports valid and valid_d).

The program written in Haskell can be converted into the code written in VHDL with the help of ForSyDe convertor, built in the system. Then the code can be analyzed in integral development environment to develop FPGA electronic designs. The program written in VHDL can be tested for work efficiency and then implemented in FPGA chip.

A simplified system model has been chosen for VHDL conversion. After the conversion procedure was over, the electronic design of the system program model was opened with the help of RTL viewer utility, built in the Quartus II development environment. In addition, digital device circuits were analyzed.

For a more detailed circuit discussion, the functions are presented at a more detailed level. The utilities that are part of ForSyDe make it possible to connect external program module ModelSim which is used to do project modeling with different input parameters. Thus, there is a possibility to do comparative analysis of the programs, which are activated by the compiler ForSyDe, and VHDL code emulator.

Once modeling has been completed, the console display shows output data that is received when the ModelSim program processed input signals. To compare the accuracy of algorithm processing the ForSyDe project was initiated with the same input data. The comparison of the obtained data shows absolute similarity and that testifies the correctness of the VHDL code compilation and correctness of its processing by Model Sim program.

The development environment Quartus II has built-in VHDL code simulator, which helps see the program work within a certain time period at the level of input/output signals. The utility enables to do simulation of the project, compiled and embedded in FPGA with a connected oscillograph.

The analysis results proved full fit of all three versions of the studied digital system that were introduced. Also the analysis results demonstrate that all the stages of the suggested methodology to model digital systems in ForSyDe environment were performed with full compliance with the input data.

V. Performance of the Metrics for Developed Projects

Performance comparison of the programs written in C and Haskell, demonstrates identity of programs execution. According to the definition, software metrics is a measure that allows to receive numerical value of some software property or its specifications. In other words, we need to receive a value, which can provide unambiguous answer whether the cores under discussion are diverse and what is their diverse value.

The analysis itself is preceded by identifying assessment method. There is no doubt that in this situation the most appropriate method is black box with fault injection, but it would require certain statistics, which, for its turn, creates necessity to have the procedures of defect control, program execution result comparison and statistics acquisition automated. This is a very labor-intensive process, which is to be considered.

A more simplified assessment method is a method of additive convolution. This method does not provide accurate results; it is based on the subjective opinion of the expert. This characteristic of the method allows some freedom in metrics

calculation accuracy, but provides the possibility to obtain certain results. [5]

The task sounds as follows: there are two IP-cores, which are to be analyzed to see their diversity. For the cores comparison, we picked up n assessment criteria, identified the diversity value assessment d_i and chose weighting factor L_i in scale of the system for each i assessment criterion. The following limits are fair:

$$d_i \in [0,1], \quad \sum_{i=1}^{n} L_i = 1. \tag{1}$$

Following the mentioned above rules, the expert made cores assessment, along with calculation of the weighting factor in compliance with (1), and provided results in Table 1.

TABLE I. DISCREPANCY ANALYSIS OF THE STUDIED DESIGNS ARCHITECTURE

#	Parameter	Nios	ForSyDe	Weighting factor	Diversity value
1	Input data	signal	Array	0,2	1
2	Program execution	sequential	parallel	0,3	1
3	Use of bus	yes	no	0,1	1
4	Use of FPGA	yes	yes	0,1	0
5	Processor capacity	16 bites	no	0,2	1
6	Use of DMA	yes	no	0,1	1

The given below formula allows to calculate core diversity factor K_d:

$$K_d = \sum_{i=1}^{n=6} L_i \cdot d_i \,. \tag{2}$$

The calculations resulted in obtaining diversity factor for Nios soft-processor cores and its ForSyDe model, which is, in according to (2), $K_d = 0,9$.

Thus, diversity has been proved. Part of the conclusion, though, is certain complexity in the development of the similar algorithm using principally different languages, as similar functions are implemented by different methods.

A. Performance comparison of the developed system and standard IP-Core

One of the focuses of the research was to compare the developed model and the program that was developed by the standard method. As it was mentioned above, the standard method to develop FPGA-based discrete input module is programming in C language. The code of C language is developed as a part of Altera Nios microprocessor emulator and is a part of FPGA electronic design in Altera Quartus II development environment.

Thus, language C was used to develop the program that corresponds to the algorithm introduced at Fig. 1. The performance comparison of the programs written in C and Haskell languages shows high similarity. Nevertheless, it has to be mentioned that there exist certain complexity for the process of development of the same algorithms in different languages, as their similar functions are provided by different methods. This fact still allows to consider the process of modeling successful.

CONCLUSIONS

The paper discusses the possibility to implement digital systems, traditionally developed in imperative languages in processors environment, in the functional programming languages.

To perform the research the methodology for modeling digital systems in ForSyDe environment is suggested. The modeling was used for FPGA-based discrete input module and the following versions of digital system implementation were studied:

- Description in Haskell language in ForSyDe environment – designated only for modeling, represents input data to obtain the second implementation version;
- Electronic design in VHDL in Altera Quartus environment;
- Code in C language, designed to use in Nios processor emulator environment; the processor is a part of electronic FPGA design.

The study proves similarity of all three versions of digital system implementation.

The research results are advised for resolution of the following practical tasks:

- SoPC functions implementation by alternative way and testing results comparison can be used for independent verification of such SoPC as parts of systems important for safety at critical facilities;
- Obtained alternative implementations of SoPC can be used to develop diverse systems at safety critical facilities. Those implementations strengthen defence-in-depth and decrease the probability of common cause failure.

Further research can be developed in the following directions:

- Formal evaluation of the discrepancy between versions developed with different technologies; such evaluation can be performed, for example, by the fault injection method;
- Study of diverse architectures of electronic designs, that include diverse IP-Cores (diverse soft processors, diverse data transfer buses, diverse structures that perform logic control functions, etc.).

REFERENCES

[1] D. Zheng, Y. Wang, Z. Xueyi, "The methods of FPGA software verification," *2011 IEEE International Conference on Computer Science and Automation Engineering*, Shanghai, 2011, pp. 86-89.

[2] L. Guan FPGA-based DSP System Verification. FPGA-based Digital Convolution for Wireless Applications, Springer, 2017, pp. 129-151.

[3] MC. Jakobs, Platzner M., Wehrheim H., Wiersema T. Integrating Software and Hardware Verification. In: Albert E., Sekerinski E. (eds) Integrated Formal Methods. IFM 2014. Lecture Notes in Computer Science, vol 8739. Springer, Cham.

[4] H. Zhu, P. Hall, J. May, Software Unit Test Coverage and Adequacy. ACM Computing Surveys, 1997.

[5] I. Sander, System Modeling and Design Refinement in ForSyDe. PhD thesis, Stockholm / Royal Institute of Technology, 2003.

[6] A.J. Field, P.G. Harrison. Functional Programming. Addison-Wesley, 1988.

[7] T. Raudvere, I. Sander, A. Jantsch, "Application and verification of local non-semantic-preserving transformations in system design," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, pp. 1091-1103, 2008.

[8] RPC Radiy FPGA-Based Safety Controller (FSC) RadICS. - Available at http://www.exida.com/SAEL/rpc-radiy-fpga-based-safety-controller-fsc-radics

How to Model Operation Threats and Climate-Weather Hazards Influence on Critical Infrastructure Safety
An Overall Approach

Krzysztof Kołowrocki
Joanna Soszyńska-Budny
Mathematics Department
Gdynia Maritime University
Gdynia, Poland

Abstract—**The paper is devoted to a comprehensive modelling of the influence of the operation process and the climate-weather change process on the safety of a critical infrastructure. Particular models of critical infrastructure safety influenced by its inside among its components and subsystems dependences and by its outside operating environment threats and climate-weather hazards are created and a reasonable perspective for their further developments and applications is marked out. A set of safety indicators for a critical infrastructure is proposed and the simplified procedures of their determination in the case of the created models of critical infrastructure safety are proposed and illustrated.**

Keywords—critical infrastructure; climate change; safety; climate impact; safety indicator;resilience indicator; modelling; prediction; optimisation

I. INTRODUCTION

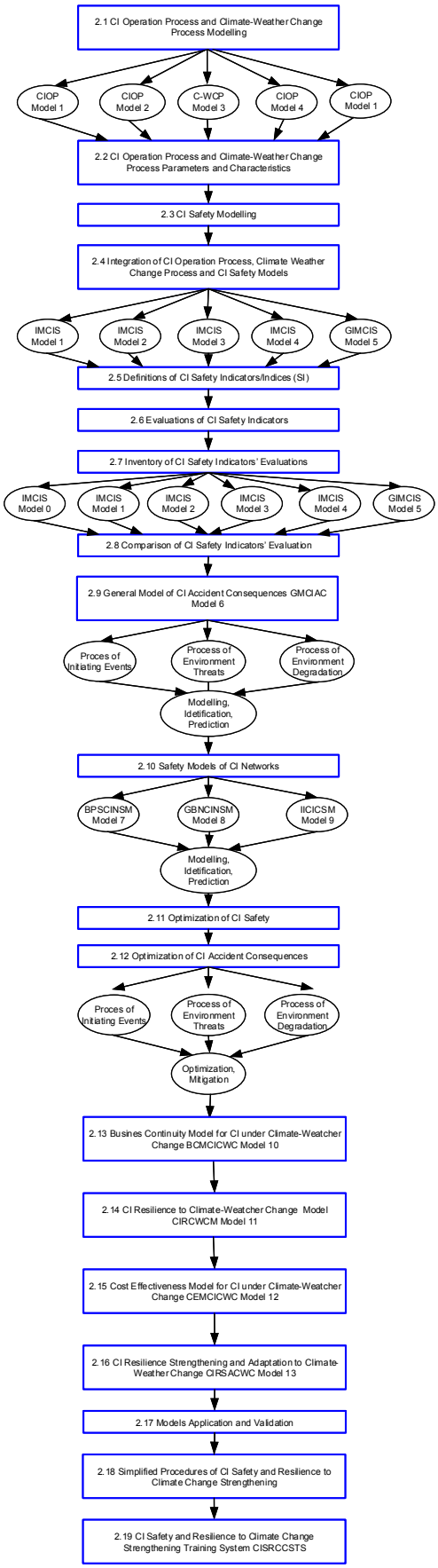
Most real complex critical infrastructures are strongly influenced by changing in time their operation conditions and the climate-weather conditions at their operating areas [1]. The time dependent interactions between the operation process related to climate-weather change process and the critical infrastructure safety structure and its components/assets safety states changing are evident features of most real critical infrastructures [4]-[5], [15]-[17]. The common critical infrastructure safety and resilience, its operation process and the climate-weather change process at its operating area analysis is of great value in the industrial practice because of often negative impacts of operating environment threats and extreme weather hazards on the critical infrastructure safety and resilience.

In the critical infrastructure safety analysis, the determination of its safety function and its risk function which graph corresponds to the fragility curve [10] and other safety and resilience indicators are crucial indices for its operators and users. These and other safety and resilience indicators can be obtained using a probabilistic approach to modelling of operation threats and extreme weather hazards impact on critical infrastructure safety [4]-[5]. To achieve the main aims of EU-CIRCLE project formulated in WP2, WP3 and WP4 Tasks, a general probabilistic approach to safety and resilience analysis of critical infrastructure is proposed. This approach was preliminarily introduced in the project deliverable D1.1 EU-CIRCLE Taxonomy [1], partly described in D3.4 and almost completely developed in the project deliverables D3.3-Part1 [4] and D3.3-Part 2 [5] for the Task3.3. Critical Infrastructure Interconnections, Task3.4. Impact Assessment Models and Task3.5. Holistic Risk assessment Propagation Model. Details of this development are given in the papers published in special issues on EU-CIRCLE project of Journal of Polish Safety and Reliability Association – JPSRA 2017, [11]-[12]. Moreover, completely developed details of the proposed probabilistic methodology concerned with critical infrastructure operation and climate-whether change modelling, prediction and data processing are given in [13] and partly developed details of this methodology concerned with critical infrastructure resilience and adaptation to climate change are given in [14].

To make the effort of the proposed approach well organized, the scheme of a general approach to safety and resilience analysis of critical infrastructure giving subsequent steps in the project research activity is presented in the next section.

II. GENERAL APPROACH TO OPERATION AND CLIMATE-WEATHER INFLUENCE ON CRITICAL INFRASTRUCTURE SAFETY AND RESILIENCE MODELLING, IDENTIFICATIION, PREDICTION AND OPTRIMISATION

An original and innovative general approach to operation process and climate-weather change process influence on critical infrastructure safety and resilience modelling and analysis has been created in the project report D3.3 [4]-[5] and in the papers included in [11]-[14].
The scheme of this approach is presented in Figure below created in [11].



In this scheme:

- the content of scheme items 2.1-2.10 is concerned with the critical infrastructure operation process and climate-weather change process at its operating area modelling, critical infrastructure safety modelling, with the integration of the designed models into a general joint models of critical infrastructure safety related to operation and climate-weather change processes and with the modelling safety of critical infrastructure networks and their accidents consequences (WP2 & WP3);
- the content of scheme items 2.11-2.16 is concerned with the critical infrastructure safety and its accident consequences optimization, business continuity, resilience and cost effectiveness modelling (WP4);
- the content of scheme item 2.17 is concerned with the case studies and the proposed models application and validation (WP6) partly developed in deliverable D6.4 [6];
- the content of scheme item 2.18 is concerned with the development of simplified procedures of critical infrastructure safety and resilience to operation and climate-weather change strengthening (WP5 & WP7);
- the content of scheme item 2.19 is directly concerned with the critical infrastructure safety and resilience to operation and climate-weather change strengthening training system (WP8).

In the first part of the proposed approach (the scheme items 2.1-2.10), starting from a simplest pure safety Model 0 without considering outside impacts on critical infrastructure defined as a multistate ageing system, the critical infrastructure and its components/assets safety functions and other safety indices like mean values and variances of lifetimes in the safety state subsets and in the particular safety states, a critical infrastructure risk function, its fragility curve, the moment of exceeding the critical safety state and intensities of ageing/degrading are defined. System components' independences are ignored and practically important modifications assuming inside critical infrastructure assets' dependences are introduced and developed. Moreover, the way of modification all considered critical infrastructure safety models into more general models for safety analysis and prediction of critical infrastructure networks and networks of critical infrastructure networks and for networks of critical infrastructure networks cascading effect analysing is proposed as well. In addition, a general approach to modelling, identification and prediction of critical infrastructure accident consequences is proposed.

Details of theoretical backgrounds of this first part of the proposed approach are given in [4] and [11].

In the second part of the proposed approach, Model 0 is joined with the critical infrastructure operation process model to create Model 1 devoted to safety modelling and prediction of critical infrastructure defined as a complex system in its operating environment that significant features are inside-system dependencies and outside-system dependencies. This is a general safety analytical model of a critical infrastructure related to its operation process, linking its multistate safety model and its operation process model and considering

variable at different operation states its safety structure and its components' safety parameters. In this model, additional safety indices typical for the critical infrastructure related to its varying in time safety structures and its components' safety parameters caused by its operation process are introduced extending the Model 0 set of safety indicators by the components and critical infrastructure conditional intensities of ageing at particular operation states and conditional and unconditional coefficients of the operation process impact on intensities of ageing. A slight generalization of Model 1 is Model 2 devoted to safety of a critical infrastructure related to its operation process including operating environment threats. It is the integrated general model of critical infrastructure linking its multistate safety model and the model of its operation process including operating environment threats and considering variable at the different operation states safety structures and their components safety parameters. Other practically significant discussed in this report critical infrastructure safety indicators are the critical infrastructure and its components intensities of degradation and the coefficients of operation process including operating environment threats influence on the critical infrastructure and its components intensities of degradation. Next, a general safety analytical Model 3 of critical infrastructure safety related to the climate-weather change process in its operating area is proposed. It is the integrated model of critical infrastructure safety, linking its multistate safety model and the model of the climate-weather change process at its operating area, considering variable at the different climate-weather states and impacted by them system components safety parameters. The conditional safety functions at the climate-weather particular states, the unconditional safety function and the risk function of the critical infrastructure at changing in time climate-weather conditions are defined. Other, practically significant, critical infrastructure safety indices introduced in the model are its mean lifetime up to the exceeding a critical safety state, the moment when its risk function value exceeds the acceptable safety level, the intensities of ageing of the critical infrastructure related to the climate-weather change process at its operating area and its components and the coefficients of the climate-weather change process impact on the critical infrastructure and its components intensities of ageing/degradation. More general Model 4 considering together the operation process and the climate-weather change process influence on the safety of a critical infrastructure, i.e. the safety analytical model of a critical infrastructure under the influence of the operation process related to climate-weather change process is proposed. It is the integrated model of a critical infrastructure safety, linking its multistate safety model and the model of its operation process related to climate-weather change process at its operating area, considering variable at the different operation and climate-weather states impacted by them the system safety structures and its components safety parameters. The conditional safety functions at the operation process related to climate-weather change process particular states, the unconditional safety function and the risk function of a critical infrastructure at changing in time operation and climate-weather conditions are defined. Other, practically significant, critical infrastructure safety indices introduced in the model are its mean lifetime up to the exceeding a critical safety state, the moment when its risk function value exceeds the acceptable safety level, the intensities of ageing of the critical infrastructure and its components impacted by the operation process related to the climate-weather change process at its operating area and the coefficients of the operation process related to the climate-weather change process impact on the critical infrastructure and its components intensities of ageing. Most general Model 5 covers the operating environment threats and climate-weather hazards influence on the safety of a critical infrastructure. A general safety analytical model of a critical infrastructure under the influence of its operation process including operating environment threats (OET) related to climate-weather change process including extreme weather hazards (EWH) is proposed. It is the integrated model of a critical infrastructure safety, linking its multistate safety model and the joint model of its operation process including OET and the climate-weather change process including EWH at its operating area, considering variable at the different operation and climate-weather states impacted by them the critical infrastructure safety structures and its components safety parameters. The conditional safety functions at the operation process including operating environment threats and climate-weather hazards particular states, the unconditional safety function and the risk function of the critical infrastructure at changing in time its operation conditions including OET and climate-weather conditions including EWH are defined. Other, practically significant, critical infrastructure safety indices introduced in the model are its mean lifetime up to the exceeding a critical safety state, the moment when its risk function value exceeds the acceptable safety level, the intensities of ageing of the critical infrastructure and its components impacted by the operation process including operating environment threats related to the climate-weather change process including extreme weather hazards and the coefficients of the operation process including operating environment threats related to the climate-weather change process including extreme weather hazards impact on the critical infrastructure and its components intensities of ageing.

These all safety indices, proposed in Models 0-5, are defined in general for any critical infrastructures varying in time their safety structures and components safety parameters influenced by changing in time operation conditions including environment threats and climate-weather conditions including climate-weather extreme weather hazards at their operating areas.

Details of theoretical backgrounds of the second part of the proposed approach (Models 6-13 presented in the scheme items 2.11-2.16) are given in [5] and [12].

After finalising tasks of scheme items 2.1-2.10, the next step can be done to perform the tasks formulated in scheme items 2.11-2.16, terminating methodological framework, where the devised risk and impact assessment framework on interconnected and interdependent critical infrastructures may be transformed into a resilience and adaptation framework. Thus, the way we should go in the research further activity is investigating and solving the problems of optimization of critical infrastructure safety (finding optimal values of safety indictors), critical infrastructure accident consequences optimisation and mitigation, critical infrastructure resilience to climate-weather change analysis and strengthening critical infrastructure resilience to climate-weather change, pointed out in the scheme of the general approach to safety and resilience analysis presented in Figure 2.1 [11]. This activity

will result in business continuity models for critical infrastructure under climate pressures elaboration, the critical infrastructure resilience indicators defining, cost-effectiveness analysis and modelling and finally in the framework for critical infrastructure adaptation to climate change creation expected to be done in the scope of WP 4 activity.

All the above Models 0-13, presented in [11]-[14], will be the basis for preparation of significantly simplified models and procedures that are very easy to use by the practitioners and operators of the critical infrastructures in their safety analysis, what is intended to be done in the reports D3.3-Part 1 [4] and D3.3-Part 2 [5] final versions. The use of these simplified procedures is intended to be presented in details in this report for real critical infrastructures.

These simplified procedures are also expected to be modified and developed for other than safety features of critical infrastructure analysis, modelling and prediction. These procedures will be the basis for preparing algorithms and computer software allowing the practitioners to apply them automatically.

All created models and based on the simplified procedures for determining the critical infrastructure safety indicators supported by suitable computer software will be very important and useful practical tools for the owners and operators of critical infrastructures. Finally, placing the software at GMU Interactive Website linked with CIRP Website and, after a clearly defined in CIRP and SimICI the plug-in mechanism(s), where new algorithms/analyses can be added, proposing its selected items for CIRP/SimICI are intended to be done.

All created models and procedures also will be used to generate the Critical Infrastructure Safety and Resilience to Climate Change Strengthening Training System (CISCCSTS) in the form of the package of training courses based on the e-learning concept partly developed in deliverable D8.8 [9].

III. Conclusions

The paper delivers procedures that allow to find the main an practically important safety characteristics of the critical infrastructures impacted by the climate-weather change process at their operation area. The safety characteristics of the critical infrastructure, using these procedures, are different from that obtained without considering the climate-weather impacts. This fact justifies the sensibility of analysing the technical critical infrastructures safety related to the climate-weather change process that improve the accuracy of their safety evaluation. Presented tools can be useful in safety evaluation of a very wide class of real technical critical infrastructures impacted by climate hazards at their operating areas that have an influence on changing their components safety parameters. The results can be interesting for safety practitioners from various industrial sectors.

Acknowledgments

The paper presents the results developed in the scope of the EU-CIRCLE project titled “A pan – European framework for strengthening Critical Infrastructure resilience to climate change” that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653824. http://www.eu-circle.eu/.

References

[1] D1.1 EU-CIRCLE Taxonomy, 2015, http://www.eu-circle.eu/.

[2] D2.2-GMU1 EU-CIRCLE Climate change related data collection for the port oil piping transportation and the maritime ferry operating at the Baltic Sea area, 2016.

[3] D2.4 EU-CIRCLE Climate Hazards, Metadata and Standards, V0.6, 2017.

[4] D3.3-Part1 EU-CIRCLE Inventory of Critical Infrastructure Impact Assessment Models for Climate Hazards-V0.4, 2017.

[5] D3.3-Part2 EU-CIRCLE Inventory of Critical Infrastructure Impact Assessment Models for Climate Hazards-V0.4, 2017.

[6] D6.4 EU-CIRCLE Case Study 2: Sea Surge and Extreme Winds at Baltic Sea Area , Scenario 1, Oil Transport in Port, Scenario 2, Chemical Spill due to Extreme Surges, Conduction, V0.4, 2017.

[7] D6.4-GMU1 EU-CIRCLE Critical Infrastructure Operation Process General Model (CIOPGM) Application to Port Oil Piping Transportation System Operation Process Related to Operating Environment Threats (OET) and Extreme Weather Hazards (EWH), Parts 1-12, 2017.

[8] D6.4-GMU2 EU-CIRCLE Critical Infrastructure Operation Process General Model (CIOPGM) Application to Maritime Ferry Operation Process Related to Operating Environment Threats (OET) and Extreme Weather Hazards (EWH), Parts1-12, 2017.

[9] D8.8 EU-CIRCLE Training Material, V0.4, 2017.

[10] P. Ben, B.P. Gouldby, M.T. Schultz, J.D. Simm and J.L. Wibowo, “Beyond the Factor of Safety: Developing Fragility Curves to Characterize System Reliability”, Report in Water Resources Infrastructure Program ERDC SR-10-1, Prepared for Headquarters, U.S. Army Corps of Engineers, Washington, 2010.

[11] K. Kołowrocki, J. Soszyńska-Budny, “Critical Infrastructure Impact Models for Operation Threats and Climate Hazards”, Part 1, Critical Infrastructure Connections, Journal of Polish Safety and Reliability Association, Special Issue on EU-CIRCLE Project, Volume 8, Number 3, 2017.

[12] Kołowrocki K., Soszyńska-Budny J., “Critical Infrastructure Impact Models for Operation Threats and Climate Hazards”, Part 2, Impact Assessment Models, Journal of Polish Safety and Reliability Association, Special Issue on EU-CIRCLE Project, Volume 8, Number 4, 2017.

[13] K. Kołowrocki, J. Soszyńska-Budny, “Critical Infrastructure Operation and Climate-Whether Change Modelling, Prediction and Data Processing”, Journal of Polish Safety and Reliability Association, Special Issue on EU-CIRCLE Project, Volume 8, Number 2, 2017.

[14] Kołowrocki K., Soszyńska-Budny J., “Critical Infrastructure Resilience and Adaptation to Climate Change”, Journal of Polish Safety and Reliability Association, Special Issue on EU-CIRCLE Project, Volume 9, Number 1, January 2018, to appear.

[15] K. Kołowrocki and J. Soszyńska-Budny, Reliability and Safety of Complex Technical Systems and Processes: Modeling – Identification – Prediction – Optimization, 1st ed. London: Springer-Verlag, 2011.

[16] K. Kołowrocki, Reliability of Large and Complex Systems, 2nd ed. London: Elsevier, 2014.

[17] A. Lauge, J. Hernantes, J.M Sarriegi, “Critical infrastructure dependencies: A holistic, dynamic and quantitative approach”, International Journal of Critical Infrastructure Protection, Vol. 8, 6-23, 2015.

Safety and Risk Prediction of Port Oil Piping Transportation System Impacted by Climate-Weather Change Process

Krzysztof Kołowrocki
Ewa Kuligowska
Joanna Soszyńska-Budny
Mateusz Torbicki
Mathematics Department
Gdynia Maritime University
Gdynia, Poland

Abstract—**The paper is concerned with an application of the simplified impact model of critical infrastructure safety related to climate-weather change process to safety and risk prediction for port oil piping transportation system operating at the variable climate-weather conditions. There are presented the identified piping system safety parameters and the climate-weather change process parameters and characteristics. Moreover, there are presented evaluated by experts the coefficients of the climate-weather impact on the piping subsystems intensities of ageing. Further, there are determined the three-state system safety function, the expected values and standard deviations of the lifetimes in the safety state subsets and in the particular safety states, the risk function, the fragility curve and other safety and resilience indicators for port oil piping transportation system.**

Keywords—safety; risk; impact; climate-weather change process; port oil piping transportation system

I. INTRODUCTION

The safety of the port oil piping transportation system is modelled in [6], [10]. The climate-weather change process for the port oil piping transportation system operating area is modelled in [3], [8]. In this paper, the safety and risk prediction of the piping system impacted by climate-weather change process is performed. To do this, we can apply the models and procedures given in [2], [4]-[5], [7], [9]. This way, having distinguished the piping safety parameters and identified the climate-weather change process, the prediction of the main piping safety characteristics and indices is performed.

II. PARAMETERS OF PORT OIL PIPING TRANSPORTATION SYSTEM SAFETY MODEL

The considered oil piping transportation system is described in [6], [10]. The system is series and consists of three subsystems:

- the subsystem S_1 composed of two pipelines, each consists of 176 pipe segments and 2 valves,
- the subsystem S_2 composed of two pipelines, each consists of 717 pipe segments and 2 valves,
- the subsystem S_3 composed of three pipelines, each consists of 360 pipe segments and 2 valves.

After considering the opinions coming from experts, taking into account the effectiveness and safety aspects of the oil pipeline transportation system operation, we fix the number of pipeline system safety states $z = 2$ and we distinguish the following three safety states [4], [6]:

- a safety state 2 – piping operation is fully safe,
- a safety state 1 – piping operation is less safe and more dangerous because of the possibility of environment pollution,
- a safety state 0 – piping is destroyed.

Moreover, we assume that there are possible transitions between the components safety states only from better to worse ones, the critical safety state of the system $r = 1$ and the system risk permitted level $\delta = 0.05$.

We fix the mean values of the port oil piping transportation subsystems lifetimes in the safety state subsets $\{1,2\}$, $\{2\}$, are as follows for the particular subsystems [4]:

- for the subsystem S_1

$$\mu(1) = 400 \text{ years}, \tag{1}$$

$$\mu(2) = 300 \text{ years}; \tag{2}$$

- for the subsystem S_2

$$\mu(1) = 140 \text{ years}, \tag{3}$$

$$\mu(2) = 100 \text{ years}; \tag{4}$$

- for the subsystem S_3

$$\mu(1) = 160 \text{ years}, \tag{5}$$

$$\mu(2) = 120 \text{ years.} \quad (6)$$

Considering that the pipeline system is a three-state ($z = 2$) series system and (1)-(6), its safety function is given by [4]

$$\boldsymbol{S}(t,\cdot) = [1, \boldsymbol{S}(t,1), \boldsymbol{S}(t,2)], \quad (7)$$

where

$$\boldsymbol{S}(t,1) = \exp[-0.0158929t], \quad (8)$$

$$\boldsymbol{S}(t,2) = \exp[-0.0216663t]. \quad (9)$$

The expected values and standard deviations of the pipeline system lifetimes in the safety state subsets {1,2}, {2}, considering (8)-(9), respectively are [4]:

$$\mu(1) = 62.92, \text{ years,} \quad (10)$$

$$\mu(2) = 46.16 \text{ years,} \quad (11)$$

$$\sigma(1) = 62.92 \text{ years,}$$

$$\sigma(2) = 46.16 \text{ years.}$$

III. Parameters and Characteristics of Climate-Weather Change Process

On the basis of the statistical data [1], it is possible to evaluate the following unknown basic parameters of the climate-weather change process [3]:

Subsystem S_1 operating area climate-weather states ($w = 6$)

- a climate-weather state c_1 – the wave height from 0 up to 2 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_2 – the wave height from 2 m up to 5 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_3 – the wave height from 5 m up to 14 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_4 – the wave height from 0 up to 2 m and the wind speed from 17 m/s up to 33 m/s,
- a climate-weather state c_5 – the wave height from 2 m up to 5 m and the wind speed from 17 m/s up to 33 m/s.
- a climate-weather state c_6 – the wave height from 5 m up to 14 m and the wind speed from 17 m/s up to 33 m/s;

Subsystems S_2 and S_3 operating areas climate-weather states ($w = 16$)

- a climate-weather state c_1 – the air temperature from -25°C up to -15°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_2 – the air temperature from -15°C up to 5°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_3 – the air temperature from 5°C up to 25°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_4 – the air temperature from 25°C up to 35°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_5 – the air temperature from -25°C up to -15°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_6 – the air temperature from -15°C up to 5°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_7 – the air temperature from 5°C up to 25°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_8 – the air temperature from 25°C up to 35°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_9 – the air temperature from -25°C up to -15°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{10} – the air temperature from -15°C up to 5°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{11} – the air temperature from 5°C up to 25°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{12} – the air temperature from 25°C up to 35°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{13} – the air temperature from -25°C up to -15°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{14} – the air temperature from -15°C up to 5°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{15} – the air temperature from 5°C up to 25°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{16} – the air temperature from 25°C up to 35°C and the soil temperature from 20°C up to 37°C.

Having the climate-weather change process parameters, we may calculate its main characteristics [3]:

- for the subsystem S_1 operating area
  the vector of the limit values of transient probabilities of the climate-weather change process $C(t)$ at the particular states c_b

$$[q_b]_{1\times 6} = [q_1, q_2, \ldots, q_6], \tag{12}$$

  where $q_1 \cong 0.839$, $q_2 \cong 0.137$, $q_3 \cong 0.005$, $q_4 \cong 0$, $q_5 \cong 0.009$, $q_6 \cong 0.010$;

- for the subsystem S_2 and S_3 operating areas
  the vector of the limit values of transient probabilities of the climate-weather change process $C(t)$ at the particular states c_b

$$[q_b]_{1\times 16} = [q_1, q_2, \ldots, q_{16}], \tag{13}$$

  where $q_1 \cong 0.001$, $q_2 \cong 0.038$, $q_3 \cong 0$, $q_4 \cong 0$, $q_5 \cong 0$, $q_6 \cong 0.867$, $q_7 \cong 0.031$, $q_8 \cong 0$, $q_9 \cong 0$, $q_{10} \cong 0.011$, $q_{11} \cong 0.052$, $q_{12} \cong 0$, $q_{13} \cong 0$, $q_{14} \cong 0$, $q_{15} \cong 0$, $q_{16} \cong 0$.

IV. Parameters of Climate-Weather Change Process Impact on Port Oil Piping Transportation System Safety

Considering the experts opinions, the coefficients of the climate-weather impact on the port oil piping transportation subsystems S_υ, $\upsilon = 1,2,3$, intensities of ageing (the coefficients of the climate-weather impact on the piping subsystems intensities of departure from the safety state subsets $\{1,2\}$, $\{2\}$) at the climate-weather change process operating area states c_b, $b = 1,2,...,w$, are as follows:

- for subsystem S_1

$$[\rho''^{(1)}(1)]^{(b)} = 1.30,\ [\rho''^{(1)}(2)]^{(b)} = 1.30,\ b = 3,4,5,6, \tag{14}$$

$$[\rho''^{(1)}(1)]^{(b)} = 1,\ [\rho''^{(1)}(2)]^{(b)} = 1,\ b = 1,2; \tag{15}$$

- for subsystem S_2

$$[\rho''^{(2)}(1)]^{(b)} = 1.10,\ [\rho''^{(2)}(2)]^{(b)} = 1.10,\ b = 1,2,...,5,8,9,12,13,...,16, \tag{16}$$

$$[\rho''^{(2)}(1)]^{(b)} = 1,\ [\rho''^{(2)}(2)]^{(b)} = 1,\ b = 6,7,10,11; \tag{17}$$

- for subsystem S_3

$$[\rho''^{(3)}(1)]^{(b)} = 1,\ [\rho''^{(3)}(2)]^{(b)} = 1,\ b = 1,2,...,16. \tag{18}$$

V. Prediction of Safety Indicators for Port Oil Piping Transportation System

Considering that the pipeline system is a three-state ($z = 2$) series system, its safety function is given by [9]

$$\boldsymbol{S}''(t,\cdot) = [1, \boldsymbol{S}''(t,1), \boldsymbol{S}''(t,2)], \tag{19}$$

where

$$\begin{aligned}\boldsymbol{S}''(t,1) = {} & 0.937936 \exp[-0.0158929t] \\ & + 0.038064 \exp[-0.0166072t] \\ & + 0.023064 \exp[-0.0166429t] \\ & + 0.000936 \exp[-0.0173572t],\end{aligned} \tag{20}$$

$$\begin{aligned}\boldsymbol{S}''(t,2) = {} & 0.937936 \exp[-0.0216663t] \\ & + 0.038064 \exp[-0.0226662t] \\ & + 0.023064 \exp[-0.0226663t] \\ & + 0.000936 \exp[-0.0236662t].\end{aligned} \tag{21}$$

The graph of the three-state pipeline system safety function is shown in Fig. 1.

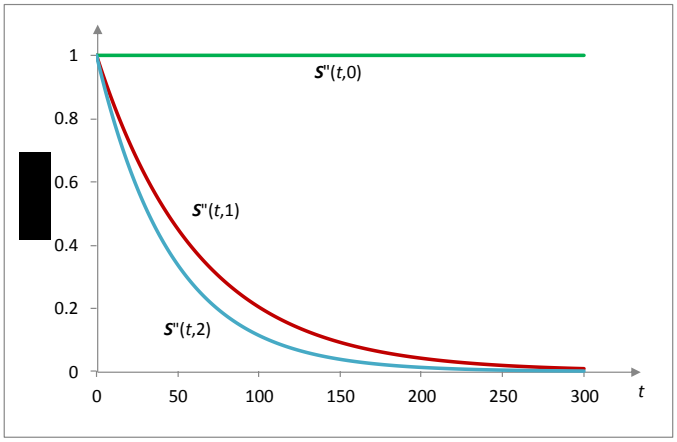


Figure 1. The graph of the port oil piping transportation system safety function coordinates

The expected values and standard deviations of the pipeline system lifetimes in the safety state subsets $\{1,2\}$, $\{2\}$, considering (20)-(21), respectively are [9]:

$$\mu''(1) = 62.75 \text{ years}, \tag{22}$$

$$\mu''(2) = 46.03 \text{ years}, \tag{23}$$

$$\sigma''(1) = 62.75 \text{ years}, \tag{24}$$

$$\sigma''(2) = 46.03 \text{ years},$$

and further, it follows that the mean values of the pipeline lifetimes in the particular safety states are:

$$\bar{\mu}''(1) = 16.72 \text{ years},$$

$$\bar{\mu}''(2) = 46.03 \text{ years}.$$

As the critical safety state is r =1, then the pipeline system risk function, is given by

$$\boldsymbol{r}''(t) = 1 - \boldsymbol{S}''(t,1), \text{ for } t \geq 0, \qquad (25)$$

where $\boldsymbol{S}''(t,1)$ is given by (20).

The graph of the risk function $\boldsymbol{r}''(t)$, also called the fragility curve of the pipeline system is shown in Fig. 2.

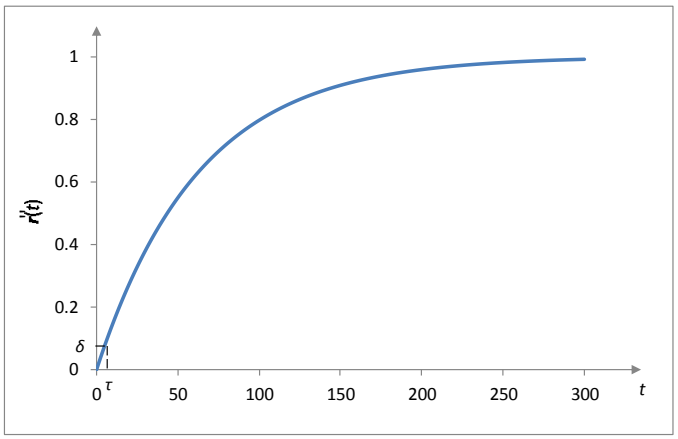


Figure 2. The graph of the fragility curve of the port oil piping transportation system

By (22) and (24), the mean and the standard deviation of the pipeline system lifetime up to exceeding critical safety state r = 1 are

$$\mu''(1) = 62.75 \text{ years},$$

$$\sigma''(1) = 62.75 \text{ years}.$$

From (25), the moment when the pipeline system risk function exceeds a permitted level δ = 0.05, is

$$\tau = \boldsymbol{r}''^{-1}(\delta) \cong 3.22 \text{ years}.$$

The pipeline system intensities of ageing according to [4] and considering (12)-(18) are:

$$\begin{aligned}\boldsymbol{\lambda}''(t,1) = \{&0.937936 \cdot 0.0158929 \exp[-0.0158929t] \\ &+ 0.038064 \cdot 0.01660719 \exp[-0.01660719t] \\ &+ 0.023064 \cdot 0.0166429 \exp[-0.0166429t] \\ &+ 0.000936 \cdot 0.01735719 \exp[-0.01735719t)\} \\ &/\{0.937936 \exp[-0.0158929t] \\ &+ 0.038064 \exp[-0.01660719t] \\ &+ 0.023064 \exp[-0.0166429t] \\ &+ 0.000936 \exp[-0.01735719t)\} \cong 0.0158929,\end{aligned}$$

$$\begin{aligned}\boldsymbol{\lambda}''(t,2) = \{&0.937936 \cdot 0.0216663 \exp[-0.0216663t] \\ &+ 0.038064 \cdot 0.0226662 \exp[-0.0226662t] \\ &+ 0.023064 \cdot 0.0226663 \exp[-0.0226663t] \\ &+ 0.000936 \cdot 0.0236662 \exp[-0.0236662t)\} \\ &/\{0.937936 \exp[-0.0216663t] \\ &+ 0.038064 \exp[-0.0226662t] \\ &+ 0.023064 \exp[-0.0226663t] \\ &+ 0.000936 \exp[-0.0236662t)\} \cong 0.0216663.\end{aligned}$$

The graphs of the intensities of ageing of the port oil piping transportation system are shown in Fig. 3.

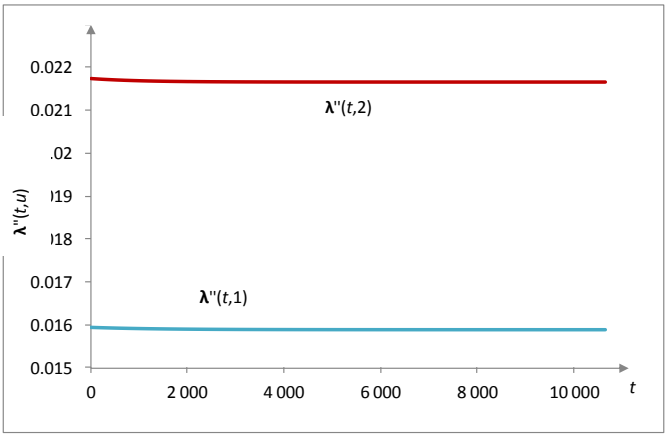


Figure 3. The graph of the intensities of ageing of the port oil piping transportation system

Considering (10)-(11), (22)-(23) and applying (29s) from [2], the coefficients of the climate-weather impact on the port oil piping transportation system are

$$\boldsymbol{\rho}''(1) = \frac{1/\mu''(1)}{1/\mu(1)} \cong \frac{1/62.75}{1/62.92} \cong 1.0027092, \qquad (26)$$

$$\boldsymbol{\rho}''(2) = \frac{1/\mu''(2)}{1/\mu(2)} \cong \frac{1/46.03}{1/46.16} \cong 1.0028242.$$

By (26), the resilience indicator, i.e. the coefficient of port oil piping transportation system resilience to climate-weather change process impact is

$$RI(t) = \frac{1}{\boldsymbol{\rho}''(1)} \cong 0.9972982 = 99.73\%.$$

VI. Conclusions

The simplified impact model of critical infrastructure safety related to climate-weather change process was applied to the safety and risk evaluation of the port oil piping transportation system operating at the variable climate-weather conditions. The predicted piping safety characteristics are different from those determined for this system operating at constant

conditions without considering climate-weather influence [6]. This fact justifies the sensibility of considering real systems' safety at the variable climate-weather conditions that is appearing out in a natural way from practice.

ACKNOWLEDGMENTS

The paper presents the results developed in the scope of the EU-CIRCLE project titled "A pan – European framework for strengthening Critical Infrastructure resilience to climate change" that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653824. http://www.eu-circle.eu/.

REFERENCES

[1] E. Jakusik, "Climate Change Related Data Collection for Port Oil Piping Transportation System and Maritime Ferry Operating Baltic Sea Areas", EU-CIRCLE Report D2.2-GMU5, 2016.

[2] K. Kołowrocki, E. Kuligowska, J. Soszyńska-Budny and M. Torbicki, "Simplified Impact Model of Critical Infrastructure Safety Related to Climate-Weather Change Process", International Conference on Information and Digital Technologies, Zilina, Slovakia, 2017, in press.

[3] K. Kołowrocki and J. Soszyńska-Budny, "Critical Infrastructure Operating Area Climate-Weather Change Process (C-WCP) Including Extreme Weather Hazards (EWH), C-WCP Model 3", EU-CIRCLE Report D3.3-GMU0, 2016.

[4] K. Kołowrocki and J. Soszyńska-Budny, "How to Model and to Analyze Operation Threats and Climate-Weather Hazards Influence on Critical Infrastructure Safety - an Overall Approach", EU-CIRCLE Report D3.3-GMU0, 2016.

[5] K. Kołowrocki and J. Soszyńska-Budny, "Integrated impact model on critical infrastructure safety related to climate-weather change process including extreme weather hazards", Summer Safety & Reliability Seminars. Journal of Polish Safety and Reliability Association vol. 8, no. 2, pp. 85-95, 2017.

[6] K. Kołowrocki and J. Soszyńska-Budny, Reliability and Safety of Complex Technical Systems and Processes: Modeling – Identification – Prediction – Optimization, 1st ed. London: Springer-Verlag, 2011.

[7] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Critical infrastructure integrated safety model related to climate-weather change process application to port oil piping transportation system operating at land Baltic seaside area", 27th ESREL Conference Proceedings, European Safety and Reliability Conference, Portoroz, Slovenia, 2017, in press.

[8] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Critical infrastructure operating area climate-weather change process including extreme weather hazards", Summer Safety & Reliability Seminars. Journal of Polish Safety and Reliability Association vol. 8, no. 2, pp. 6-14, 2017.

[9] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Integrated Model of Critical Infrastructure Safety (IMCIS) Related to Climate-Weather Change Process Including Extreme Weather Hazards (EWH), IMCIS Model 3", EU-CIRCLE Report D3.3-GMU0, 2016.

[10] K. Kołowrocki, Reliability of Large and Complex Systems, 2nd ed. London: Elsevier, 2014.

Safety and Risk Prediction of Baltic Oil Terminal Critical Infrastructure Impacted by Climate-Weather Change Process

Krzysztof Kołowrocki
Ewa Kuligowska
Joanna Soszyńska-Budny
Mateusz Torbicki
Mathematics Department
Gdynia Maritime University
Gdynia, Poland

Abstract—**The paper is concerned with an application of the simplified impact model of critical infrastructure safety related to climate-weather change process to safety and risk prediction for Baltic Oil Terminal Critical Infrastructure operating at the variable climate-weather conditions. There are presented the identified oil terminal safety parameters and the climate-weather change process parameters and characteristics. Moreover, there are presented evaluated by experts the coefficients of the climate-weather impact on the oil terminal assets intensities of ageing. Further, there are determined the three-state critical infrastructure safety function, the expected values and standard deviations of the lifetimes in the safety state subsets and in the particular safety states, the risk function, the fragility curve and other safety and resilience indicators for the Baltic Oil Terminal Critical Infrastructure.**

Keywords—safety; risk; impact; climate-weather change process; Baltic Oil Terminal Critical Infrastructure

I. INTRODUCTION

The safety of the Baltic Oil Terminal Critical Infrastructure is modelled in [6], [10]. The climate-weather change process for the Baltic Oil Terminal operating area is modelled in [3], [8]. In this paper, the safety and risk prediction of the oil terminal impacted by climate-weather change process is performed. To do this, we can apply the models and procedures given in [2], [4]-[5], [7], [9]. This way, having distinguished the oil terminal assets' safety parameters and identified the climate-weather change process, the prediction of the main oil terminal safety characteristics and indices is performed.

II. PARAMETERS OF BALTIC OIL TERMINAL CRITICAL INFRASTRUCTURE SAFETY MODEL

The considered Baltic Oil Terminal is described in [6], [10]. Its main technical assets are:

- A1 – port oil piping transportation system,
- A2 – internal pipeline technological system,
- A3 – supporting pump station,
- A4 – internal pump system,
- A5 – port oil tanker shipment terminal,
- A6 – loading railway carriage station,
- A7 – loading road carriage station,
- A8 – unloading railway carriage station,
- A9 – oil storage reservoir system.

After considering the opinions coming from experts, taking into account the effectiveness and safety aspects of the oil terminal assets operation, we fix for all of them the number of critical infrastructure safety states $z = 2$ and we distinguish their following three safety states [4], [6]:

- a safety state 2 – an asset and the oil terminal critical infrastructure operation is fully safe,
- a safety state 1 – an asset and the oil terminal critical infrastructure is less safe and more dangerous because of the possibility of environment pollution,
- a safety state 0 – an asset and the oil terminal critical infrastructure is destroyed.

Moreover, we assume that there are possible the transitions between the assets and the oil terminal critical infrastructure safety states only from better to worse ones, the critical safety state of an asset and the oil terminal critical safety state $r = 1$ and the critical infrastructure risk function permitted level $\delta = 0.05$.

We fix the mean values of the particular assets, the lifetimes in the safety state subsets $\{1,2\}$, $\{2\}$, calculated on the basis of safety data coming from or evaluated approximately by experts, are as follows [4]:

- for the asset A1, subsystem S_1

$$\mu^{(1)}(1) = 400 \text{ years}, \tag{1}$$

$$\mu^{(1)}(2) = 300 \text{ years}; \tag{2}$$

- for the asset A1, subsystem S_2

$$\mu^{(1)}(1) = 140 \text{ years}, \quad (3)$$

$$\mu^{(1)}(2) = 100 \text{ years}; \quad (4)$$

- for the asset A1, subsystem S_3

$$\mu^{(1)}(1) = 160 \text{ years}, \quad (5)$$

$$\mu^{(1)}(2) = 120 \text{ years}; \quad (6)$$

- for the assets A2-A9:

$$\mu^{(2)}(1) = 80 \text{ years}, \quad (7)$$

$$\mu^{(2)}(2) = 50 \text{ years}. \quad (8)$$

Considering that the oil terminal critical infrastructure is a three-state ($z = 2$) series system, its safety function is given by [4]

$$\boldsymbol{S}(t,\cdot) = [1, \boldsymbol{S}(t,1), \boldsymbol{S}(t,2)], \quad (9)$$

where

$$\boldsymbol{S}(t,1) = \exp[-0.1158929t], \quad (10)$$

$$\boldsymbol{S}(t,2) = \exp[-0.1816663t]. \quad (11)$$

The expected values and standard deviations of the oil terminal critical infrastructure lifetimes in the safety state subsets {1,2}, {2}, considering (10)-(11), respectively are [4]:

$$\mu(1) = 8.63 \text{ years}, \quad (12)$$

$$\mu(2) = 5.50 \text{ years}, \quad (13)$$

$$\sigma(1) = 8.63 \text{ years},$$

$$\sigma(2) = 5.50 \text{ years}.$$

III. Parameters and Characteristics of Climate-Weather Change Process

On the basis of the statistical data [1], it is possible to evaluate the following unknown basic parameters of the climate-weather change process [3]:

Asset A1 (subsystem S_1) operating area climate-weather states ($w = 6$)

- a climate-weather state c_1 – the wave height from 0 up to 2 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_2 – the wave height from 2 m up to 5 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_3 – the wave height from 5 m up to 14 m and the wind speed from 0 m/s up to 17 m/s,
- a climate-weather state c_4 – the wave height from 0 up to 2 m and the wind speed from 17 m/s up to 33 m/s,
- a climate-weather state c_5 – the wave height from 2 m up to 5 m and the wind speed from 17 m/s up to 33 m/s.
- a climate-weather state c_6 – the wave height from 5 m up to 14 m and the wind speed from 17 m/s up to 33 m/s;

Assets asset A1 (subsystems S_2 and S_3) operating area and the assets A2-A9 operating areas climate-weather states ($w = 16$)

- a climate-weather state c_1 – the air temperature from -25°C up to -15°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_2 – the air temperature from -15°C up to 5°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_3 – the air temperature from 5°C up to 25°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_4 – the air temperature from 25°C up to 35°C and the soil temperature from -30°C up to -5°C,
- a climate-weather state c_5 – the air temperature from -25°C up to -15°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_6 – the air temperature from -15°C up to 5°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_7 – the air temperature from 5°C up to 25°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_8 – the air temperature from 25°C up to 35°C and the soil temperature from -5°C up to 5°C,
- a climate-weather state c_9 – the air temperature from -25°C up to -15°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{10} – the air temperature from -15°C up to 5°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{11} – the air temperature from 5°C up to 25°C and the soil temperature from 5°C up to 20°C,

- a climate-weather state c_{12} – the air temperature from 25°C up to 35°C and the soil temperature from 5°C up to 20°C,
- a climate-weather state c_{13} – the air temperature from -25°C up to -15°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{14} – the air temperature from -15°C up to 5°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{15} – the air temperature from 5°C up to 25°C and the soil temperature from 20°C up to 37°C,
- a climate-weather state c_{16} – the air temperature from 25°C up to 35°C and the soil temperature from 20°C up to 37°C.

Having the climate-weather change process parameters, we may calculate its main characteristics [3]:

- for the asset A1 (subsystem S_1) operating area
  the vector of the limit values of transient probabilities of the climate-weather change process $C(t)$ at the particular states c_b

$$[q_b]_{1\times 6} = [q_1, q_2, \ldots, q_6], \tag{14}$$

  where $q_1 \cong 0.839$, $q_2 \cong 0.137$, $q_3 \cong 0.005$, $q_4 \cong 0$, $q_5 \cong 0.009$, $q_6 \cong 0.010$;

- for the asset A1 (subsystems S_2 and S_3) operating area and the assets A2-A9 operating areas
  the vector of the limit values of transient probabilities of the climate-weather change process $C(t)$ at the particular states c_b

$$[q_b]_{1\times 16} = [q_1, q_2, \ldots, q_{16}], \tag{15}$$

  where $q_1 \cong 0.001$, $q_2 \cong 0.038$, $q_3 \cong 0$, $q_4 \cong 0$, $q_5 \cong 0$, $q_6 \cong 0.867$, $q_7 \cong 0.031$, $q_8 \cong 0$, $q_9 \cong 0$, $q_{10} \cong 0.011$, $q_{11} \cong 0.052$, $q_{12} \cong 0$, $q_{13} \cong 0$, $q_{14} \cong 0$, $q_{15} \cong 0$, $q_{16} \cong 0$.

IV. Parameters of Climate-Weather Change Process Impact on Baltic Oil Terminal Critical Infrastructure Safety

Considering the experts opinions, the coefficients of the climate-weather impact on the oil terminal critical infrastructure assets intensities of ageing (the coefficients of the climate-weather impact on the oil terminal critical infrastructure assets intensities of departure from the safety state subsets {1,2}, {2}) at the climate-weather change process operating area states c_b, $b = 1,2,...,w$, are as follows:

- for the asset A1 (subsystem S_1)

$$[\rho''^{(1)}(1)]^{(b)} = 1.30,\ [\rho''^{(1)}(2)]^{(b)} = 1.30,\ b = 3,4,5,6, \tag{16}$$

$$[\rho''^{(1)}(1)]^{(b)} = 1,\ [\rho''^{(1)}(2)]^{(b)} = 1,\ b = 1,2; \tag{17}$$

- for the asset A1 (subsystem S_2) and the assets A2-A9:

$$[\rho''^{(i)}(1)]^{(b)} = 1.10,\ [\rho''^{(i)}(2)]^{(b)} = 1.10,\ b = 1,2,...,5,8,9,12,13,...,16,\ i = 1,2,...,9, \tag{18}$$

$$[\rho''^{(i)}(1)]^{(b)} = 1,\ [\rho''^{(i)}(2)]^{(b)} = 1,\ b = 6,7,10,11,\ i = 1,2,...,9; \tag{19}$$

- for the asset A1 (subsystem S_3):

$$[\rho''^{(1)}(1)]^{(b)} = 1,\ [\rho''^{(1)}(2)]^{(b)} = 1,\ b = 1,2,...,16. \tag{20}$$

V. Prediction of Safety Indicators for Baltic Oil Terminal Critical Infrastructure

Considering that the oil terminal is a three-state ($z = 2$) series system, its safety function is given by [9]

$$\boldsymbol{S}''(t,\cdot) = [1, \boldsymbol{S}''(t,1), \boldsymbol{S}''(t,2)], \tag{21}$$

where

$$\begin{aligned}\boldsymbol{S}''(t,1) = {} & 0.937936 \exp[-0.1158929t] \\ & + 0.023064 \exp[-0.1166429t] \\ & + 0.038064 \exp[-0.1266072t] \\ & + 0.000936 \exp[-0.1273572t],\end{aligned} \tag{22}$$

$$\begin{aligned}\boldsymbol{S}''(t,2) = {} & 0.937936 \exp[-0.1816663t] \\ & + 0.023064 \exp[-0.1826662t] \\ & + 0.038064 \exp[-0.1986663t] \\ & + 0.000936 \exp[-0.1996662t].\end{aligned} \tag{23}$$

The graph of the three-state oil terminal safety function is shown in Fig. 1.

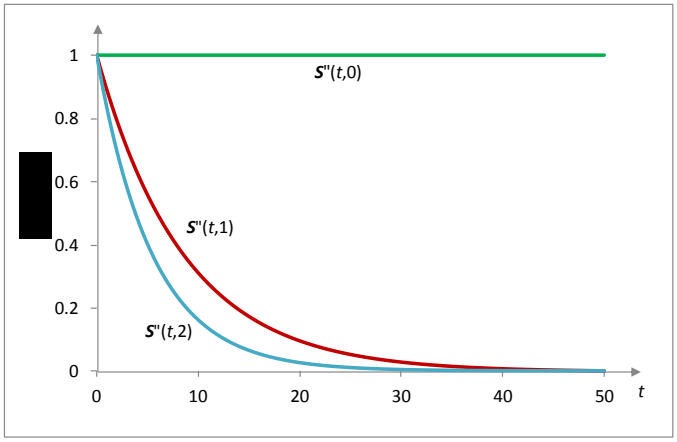


Figure 1. The graph of the Baltic Oil Terminal Critical Infrastructure safety function coordinates

The expected values and standard deviations of the oil terminal lifetimes in the safety state subsets {1,2}, {2}, considering (22)-(23), respectively are [9]:

$$\mu''(1) = 8.60 \text{ years}, \tag{24}$$

$$\mu''(2) = 5.49 \text{ years}, \tag{25}$$

$$\sigma''(1) = 8.60 \text{ years}, \tag{26}$$

$$\sigma''(2) = 5.49 \text{ years},$$

and further, it follows that the mean values of the oil terminal lifetimes in the particular safety states are:

$$\overline{\mu}''(1) = 3.11 \text{ years},$$

$$\overline{\mu}''(2) = 5.49 \text{ years}.$$

As the critical safety state is r =1, then the oil terminal risk function, is given by

$$\boldsymbol{r}''(t) = 1 - \boldsymbol{S}''(t,1), \text{ for } t \geq 0, \tag{27}$$

where $\boldsymbol{S}''(t,1)$ is given by (22).

The graph of the risk function $\boldsymbol{r}''(t)$ of the oil terminal is shown in Fig. 2.

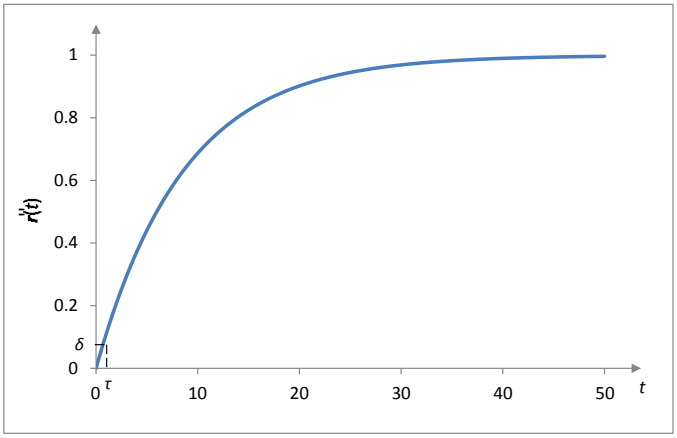


Figure 2. The graph of the risk function (the fragility curve) of the Baltic Oil Terminal Critical Infrastructure

By (24) and (26), the mean and the standard deviation of the oil terminal lifetime up to exceeding critical safety state $r = 1$ are

$$\mu''(1) = 8.60 \text{ years},$$

$$\sigma''(1) = 8.60 \text{ years}.$$

From (27), the moment when the oil terminal risk function exceeds a permitted level $\delta = 0.05$, is

$$\tau = \boldsymbol{r}''^{-1}(\delta) \cong 0.44 \text{ year}.$$

The oil terminal intensities of ageing according to [4] are:

$$\begin{aligned}\boldsymbol{\lambda}''(t,1) = \{&0.937936 \cdot 0.1158929 \exp[-0.1158929t] \\ &+ 0.023064 \cdot 0.1166429 \exp[-0.1166429t] \\ &+ 0.038064 \cdot 0.1266072 \exp[-0.1266072t] \\ &+ 0.000936 \cdot 0.1273572 \exp[-0.1273572t)\} \\ &/\{0.937936 \exp[-0.1158929t] \\ &+ 0.023064 \exp[-0.1166429t] \\ &+ 0.038064 \exp[-0.1266072t] \\ &+ 0.000936 \exp[-0.1273572t)\} \cong 0.1158929,\end{aligned}$$

$$\begin{aligned}\boldsymbol{\lambda}''(t,2) = \{&0.937936 \cdot 0.1816663 \exp[-0.1816663t] \\ &+ 0.023064 \cdot 0.1826662 \exp[-0.1826662t] \\ &+ 0.038064 \cdot 0.1986663 \exp[-0.1986663t] \\ &+ 0.000936 \cdot 0.1996662 \exp[-0.1996662t)\} \\ &/\{0.937936 \exp[-0.1816663t] \\ &+ 0.023064 \exp[-0.1826662t] \\ &+ 0.038064 \exp[-0.1986663t] \\ &+ 0.000936 \exp[-0.1996662t)\} \cong 0.1816663.\end{aligned}$$

The graphs of the intensities of ageing of the oil terminal are shown in Fig. 3.

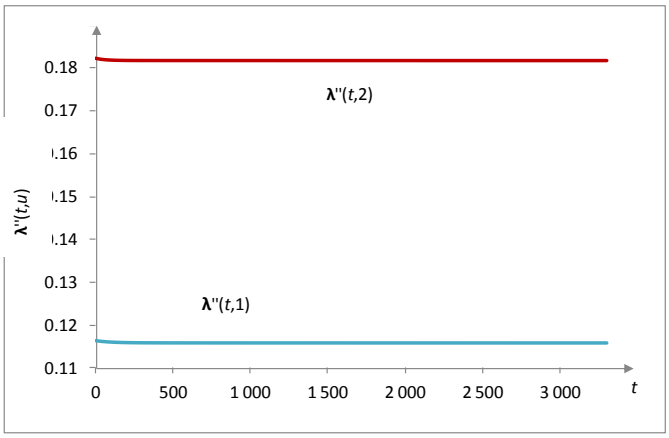


Figure 3. The graphs of the intensities of ageing of the Baltic Oil Terminal Critical Infrastructure

Considering (12)-(13), (24)-(25) and applying (29) from [2], the coefficients of the climate-weather impact on the Baltic Oil Terminal Critical Infrastructure are:

$$\rho''(1) = \frac{1/\mu''(1)}{1/\mu(1)} \cong \frac{1/8.60}{1/8.63} \cong 1.0034884, \qquad (28)$$

$$\rho''(2) = \frac{1/\mu''(2)}{1/\mu(2)} \cong \frac{1/5.49}{1/5.50} \cong 1.0018215.$$

By (28), the resilience indicator, i.e. the coefficient of port oil piping transportation system resilience to climate-weather change process impact is

$$RI(t) = \frac{1}{\rho''(1)} \cong 0.9965238 = 99.65\%.$$

VI. Conclusions

The simplified impact model of critical infrastructure safety related to climate-weather change process was applied to the safety and risk evaluation of the Baltic Oil Terminal Critical Infrastructure operating at the variable climate-weather conditions. The predicted oil terminal safety characteristics are different from those determined for this system operating at constant conditions without considering climate-weather influence [6]. This fact justifies the sensibility of considering real systems' safety at the variable climate-weather conditions that is appearing out in a natural way from practice.

Acknowledgments

The paper presents the results developed in the scope of the EU-CIRCLE project titled "A pan – European framework for strengthening Critical Infrastructure resilience to climate change" that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653824. http://www.eu-circle.eu/.

References

[1] E. Jakusik, "Climate Change Related Data Collection for Port Oil Piping Transportation System and Maritime Ferry Operating Baltic Sea Areas", EU-CIRCLE Report D2.2-GMU5, 2016.

[2] K. Kołowrocki, E. Kuligowska, J. Soszyńska-Budny and M. Torbicki, "Simplified Impact Model of Critical Infrastructure Safety Related to Climate-Weather Change Process", International Conference on Information and Digital Technologies, Zilina, Slovakia, 2017, in press.

[3] K. Kołowrocki and J. Soszyńska-Budny, "Critical Infrastructure Operating Area Climate-Weather Change Process (C-WCP) Including Extreme Weather Hazards (EWH), C-WCP Model 3", EU-CIRCLE Report D3.3-GMU0, 2016.

[4] K. Kołowrocki and J. Soszyńska-Budny, "How to Model and to Analyze Operation Threats and Climate-Weather Hazards Influence on Critical Infrastructure Safety - an Overall Approach", EU-CIRCLE Report D3.3-GMU0, 2016.

[5] K. Kołowrocki and J. Soszyńska-Budny, "Integrated impact model on critical infrastructure safety related to climate-weather change process including extreme weather hazards", Summer Safety & Reliability Seminars. Journal of Polish Safety and Reliability Association vol. 8, no. 2, pp. 85-95, 2017.

[6] K. Kołowrocki and J. Soszyńska-Budny, Reliability and Safety of Complex Technical Systems and Processes: Modeling – Identification – Prediction – Optimization, 1st ed. London: Springer-Verlag, 2011.

[7] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Critical infrastructure integrated safety model related to climate-weather change process application to port oil piping transportation system operating at land Baltic seaside area", 27th ESREL Conference Proceedings, European Safety and Reliability Conference, Portoroz, Slovenia, 2017, in press.

[8] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Critical infrastructure operating area climate-weather change process including extreme weather hazards", Summer Safety & Reliability Seminars. Journal of Polish Safety and Reliability Association vol. 8, no. 2, pp. 6-14, 2017.

[9] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, "Integrated Model of Critical Infrastructure Safety (IMCIS) Related to Climate-Weather Change Process Including Extreme Weather Hazards (EWH), IMCIS Model 3", EU-CIRCLE Report D3.3-GMU0, 2016.

[10] K. Kołowrocki, Reliability of Large and Complex Systems, 2nd ed. London: Elsevier, 2014.

An Approach to Safety Prediction of Critical Infrastructure Impacted by Climate-Weather Change Process

Krzysztof Kołowrocki
Ewa Kuligowska
Joanna Soszyńska-Budny
Mateusz Torbicki
Mathematics Department
Gdynia Maritime University
Gdynia, Poland

Abstract— **The paper is devoted to the climate influence on the safety of a critical infrastructure defined as a complex system in its operating environment that in the case of its degradation have significant destructive influence on the health, safety and security, economics and social conditions of large human communities and territory areas. The method based on the joint model linking a multistate approach to critical infrastructure safety with a semi-Markov modelling of the climate-weather change process at the critical infrastructure operation area is proposed to the safety analysis and prediction of critical infrastructures impacted by the climate hazards.**

Keywords— critical infrastructure; climate change; safety; climate impact; safety indicator; prediction

I. INTRODUCTION

Most real critical infrastructures are strongly influenced by changing in time the climate-weather conditions at their operating area that increasing their degradation/ageing. The time dependent interactions between the climate-weather states varying at the critical infrastructure operating area and the critical infrastructure components/assets safety states changing are evident features of most real technical critical infrastructures [3]. The common analysis of critical infrastructures safety and the climate-weather change at their operating area is of great value in the industrial practice. The convenient tools for analyzing this problem are the critical infrastructure multistate safety modelling [7], [10], [13]-[16] commonly used with the semi-Markov modelling [1], [4], [6], [10] the climate-weather change process at the critical infrastructure operating area, leading to the construction of the joint general safety model of the critical infrastructure related to climate-weather change process at its operating area [8], [11]-[12]. In everyday practice, there are needed the tools that could be applied to evaluating the climate-weather hazards influence on the safety characteristics of a critical infrastructure defined as a complex system in its operating environment that significant features are inside-system dependencies and outside-system dependencies, that in the case of its degradation have significant destructive influence on the health, safety and security, economics and social conditions of large human communities and territory areas [3], [9]. In the safety analysis of the critical infrastructure impacted by climate hazards, the determination of its safety function and its risk function, which graph corresponds to the fragility curve [2], are crucial indicators/indices for safety practitioners. Other practically significant critical infrastructure safety indices defined in the paper are its mean lifetime up to the exceeding a critical safety state, the moment when its risk function value exceeds the acceptable safety level, the critical infrastructure intensity of ageing/degradation, the coefficient of climate-weather change process impact on critical infrastructure intensities of ageing and the coefficient of critical infrastructure resilience to climate-weather change process impact. The knowledge of these critical infrastructure safety indicators is of great value in the industrial practice. Thus, there are needed the tools for finding the critical infrastructure safety and resilience indicators and the procedures allowing for changing the critical infrastructure features, leading to the strengthening its resilience to climate change [12].

In the paper, to satisfy those needs, the simplified procedures based on the authors' theoretical results [7], [10] developments are proposed directly to users dealing with, and assuring critical infrastructure safety in everyday practice.

II. MULTISTATE SYSTEM SAFETY ANALYSIS

Multistate approach to safety analysis is presented in [10].

III. CRITICAL INFRASTRUCTURE SAFETY INDICATORS

We assume that the changes of the climate-weather change process $C(t)$ states at the critical infrastructure operating area have an influence on its multistate components/assets E_i, $i = 1,2,...,n$, safety. Consequently, we denote the system multistate component E_i, $i = 1,2,...,n$, conditional lifetime in the safety state subset $\{u, u+1,...,z\}$ while the climate-weather change process $C(t)$ at the system operating area is at the state c_b, $b = 1,2,...,w$, by $T''^{(b)}_i(u)$ and its conditional safety function by the vector

$$[S''_i(t,\cdot)]^{(b)} = [1, [S''_i(t,1)]^{(b)},..., [S''_i(t,z)]^{(b)}],$$
$$t \in <0,\infty),\ b = 1,2,...,w,\ i = 1,2,...,n, \quad (1)$$

with the coordinates defined by

$$[S''_i(t,u)]^{(b)} = P(T''^{(b)}_i(u) > t \mid C(t) = c_b),$$

for $t \in <0,\infty)$, $u = 1,2,...,z$, $b = 1,2,...,w$.
The safety function $[S''_i(t,u)]^{(b)}$ is the conditional probability that the component E_i lifetime $T''^{(b)}_i(u)$ in the safety state subset $\{u, u+1,...,z\}$ is greater than t, while the climate-weather change process $C(t)$ at the system operating area is at the state c_b, $b = 1,2,...,w$.
In the case, the system components E_i, $i = 1,2,...,n$, at the climate-weather change process $C(t)$ at the system operating area states c_b, $b = 1,2,...,w$, have the exponential safety functions, the coordinates of the vector (1) are given by

$$[S''_i(t,u)]^{(b)} = P(T''^{(b)}_i(u) > t \mid C(t) = c_b) = \exp[-[\lambda''_i(u)]^{(b)}t],$$
$$t \in <0,\infty),\ b = 1,2,...,w,\ i = 1,2,...,n, \quad (2)$$

Existing in (2) the intensities of ageing of the system components E_i, $i = 1,2,...,n$, (the intensities of the system components E_i, $i = 1,2,...,n$, departure from the safety state subset $\{u, u+1,...,z\}$) at the climate-weather change process $C(t)$ at the system operating area states c_b, $b = 1,2,...,w$, i.e. the coordinates of the vector

$$[\lambda''_i(\cdot)]^{(b)} = [0, [\lambda''_i(1)]^{(b)},..., [\lambda''_i(z)]^{(b)}],$$
$$b = 1,2,...,w,\ i = 1,2,...,n, \quad (3)$$

are given by

$$[\lambda''_i(u)]^{(b)} = [\rho''_i(u)]^{(b)} \cdot \lambda_i(u),$$
$$u = 1,2,...,z,\ b = 1,2,...,w,\ i = 1,2,...,n, \quad (4)$$

where $\lambda_i(u)$ are the intensities of ageing of the system components E_i, $i = 1,2,...,n$, (the intensities of the system components E_i, $i = 1,2,...,n$, departure from the safety state subset $\{u, u+1,...,z\}$) without climate-weather change impact, i.e. the coordinate of the vector

$$\lambda_i(u) = [0, \lambda_i(1), \ldots, \lambda_i(z)],\ i = 1,2,...,n,$$

and

$$[\rho''_i(u)]^{(b)},\ u = 1,2,...,z,\ b = 1,2,...,w,\ i = 1,2,...,n,$$

are the coefficients of climate-weather impact on the system components E_i, $i = 1,2,...,n$, intensities of ageing (the coefficients of climate-weather impact on critical infrastructure component E_i, $i = 1,2,...,n$, intensities of departure from the safety state subset $\{u, u+1,...,z\}$) at the climate-weather change process operating area states c_b, $b = 1,2,...,w$, i.e. the coordinate of the vector

$$[\rho''_i(\cdot)]^{(b)} = [0, [\rho''_i(1)]^{(b)},..., [\rho''_i(z)]^{(b)}],$$
$$b = 1,2,...,w,\ i = 1,2,...,n. \quad (5)$$

The system component safety function (1), the system components intensities of ageing (3) and the coefficients of the climate-weather impact on the system components intensities of ageing (5) are main system component safety indices.
Further, we denote the critical infrastructure conditional lifetime in the safety state subset $\{u, u+1,...,z\}$ while the climate-weather change process $C(t)$ at the critical infrastructure operating area is at the climate-weather state c_b, $b = 1,2,...,w$, by $T''^{(b)}(u)$ and the conditional safety function (SI1) of the critical infrastructure by the vector [10]

$$[\boldsymbol{S}''(t,\cdot)]^{(b)} = [1, [\boldsymbol{S}''(t,1)]^{(b)},..., [\boldsymbol{S}''(t,z)]^{(b)}],$$

with the coordinates defined by

$$[\boldsymbol{S}''(t,u)]^{(b)} = P(T''^{(b)}(u) > t \mid C(t) = c_b),$$
$$\text{for } t \in <0,\infty),\ u = 1,2,...,z,\ b = 1,2,...,w.$$

The safety function $[\boldsymbol{S}''(t,u)]^{(b)}$ is the conditional probability that the critical infrastructure lifetime $T''^{(b)}(u)$ in the safety state subset $\{u, u+1,...,z\}$ is greater than t, while the climate-weather change process $C(t)$ is at the climate-weather state c_b, $b = 1,2,...,w$.
Further, we assume that the critical infrastructure has the exponential conditional safety function (SI1), i.e.

$$[\boldsymbol{S}''(t,u)]^{(b)} = \exp[-[\boldsymbol{\lambda}''(u)]^{(b)}t],$$
$$t \in <0,\infty),\ [\boldsymbol{\lambda}''(u)]^{(b)} \geq 0,\ u = 1,2,...,z.$$

Under this assumption, the mean lifetime of the critical infrastructure in the safety state subset $\{u, u+1,...,z\}$, is given by

$$\mu''_b(u) = \frac{1}{[\boldsymbol{\lambda}''(u)]^{(b)}},\ u = 1,2,...,z,\ b = 1,2,...,w, \quad (6)$$

We denote the critical infrastructure unconditional lifetime in the safety state subset $\{u, u+1,...,z\}$ by $T''(u)$ and the unconditional safety function of the critical infrastructure (SI1) by the vector

$$\boldsymbol{S}''(t,\cdot) = [1, \boldsymbol{S}''(t,1),..., \boldsymbol{S}''(t,z)], \quad (7)$$

with the coordinates defined by

$$\boldsymbol{S}''(t,u) = P(T''(u) > t), \text{ for } t \in <0,\infty),\ u = 1,2,...,z. \quad (8)$$

In the case when the critical infrastructure operation time C is large enough, the coordinates (8) of the unconditional safety function (SI1) of the critical infrastructure defined by (7) are given by

$$\boldsymbol{S}''(t,u) \cong \sum_{b=1}^{w} q_b \exp[-[\boldsymbol{\lambda}''(u)]^{(b)}t], \text{ for } t \in <0,\infty),\ u = 1,2,...,z, \quad (9)$$

where q_b, $b = 1,2,...,w$, are the climate-weather change process $C(t)$ limit transient probabilities (C-WCPC1) defined in [11].

The mean value of the critical infrastructure unconditional lifetime $T''(u)$ in the safety state subset $\{u, u+1,...,z\}$ is given by [4], [10]

$$\mu''(u) \cong \sum_{b=1}^{w} q_b \mu''_b(u), \; u = 1,2,...,z, \tag{10}$$

where $\mu''_b(u)$ are the mean values of the critical infrastructure conditional lifetimes $T''^{(b)}(u)$ in the safety state subset $\{u, u+1,...,z\}$ at the climate-weather state c_b, $b = 1,2,...,w$, given by (10) and q_b are defined in [12].
Moreover, according to [10], if r is the critical safety state, then the critical infrastructure risk function (SI2)

$$\boldsymbol{r}''(t) = P(S''(t) < r \mid S''(0) = z) = P(T''(r) \le t), \; t \in \langle 0,\infty), \tag{}$$

defined as a probability that the critical infrastructure is in the subset of safety states worse than the critical safety state r, $r \in \{1,...,z\}$ while it was in the safety state z at the moment $t = 0$ is given by [10], [12]

$$\boldsymbol{r}''(t) = 1 - \sum_{b=1}^{w} q_b \exp[-[\boldsymbol{\lambda}''(r)]^{(b)} t], \; t \in \langle 0,\infty). \tag{11}$$

The critical infrastructure safety function (SI1), the critical infrastructure risk function (SI2) and its graph called the critical infrastructure fragility curve (SI3) are main critical infrastructure safety indicators (SI).
Other practically useful critical infrastructure safety factors are:
- the mean value of the unconditional critical infrastructure lifetime $T''(r)$ up to the exceeding the critical safety state r (SI4) given by

$$\mu''(r) \cong \sum_{b=1}^{w} q_b \mu''_b(r), \tag{12}$$

where $\mu''_b(r)$ are the mean values of the critical infrastructure conditional lifetimes $T''^{(b)}(r)$ in the safety state subset $\{r, r+1,...,z\}$ at the climate-weather state c_b, $b = 1,2,...,w$, according to (6), given by

$$\mu''_b(r) = \frac{1}{[\boldsymbol{\lambda}''(r)]^{(b)}}, \; b = 1,2,...,w, \tag{13}$$

and q_b are defined in [11];
- the standard deviation of the critical infrastructure lifetime $T''(r)$ up to the exceeding the critical safety state r (SI5) given by

$$\sigma''(r) = \sqrt{n''(r) - [\mu''(r)]^2}, \tag{14}$$

where

$$n''(r) = 2 \int_0^{\infty} t \, \boldsymbol{S}''(t,r) dt, \tag{15}$$

where $\boldsymbol{S}''(t,r)$ is given by (9) for $u = r$ and $\mu''(r)$ is given by (12);
- the moment τ'' when the critical infrastructure risk function exceeds a permitted level δ (SI6) given by

$$\tau'' = \boldsymbol{r}''^{-1}(\delta),$$

where $\boldsymbol{r}''^{-1}(t)$ is the inverse function of the risk function $\boldsymbol{r}''(t)$ given by (11).
Other critical infrastructure safety indices are:
- the intensities of ageing (degradation) of the critical infrastructure impacted by the climate-weather change process /the intensities of critical infrastructure departure from the safety state subset $\{u, u+1,...,z\}$ impacted by the climate-weather change process (SI7), i.e. the coordinates of the vector

$$\boldsymbol{\lambda}''(t,\cdot) = [0, \boldsymbol{\lambda}''(t,1),..., \boldsymbol{\lambda}''(t,z)], \; t \in \langle 0,\infty),$$

where

$$\boldsymbol{\lambda}''(t,u) = [\sum_{b=1}^{w} [q_b [\boldsymbol{\lambda}''(u)]^{(b)} \exp[-[\boldsymbol{\lambda}''(u)]^{(b)} t]] / [\sum_{b=1}^{w} [q_b \exp[-[\boldsymbol{\lambda}''(u)]^{(b)} t]], \; t \in \langle 0,\infty), \; u = 1,2,...,z; \tag{16}$$

- the coefficients of the climate-weather change process impact on the critical infrastructure intensities of ageing /the coefficients of the climate-weather change process impact on critical infrastructure intensities of departure from the safety state subset $\{u, u+1,...,z\}$) (SI8), i.e. the coordinates of the vector

$$\boldsymbol{\rho}''(t,\cdot) = [0, \boldsymbol{\rho}''(t,1),..., \boldsymbol{\rho}''(t,z)], \; t \in \langle 0,\infty),$$

where

$$\boldsymbol{\lambda}''(t,u) = \rho''(t,u) \cdot \boldsymbol{\lambda}^0(t,u), \; t \in \langle 0,\infty), \; u = 1,2,...,z, \tag{17}$$

and $\boldsymbol{\lambda}^0(t,u)$ are the intensities of ageing of the critical infrastructure (the intensities of the critical infrastructure departure from the safety state subset $\{u, u+1,...,z\}$) without of climate-weather change process impact, i.e. the coordinate of the vector

$$\boldsymbol{\lambda}^0(t,\cdot) = [0, \boldsymbol{\lambda}^0(t,1),..., \boldsymbol{\lambda}^0(t,z)], \; t \in \langle 0,\infty),$$

Additionally, we define the critical infrastructure resilience indicator (RI), i.e. the coefficient of critical infrastructure resilience to climate-weather change process impact

$$RI(t) = 1 / \boldsymbol{\rho}''(t,r), \; t \in \langle 0,\infty), \tag{18}$$

where $\boldsymbol{\rho}''(t,r)$ is the coefficients of the climate-weather change process impact on the critical infrastructure intensity of ageing $\boldsymbol{\lambda}''(t,r)$, i.e. the coefficients of the climate-weather change process impact on critical infrastructure intensities of departure from the safety state subset $\{r, r+1,...,z\}$ of states not worse than the critical safety state r.

IV. CONCLUSIONS

The paper delivers procedures that allow to find the main an practically important safety characteristics of the critical infrastructures impacted by the climate-weather change process at their operation area. The safety characteristics of the critical infrastructure, using these procedures, are different from that obtained without considering the climate-weather impacts. This fact justifies the sensibility of analysing the technical critical infrastructures safety related to the climate-weather change process that improve the accuracy of their safety evaluation. Presented tools can be useful in safety evaluation of a very wide class of real technical critical infrastructures impacted by climate hazards at their operating areas that have an influence on changing their components safety parameters. The results can be interesting for safety practitioners from various industrial sectors.

ACKNOWLEDGMENTS

The paper presents the results developed in the scope of the EU-CIRCLE project titled “A pan – European framework for strengthening Critical Infrastructure resilience to climate change” that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653824. http://www.eu-circle.eu/.

REFERENCES

[1] T. Aven, “Reliability evaluation of multistate systems with multistate components”, IEEE Transactions on Reliability, 34, pp. 473-479, 1985.

[2] P. Ben, B.P. Gouldby, M.T. Schultz, J.D. Simm and J.L. Wibowo, “Beyond the Factor of Safety: Developing Fragility Curves to Characterize System Reliability”, Report in Water Resources Infrastructure Program ERDC SR-10-1, Prepared for Headquarters, U.S. Army Corps of Engineers, Washington, 2010.

[3] “EU-CIRCLE Taxonomy”, EU-CIRCLE Project Report D1.1, 2015.

[4] F. Grabski, Semi-Markov Processes: Application in System Reliability and Maintenance, Elsevier, 2014.

[5] S. Helvacioglu and M. Insel, “Expert system applications in marine technologies”, Ocean Engineering, 35, 11,12, pp. 1067-1074, 2008.

[6] D. Klabjan and D. Adelman, “Existence of optimal policies for semi-Markov decision processes using duality for infinite linear programming”, Siam J Contr Optim 44, 6, pp. 2104-2122, 2006.

[7] K. Kołowrocki, Reliability of Large and Complex Systems, 2nd ed. London: Elsevier, 2014.

[8] K. Kołowrocki and J. Soszyńska-Budny, “How to model and to analyse operation threats and climate-weather hazards influence on critical infrastructure safety – An overall approach”, EU-CIRCLE Report D3.3-GMU0, 2016.

[9] K. Kołowrocki and J. Soszyńska-Budny, “Introduction to safety analysis of critical infrastructures”, Journal of Polish Safety and Reliability Association, Summer Safety and Reliability Seminars 3, pp. 73-88, 2012.

[10] K. Kołowrocki and J. Soszyńska-Budny, Reliability and Safety of Complex Technical Systems and Processes: Modeling – Identification – Prediction – Optimization, 1st ed. London: Springer-Verlag, 2011.

[11] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, “Critical Infrastructure Operating Area Climate-Weather Change Process (C-WCP) Including Extreme Weather Hazards (EWH), C-WCP Model 3”, EU-CIRCLE Report D3.3-GMU0, 2016.

[12] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, “Integrated Model of Critical Infrastructure Safety (IMCIS) Related to Climate-Weather Change Process, IMCIS Model 3”, EU-CIRCLE Report D3.3-GMU0, 2016.

[13] A. Lisnianski and G. Levitin, Multi-State System Reliability. Assessment, Optimization and Applications, World Scientific Publishing Co. Pte. Ltd., 2003.

[14] J. Xue, “On multi-state system analysis. IEEE Transactions on Reliability”, 34, pp. 329-337, 1985.

[15] J. Xue and K. Yang, “Dynamic reliability analysis of coherent multi-state systems”, IEEE Transactions on Reliability, 4, 44, pp. 683-688, 1995.

[16] J. Xue and K. Yang. “Symmetric relations in multi-state systems”, IEEE Transactions on Reliability, 4, 44, pp.689-693, 1995.

Simplified Impact Model of Critical Infrastructure Safety Related to Climate-Weather Change Process

Krzysztof Kołowrocki
Ewa Kuligowska
Joanna Soszyńska-Budny
Mateusz Torbicki
Mathematics Department
Gdynia Maritime University
Gdynia, Poland

Abstract—***The paper is devoted to the critical infrastructure safety impacted by climate-weather change. A simplified safety model of the critical infrastructure influenced by the climate-weather change process at its operating area is proposed. This model is linking critical infrastructure multistate safety model and the model of the climate-weather change process at its operating area. The conditional safety functions at the climate-weather particular states and the unconditional safety function of the critical infrastructure at changing in time climate-weather conditions, the mean value and the variance of the critical infrastructure unconditional lifetime and other safety indicators are defined.***

Keywords—safety, critical infrastructure, climate-weather change process

I. INTRODUCTION

Changing in time climate-weather conditions have impact on most real complex technical systems including critical infrastructures defined in [1] at their operating areas. The analysis of the critical infrastructure safety and the climate-weather change at its operating area is very important in the industrial practice because of destructive impacts of some extreme weather hazards on the critical infrastructure safety. The multistate critical infrastructures safety modelling used with the semi-Markov modeling of the climate-weather change processes at their operating areas is a practical approach to analyzing this problem. This approach leads to the construction of the joint safety model of the critical infrastructure related to the climate-weather change process at its operating area.

II. CRITICAL INFRASTRUCTURE SAFETY WITHOUT CLIMATE-WEATHER IMPACT

We consider the critical infrastructure composed of n, $n \in N$, subsystems/assets S_i, $i = 1,2,\ldots, n$, and we assume that it has z, $z \in N$, different safety states. The analysis of the critical infrastructure safety could be prepared only after introducing all its safety parameters. First, we have to define the safety states of the critical infrastructure and its assets, to fix their critical safety state r, $r = 1,2,\ldots, z$, and the critical infrastructure risk function permitted level δ.

Next, we have to receive from the critical infrastructure operators the mean values $\mu_i(u)$, $i = 1,2,\ldots, n$, of the subsystems lifetimes in the safety state subsets $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, without considering impact of the climate-weather change.

Thus, assuming that subsystems S_i, $i = 1,2,\ldots, n$, have the exponential safety functions, they are given by the vector

$$S_i(t, \cdot) = [1, S_i(t, 1), \ldots, S_i(t, z)], \; t \geq 0, \; i = 1,\ldots, n, \qquad (1)$$

with particular coordinates

$$S_i(t, u) = P(T_i(u) > t) = \exp[-\lambda_i(u)t] = \exp[-\frac{t}{\mu_i(u)}], \quad t \geq 0, \; i = 1,\ldots, n, \; u = 1,2,\ldots, z, \qquad (2)$$

where $T_i(u)$ and $\lambda_i(u)$, $i = 1,\ldots, n$, $u = 1,2,\ldots, z$, respectively denote the conditional lifetime of the subsystem S_i in the safety state subset $\{u, u+1,\ldots, z\}$ and the intensities of the subsystem S_i departure from the safety state subset $\{u, u+1,\ldots, z\}$.

For simplicity, we consider particularly only two types of the critical infrastructure structures: a series critical infrastructure and a parallel critical infrastructure.

Hence, marking by $T(u)$, $u = 1,2,\ldots, z$, the unconditional critical infrastructure lifetime in the safety state subset $\{u, u+1,\ldots, z\}$ and the unconditional critical infrastructure safety function by the vector

$$\boldsymbol{S}(t, \cdot) = [1, \boldsymbol{S}(t, 1), \ldots, \boldsymbol{S}(t, z)], \; t \geq 0, \qquad (3)$$

with the coordinates defined by

$$\boldsymbol{S}(t, u) = P(T(u) > t), \; t \geq 0, \; u = 1, \ldots, z, \qquad (4)$$

where particular we have:
- for a series critical infrastructure

$$\boldsymbol{S}(t, u) = \prod_{i=1}^{n} \exp[-\frac{1}{\mu_i(u)} t] = \exp[-t \cdot \sum_{i=1}^{n} \frac{1}{\mu_i(u)}],$$

$$t \geq 0,\ u = 1,\ldots, z, \tag{5}$$

- for a parallel critical infrastructure

$$\boldsymbol{S}(t, u) = 1 - \prod_{i=1}^{n} [1 - \exp[-\frac{1}{\mu_i(u)} t]] =$$

$$\sum_{j=1}^{n} [\sum_{\substack{i_1,\ldots i_j=1, \\ i_1<\ldots<i_j}}^{n} (-1)^{j+1} \exp[-(\sum_{k=1}^{j} \frac{1}{\mu_{i_k}(u)}) t]],\ t \geq 0,\ u = 1,\ldots, z. \tag{6}$$

Moreover, the critical infrastructure unconditional lifetime $T(u)$ mean value in the safety state subset $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, can be calculated using the formula

$$\mu(u) = \int_0^{\infty} \boldsymbol{S}(t,u) dt,\ u = 1,2,\ldots,z, \tag{7}$$

where $\boldsymbol{S}(t,u),\ u = 1,2,\ldots,z,$ are given by (3)-(6) and particularly:
- for a series critical infrastructure

$$\mu(u) = (\frac{1}{\mu_1(u)} + \ldots + \frac{1}{\mu_n(u)})^{-1},\ u = 1,2,\ldots,z, \tag{8}$$

- for a parallel critical infrastructure

$$\mu(u) = \sum_{j=1}^{n} [\sum_{\substack{i_1,\ldots i_j=1, \\ i_1<\ldots<i_j}}^{n} (-1)^{j+1} (\sum_{k=1}^{j} \frac{1}{\mu_{i_k}(u)})^{-1}],\ u = 1,2,\ldots, z, \tag{9}$$

Further, we can evaluate the critical infrastructure intensities of ageing without climate-weather change impact

$$\boldsymbol{\lambda}^{\boldsymbol{0}}(t,u) = \frac{-\frac{d\boldsymbol{S}(t,u)}{dt}}{\boldsymbol{S}(t,u)},\ t \geq 0,\ u = 1,2,\ldots, z, \tag{10}$$

where in particular:
- for a series critical infrastructure

$$\boldsymbol{\lambda}^{\boldsymbol{0}}(t,u) = \sum_{i=1}^{n} \frac{1}{\mu_i(u)},\ t \geq 0,\ u = 1,2,\ldots, z, \tag{11}$$

- for a parallel critical infrastructure

$$\boldsymbol{\lambda}^{\boldsymbol{0}}(t,u) = (\sum_{j=1}^{n} [\sum_{\substack{i_1,\ldots i_j=1, \\ i_1<\ldots<i_j}}^{n} (-1)^{j+1} (\sum_{k=1}^{j} \frac{1}{\mu_{i_k}(u)}) \cdot$$

$$\cdot \exp[-(\sum_{k=1}^{j} \frac{1}{\mu_{i_k}(u)}) t]]) / (1 - \prod_{i=1}^{n} [1 - \exp[-\frac{1}{\mu_i(u)} t]]),\ t \geq 0,$$

$$u = 1,2,\ldots, z. \tag{12}$$

III. CLIMATE-WEATHER CHANGE PROCESS AT CRITICAL INFRASTRUCTURE OPERATING AREA

To model and identify the climate-weather change process at critical infrastructure operating area we need experts opinions to distinguish climate-weather parameters having negative impact on the considered critical infrastructure safety. Next, we have to collect the climate-weather change data from measurement points near the critical infrastructure operating area and perform steps and procedures from [4]-[5] to identify the climate-weather change process at the fixed area.

In this approach, we assume that the climate-weather change at the critical infrastructure operating area is described by κ, $\kappa \in N$, different climate-weather change processes $C^b(t)$, $t \geq 0$, $b = 1,2,\ldots, \kappa$. Each of them is taking w^b, $w^b \in N$, $b = 1,2,\ldots, \kappa$, different climate-weather states c_l^b, $l = 1,2,\ldots, w^b$, in the critical infrastructure operating area. Moreover, we can define them as semi-Markov processes with discrete states from sets $\{c_1^b, c_2^b, \ldots, c_{w^b}^b\}$, $b = 1,2,\ldots, \kappa$.

We describe the climate-weather change process $C^b(t)$, $b = 1,2,\ldots, \kappa$, by:

- the vector $[q_l^b(0)]_{1 x w^b}$, of the initial probabilities $q_l^b(0) = P(C^b(0) = q_l^b)$, $b = 1,\ldots, w^b$, of the process $C^b(t)$ staying at particular states c_l^b at the moment $t = 0$;
- the matrix $[q_{lk}^b]_{w^b x w^b}$ of the probabilities of transitions q_{lk}^b, $l, k = 1,\ldots, w^b$, $l \neq k$, of the process $C^b(t)$ from the states c_l^b to c_k^b;
- the matrix $[C_{lk}^b(t)]_{w^b x w^b}$ of the conditional distribution functions $C_{lk}^b(t)$, $l, k = 1,\ldots, w^b$, of the conditional sojourn times C_{lk}^b at the state c_l^b when its next state is c_k^b.

Next, we have to determine limit values of the process $C^b(t)$, $b = 1,2,\ldots, \kappa$, transient probabilities at the particular climate-weather states

$$q_l^b = \lim_{t \to \infty} q_l^b(t) = \lim_{t \to \infty} \mathrm{P}(C^b(t) = c_l^b),\ l = 1,\ldots, w^b, \tag{13}$$

using procedure from [4]. In [5] we assume that above limit values are coming from experts.

IV. CRITICAL INFRASTRUCTURE SAFETY WITH CLIMATE-WEATHER IMPACT

We consider that the single subsystem S_i, $i = 1,2,\ldots, n$, is affected only by the one climate-weather change process $C^b(t)$, $b = 1,2,\ldots, \kappa$, and each of processes $C^b(t)$ have influence on at least one subsystem S_i. The set of subsystems S_i, $i = 1,2,\ldots, n$, impacted by the process $C^b(t)$, $b = 1,2,\ldots, \kappa$, is denoted by $\{ S_{i_1^b}, \ldots S_{i_{k_b}^b} \}$, where

$$\{i_1^b, i_2^b, \ldots i_{k_b}^b\} \subseteq \{1,2,\ldots n\}, \; k_1 + k_2 + \ldots + k_\kappa = n. \quad (14)$$

To evaluate the critical infrastructure safety with considering climate-weather change impact is necessary to ask experts what climate-weather states c_l^b, $l = 1,2,\ldots, w^b$, $b = 1,2,\ldots, \kappa$, have destructive influence on the critical infrastructure operating at fixed area and determine with them values of coefficients $[\rho''^{(b)}_j(u)]^{(l)}$, $[\rho''^{(b)}_j(u)]^{(l)} \geq 1$, $u = 1,2,\ldots, z$, $j = i_1^b, i_2^b, \ldots i_{k_b}^b$, of the climate-weather impact on subsystems S_j intensities of ageing at the climate-weather change process operating area states c_l^b.

We denote the conditional lifetimes of subsystems $S_{i_1^b}, \ldots S_{i_{k_b}^b}$, in the safety state subset $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, while the climate-weather change process $C^b(t)$, $b = 1,2,\ldots, \kappa$, at the critical infrastructure operating area is at the climate-weather state c_l^b, $l = 1,2,\ldots, w^b$, by $[T''^{(b)}_j]^{(l)}(u)$, $j = i_1^b, i_2^b, \ldots i_{k_b}^b$.

Next, we assume that the coordinates of the vector

$$[S''^{(b)}_j(t,\cdot)]^{(l)} = [1, [S''^{(b)}_j(t,1)]^{(l)}, \ldots [S''^{(b)}_j(t,z)]^{(l)}], \; t \geq 0, \quad (15)$$

$j = i_1^b, i_2^b, \ldots i_{k_b}^b$, $l = 1,2,\ldots, w^b$, $b = 1,2,\ldots, \kappa$, are given by

$$[S''^{(b)}_j(t,u)]^{(l)} = P([T''^{(b)}_j]^{(l)}(u) > t \,|\, C^{(b)}(t) = c_l^b) = \\ = \exp[-[\rho''^{(b)}_j(u)]^{(l)} \lambda_j(u) t], \; t \geq 0, \quad (16)$$

$u = 1,2,\ldots, z$, $j = i_1^b, i_2^b, \ldots i_{k_b}^b$, $l = 1,2,\ldots, w^b$, $b = 1,2,\ldots, \kappa$, where $\lambda_j(u)$, $u = 1,2,\ldots, z$, $j = i_1^b, i_2^b, \ldots i_{k_b}^b$, are the intensities of the subsystems S_i departure from the safety state subset $\{u, u+1,\ldots, z\}$ without of the climate-weather change influence.

Further, we mark the unconditional critical infrastructure lifetime in the safety state subset $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, by $T''(u)$ and the unconditional critical infrastructure safety function by the vector

$$\boldsymbol{S}''(t,\cdot) = [1, \boldsymbol{S}''(t,1), \ldots, \boldsymbol{S}''(t,z)], \; t \geq 0, \quad (17)$$

with the coordinates

$$\boldsymbol{S}''(t,u) = P(T''(u) > t), \; t \geq 0, \; u = 1,2,\ldots, z, \quad (18)$$

If the critical infrastructure operation time θ is large enough, the coordinates (18) of the unconditional critical infrastructure safety function are given by

- for a series critical infrastructure,

$$\boldsymbol{S}''(t,u) \cong \prod_{b=1}^{\kappa} \Big(\sum_{l=1}^{w^b} q_l^b \Big(\prod_{j=1}^{k_b} \exp[-\lambda_{i_j^b}(u) [\rho''^{(b)}_{i_j^b}(u)]^{(l)} t] \Big) \Big), \quad (19)$$

- for a parallel critical infrastructure,

$$\boldsymbol{S}''(t,u) \cong 1 - \prod_{b=1}^{\kappa} \sum_{l=1}^{w^b} q_l^b \prod_{j=1}^{k_b} (1 - \exp[-\lambda_{i_j^b}(u) [\rho''^{(b)}_{i_j^b}(u)]^{(l)} t]), \quad (20)$$

where $t \geq 0$, $u = 1,2,\ldots, z$, and q_l^b, $b = 1,2,\ldots, \kappa$, $l = 1,\ldots, w^b$, are limit transient probabilities of the climate-weather change process $C^b(t)$ at the critical infrastructure operating area at the state c_l^b given by (13).

V. CRITICAL INFRASTRUCTURE SAFETY INDICATORS

The critical infrastructure unconditional lifetime $T''(u)$ mean value and the variance in the safety state subset $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, can be obtained from formulae [5], [8]

$$\mu''(u) = \int_0^\infty \boldsymbol{S}''(t,u) dt, \; u = 1,2,\ldots, z, \quad (21)$$

$$\sigma''^2(u) = 2 \int_0^\infty t \, \boldsymbol{S}''(t,u) dt - [\mu''(u)]^2, \; u = 1,2,\ldots, z, \quad (22)$$

where $\boldsymbol{S}''(t,u)$, $u = 1,2,\ldots, z$, are given by (17)-(18). Next, the unconditional critical infrastructure lifetimes mean values in particular safety states using (21) are can be evaluated from

$$\bar{\mu}''(z) = \mu''(z), \\ \bar{\mu}''(u) = \mu''(u) - \mu''(u+1), \; u = 0,1,\ldots z-1. \quad (23)$$

Moreover, the critical infrastructure risk function

$$\boldsymbol{r''}(t) = P(\boldsymbol{S''}(t) < r \,|\, \boldsymbol{S''}(0) = z) = P(T''(r) \leq t), \; t \geq 0, \quad (24)$$

is defined as a probability that the critical infrastructure is in the subset of safety states worse than the critical safety state r, $r = 1,2,\ldots, z$, while it was in the safety state z at the moment $t = 0$. Using the coordinate of the critical infrastructure

unconditional safety function given by (18) for $u = r$, the critical infrastructure risk function is given by

$$\boldsymbol{r''}(t) = 1 - \boldsymbol{S''}(t, r), \; t \geq 0. \tag{25}$$

Further, if the inverse function of the risk function $\boldsymbol{r''}(t)$ exists, we can calculate the moment τ the critical infrastructure risk function exceeds a permitted level δ using formula

$$\tau = (\boldsymbol{r''})^{-1}(\delta). \tag{26}$$

Remaining often used indicators are:
- the critical infrastructure intensities of ageing related to the climate-weather change impact

$$\boldsymbol{\lambda''}(t,u) = \frac{-\frac{d\boldsymbol{S''}(t,u)}{dt}}{\boldsymbol{S''}(t,u)}, \; t \geq 0, \; u = 1,2,\ldots, z, \tag{27}$$

- the coefficients of the climate-weather change impact on the critical infrastructure intensities of ageing

$$\boldsymbol{\rho''}(t,u) = \frac{\boldsymbol{\lambda''}(t,u)}{\boldsymbol{\lambda^0}(t,u)}, \; t \geq 0, \; u = 1,2,\ldots, z, \tag{28}$$

where $\lambda^0(t,u)$, $t \geq 0$, $u = 1,2,\ldots, z$, are the intensities of ageing of the critical infrastructure without of climate-weather impact,
- the coefficients of the climate-weather impact on the critical infrastructure without considering varying in time

$$\boldsymbol{\rho''}(u) = \frac{1/\mu''(u)}{1/\mu(u)}, \; u = 1,2,\ldots, z, \tag{29}$$

where $\mu''(u)$, $\mu(u)$, $u = 1,2,\ldots, z$, are respectively the critical infrastructure unconditional lifetime mean values in the safety state subset $\{u, u+1,\ldots, z\}$, $u = 1,2,\ldots, z$, with and without considering the climate-weather impact on the safety of the critical infrastructure given by (7) and (21),
- the coefficient of the critical infrastructure resilience to climate-weather change process impact

$$\boldsymbol{Rl''}(u) = 1/\boldsymbol{\rho''}(u), \; u = 1,2,\ldots, z, \tag{30}$$

where $\boldsymbol{\rho''}(u)$, $u = 1,2,\ldots, z$, are the coefficients of the climate -weather impact on the critical infrastructure given by (29).

VI. Conclusions

The proposed method of determination the critical infrastructure safety and other safety indicators with consideration of influence the climate-weather change process at its operating area is a very convenient when the real critical infrastructures and climate-weather change processes related to its operating areas are considered. Examples of the proposed method application to real critical infrastructures, i.e. the port oil piping transportation system and the Baltic oil terminal, are given in [2]-[3].

Acknowledgments

The paper presents the results developed in the scope of the EU-CIRCLE project titled “A pan – European framework for strengthening Critical Infrastructure resilience to climate change” that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653824, http://www.eu-circle.eu/.

References

[1] K. Kołowrocki, E. Kuligowska, J. Soszyńska-Budny and M. Torbicki, “An Approach to Safety Prediction of Critical Infrastructure Impacted by Climate-Weather Change Process”, International Conference on Information and Digital Technologies, Zilina, Slovakia, 2017, in press.

[2] K. Kołowrocki, E. Kuligowska, J. Soszyńska-Budny and M. Torbicki, “Safety and Risk Prediction of Critical Infrastructure Impacted by Climate-Weather Change Process with Application to Baltic Oil Terminal Critical Infrastructure”, International Conference on Information and Digital Technologies, Zilina, Slovakia, 2017, in press.

[3] K. Kołowrocki, E. Kuligowska, J. Soszyńska-Budny and M. Torbicki, “Safety and Risk Prediction of Port Oil Piping Transportation System Impacted by Climate-Weather Change Process”, International Conference on Information and Digital Technologies, Zilina, Slovakia, 2017, in press.

[4] K. Kołowrocki and J. Soszyńska-Budny, “Critical Infrastructure Operating Area Climate-Weather Change Process (C-WCP) Including Extreme Weather Hazards (EWH), C-WCP Model 3”, EU-CIRCLE Report D3.3-GMU0, 2016.

[5] K. Kołowrocki and J. Soszyńska-Budny, “How to Model and to Analyze Operation Threats and Climate-Weather Hazards Influence on Critical Infrastructure Safety - an Overall Approach”, EU-CIRCLE Report D3.3-GMU0, 2016.

[6] K. Kołowrocki and J. Soszyńska-Budny, “Integrated impact model on critical infrastructure safety related to climate-weather change process including extreme weather hazards”, Summer Safety & Reliability Seminars. Journal of Polish Safety and Reliability Association vol. 8, no. 2, pp. 85-95, 2017.

[7] K. Kołowrocki and J. Soszyńska-Budny, Reliability and Safety of Complex Technical Systems and Processes: Modeling – Identification – Prediction – Optimization, 1st ed. London: Springer-Verlag, 2011.

[8] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, “Critical infrastructure integrated safety model related to climate-weather change process application to port oil piping transportation system operating at land Baltic seaside area”, 27th ESREL Conference Proceedings, European Safety and Reliability Conference, Portoroz, Slovenia, 2017, in press.

[9] K. Kołowrocki, J. Soszyńska-Budny and M. Torbicki, “Integrated Model of Critical Infrastructure Safety (IMCIS) Related to Climate-Weather Change Process Including Extreme Weather Hazards (EWH), IMCIS Model 3”, EU-CIRCLE Report D3.3-GMU0, 2016.

Usability Evaluation of Web-Based GIS by means of a Model

Jitka Komarkova, Pavel Sedlak, Jakub Habrman, Ivana Cermakova
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
{jitka.komarkova, pavel.sedlak, ivana.cermakova}@upce.cz

***Abstract*—Spatial information is required by many users to support their decision-making. Application that provide spatial information may be quite complicated so evaluation of their usability is very important. Usability evaluation can help to improve their design or how to choose better application. Many various methods have been proposed. The paper describes a case study which uses combination of usability user testing and NGOMSL model to evaluate usability of chosen Web-based applications by means of calculating a utility. Case study shows that very similar results are obtained by both methods. Average time necessary for usability evaluation is shorter in the case of NGOMSL model.**

Keywords—Usability; NGOMSL; user testing; Web-based GIS

I. INTRODUCTION

Usability of a user interface is the only quality characteristics focused directly on users and on ability of an application to meet users' requirements and needs. Usability is partly overlapping accessibility [15]:

- 'Pure accessibility' problems only affect disabled people
- 'Pure usability' problems only affect non-disabled people
- 'Universal usability' problems affect both disabled and nondisabled people.

There are two basic ways of usability evaluation [5]: formative and summative. Formative evaluation methods are used during the design stage (before releasing the final product). Formative evaluation is focused on identification of usability problems that should be solved during the design. It is recommended to use a combination of expert-based and user-based inspection methods for evaluation. Summative evaluation is focused on a final product or on a comparison of competitive design alter-natives [5].

Various usability evaluation and testing methods have been used in many studies to improve quality or to compare products. Lately, there can be seen an attempt to improve and enrich methods, e.g. by means of fuzzy logic [7], by involving cognitive modelling [12] or by utilization of eye-tracking [11]. Today, the term usability is being replaced by terms quality in use or user experience, which adds emotions of users. A new instrument allowing measurement of satisfactions of video games users can be given as an example. It aims at measuring of user satisfaction and gaming experience [16].

The main aim of the paper is utilization of a model for usability evaluation along with a user testing to identify problems in usability. Based on the results, used methods of usability evaluation will be evaluated from the point of view of obtained results. The paper describes a case study evaluating a universal usability of three chosen Web-based GIS applications. NGOMSL model and user testing are used as usability evaluation methods and their results are compared.

Structure of the paper is as follows: the second chapter describes used data and methods. Next, hypotheses and research questions are stated. Next chapter describes the proposed procedure. The following chapter provides answers to the research questions and hypotheses. Conclusion follows.

II. USABILITY EVALUATION AND TESTING BASED ON MODELS

Many various factors can influence usability and many factors have to be taken into account while making design easier and more usable. Analytical models can help in both design and usability evaluation because of complexity of the problem of usability.

Task models describe how activities can be performed to meet user's goals by means of an application. Task models allow designers to describe both provided functions and interaction between user and user interface. Thus, importance of task models has been widely recognized for a long time [14].

Importance of goal-oriented modelling was emphasized by [2]. Goal-oriented models should describe users' knowledge necessary to operate a computer and actions, which have to be done by users to reach the goal. Models for usability evaluation should describe the following aspects [2]:

- External tasks to be completed by users with help of software
- User knowledge necessary to use computer
- User performance necessary to delegate tasks to computer, both mental and physical

- The computer system as tool to support user and as a result of design

Utilization of a software means that there is an interaction between a user and a software going through a user interface. An interaction evaluation model has to identify the following facts [3]:

- Accessibility conformance by an objective-oriented observation
- All types of accessibility and usability problems by testing a mixed panels of disabled and non-disabled users with a subjective-oriented analysis
- User satisfaction in order to complete the subjective-oriented observation.

Various analytical models have been proposed for usability evaluation: some of them are based on GOMS method, e.g. NGOMSL, CPM, CMN or KLM; another is based on PUM method. Utilization of natural language belongs to NGOMSL advantages [6]. GOMS analysis was used to evaluate user interface for disabled users [17] and for visually impaired people in a modified form of GOMS as well [19] so it can be understood as a robust tool.

Several software tools have been developed to support utilization of GOMS, e.g. QGOMS, CATHCI and GLEAN3 [1], [8].

Late, cognitive modelling was introduced as a new approach how to improve usability evaluation. Cognitive modelling has evolved from task-analysis, e.g. GOMS, to better predict human performance. The study [12] showed a suitability of cognitive modelling for improvement of usability of a user interface.

III. Case Study

The whole case study is focused on usability evaluation of chosen Web-based GIS applications of three regional authorities in the Czech Republic. The whole study is designed in a qualitative way, i.e. to identify serious problems in usability of the applications and to evaluate changes since the last evaluation. The study is designed to this purpose and it is split into 7 phases – see Figure 1. Existence of suitable and reasonably expensive procedures of usability evaluation of Web-based GIS is important because they are often outsourced. Quality monitoring of outsourced services is very important and requires suitable measures [17]. Consequently, usability of applications can influence success of e-Government services [10] from the point of view of acceptance of applications by citizens.

A. Phase 1 – Aim Definition and Choice of Evaluated Applications

Aim definition: to identify problems in usability of chosen Web-based GIS applications of regional authorities in the Czech Republic. As far as there are 14 authorities in the country and several studies were done previously, e.g. [9], it was decided to include only three of them in this case to see changes in design of applications. It was decided to include usability evaluation based on model because of scientific character of the study to allow identification of benefits of this type of methods.

Choice of applications for evaluation: at first, list of used software solutions is provided. Each software solution can provide similar applications so duplicate applications can be excluded.

Software solutions used to run Web-based GIS of particular regional authorities:

- ArcGIS Server: Jihocesky, Zlinsky, Stredocesky, Plzensky, Ustecky
- Hydrosoft Veleslavin: Pardubicky, Kralovehradecky
- Geocortex Essentials: Moravskoslezsky
- GeoMedia WebMap: Jihomoravsky
- Vars Brno: Karlovarsky, Vysocina
- T-mapy: Praha
- Maps only for download: Liberecky, Olomoucky

Next, a multi-criteria decision-making choice is done using the following criteria (all criteria are minimization ones):

- C1: Direct accessibility of administrative units map from the main menu
- C2: Similar range of functions of data which allows to use similar scenario for usability evaluation
- C3: Tools accessibility
- C4: Demandingness on users
- C5: Uniqueness of an application (its user interface)

An expert (one of authors) evaluated all applications and assigned points to all of them. The following application are chosen for usability evaluation during this case study:

- Pardubicky kraj (Hydrosoft Veleslavin)
- Moravskoslezsky kraj (Geocortex Essentials)
- Stredocesky kraj (ArcGIS Server)

They are based on different software solution including ArcGIS for Server, which is the most often used solution.

Usability evaluation is focused on the following common functions: scale; search tools; overview map; available layers; distance measurement tool; retrieval of an attribute information about a chosen feature; zooming and panning of a map; availability of information about an application.

Figure 1. Usability Evaluation Procedure (source: authors based on [4])

B. Phase 2 – Choice of Methods

The study aims to involving some representatives of users and wants to minimize costs so NGOMSL model is chosen. Main reasons of this choice are: a short time necessary to build the model; simplicity and understandability for participants; utilization of natural language; and suitability for evaluation of usability of Web-based GIS.

To verify results, a suitable method must be chosen which provides comparable results for reasonable costs. Time necessary to run particular operations is a key measure of NGOMSL so a method for results verification must provide time spent by users to fully or partly finish given tasks. Usability user testing is used because of the stated reasons as a verification method in this case.

C. Phase 3: Choice of Participants

Aim of the study is to identify serious problems in usability so the study follows Nielsen's approach – 3 to 5 participants are able to reveal about 85 % of problems in usability [13].

Choice of participants according to their availability belongs to often used methods. A representative set of citizens would be more interesting from the statistical point of view but this approach is quite expensive. So, some students of the faculty are used as participants in the case of this study. Only students before attending any GIS class are included into study. To speed up study progress students are involved in two roles: participants and evaluator's assistants. Amount of participants (and used order) must prevent an effect of learnability as well. It results into necessity of slightly more participants. In total, 12 participants are included to allow evaluation of all three web site by means of both methods in various orders.

D. Phase 4: Necessary Tools and Equipment Preparation

A computer lab with 21 seats and computers is used as a simple test-room. Thanks to its arrangement, 6 participants could obtain introductory information at the same time and they could simultaneously evaluate the applications. Because of a size of the room, they did not disturb each other. There are no requirements on a special room arrangement and on a video-recording.

Used PC: Dell Optiplex 380 Desktop with Intel Pentium Dual Core E5500 2.80 Ghz CPU, 2GB DDR3 RAM and Win 7 with SP1, 32-bit.

E. Phase 5: Model Proposal, Scenario Definition and Verification

The following main functions are included into usability evaluation: utilization of a scalebar, finding a feature by means of available search tools, utilization of data layers, distance measurement, identification of a feature and retrieval of attribute information about it, basic control functions: pan, zoom in, zoom out, and findability of an application by means of Web searching engine (e.g. Google).

The above listed aims are turned into set of particular activities to collect both performance and subjective (the last five items) data [4]:

- Findability of an application by means of Web searching engine
- Clearness of arrangements of main menu after starting Web-based GIS application
- Adjustment of scalebar/scale of maps
- Finding a required tool
- Utilization of particular layers (turning them on and off)
- Finding a required feature in map
- Utilization of zooming to display required area of interest
- Distance measurement between features
- Retrieval of attribute information about a feature
- Understandability of cartographic symbols
- Pleasance of a user interface
- Response time of an application
- Design (looking) of an application and its compliance with design of a "general" web site of a regional authority
- Necessity of plug-ins installation

Scenario for **user testing** contained 18 tasks in the beginning. Some tasks were excluded because of their difficulty or misunderstanding. On the other hand, there was added one task focusing on search tools. Some tasks were re-formulated and the Excel forms with tasks were prepared. The final list, after pilot testing, contains 14 tasks [4]:

1. Use Google to find the Web-based GIS application and its starting page. Choice of keywords is up to you
2. Find a map of administrative division of the region and open it
3. Set a scale to approx. 1 : 200 000
4. Find a tool "Undo/Back" for one step back action within map tools

5. Display layer containing municipalities with extended competence and turn all other layers off
6. Use any way to display name of one of municipalities with extended competence at a scale approx. 1 : 500 000
7. Find and use a tool to display full extent of the region (display all municipalities of the region)
8. Measure a direct distance between any two municipalities with extended competence
9. Use a suitable tool to retrieve an information about features – find a code of a regional town
10. Turn on an orthophoto map
11. Find a legend for the layer of municipalities with extended competence
12. Turn on an overview map (a small map placed at the bottom) and use it to center a big map to the regional town at a scale approx. 1 : 500 000
13. Display names of streets in the regional town
14. Use a search tool to find regional town in the map

NGOMSL model is proposed in accordance with user testing scenario. Tasks are transformed into particular activities.

Task 1

1.1 Type search keywords into Google field to start searching for starting page of the application

1.2 Enter the main menu of the application

1.3 Document tasks fulfilment

Task 2

2.1 Find a link to open a map of administrative division of the region

2.2 Enter of the GIS application, namely map of administrative division, by means of the found link

2.3 Document tasks fulfilment

Task 3

3.1 Find a scale of the map

3.2 Set a scale of the map to 1 : 20 000. If there is no scale, set 1 km at scalebar

3.3 Document tasks fulfilment

Task 4

4.1 Find a toolbar containing tools to control the map (pan, zoom, refresh functions)

4.2 Find a tool “Undo/Back”

4.3 Document tasks fulfilment

Task 5

5.1 Find a list of available layers

5.2 Turn on the layer containing municipalities with extended competence

5.3 Turn off all other layers

5.4 Document tasks fulfilment

Task 6

6.1 Pan the map to focus it on any of municipalities with extended competence

6.2 Set a scale of the map to 1 : 50 000. If there is no scale, set 2 km at scalebar

6.3 Display name of the above focused municipality with extended competence

6.4 Document tasks fulfilment

Task 7

7.1 Find tools for zooming in and out and tools for changing a size of the map

7.2 Use one of the tools to display all municipalities with extended competence in the map

7.3 Document tasks fulfilment

Task 8

8.1 Find a tool for distance measurement

8.2 Measure distance between two chosen municipalities with extended competence

8.3 Document tasks fulfilment

Task 9

9.1 Find a tool for obtaining descriptive information about features

9.2 Find a code of the regional town

9.3 Document tasks fulfilment

Task 10

10.1 Find a tool for changing background maps

10.2 Switch a background map to orthophoto map

10.3 Document tasks fulfilment

Task 11

11.1 Find a legend for all layers

11.2 Find a symbol used for municipalities with extended competence

11.3 Document tasks fulfilment

Task 12

12.1 Find a smaller map on the screen – an overview map

12.2 Set a scale of the map to 1 : 50 000. If there is no scale, set 2 km at scalebar

12.3 Use the overview map to display name of the regional town

12.4 Document tasks fulfilment

Task 13

13.1 Find a tool to display street names

13.2 Find name of a street in map window

13.3 Document tasks fulfilment

Task 14

14.1 Find a suitable tool to search for names of municipalities

14.2 Find and display regional town in the map by means of this tool

14.3 Document tasks fulfilment

Total performance time is calculated as a sum of: (time of NGOMSL tasks * 0.1 s), operators, mental operators, and response time of system. For example, an average time for operator 'find' is 18.6 s.

The last step of preparation of both usability user testing scenario and NGOMSL model is pilot testing. Results concerning scenario for user testing are mention above. Pilot testing is used as a source of one more important information for the next step: a maximum time available for users to try to solve tasks. After reaching deadline, the task is marked as not fulfilled. The deadline is set as tree-times longer time than time necessary for pilot testing but no longer than 300 s (5 min) to keep reasonable times. E.g. for distance measurements, the pilot testing time was 47 s, so deadline is set to 141 s. [4]

F. Evaluation Planning

Three Web-based GIS applications represent subjects of evaluation. Two methods are chosen for the evaluation so evaluation must be planned in a way, which prevents learnability effect. The evaluation plan is described in the Table 1

TABLE I. PLAN FOR USABILITY EVALUATION

	Participants	Order of Evaluated Applications		
		Pardubicky	Moravsko-slezsky	Stredoces-ky
User Test.	Participant 1 and 2	1	2	3
	Participant 3 and 4	3	1	2
	Participant 5 and 6	2	3	1
NGOMSL model	Participant 8 and 9	1	2	3
	Participant 7 and 11	3	1	2
	Participant 10 and 12	2	3	1

G. Evaluation Itself

Usability user testing is done in couples. A participant fulfils tasks, an inquirer measures times necessary to finish tasks. In the case of exceeded time, he stops participant. Participant makes printscreens and enters answers and times into Excel sheet (see Fig. 2.) Next, participant fulfils subjective evaluation of all application in Excel. Scale from 1 (the best) to 5 (the worst – fully unsatisfied) is used. As the last step, participant is asked to compare importance of evaluated functions (by pairwise comparison).

Similar approach is used for NGOMSL model based evaluation. All tasks and questionnaires are provided in Excel.

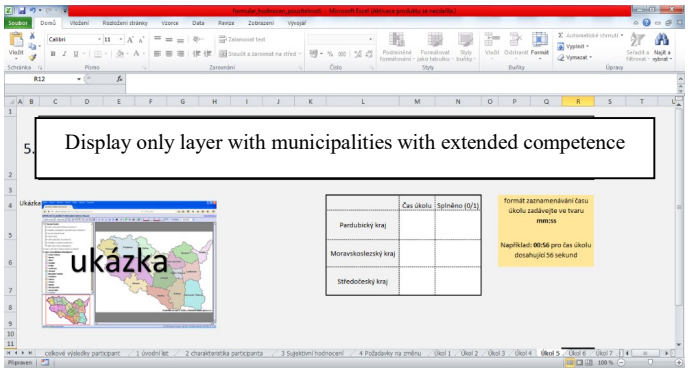


Figure 2. Environment for Usability User Testing (source: [4])

H. Results Interpretation

1) Usability User Testing

At first, times necessary to finish particular tasks must be summarized and an average is calculated. An average time necessary to evaluate one application is 11 min 36 s in this case study. The shortest time was necessary to finish task 1 (27.9 s), the longest time was necessary to finish task 13 (93.5 s)

Next, a multi-criteria approach is used to set weights for all particular criteria to allow to calculate a utility at the end. Evaluation obtained from participants by pairwise comparison is used in this step. Weight of the group of subjective criteria is 0.5, weight of the group of performance criteria is 0.5 as well. Then, a utility per user per application is calculated and finally, an average utility per application is calculated. The final utility is as follows: Moravskoslezsky – 0.33 (see Fig. 3); Pardubicky – 0.19 (see Fig. 4); Stredocesky – 0.18 (see Fig. 5).

2) NGOMSL Model

Time necessary for evaluation is slightly shorter than in case of usability user evaluation. Time calculated in advance is 9 min 1 s. In reality, the following times were necessary: 7 m 31 (Moravskoslezsky), 9 min 22 s (Stredocesky), 9 min 28 s (Pardubicky) in comparison with above mentioned 11 min. [4]

I. Summary

Both methods provided very similar results concerning necessary times and identified problems in usability. Distance measurements (66.7 s in user testing and 62 s in NGOMSL), search tools (64.8 s in user testing and 30.8 in NGOMSL) and

displaying names of streets (93.5 s in user testing and 76 s in NGOMSL) can be given as the most serious problems in usability [4]. In fact, they belonged to the identified problems in usability during previous study as well [9]. The highest number of problems in usability was found in the case of Pardubicky region application (see Fig. 4), which has not been change since the previous study. Its most serious problem in usability is utilization of Java applet, which in fact makes utilization of this application impossible because it requires users to allow dangerous plugins in web-browser. Setting a map scale to approx. 1 : 200 000 was the easiest task (29.4 s in user testing and 16 s in NGOMSL) [4].

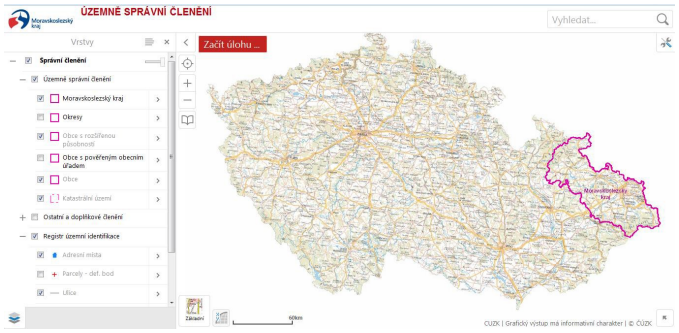


Figure 3. Web-Based Application of Moravskoslezsky region

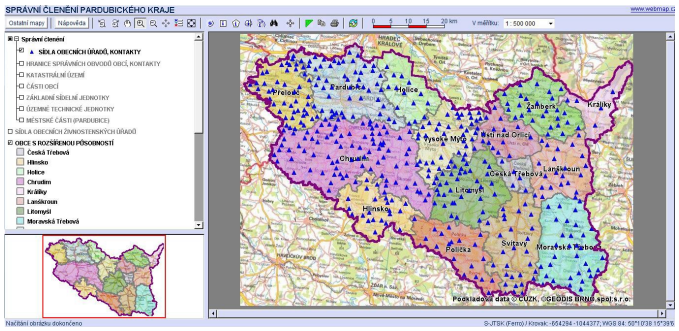


Figure 4. Web-Based Application of Pardubicky region

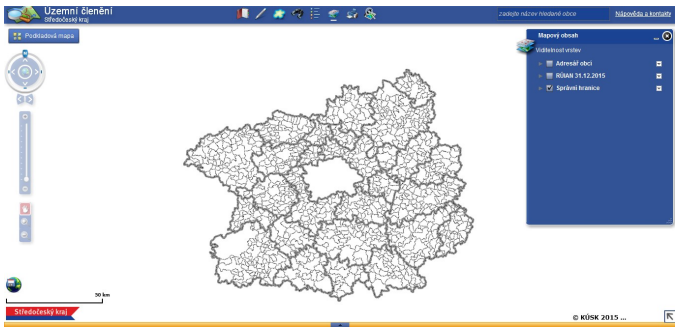


Figure 5. Web-Based Application of Strdocesky region

IV. Conclusion

Usability evaluation can significantly help designers to design a high-quality user interfaces to support users and their work. It is very important especially in the case of Web-based GIS because they are designed for end-users without GIS knowledge as a part of e-Government service. It means that application must prevent users to make mistakes. Usability user testing includes representatives of users which is very important on one side but it is very time and costs demanding on the other side. Combination of more methods can bring results in a more efficient way.

The case study shows that NGOMSL model can be successfully used for usability evaluation. It is able to provide similar results in a shorter time in comparison with a traditional usability user testing.

Acknowledgment

This research is supported by University of Pardubice, SGS_2017_17 project.

References

[1] L.K. Baumeister, B.E. John, M.D. Byrne, “A comparison of tools for building GOMS models”, Proceedings of the CHI 2000 Conference on Human factors in computing systems, The Hague, The Netherlands, April 1-6. New York: ACM, 502-509, 2000.

[2] G. De Haan, G.C. Van Der Veer, and J.C. Van Vliet, “Formal modelling techniques in human-computer interaction” [online]. Acta Psychologica 78, pp. 27-67, 1991. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.155.3385&rep=rep1&type=pdf.

[3] S. Federici, S. Borsci, “Usability evaluation: models, methods, and applications” [online]. J.H. Stone, M. Blouin, Eds. International Encyclopedia of Rehabilitation. URL: http://cirrie.buffalo.edu/encyclopedia/en/article/277/

[4] J. Habrman, “Usability evaluation of Web GIS applications based on model”, Master thesis, Pardubice: University of Pardubice, 2016 [Available in Czech only, title “Hodnocení použitelnosti webových GIS aplikací založené na modelu”].

[5] H.R. Hartson, T.S. Andre, and R.C. Williges, “Criteria for Evaluating Usability Evaluation Methods”, International Journal of Human–Computer Interaction 15(1), pp. 145–181, 2003.

[6] M. Hub, and B., Musilová, “Comparison of usability evaluation of public administration webpages by user testing and by analytical models”. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration, 23 (37), pp. 39-47, 2016.

[7] M. Hub, and M. Zatloukal, “Usability Evaluation of Selected Web Portals”. Recent advantages in applied informatics and communication: Proceedings of the 9th WSEAS International Conference on APPLIED INFORMATICS AND COMMUNICATIONS (AIC'9). 1st edition. Moscow: WSEAS Press, pp. 259-264, 2009.

[8] D. E. Kieras, S. D. Wood, K. Abotel, and A. Hornof, “GLEAN: a computer-based tool for rapid GOMS model usability evaluation of user interface designs”, UIST (User Interface Software and Technology): Proceedings of the ACM Symposium. pp. 91-101. ACM, 1995.

[9] J. Komárková, M. Jedlička and M. Hub, “Web-based Geographic Information Systems and their Usability”, Recent advances in applied computer science. Athens: WSEAS Press, pp. 97-102, 2009.

[10] H. Kopáčková, “The issue of measuring e-government success in context of the Initiative 202020”, Proceedings of the 21th International Conference Current Trends in Public Sector Research. Brno: Masarykova univerzita, pp. 41-49, 2017.

[11] V. Kubelka, and Z. Dobesova, “Eye-Tracking Testing of GIS Interfaces”, Proceedings of 15th International Multidisciplinary Scientific GeoConference Volume I, STEF92 Technology Ltd. Sofia, Bulgaria, pp. 585-592, 18-24 June 2015.

[12] X. Li, änd M. Gunal, “Exploring cognitive modelling in engineering usability design”, Journal of Engineering Design 23 (2), pp. 77-97, 2012.

[13] J. Nielsen, *Why You Only Need to Test with 5 Users [online],* URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

[14] F. Paternò, “Model-Based Design and Evaluation of Interactive Applications”, Springer, 2012.

[15] H. Petrie, and O. Kheir, “The relationship between accessibility and usability of websites”, Proceedings of the SIGCHI conference on Human factors in computing systems. San Jose (CA): ACM. pp. 397-406, 2007.

[16] M.H. Phan, J.R. Keebler, and B.S. Chaparro, “The Development and Validation of the Game User Experience Satisfaction Scale (GUESS)”, HUMAN FACTORS 58(8), pp. 1217-1247, 2016.

[17] M. Schrepp, “GOMS analysis as a tool to investigate the usability of web units for disabled users”, Universal Access in the Information Society 9 (1), pp. 77-86, 2010.

[18] S. Simonova, “Identification of IT-Service Metrics for a Business Process when Planning a Transition to Outsourcing”, In International Conference on Information and Digital Technologies (IDT) 2016, pp. 274 - 279, 5-7 July 2016.

[19] H. Tonn-Eichstädt, “Measuring website usability for visually impaired people – A modified GOMS analysis”, ACM SIGACCESS Conference on Assistive Technology. New York: ACM Press, pp. 55–62, 2005.

Smart City Concept as Socio-Technical System

Hana Kopackova, Petra Libalova
Institute of System Engineering and Informatics
University of Pardubice - FES
Pardubice, Czech Republic
hana.kopackova@upce.cz

***Abstract* — Smart City is a very popular concept as the advancement in technology bring new possibilities how to make life in cities easier and more comfortable. Whereas the technical part of Smart City concept is highly discussed in research articles, impact on social part of this system is not in the center of interest. This article aims to contribute to the complex view of this concept. First we discuss different concepts close to Smart City and based on literature review to evaluate involvement into technical or social research areas. Then we use Leavitt's model adapted for Smart Cities to discuss all interconnected elements. Last part describe plans for Pardubice Smart City in the view of Leavitt's model.**

Keywords — smart city; social system; technical system; Leavitt; technology

I. INTRODUCTION

The concept of Smart City comes as the response to two independent phenomena – necessity to solve problems of urbanization and possibility given by the development of new technologies. As such, technology can serve as driver to solve old problems by new ways, and also as an innovation that brings new options how to enhance sustainable development. Smart City can be viewed also as an innovation because it brings new technology, new processes, new participatory policy, new management methods, etc. In sum, Smart City represents changes in governance and daily life.

Smart City concept is not new if we think in the dimension of alignment between technology and city governance. Different other concepts have appeared in the past. Some of them replacing one another and some coexisting in the same time and space. More detailed discussion about different concepts will follow in next chapter.

Definitions of Smart City are not unified instead we can find many different definitions, models and frameworks. We introduce four representative examples of definitions below.

"A Smart City places people at the center of development, incorporates Information and Communication Technologies into urban management, and uses these elements as tools to stimulate the design of an effective government that includes collaborative planning and citizen participation. By promoting integrated and sustainable development, Smart Cities become more innovative, competitive, attractive, and resilient, thus improving lives."[1]

"*A smart sustainable city is an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, environmental as well as cultural aspects."*[11]

"What makes a city a smart city is its use of Smart Computing to deliver its core services to the public in a remarkably efficient manner."[28]

"Smart city is an instrumented, interconnected and intelligent city. Instrumentation enables the capture and integration of live real-world data through the use of data-acquisition systems. Interconnection means integration of obtained data. Intelligence cover analysis of this interconnected information that must yield new insights to drive decisions and actions that improve process outcomes or system, organization, and industry value chains."[10]

When comparing those definitions we can see that all of them include two systems that must be aligned together in order to get real Smart City – social system (eg. people, organizations, social structure, government) and technical system (eg. hardware, software, networks, intelligence methods). The difference lies in the emphasis of each system. In chapter three, we discuss the concept of Smart City from the perspective of socio-technical system.

Last part of paper includes case study of city Pardubice with focus on innovative potential and political issues. Political issues hidden in development of Smart Cities can represent threats to otherwise very successful concept. Insufficient openness of government decisions, suspicion of corrupt behavior, political pressures and insufficient transparency can destroy all efforts to make city smarter.

II. SMART CITY AND IT'S CLOSE CONCEPTS

While the term Smart City is very popular in these days, so that it is almost buzz word, concept of implementation of information technologies into city governance and everyday life in the city is not new. First attempts to digitize the city date back to the beginning of the millennium with different concepts: Wired City, Digital City, U-City, Information City, Efficient City, Intelligent City or Cyber City. Although research articles [17], [26], [6], [7] mention this broad spectrum of concepts, referring to mutual connectedness, complicated usage and confusion in definitions, to our knowledge, no research paper evaluated popularity of each concept and focus on technical or social part of this socio-technical system. The aim of this chapter is to clarify this issue by providing literature review. Search for papers was held in March 2017 at Web of Science search engine. Searched phrases were: "smart city", "digital city", intelligent city", "U-

city", "efficient city", "cyber city", "information city", and "wired city".

Figure 1 shows results of search. It is quite obvious that the concept of Smart City is the most popular with 1972 hits. Second popular concept is Digital City. It is popular even now, when the term Smart City is highly used. Since 2004, each year have brought about 20 hits for Digital City. This is the example of terms coexisting together. Other terms were used only for some time or by some authors.

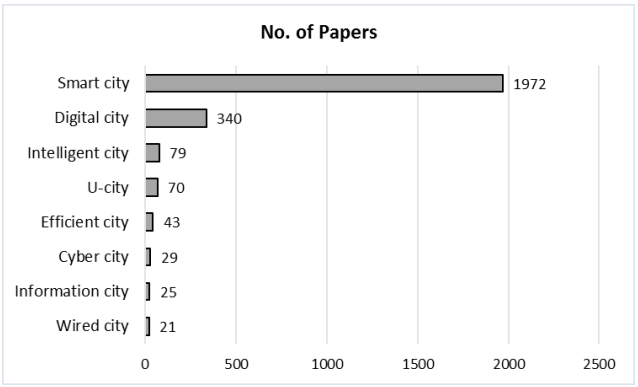


Figure 1. Results of search for Smart City concepts

Popularity of terms was the first studied topic. Then we focused on research areas in which those papers were published. If the Smart City concept cover two systems, social and technical, then we should find articles included in both research areas. As technical research area were classified those categories: Computer Science, Engineering, Telecommunications, and Automation Control Systems. As social research area were classified those categories: Urban Studies, Business Economics, Social Sciences Other Topics, Public Administration, and Operations Research Management Science. Table 1 shows results given in percentages of number of papers. Results speak clearly, most scientific papers are categorized into Computer Science category and Engineering whereas topics from social research area are represented significantly less.

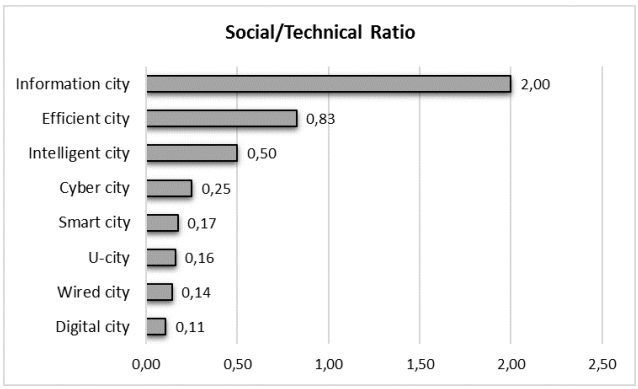


Figure 2. Classification of research papers – social to technical research areas

Socio/technical ratio for each term is depicted in figure 2, showing that only Information City concept was aimed more at the social context. Undervaluation of social dimension of Smart Cities was also mentioned in Nam and Pardo [17], however, this situation is awkward in the view of critical success factors of information systems that emphasize the importance of social dimension (eg. planning, leadership, communication, user readiness). Moreover, each change in technology invoke changes in social system as both are interconnected, that is why the social part of socio-technical system should not be neglected. Next chapter will discuss both parts of socio-technical system and their connectedness in more detail.

TABLE 1. RESEARCH AREAS OF STUDIED PAPERS

		TECHNICAL RESEARCH AREAS (%)				SOCIAL RESEARCH AREAS (%)					
	No. of papers	COMPUTER SCIENCE	ENGINEERING	TELECOMMUNICATIONS	AUTOMATION CONTROL SYSTEMS	URBAN STUDIES	BUSINESS ECONOMICS	SOCIAL SCIENCES OTHER TOPICS	PUBLIC ADMINISTRATION	OPERATIONS RESEARCH MANAGEMENT SCIENCE	SOCIAL/TECHNICAL RATIO
Smart city	1972	48,23	38,39	16,94	2,69	5,27	4,97	3,14	2,94	2,18	0,17
Digital city	340	47,35	35,29	9,12	4,41	0,00	6,47	3,82	0,00	0,00	0,11
Intelligent city	79	26,58	18,99	6,33	6,33	6,33	12,66	5,06	5,06	0,00	0,50
U-city	70	65,71	35,71	20,00	2,86	2,86	7,14	0,00	2,86	7,14	0,16
Efficient city	43	16,28	30,23	6,98	0,00	13,95	13,95	6,98	4,65	4,65	0,83
Cyber city	29	41,38	20,69	6,90	0,00	10,34	0,00	0,00	6,90	0,00	0,25
Information city	25	20,00	4,00	0,00	0,00	12,00	12,00	0,00	24,00	0,00	2,00
Wired city	21	4,76	19,05	9,52	0,00	0,00	4,76	0,00	0,00	0,00	0,14

III. SMART CITY VIEWED AS SOCIO-TECHNICAL SYSTEM

Smart City is sometimes called system of systems [13], [8], [19], [24], or even an organic whole. These statements highlight the importance of system thinking when conceptualizing Smart City. Mostly this system view is used to describe functions of Smart City. For example definition in [9] describe six key fields of urban development: smart economy, smart mobility, smart environment, smart people, smart living, and smart governance. According to [8] Smart Cities "… are based on a number of core systems composed of different networks, infrastructures and environments related to their key

functions: city services, citizens, business, transport, communication, water and energy". Kanter and Litow [13] encompass city's systems as transportation, commerce, food, energy, safety, education, health care that are organically fused.

Unlike presented definitions, we do not try to find original system view, instead we will use traditional socio-technical model, proposed by Leavitt in 1965. Leavitt's model [16] of organizational change comprises four concepts tightly connected to each other – people, task, structure, and technology. Three concepts being from social area (people, task, structure) and one from technical (technology). The model explains that the change in one part of system shifts the stability of the organization, which influences or even may have a negative impact on other parameters.

Task is the purpose of organization, why the organization exists (e.g. produce goods, provide a service, teach students, serve public). The concept of people covers all humans in the organization who contribute to the realization of task. Structure comprises everything what defines the organization - formal and informal structures, processes, ways of communication, and so on. The last concept is technology – all tools, machinery, information technology, mobile technology, etc. Leavitt's model explains why introduction of new technology may not produce the desired effect due to the connectedness with other elements. In figure 3 are depicted all parts described above with addition of environment concept.

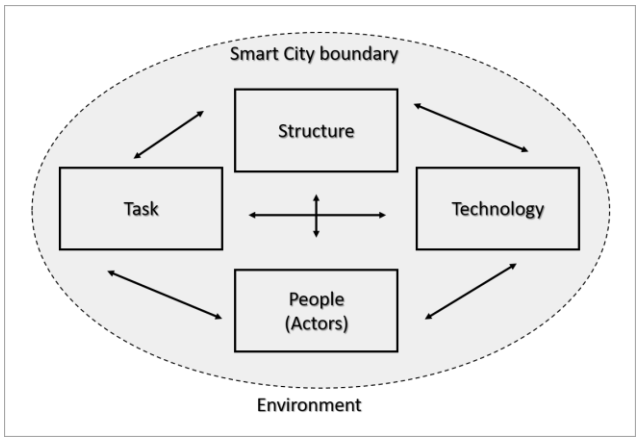


Figure 3. Adaptation of Leavitt's model [16] for Smart City

IV. RESULTS – APPLICATION OF LEAVITT'S MODEL

Now we would like to explain each part of model in relation to Smart City concept. When city starts some smart initiative it brings changes to the daily life of its citizens and the way the city is governed. In the rest of this chapter we will first introduce contemporary technology suitable for smart city initiatives. Then we will track changes invoked in social parts of the model.

A. Technology

The development of information and communication technologies is so fast, that enumeration of available technologies would be obsolete even at the time of publication and also too expansive. Instead we use categories of usage based on [3] and examples in these categories.

Network infrastructure – represents fundamental technology level necessary to support digital applications. Existence and availability of fixed or mobile broadband internet networks is necessary to send and receive data. Examples in this category are cable transmission, optical fiber, and wireless networks (Wi-Fi, 3G, 4G, or radio).

Sensor devices – monitor and capture data and send them through network infrastructure to the control centers. Multiple sensors can be attached to an object or device. Sensors along with the network infrastructure form the basis in the construction of a Smart City. Internet of Things [27], [22], [29] is the global term including physical devices, vehicles, buildings and other objects that can be sensed or controlled remotely across existing network infrastructure. Examples of sensor devices are motion sensors on the streets, cameras at intersections, smart sensors in lampposts, detection of rubbish levels in containers, monitoring of parking spaces, energy consumption monitoring and management, or security systems that allow facial recognition of people in the crowd.

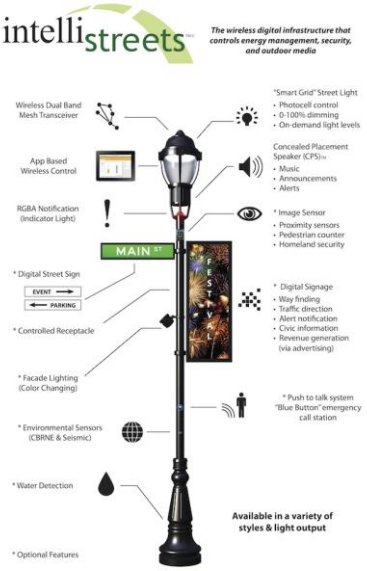


Figure 4. Intelligent lamppost [4]

Communication interfaces – are collaborative platforms (mobile applications, web portals, or services) used to send and receive data within participatory governance framework. Using of such technologies enhance openness and transparency of government. Examples of such platforms are community networks, newsgroups, online forums, e-mail lists, pools, voting tools, or mobile apps for reporting on the city's infrastructure problems and public safety risks.

Control centers – are places where all gathered data are processed, analyzed, and stored. The main objective of control center is to allow quick and effective decision-making by system's operators, therefore special methods to control large amount data, called Big Data, are necessary. Integrated Control

Centre should include for example video surveillance system, video management software, video analytics, decision support system based on GIS, security and emergency management system, or workforce planner.

Figure 5. Control center in Rio de Janeiro [23]

Described technology classification for Smart Cities is temporal, we expect that in future will appear new categories as the development will continue. Nevertheless, requirements [24] on the quality of technology will remain the same. There are two types of technology quality – inherent and emerging. At first, hardware, software and network have quality dimensions inherent to its parts (e.g. reliability, compatibility, security, maintainability, etc.). We can measure them on each part independently without integration together. Secondly, technology has quality dimensions emerging during the use (effectiveness, efficiency, satisfaction, freedom from risk, and context coverage), which suppose integration of all parts together. Only then will these quality dimensions appear. Technology quality is fundamental prerequisite for the success of Smart Cities but it is not the only one. In accordance with changes of technology must come changes in social part of Smart City system as described in Leavitt's model.

B. Task

This part of model address the problem of scope, functions that city performs. Main function of traditional city is to be safe, livable place that offer its citizens services provided by city government. Essential services cover:

- city administration (planning and monitoring, financial management, legislation)
- management of utilities (water supply and waste removal),
- economic development (support of business growth by eg. special loans, tax incentives, strategic zoning),
- public safety (eg. fire and rescue services, police, emergency services, animal control),
- management of public buildings (provide and maintain public housing, maintain public buildings),
- city transport (construction, operation and maintenance of infrastructure, public transportation, traffic management and support to non-motorised transport),
- health services (provision of medical care in hospitals, emergency medical services, identifying and monitoring public health issues),
- education (funding for public schools, setting the strategy and education policy),
- environment protection (reducing climate impact, efficient energy use, development and maintenance of green spaces).

If we know what functions a traditional city should provide, we can look at how these functions will be affected by the introduction of new technologies. What functions would Smart City perform as the response to changes in technology? The answer cover two parts – firstly, what types of services cities try to offer in order to be smart and secondly, how these services will be delivered. Both problems are highly studied. This paper will follow division by Washburn et al. [28] and explain what changes in particular type of service can happen with smart technology.

Smart transportation – reduce traffic congestion while encouraging the use of public transportation. Promotion of public transportation and cycling reduce number of cars. Electric cars, especially when shared make transportation greener. Mobile applications with actual traffic information, closures and restrictions help drivers spend less time driving.

Smart utilities – deliver only as much energy or water as is required while reducing waste. A smart utility infrastructure make existing systems efficient whereas finding new ways of producing and delivering water, gas, and electricity, which ensure green and sustainable development.

Smart education – cover increase in access and enrollment ratio, improvement in quality, and reduction of costs. Data literacy is now part of necessary education. It means the ability to use data as part of complex reasoning and to accurately critique and interpret data to ask and answer meaningful questions. Smart education use digital content and collaboration technologies.

Smart healthcare – increases the availability of more rapid and accurate diagnosis. Patients get online access to their electronic health enabling them to control provided care. Sharing of data among healthcare providers simplifies diagnosis of disease and prevents multiple prescribing of drugs. The communication platform enables quick response to emergency services. New technologies make it possible to get more accurate information about individual risks that can change involvement in healthcare insurance. Videoconferencing technologies enable remote medical care for those who can't travel to hospitals.

Smart public safety – uses real-time information to respond immediately to emergencies and threats. Police, fire, and other public safety personnel need to respond more quickly to

emergency situations, which needs effective and smart emergency management solutions. Data collected from city cameras can serve as the source of real-time information to fire and police departments. Social networks can be used for communication with citizens by push messages or citizens can provide information about the current situation and thereby help public safety personnel manage the situation. Moreover, to manage emergency situation, data must be analyzed to provide the decision maker with notifications, alarms and suggestions to support the decisions. Data about emergency situation should be also stored even when the situation disappear in order to make plans for possible future emergency situation. There has been a significant progress in sensing for security applications and in the analysis and processing of sensor data. The third generation of surveillance systems addresses topics such as multi-sensor environments, video surveillance, audio surveillance, wireless sensor networks, distributed intelligence and awareness, and architecture and middleware. An attempt should be made to construct a multi-sensor distributed intelligent surveillance system that functions at a relatively high level, capturing alerting situations with a very low false alarm rate.

Smart real estate – reduce operating costs, increase the value, and improve occupancy rates. Smart public buildings save public money and smart residential homes make living in the city convenient for citizens.

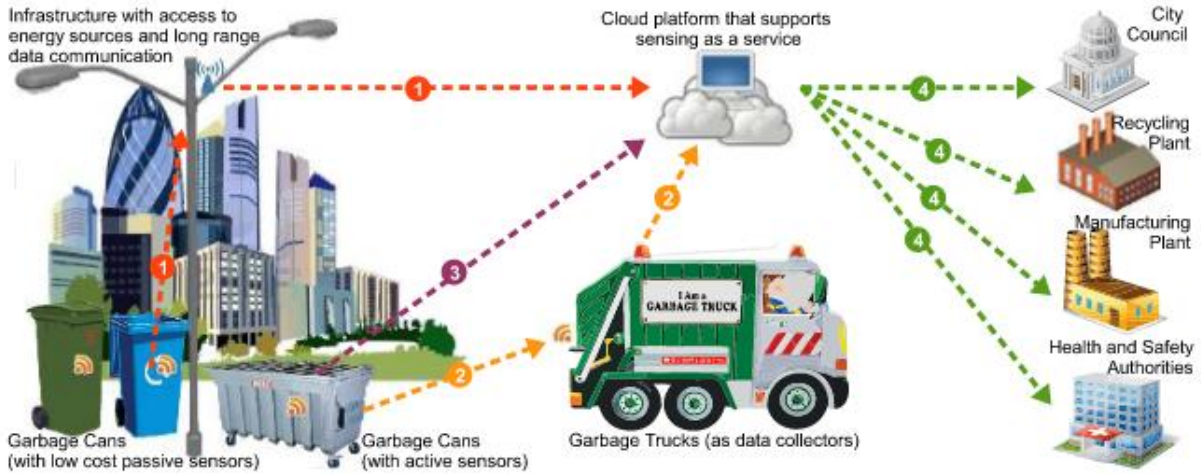


Figure 6. Example of smart city innovation in waste management [21]

Example of smart city innovation in waste management can be seen in figure 6. Garbage cans with low cost passive sensors can use nearby infrastructure (eg. lamppost) to send data to the cloud. Garbage cans with active sensors can use at least two ways for sharing data: garbage trucks as infrastructure or 3G wireless mobile telecommunications technology for direct communication. Cloud platform that supports sensing as service [1], [14], [17] then plays crucial role in sharing information with consumers (city council, recycling plant, manufacturing plant or health and safety authorities).

New technologies can bring improvement in existing categories of services as they were described here (changing how question) or even form new category of service (changing what question). Goals should be then moved to the higher level to obtain most possible benefits and to cover the costs of technology [2]. On the other hand, urbanization can induce new unexpected problems that would call for new functions and those functions should be supported by technological changes.

C. Structure

Structure in Leavitt's model represents hierarchy, formal structure, communication channels, and informal structure. Using this model to describe changes from traditional city to Smart City would need broader understanding of structure. Figure 7 show three most important changes in structure: participatory government, whole-of-government, and open government.

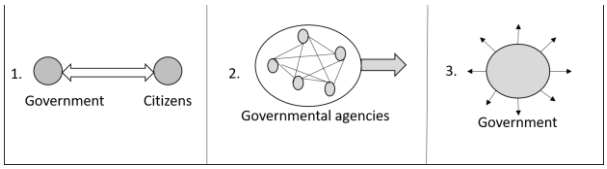


Figure 7. Changes in structure invoked by technology changes (participatory government, whole-of-government, and open government)

First difference between traditional and Smart City can be found in assigning of roles. Traditional structure expect that citizens engage in public affairs only at election time and then elected representatives take care of the city with minimal interference of citizens. Information technologies brought possibility of higher citizen participation with minimal demands and costs. City governments can use information technologies to engage citizens into decision-making in many ways. Crowdsourcing can be for example used for identification of problems and requirements. Issue tracking systems that collect, aggregate, and follow up on citizen-

identified problems are increasingly available. Crowdsourcing can be used also for prioritization and finding solution. Online forums and voting systems give citizens possibility to comment on current events and vote for some alternative. However, not all citizen participation must be efficient, Economic and Social Council of United Nations defined conditions under which the participatory governance can be effective.

"Participation is regarded as "effective" when it yields greater influence for ordinary people, especially the poor and socially excluded. That influence can be over government actors, politicians and bureaucrats, and their own destinies. In order to be "effective", participation must not only reveal the preferences of ordinary people but also enable those preferences to shape outcomes. There must be processes and forums, formal or informal, through which they can voice their concerns and affect decisions. Ordinary citizens should be sufficiently confident and capable of contacting and obtaining responses from bureaucrats, elected representatives and other public agents. They should be able to have an impact by lobbying or demonstrating collectively. When those conditions are manifest, "effective" participation and empowered participatory governance is a reality."[5]

Second difference between traditional and Smart City is anchored in organizational change of government bodies. Today's term whole-of-government suggests that the main change is that government departments and agencies will be working together as a team for the desired outcome. Smarter government means communication, collaboration, and coordination across departments to be more effective and to be more citizen-centric. Necessary condition for this change is sharing and integration of information and knowledge. Changes in technology, especially network infrastructure enhance this change but it must be followed by reengineering of processes.

Third possible change that brings new technology is transparency and openness of government. Government agencies collect large amount of data that serve as the source for decision-making. But some of those data can be used as open data [12], which anyone can access, use or share. By opening and sharing of information, Smart Cities can become more transparent. O'Hara [20] identifies three ways, how transparency given by open data enhance trust in government and community.

a) Citizen can see what the government is doing, and how it is doing. More information about contracts and procurement processes make government more trustworthy.

b) Accessible information about community rise the trust in community that support interactions and collaboration.

c) Data available for service providers, encourage the development of innovative services for citizens.

Especially first way of trust promotion is important in development of Smart City initiatives. Citizens' lack of confidence in the transparency of contracts, can postpone or stop all smart initiatives.

D. People

In Leavitt's model are people considered as employees that will be affected by the change in technology, that's why they have to be prepared for this change in order to get expected benefits. In this paper we broaden this perspective to cover users, administrators and leaders. Group of users is not coherent in their requirements even in capabilities. Users are persons really working with the system, using it to accomplish some tasks. It can be citizen searching for free parking lot, or employee working at control center, researcher using open data to make predictions about public housing, city manager evaluating power consumption, and so on. It is possible that changes in technology can simplify the work of one group of users, while other group's work increases. Success of Smart City initiatives at the user level determines individual acceptance, especially perceived usefulness and perceived ease of use. Successful Smart City initiatives give users possibility to accomplish tasks to their satisfaction. Readiness for users means to be able and willing to use new system. Ability to use new system depends on overall computer literacy and the level of training. This is just the matter of money and time, while willingness is the state of mind and as such, it is more demanding to influence. Change management methods are used for overcoming the resistance.

Introduction of new technology to the city life needs some administrators ensuring that the information technologies are functional and up to date, securing access only to authorized persons. IT staff should be qualified to administer and operate new technology. From their point of view is successful such Smart City technology that needs minimum interference during operation - it is secure, reliable, easily maintainable and portable. Government agencies can provide administrators on their own or they can use outsourcing.

Last group of people involved in the change of the city from traditional to smart concept are leaders. Those people who are visionaries, who can clearly specify goals, have realistic expectations, are able to motivate others and push the change through difficult moments. Those people are absolutely necessary at the beginning of any Smart City project but they are also affected by the result of change. If the result is satisfying, their motivation and energy is enhanced and they are able to continue the changes.

V. PARDUBICE CASE STUDY

In the last part of our paper we introduce strategy for Pardubice Smart City and discuss planed changes from the perspective of Leavitt's model. Representatives of the Statutory City of Pardubice have officially signed up for the Smart City (Intelligent City) concept in April 2016. Representatives declare that the concept of a smart city makes it possible for Pardubice residents to ensure that the functioning of their urban ecosystem is as simple, environmentally friendly and energy-efficient as possible thanks to modern technologies. As a partner for the preparation of a smart concept, Pardubice chose the Pardubice organization Smart City Point.

TABLE 2. THE CONCEPT OF PARDUBICE SMART CITY FROM THE PERSPECTIVE OF LEAVITT'S MODEL

Domain	Task	Technology	People	Structure
Mobility	Places for recharging of electric vehicles	Recharging stations	Higher convenience in using electric cars. Knowledge of location of recharging stations.	No
	Intelligent parking	Sensors for monitoring free parking lots, network infrastructure, centre for gathering and processing of such information, and user application to receive this information.	Possibility to easily find place for parking. Usage of special application.	Open government
	Intelligent parking	Roofing selected parking spaces and installation of photovoltaic elements.	Reducing the thermal load of vehicles in the summer.	No
	Electromobility - public transportation	Electro bus	Lower air pollution in the downtown can increase pedestrian traffic.	No
	Electromobility - individual transportation	No	Use of dedicated BUS lanes for electric vehicles. Possible discounts for parking.	No
	Sharing of bikes	Secure bike sharing stations	Possibility of cycling without having own bike (especially useful for visitors). Knowledge of location of sharing stations and method of usage.	No
	Places for safety bike storage	Special buildings - biketowers	Secure place to save bike. Knowledge of location of biketowers and method of usage.	No
	Sharing of cars	Electric cars, vehicle stations with recharging equipment	Possibility to borrow car. Knowledge of location of vehicle stations and method of usage.	No
Traffic	Provision of actual traffic information	Traffic cameras, network infrastructure, centre for gathering and processing of such information, and user application to receive this information.	Convenient way of transportation without being stuck in traffic jam. Usage of special application.	Open government
	Provision of actual traffic closures and restrictions	User application to receive this information.	Convenient way of transportation. Usage of special application.	Open government
	Intelligent bus stop	Bus stop equipped with recharging module for smartphones, actual timetable, communication module.	Citizens get actual information about time of departure or delay and during waiting can recharge their smartphones. Communication module can be used in emergency.	No
Energy	Alternative source of energy	Roofing of selected parking spaces and installation of photovoltaic elements, photothermic elements for city aqua park, heat pumps in selected public buildings.	Saved public money can be used for other services. Higher citizens trust in city energy self-sufficience.	No
	Saving energy for public lighting	Intelligent lampposts	Saving money, mobile lampposts can be moved when necessary, additional services (wifi, communication module).	No
	Power dispatching	Sensors of power consumption installed in public buildings, control centre for gathering and processing of such information	Saving money, quick detection of damage.	Whole-of-government
	Consulting and advisory service	No	Advisory activity during new construction and reconstruction. Information on subsidies.	Open government
Services	Intelligent waste management	Intelligent dustbins	Dustbins ready for use.	No
	Utilization of bank debit cards for payments in public transportation	Card reader	The user does not need a special card. Into accounts once per day, with multiple paths, the system selects the optimal tariff.	No
	Data gathering from citizens	Application for data gathering, control centre for processing of such information	Citizens can report and map problems, suggest missing services. Usage of application.	Participatory government
Socio-cultural	Provision of information about barrier-free routes	User application to receive this information.	Citizens can search for optimal route without barriers. Usage of application.	Open government
	Reservation of tickets	User application for reservations.	Citizens can reserve tickets for cultural or sport events through city application. Usage of application.	No
Information technology	Smart City Application	Mobile application integrating all applications offered by Pardubice city	Concentration of all available information in one application. Usage of application.	Whole-of-government, Open government

The concept of Pardubice Smart City [14] cover six domains: mobility, traffic, energy, IT, services, and socio-cultural domain. All domains are introduced in table 2. For each task presented in concept we tried to find necessary changes in technology, people and structure. Changes in people element of model are described only from user perspective. There is first explained what benefits it will bring to users and then what demands are placed on them. All changes will affect administrators of newly introduced services demanding new technical and also communication skills that is why it is not particularly mentioned in table.

Changes in structure are visible only for some tasks as the rest do not need any change in structure. But we must mention one structure change that affect almost all tasks. It is incorporation of the organization Smart City Point that will cooperate on preparation of particular plans and participate on realization. This change can be very useful as this organization puts together representatives of the industry and the academic sphere. However, it can be also source of problems when this organization is seen as nontransparent. Recently, this organization is being discussed because of a connection with the political party ANO. If smart city projects are to be successful then there should be no doubts about the transparency of their investments as happened with the Open Card project.

VI. CONCLUSION

Smart City initiatives change the way of daily life in cities and even the way how the city is governed. New technologies brought possibility to offer new services or make existing services more effective. As the focus of research community is concentrated much more on technical issues, we tried to show the need for different perspective. Technology is only the tool that can be as successful as how well the city leaders can estimate the needs of citizens and impact of changes on the structure. Rapid development of Smart City initiatives also brought necessity to evaluate those initiatives and make comparisons. British standards institution in 2014 prepared Smart city framework – Guide customer service to establishing strategies for smart cities and communities. This guide defines key governance and delivery processes and for each process defines context, need, recommendation and linkages. UrbanTide prepared overview of the Smart Cities Maturity Model together with self-assessment tool. We can expect that forms of evaluation will increase in common with the development of Smart City initiatives.

ACKNOWLEDGMENT

The paper has been completed with the kind support of SGS_2017_17 project of Faculty of economics and Administration, University of Pardubice.

REFERENCES

[1] A Alamri, W. S. Ansari, M. M. Hassan, M. S. Hossain, A. Alelaiwi, and M. A. Hossain, "A survey on sensor-cloud: architecture, applications, and approaches," International Journal of Distributed Sensor Networks, 9(2), 2013.

[2] R. Bilkova and R. Machova, "Affect e-commerce services the use of services offered by eGovernment?," In Proceedings of the International Conference on Information and Digital Technologies 2015. New York: IEEE (Institute of Electrical and Electronics Engineers), 2015. pp. 39-43.

[3] M. Bouskela, et al., "The Road toward Smart Cities- Migrating from Traditional City Management to the Smart City" Inter-American Development Bank (IDB), 2016. Online https://publications.iadb.org/bitstream/handle/11319/7743/The-Road-towards-Smart-Cities-Migrating-from-Traditional-City-Management-to-the-Smart-City.pdf

[4] G. Clifford, "Streetlights Can Do That? An Entrepreneur Creates Smarter Cities," Wired. 11.7.2012 Online https://www.wired.com/2012/11/streetlights-smart-cities/

[5] Committee of Experts on Public Administration, "Participatory governance and citizens' engagement in policy development, service delivery and budgeting," New York, 10-13 April 2007 http://unpan1.un.org/intradoc/groups/public/documents/un/unpan025375.pdf

[6] R. P. Dameri, "Searching for smart city definition: a comprehensive proposal," International Journal of Computers & Technology, 11(5), 2013.

[7] R. P. Dameri and A. Cocchia, "Smart city and digital city: Twenty years of terminology evolution," In: X. Conference of the Italian Chapter of AIS, ITAIS, 2013, pp. 1-8.

[8] S. Dirks and M. Keeling, "A Vision of Smarter Cities: How Cities Can Lead the Way into a Prosperous and Sustainable Future," Somers, NY: IBM Global Business Services, 2009. Online http://www-05.ibm.com/cz/gbs/study/pdf/GBE03227USEN.PDF

[9] R. Giffinger and H. Gudrun, "Smart cities ranking: an effective instrument for the positioning of the cities?," ACE: Architecture, City and Environment, 4(12), 2010, pp. 7-26.

[10] C. Harrison et al., "Foundations for smarter cities." IBM Journal of Research and Development, 54(4), 2010, pp. 1-16.

[11] ITU-T Study Group 5 (FG-SSC). "Agreed definition of a Smart Sustainable City". Online http://www.itu.int/en/ITU-T/focusgroups/ssc/Pages/default.aspx

[12] E. W. Johnston and D. L. Hansen, "Design lessons for smart governance infrastructures," in Transforming American governance: Rebooting the public square, 2011, pp. 197-212.

[13] R. M. Kanter and S. S. Litow, "Informed and interconnected: A manifesto for smarter cities," Harvard Business School General Management Unit Working Paper, 09-141, 2009. Online http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1420236.

[14] J. King, R. Bose, H. I. Yang, S. Pickles and A. Helal, "Atlas: A service-oriented sensor platform: Hardware and middleware to enable programmable pervasive spaces," in local computer networks, proceedings 2006 31st IEEE conference on 2006, (pp. 630-638). IEEE.

[15] Koncepce Smart City města Pardubic. Odbor rozvoje a strategie. http://www.cirihk.cz/files/ppt/chvojka-smartcity-2016-10-konf.pdf

[16] H. Leavitt, "Applied organizational change in industry: Structural, technological and humanistic approaches," Handbook of organizations, 1965, pp. 1144-1170.

[17] M. Lnenicka and J. Komarkova, "The Impact of Cloud Computing and Open (Big) Data on the Enterprise Architecture Framework," in Proceedings of the 26th International Business Information Management Association Conference. Norristown: International Business Information Management Association-IBIMA, 2015. pp. 1679-1683.

[18] T. Nam, and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in: Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times. ACM, 2011, pp. 282-291.

[19] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges," Computer, 44(6), 2011, pp. 32-39.

[20] K. O'Hara, "Transparency, open data and trust in government: shaping the infosphere," in Proceedings of the 4th annual ACM web science conference, 2012, pp. 223-232.

[21] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Sensing as a service model for smart cities supported by internet of things," in Transactions on Emerging Telecommunications Technologies, 25(1), 2014, pp. 81-93.

[22] G. Santucci, "The Internet of Things: Between the Revolution of the Internet and the Metamorphosis of Objects" in European Commission Community Research and Development Information Service. Oline https://pdfs.semanticscholar.org/adb7/03eb4c53ccba53a8973fbff2f30563363a58.pdf

[23] R. Sennet, "The Stupefying Smart Cit," Urban Choreography. 10.6.2013 Online https://urbanchoreography.net/2013/06/10/richard-sennet-the-stupefying-smart-city/

[24] S. Simonova and A. Hudec, "Enterprise content management based on identified requirements", In International Conference on Information and Digital Technologies (IDT), 2015, pp. 324 - 329, 7-9 July 2015 DOI: 10.1109/DT.2015.7222991.

[25] O. Söderström, T. Paasche, and F. Klauser, "Smart cities as corporate storytelling," City, 18(3), 2014, pp. 307-320.

[26] K. SU, J. LI, and H. FU, "Smart city and the applications," in 2011 International Conference on IEEE - Electronics, Communications and Control (ICECC), 2011. pp. 1028-1031.

[27] O. Vermesan and P. Friess, "Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems," Aalborg, Denmark: River Publishers, 2013. Online http://www.internet-of-things-research.eu/pdf/Converging_Technologies_for_Smart_Environments_and_Integrated_Ecosystems_IERC_Book_Open_Access_2013.pdf ISBN 978-87-92982-96-4.

[28] D. Washburn, et al., "Helping CIOs Understand "Smart City" Initiatives: Defining the Smart City, Its Drivers, and the Role of the CIO," Cambridge, MA: Forrester Research, Inc., 2010. Online http://public.dhe.ibm.com/partnerworld/pub/smb/smarterplanet/forr_help_cios_und_smart_city_initiatives.pdf.

[29] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," IEEE Internet of Things Journal, 1(1), pp. 22–32.

Using Logistic Regression for Assessing the Probability of Serious Postoperative Complications after Colorectal Operations in Geriatric Patients

Veronika Kubíčková
Department of Applied Mathematics
VŠB – Technical University of Ostrava
Ostrava, Czech Republic
Email: veronika.kubickova.st@vsb.cz

MUDr. Lubomír Martínek
Department of Surgery
Motol University Hospital
Prague, Czech Republic

***Abstract*—This paper focuses on an analysis of data from elderly patients (aged 65 and older) who underwent an elective operation of colon or rectum in the years 2001-2009 at the Department of Surgery of the University Hospital Ostrava. The main purpose is to determine the main risk factors leading to serious postoperative complications or even death in consequence of the operation. Our results will serve as a foundation for creating a reliable predictor model that will assess the overall health of the patients and the potential benefits (and risk) of the operation, leading to a reduction in incidence of serious postoperative complications of the operation and an improvement in the quality of life for the seniors. For the primary analysis we used the logistic regression methodology.**

***Index Terms*—colorectal operations, assessing risks, logistic regression**

I. Introduction

This paper focuses on the analysis of medical data of 610 elderly (aged 65 and over) patients that underwent the elective operation of either colon or rectum during the years 2001-2009. The data has been used for a number of similar successfully solved problems, such as the comparison of risk of the laparoscopic and open surgery (for the full text please see [1]). The aim of the analysis is to find the most significant factors that influence the occurrence of serious postoperative complications following the operation (the serious postoperative complications were defined as those obtaining Grade III or higher in the Clavien-Dindo classification - see [2]), and especially the mortality rate. The factors (covariates) analysed include the patients' characteristics, their diagnoses, and the details on the operative procedure.

The logistic regression was used for the analysis of the problem. First, we considered the occurrence of serious postoperative complications (including death) as the response variable. After that, we narrowed our focus and restricted the output variable only to the mortality rate, i.e. whether or not the subject died as a result of the operation.

II. Logistic Regression

A. Introduction

This section draws from [3, Chapters 1-5].

Logistic regression, as any other regression method, is aimed on finding a model that, firstly, best describes the relationship between a set of explanatory variables (covariates) and an outcome variable, and secondly, is as economical (in the way of reducing the number of covariates) as possible. The important difference compared to the most common – linear – regression, and also the reason why the logistic regression model is so widely used, is that the outcome variable is *dichotomous* which means it can be used on a whole spectrum of problems which the linear model is not suitable for.

Let us suppose that Y denotes the binary outcome variable, $\mathbf{X} = (X_1, \ldots, X_p)$ the vector of explanatory variables (with y and $\mathbf{x} = (x_1, \ldots, x_p)$ denoting the specific values of the response variable and the covariates). Then the expression $\mathsf{E}(Y|\mathbf{x})$ describes the mean value of the outcome variable given the values of the independent variables $\mathbf{x}$. For the purposes of the logistic regression (and to make the notation easier), we will denote the conditional mean $\mathsf{E}(Y|\mathbf{x})$ as $\pi(\mathbf{x})$.

Note: The outcome variable Y very often describes an occurrence of an event (such as a serious post-operational complication or death in our case), coded 1 if the event occurred and 0 if it did not. The conditional mean $\pi(\mathbf{x}) = \mathsf{E}(Y|\mathbf{x})$ then denotes the probability of the event.

The model we use in the logistic regression is based on the *logistic function*:

$$\pi(\mathbf{x}) = \frac{e^{\beta_0 + \sum_i \beta_i x_i}}{1 + e^{\beta_0 + \sum_i \beta_i x_i}}, \tag{1}$$

where β_j, $j = 0, \ldots p$ are the estimated coefficients.

The focus of the logistic regression analysis is the *logit*

transformation of $\pi(\mathbf{x})$:

$$g(\mathbf{x}) = \ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \sum_i \beta_i x_i. \qquad (2)$$

The function $g(x)$, also called the *logit function*, has some useful properties, such as being linear in the parameters and being continuous (on a connected set).

For further analysis (specifically for the estimation of the unknown parameters) we need to make an assumption about the conditional distribution of the response variable. We further assume that an observation of the outcome variable can be expressed as $y = \pi(\mathbf{x}) + \varepsilon$ (where ε is usually called the *error* and describes the deviation of the observation from the conditional mean). Unlike in linear regression, where we suppose that ε (and thus, the conditional distribution of Y given $\mathbf{x}$) follows the normal distribution, in logistic regression we assume that ε follows a binomial distribution, and therefore the conditional distribution of Y follows a binomial distribution with probability equalling the conditional mean $\pi(\mathbf{x})$.

B. Fitting the Model

Let us assume that we have a sample of n independent observations $(\mathbf{x}_i, y_i)$. Fitting the logistic regression model (1) means obtaining estimate of the vector of the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. For this goal a classic tool, the *maximum likelihood estimation*, is used, resulting in these $p+1$ likelihood equations:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0, \qquad (3)$$

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0, \qquad j \in \{1, \ldots, p\}. \qquad (4)$$

The solution of this equation system is not straightforward (as is the case for the linear regression) and cannot be reached analytically. A special software is needed which should be available in most statistical packages and programs. We will use (as discussed later) the statistical software `R`. Let $\hat{\boldsymbol{\beta}}$ denote the solution of the likelihood equations acquired by any suitable statistical software and $\hat{\pi}(\mathbf{x}_i)$ the fitted values for the model.

With the use of maximum likelihood estimation we also compute the estimate of the variance matrix of the estimator vector $\hat{\boldsymbol{\beta}}$ that is used for evaluating the significance of the individual parameters. The details on the theory of how this matrix is computed (again, with the use of a statistical software) can be found in [3, Chapter 2.3].

C. Model Building – Variable Selection

This chapter briefly deals with model building strategies for logistic regression, the selection of variables and the scaling of continuous covariates.

1) Step 1 – Univariate Analyses: First, each of the variables should be analysed separately. For continuous variables, the best method is to fit a univariable logistic regression model and test for the significance of the resultant coefficient. For discrete variables the analysis of a contingency table of the outcome variable versus the k levels of the covariate can be done (with the likelihood ratio chi-square test with $k-1$ degrees-of freedom, or the Pearson chi-square test that is asymptotically equal, as the decision tool). Separate logistic regression models can also be fitted for the discrete variables.

2) Step 2 – Fit of the Multivariable Model: With the univariate analyses completed, we select variables for the multivariable analysis. There are many schools of thoughts regarding this initial selection, from taking all the covariates (if the data is sufficient to support such analysis) to various mechanical procedures, such as stepwise or best subsets selection. We have used a general recommendation to include any variable that had a p-value < 0.25 in the univariate analysis as well as every variable of known clinical importance.

3) Step 3 – Reducing the Number of Covariates: After the preliminary multivariable model is fit we need to check for the significance of each variable included, by examining the Wald statistic of each estimated coefficient. Variables not contributing to the model should be removed and a new model fit. As well as checking the Wald statistics, we should also ascertain that the estimated coefficients have not changed much in magnitude (that could mean that we eliminated a variable that adds an important adjustment to the variable remaining in the model). We also compare the mew model to the old one using the likelihood ratio test. We continue the process of eliminating variables and refitting the model until it seems we have included all the important variables and excluded only the clinically and/or statistically unimportant ones.

4) Step 4 – Scaling the Continuous Variables: Up until now we have assumed linearity of the model in the continuous variables, which is not necessarily true. One of the analytic approaches regarding the process of finding the most proper form is called *method of fractional polynomials* and routines for this process are available in most statistical softwares. In short, instead of assuming that the logit is linear in this particular covariate, we assume that the relationship takes on the form of a sum of power functions $\sum_{j=1}^{J} F_j(x)\beta_j$. Even though we could use any number of functions, in most applied settings we can find adequate transformation if we use one or two functions ($J = 1/2$). For a detailed description of the method kindly see [3, Chapter 4.2].

5) Step 5 – Checking for Interactions: The final step is to consider possible interactions among the variables in the model. We say that two covariates interact if the effect of one covariate is not constant over the levels of the other. First, we should inspect the clinical plausibility of the possible interacting couples. We should make a list of such pairs of covariates that have some scientific basis to interact with each

other. When we have finished that, we add the interaction terms (in the form of arithmetic product of the respective variables) one by one into the model (the product of *Step 4* and assess their significance using likelihood ratio test.

III. Computer tool

We have used the statistical language R, specifically the RkWard GUI (version 0.6.5; for source please see [4]). We will briefly mention the commands and routines we used for the whole process of building the model.

1) Steps 1 and 2 – Fitting a Model: The command for fitting a logistic regression model is included in the basic packages of R. We use the command for generalized linear models, glm, and specify the class in the argument family (detailed directions with an example can be found in [5]):

```
>glm(formula = covariate1 + ... ~ response,
family = binomial(logit), data = your.data.frame)
```

2) Step 3 – Checking the Model: The user will find the necessarry information – the estimated coefficients and their Wald statistics – if they call summary of the fitted model.

For the likelihood ratio test of the submodels there are two possibilities: anova command in the base and lrtest in the package lmtest. The use of both commands (including the commands that download the necessary package and attach the functions it contains) is here:

```
>anova(larger_model, smaller_model, test ="Chisq"),
```

```
>install.packages("mfp")
>library(lmtest)
>lrtest(larger_model, smaller_model).
```

3) Step 4 – Scaling – the Fractional Polynomials: For the fractional polynomials method for scaling of continuous variables we have used the R package mfp, by the authors Gareth Ambler and Axel Benner (the documantation for the package can be downloaded from [6]). We can download the package and attach the functions it contains by analogous commands to these shown above. The package contains two main functions. The first command is fp(x, df = 4, alpha=NA, select=NA,...), that transforms the quantitative variable x into a fractional polynomial object used for the actual procedure. The arguments for the function are the number of degrees of freedom, and selection levels (described below) that the user can specify individually for every covariate if it is desirable. If not specified, the general levels specified for the procedure will be used.

The function that fits the model including the fractional polynomial procedure (for specified covariates) is this:

```
>mfp(formula, data, family = gaussian, alpha=0.05,
select = 1, keep = NULL, ...)
```

The arguments shown mean:

- formula a formula object in a standard form (with the response variable on the left of), the candidates for the fractional polynomials are indicated by fp (are included as a fp() function),
- data the data frame containing the variables,
- family specifies the type of regression. Supported families are: gaussian, binomial, poisson, Gamma, inverse.gaussian and quasi, as well as Cox (that requires additional arguments in the function as well as a specific input of the formula, for details please see [6]).
- alpha sets the fractional polynomials selection level for all the predictors (values for individual variables can be set in the appropriate fp function),
- select sets the variable selection level for all the predictors,
- keep keeps one or more variable in the model. The selection level of these variables will be set to 1.

IV. Analysis

A. Data Description

The analysed cohort consists of 609 subjects, 103 of which suffered serious postoperative complications, in 35 cases it meant death, in the other 68 cases it meant other, non-lethal, type of complication (classified as Grade III or IV in the Clavien-Dindo classification of surgical complications).

Following variables were taken into account:

- group – polynomous variable denoting the type of the operation the subject underwent (laparoscopic, open or a conversion – the operation was started as laparoscopic but was changed to open due to some complications during the operation itself)
- age – continuous variable describing the age of the subject at the time of the operation (rounded to years)
- sex – dichotomous variable denoting the sex of the subject
- BMI – continuous variable describing the Body Mass Index of the subject
- Dg – polynomous variable describing the diagnosis leading to the operation in question (we distinguish only the cancer of the rectum, the cancer of any other type and other, noncancerous diagnoses)
- ASA – polynomous variable, the result of the physical status classification system that assesses the fitness of the patients (ranging from 1 – a healthy person – to 6 a brain-dead person whose organs are being donated; practically the number in our cohort doesn't exceed 4, a patient with a severe systemic disease threatening their life),
- a set of dichotomous variables denoting whether or not the subject suffered from various types of problems: CHD (coronary artery disease), arrhythmia (cardiac arrhythmia), hypertesion, cerebrovascular, pulmonary, DM (diabetes mellitus), renal (kidney diseases) and hepatic (hepatic diseases).

TABLE I
THE DISTRIBUTION OF THE CONTINUOUS COVARIATES (COLUMNS 3,4 AND 6 INDICATE APPROPRIATE QUANTILES).

Variable	min	0.25	0.5	mean	0.75	max
age	65.0	68.0	73.0	73.21	78.0	97.0
BMI	15.4	24.0	26.3	26.76	29.4	42.6
No.prev.OP	0.0	0.0	1.0	1.2	0.71	5.0
length.OP	20.0	90.0	140.0	143.5	180.0	400.0

TABLE II
THE DISTRIBUTION OF THE COVARIATES FOR THE VARIOUS DISEASES.

Variable	no	yes	Variable	no	yes
CHD	300	309	pulmonary	499	110
arrhythmia	492	117	DM	419	190
hypertension	184	425	renal	573	36
cerebrovascular	531	78	hepatic	595	14

TABLE III
THE DISTRIBUTION OF THE OTHER DISCRETE COVARIATES.

Variable	1	2	3	4
ASA	4	253	312	40
Variable	rectum	other	benign	
Dg	198	348	63	
Variable	lpsc	open	conversion	
group	318	266	25	
Variable	male	female		
sex	375	234		
Variable	no	yes		
perop.comp	556	53		

- previous.OP – dichotomous variable indicating whether or not the subject had undergone an operation before,
- No.prev.OP – discrete variable describing the number of previous operations,
- length.OP – continuous variable describing the length of the operation (rounded to minutes),
- perop.comp. – alternative variable that indicates whether or not there were complications during the operation,
- goal – polynomous variable defined only for cancerous patients, denoting the aim of the operation – either curative (trying to cure the disease) or palliative (trying only to control the symptoms),
- stage – polynomous variable defined only for cancerous patients, denoting the stage of the cancer (ranging from 0 to 4).

B. The Incidence of Serious Postoperative Complications

We started with analysing the simple models for each of the variables separately and tested the significance of the coefficients, as well as analysing the simple contingency tables in the case of discrete covariates. We set up a limit for the *potentially significant variable* to be the p-value of 0.25. The following covariates appeared significant: cerebrovascular (at the 0.05 significance level), group, lenght.OP, perop.comp (significance level 0.1) and sex and ASA (significance level 0.25).

TABLE IV
LOGISTIC MODEL FOR THE INCIDENCE OF SERIOUS POSTOPERATIVE COMPLICATIONS

Variable	$\hat{\beta}$	OR	p-value	90% conf. int. for OR
cerebr. diseases	0.609	1.839	0.033	$[1.148, 2.946]$
length of op. [min]	0.003	1.003	0.059	$[1.0004, 1.0054]$

TABLE V
LOGISTIC MODEL FOR THE INCIDENCE OF SERIOUS POSTOPERATIVE COMPLICATIONS II

Variable	$\hat{\beta}$	OR	p-value	90% conf. int. for OR
cerebr. diseases	0.631	1.879	0.033	$[1.173, 3.011]$
perop. comp.	0.652	1.919	0.059	$[1.108, 3.324]$

Subsequently we created a model containing all of the covariates deemed potentially statistically significant (without interactions), as well as the variables indicated as of special clinical interest by the surgeons, specifically group and Dg. This is the result as produced by the summary function of the fitted model inthe statistic software R:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.724105   0.586990  -2.937  0.00331 **
cerebrovascular1  0.630202   0.289144   2.180  0.02929 *
perop.comp        0.477446   0.383475   1.245  0.21311
grouplpsc        -0.090508   0.525227  -0.172  0.86318
skupinaopen      -0.410179   0.527222  -0.778  0.43657
sexf             -0.222329   0.232294  -0.957  0.33852
length.OP         0.002002   0.001622   1.235  0.21701
---
Signif.:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the only statistically significant variable in this model is the variable coding the occurrence of cerebrovascular diseases. Then we gradually tried omitting some of the covariates showing the weakest link to the response variable. We tested the suitability of the submodels with the likelihood ratio test function mentioned in Section III-2.

The model building process resulted in two models equally suitable for describing the data, one with the covariates cerebrovascular and perop.comp, and the other with cerebrovascular again and length.OP. The results for these models can be seen in the Tables IV and V.

C. Mortality Incidence

In the case of mortality as the response variable, we proceeded in the same way as described in the previous section. Regarding the individual univariate analyses, following variables appeared to be significant: length.OP (level of significance 0.05), cerebrovascular (significance level 0.1) and perop.comp, group, Dg, ASA, goal and stage (significance level 0.25).

TABLE VI
LOGISTIC MODEL FOR THE INCIDENCE OF SERIOUS POSTOPERATIONAL COMPLICATIONS

Variable	$\hat{\beta}$	OR	p-value	95% conf. int. for OR
Dgother	-0.593	0.553	0.236	[0.243, 1.260]
Dgrectum	-1.029	0.357	0.069	[0.141, 0.906]
ASA3	0.170	1.185	0.661	[0.627, 2.242]
ASA4	1.336	3.802	0.023	[1.446, 10.000]
length.OP	0.007	1.007	0.008	[1.003, 1.011]

Only three of these covariates were significant in the model including all the covariates – the length of the operation (at the 0.05 confidence level) and the diagnosis of cancer of rectum and ASA level 4 (at the 0.1 significance level), as can be seen below:

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.171657   0.947865  -3.346  0.00082 ***
grouplpsc        -0.097007   0.760329  -0.128  0.89848
groupopen        -0.419444   0.764566  -0.549  0.58328
Dgother          -0.550455   0.507682  -1.084  0.27825
Dgrectum         -1.031573   0.579479  -1.780  0.07505 .
ASA3              0.081468   0.404937   0.201  0.84055
ASA4              1.093736   0.620548   1.763  0.07798 .
cerebrovascular1  0.590156   0.457466   1.290  0.19703
length.OP         0.005905   0.002664   2.217  0.02665 *
perop.comp1       0.447883   0.600452   0.746  0.45572
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the search for the most suitable model we proceeded in the same manner as in the previous section. The final model was chosen not only by the strict rule of likelihood ratio test but the potential clinical significance was also taken into account (as recommended by [3]). The final model contains the covariates describing the diagnosis, ASA, and the length of the operation. Details can be found in the Table VI.

V. DISCUSSION

The increasing population size of the elderly is affecting all developed countries, thus becoming one of the most important and most debated social issues. Especially the increasing incidence of malignant tumours in elderly patients will result in a dramatic increase in geriatric patients scheduled for surgery. In the Czech Republic, $33-60\%$ of all new cancers are diagnosed in patients over the age of 70 years ([7]). Elderly population is heterogeneous and consists of patients fully capable of undergoing operations (fit) to patients at high operative risk (frail). The ageing process is individual and physiological age does not correspond to the calendar age.

In surgery, the optimal treatment should be determined by the objective prediction of the operative risk. Geriatric patients should be divided into three groups: those capable of undergoing standard treatment, those who require the treatment to be adapted and those requiring palliative or symptomatic treatment. The ultimate goal is to achieve a reduction in mortality, morbidity and improve quality of life for seniors. Therefore we need to assess the overall health and wellbeing of each patient before the operation so that we can decide which of the three approaches would be the best for each individual. In this paper we tried to find logistic regression models that would indicate the variables with the statistically significant influence on the mortality and morbidity rates as well as assess the influence numerically.

A. Morbidity Rate (The Incidence of Serious Postoperative Complications)

The variable selection in this case resulted in two similarly significant models. One containing the variables indicating the presence of a cerebrovascular disease in the subject and occurrence of perioperative complications during the operation, and the other containing the cerebrovascular disease incidence as well, along with the continuous variable marking the length of the operation.

The latter was chosen for the ensuing discussion because it contains a variable, the length of the operation, that can be influenced before the operation starts (by choosing the technique or the scope of the operation), whereas the link with the variable `perop.comp` does not contribute to the decision making before the operation as it is an event not likely to be predicted or influenced before the operation starts.

The chosen model contains the variable indicating whether or not the subject suffers from a cerebrovascular disease (OR $= 1.84$ with 90% confidence interval $I = [1.15, 2.95]$, p-value 0.033) and the variable that registers the length of the operation in minutes (OR $= 1.003$ with 90% confidence interval $I = [1.0003, 1.005]$, p-value 0.058).

The subjects falling into the group with `cerebrovascular` $= 1$ are the patients with irreversible quantitative or qualitative consciousness disorders or handicaps usually caused by a cerebral stroke with serious neurological disability (movement disorder, speech disorder, etc.). From the clinical point of view a minimal level of cooperation in the postoperative period, including during rehabilitation, can be expected in these patients. Their perception of the world around them is usually greatly limited with no chance to improve and the decision to modify the surgical treatment of these patients is most certainly appropriate.

The length of the operation is a factor that is commonly discussed in the surgical field as a potential risk factor with regards to the postoperative complications. Special attention in this context is given to the subgroup of seniors. The conclusions so far have not been consistent (for example [8], [9]). The obtained results are important from the clinical point of view because, as mentioned previously, the length of the operation is related to the considered scope of the operation and is therefore a parameter that can be influenced by the surgeon.

B. Mortality Rate

For the case of mortality as the response variable a model containing following covariates was found (we will mention only those levels of multinomous variables that were assessed as being significant): the ASA Grade IV (OR $= 3.8$ compared to Grade II or lower, $I = [1.45, 10.00]$, p-value 0.02), the diagnosis of cancer (OR = 0.36 compared to benign diagnoses, $I = [0.141,\ 0.906]$, p-value 0.07) and the length of the operation in minutes (OR = 1.007, $I = [1.003,\ 1.011]$, p-value 0.01).

As regards the length of the operation, the same applies as in the previous section.

The ASA classification is used by the anaesthesiologists to assess the fitness of patients before surgery (for the general anaesthesia). Assumptions are (e.g. in [10]) that there is a close association between the risk brought by general anaesthesia and the risk of postoperative complications including death. The results of our analysis confirm such expectations. ASA is not a parameter that could be influenced before the operation but it can serve as an indicator of the seriousness of the patient's condition and his limited chances for survival.

The results for the rectal carcinoma are quite interesting as the surgery of rectal carcinoma represents the most technically challenging part of the colorectal surgery and as such it is usually encumbered by high rates of mortality and morbidity. The fact that in our cohort the patients operated for rectal carcinoma had lower risk of death (the odds ratio was 0.36 compared to patients with a benign problem) is surprising but not unprecedented. For example [11] explain similar results by the fact that the most demanding operations are usually carried out by the most experienced surgeons, which might contribute to the lower morbidity rate.

ACKNOWLEDGMENT

This work was supported by the FEECS VŠB – Technical University of Ostrava (Project No. SP2017/56).

REFERENCES

[1] Jahoda, P., Briš, R., Běloch, M. & Martínek L. (2016): Decision Support System Minimizing the Risk of Post-operation Complications after Surgeries. Health and Technology. 2016, 6: 149-156. doi:10.1007/s12553-016-0133-7. Springer Verlag.

[2] Dindo, D., Demartines, N., Clavien, P.A. (2004): Classification of Surgical Complications. A New Proposal With Evaluation in a Cohort of 6336 Patients and Results of a Survey. Annals of Surgery. 2004, 240: 205-213.

[3] Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Second Edition. John Wiley & Sons, Inc. 2000.

[4] Friedrichsmeier, T. et al.: RKWard [Computer software]. Retrieved from https://rkward.kde.org/, [downloaded 1.3.2017].

[5] King, W.B.: Logistic Regression [online]. Revised February 2016. Available from https://ww2.coastal.edu/kingw/statistics/R-tutorials/logistic.html, [sourced 18.3.2017].

[6] Ambler, G., Benner, A.: Package "mfp" [online]. Version 1.5.2, September 2015. Available from https://cran.r-project.org/web/packages/mfp/mfp.pdf, [downloaded 18.3.2017].

[7] Petera, J., Dusek, L. (2012): Cancer in the Elderly. Reports of Practical Oncology and Radiotherapy. 2012, 18(1): 1-5.

[8] Konishi, T., Watanabe, T., Kishimoto, J., et al. (2006): Risk Factors for Anastomotic Leakage after Surgery for Colorectal Cancer: Results of Prospective Surveillance. Journal of the American College of Surgeons. 2006, 20: 439-444.

[9] Mäkelä, J.T., Kiviniemi, H., Laitinen, S. (2003): Risk Factors for Anastomostic leakage anfter Left-Sided Colorectal Ressecton with Rectal Anastomosis. Diseases of the Colon & Rectum. 2003, 46: 653-670.

[10] Wolters, U., et al. (1996): ASA Classification and Perioperative Variables Predictors of Postoperative Outcome. British Journal of Anaesthesia. 1996, 77(2): 217-222.

[11] Jannasch, O., Klinge, T., Otto, R., et al.: Risk Factors, Short and Long Term Outcome of Anastomotic Leaks in Rectal Cancer. Oncotarget. 2015, 6: 36884-3693.

Multiple-Valued Logic in Analysis of Critical States of Multi-State System

Miroslav Kvassay, Jan Rabcan, Patrik Rusnak
Department of Informatics
University of Zilina
Zilina, Slovakia
miroslav.kvassay@fri.uniza.sk, jrabcanj@gmail.com, pattrik.rusnak@gmail.com

***Abstract*—Multi-State System (MSS) is mathematical model in reliability engineering that allows representing investigated system/object with some performance levels from perfect working to complete failure. New approach for the analysis of MSS based on mathematical methods of Multiple-Valued Logic is proposed in this paper, In particular, the method for the calculation of Critical System States is developed.**

Keywords—Multi-State System (MSS); structure function; reliability; critical system states

I. INTRODUCTION

The typical mathematical model in the reliability analysis allows analysis of two principal states of investigated object as "work" and "failure" [1]. This mathematical model was named *Binary-State System* (BSS). Special mathematical approaches are developed for the qualitative and quantitative evaluations based on such mathematical model. One of these mathematical approaches use mathematical background of Boolean algebra [1, 2]. BSS is effective mathematical representation for the investigation of system failure, it allows developing of scenarios of system failure as qualitative analysis [1, 3], calculating of the set of different indices and measures that characterize the system behavior [4]. Last reviews in reliability engineering [5, 6, 7] show that the analysis of failure is not sufficient complex system. The analysis has to include investigation of states that are preceding to failure. In this case the mathematical model must include description of system failure and some performance levels pf the working states. Such model was introduced in reliability engineering in papers [8, 9] and named as *Multi-State System* (MSS). This mathematical model permits to indicate some of investigated system performance levels from complete failure to perfect work, for example, as “failure”, “partly work” and “work”.

Some mathematical approaches were proposed for estimation of MSS. According [10] these approaches are based on following four mathematical backgrounds: extensions of Boolean methods, stochastic processes, universal generating function, and Monte Carlo simulation. The approach based on the extension of Boolean methods needs representation of investigated system in form of structure function [11, 12, 13].

The extension of Boolean methods into the MSS analysis is nature step. But the MSS structure function can’t be interpreted as Boolean function. Therefore, some works, e.g. [11, 13, 14], have tried to transform the structure function of a MSS into a Boolean function using Boolean algebra with restrictions [14]. Such transformation allows us to use methods of Boolean algebra in reliability analysis of MSSs. However, this transformation can result in increase of model complexity. To avoid these complications, other authors have proposed using *Multiple-Valued Logic* (MVL) in the MSS reliability analysis since there exists some correlation between the MSS structure function and MVL function [10, 15, 16]. MVL is nature extension of Boolean methods in MSS reliability assessment. Some theoretical aspects of MVL application for the estimation of MSS have been developed in papers [16, 17, 18]. In these papers new methods for calculation reliability indices and measures based on application of MVL are proposed. In this paper the method for evaluation of critical states of MSS are developed based on mathematical methods of MVL.

The important problem in the MSS assessment is analysis of boundary (critical) states. The minimal cut/path sets as one of type of critical states of MSS are considered in [9, 10]. Critical states named Lower (Upper) Boundary Points are investigated in [13]. Authors of paper [19] shown that different types of critical states of MSS can be calculated and analysed based one mathematical approach that is Logical Differential Calculus. In this paper this investigation is continued and new method for qualitative analysis of critical state of MSS is considered based on example of two-engine jet, that has been described in [11]. These analysis is implemented by *Direct Partial Logic Derivative* (DPLD) introduced in [20]. Use of DPLDs in reliability analysis of MSSs has been considered in several works, e.g. [17, 19]. In these papers, it has been shown that DPLDs are useful in finding of critical states for different condition of MSS.

In this paper, critical states of MSS are considered and algorithms for their calculation based on mathematical approach of MVL are developed. The critical states are used in qualitative and quantitative methods for the estimation of MSS. The methods for qualitative and quantitate analysis are illustrate by simple example.

This work was partly supported by the grants of VEGA 1/0038/16 and VEGA 1/0354/17.

II. Multi-State System

MSS permits to consider more that only two states in behaviour of system reliability or availability. All MSS properties as reliability, availability and system states can be considered as general conception that is called "performance level" [10]. System performance level changes can be represented in more details by MSS. But at the present time there are not a lot of effective methods and algorithms to calculate different indices and measures of MSS [5]. Therefore the development new methods for MSS qualitative and quantitative analysis are actual problem in reliability analysis.

One of approaches for MSS analysis is based on MSS representation by a structure function. The structure function is defines univalent correlation of a system performance level and components states. Need to say that the structure function of BSS is Boolean function as a rule [2, 4]. The structure function method extension for the investigation of MSS is based on Boolean logic mostly [11, 14, 21]. In this case a system with some performance levels and some states of components is transformed into set of systems with two performance levels and two states of components. It means representation of one MSS by set of BSSs: every performance level of MSS is considered separately and coding by 0 and 1 for all possible components states [11, 14]. BSS of fixed performance level for components states has value 1 if this performance level confirms with equal components states of MSS.

There is alternative way for MSS analysis that is based on the use of methods of MVL for the analysis of MSS. Such methods have been mentioned in [10, 15]. In papers [16, 17, 18] theoretical aspects for the application of MVL in MSS reliability analysis have been considered.

A. Structure Function of Multi-State System

Consider a system of n components in stationary state or in fixed time point. This assumption allows defining MSS structure function as a time-independent function [1, 19]. The i-th system component state is denoted as x_i ($i = 1, 2, \ldots, n$). All system components states are defined by the vector states $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Assume equal number of the system performance levels and number of every component states. Each component in such mathematical representation has m states that are indicated as 0 for the complete fault and as m-1 for perfect functioning. The system has m performance level too: from the complete failure (it is 0) to the perfect functioning (it is m-1). The system performance levels depend on components states and this dependency is defined by the structure function $\phi(\boldsymbol{x})$ identically:

$$\phi(x_1,\ldots, x_n) = \phi(\boldsymbol{x}): \{0, 1, \ldots, m\text{-}1\}^n \rightarrow \{0, 1, \ldots, m\text{-}1\}. \quad (1)$$

The definition of the structure function (1) agrees with the definition of a MVL function [2] and the mathematical approaches of MVL can be used in analysis of MSS represented by the structure function (1).

For example, in paper [11] the twin-engine jet reliability analysis is implemented as MSS estimation based on extension of Boolean algebra for multi-state case. Let us, consider the representation of twin-engine jet analysed in [11] by the structure function (1). This jet is interpreted as the MSS by space of components states {0 = failed, 1 = half power, 2 = full power}, and MSS state space {0 = crash, 1 = land on foamed runway, 2 = land normally}. According to description in [11]:

- It can land normally if one engine is at full power and the other engine is at half power.
- It can land on a foamed runway if one engine is at full power or if both engines are at half power.
- It will crash if one engine is at half power and the other engine is failed.

The function of the jet can be interpreted as MVL function of two variables ($n = 2$) with three values ($m = 3$) (Table I).

TABLE I. Structure Function of Twin-Engine Jet

x_1	x_2	$\phi(\boldsymbol{x})$
0	0	0
0	1	0
0	2	1
1	0	0
1	1	1
1	2	2
2	0	1
2	1	2
2	2	2

The representation of investigates system in form of the structure function allows calculating indices and measures for estimation of this system reliability/availability. Let us consider the calculation of fundamental indices of availability and unavailability based on structure function.

B. Availability and Unavailability of Multi-State System

The system availability and unavailability are defined as a probability of system functioning and failure and for BSS these measures are calculated based on the structure function as:

$$A = \Pr\{\phi(\boldsymbol{x}) = 1\} \text{ and } U = \Pr\{\phi(\boldsymbol{x}) = 0\}. \quad (2)$$

Availability of MSS must be considered for some different performance levels and the availability (5) can be transformed into two types of measures for MSS [10, 15]: system availability and probability of system performance level. The probability of system performance level is defined for every performance level as:

$$A_j = \Pr\{\phi(\boldsymbol{x}) = j\}, j = 1,\ldots, m\text{-}1. \quad (3)$$

Note that the system unavailability can be considered as failed performance level:

$$U = A_0 = \Pr\{\phi(\boldsymbol{x}) = 0\}. \quad (4)$$

MSS availability is defined as follows [10, 19]:

$$A(j) = \Pr\{\phi(\boldsymbol{x}) \geq \mathrm{j}\},\ \mathrm{j} = 1,\ldots,\ m\text{-}1 \tag{5}$$

and this measure can be calculated based on probabilities of system performance levels A_j (3):

$$A(j) = \sum_{q=j}^{m-1} A_q\ ,\ j = 1,\ ...,\ m\text{-}1. \tag{6}$$

The probabilities of the system performance levels according (3) and (4) are initial measure that allows computing the system availability and unavailability. Therefore consider the calculation of this measure primarily. In papers [10, 12, 13] authors shown that any system state j (j = 1,…, m -1) for fixed components state vector $\boldsymbol{x}$ = (x_1, …, x_n) of MSS can be calculated as the product of probabilities of components states:

$$p_{i,s_i} = \Pr\{x_i = s_i\}\ ,\ s_i = 0,\ \ldots,\ m\text{-}1. \tag{7}$$

The system availability for performance level j is sum of probabilities of all possible states for the performance level j. In terms of structure function it means that the system availability (7) can be calculate as the sum of probabilities of all values j of the structure function $\phi(\boldsymbol{x})$ because components state vectors represent x are mutually exclusive events [1]:

$$A_j = \sum_{\phi(\boldsymbol{x})=j} p_{1,s_1} \cdot\ldots\cdot p_{n,s_n} \tag{8}$$

For example, consider a twin-engine jet [11] availability and unavailability. The MSS structure function of this object is defined in Table I. According to (4) and (8) the MSS unavailability is:

$$\mathrm{U} = \mathrm{p}_{1,0}\cdot p_{2,0} + p_{1,0}\cdot p_{2,1} + p_{1,1}\cdot p_{2,0}, \tag{9}$$

and probabilities of performance levels “1” and “2” according to (3) and (8) are:

$$A_1 = p_{1,0}\cdot p_{2,2} + p_{1,1}\cdot p_{2,1} + p_{1,2}\cdot p_{2,0}, \tag{10}$$

$$A_2 = p_{1,1}\cdot p_{2,2} + p_{1,2}\cdot p_{2,1} + p_{1,2}\cdot p_{2,2}. \tag{11}$$

Suppose that this system has equal component probabilities that are defined as: $p_{1,0} = p_{2,0}$ =0.1, $p_{1,1} = p_{2,1}$ =0.2 and $p_{1,2} = p_{2,2}$ =0.7. The system unavailability for this data is U = 0.05, probabilities of the system performance levels are A_1 = 0.18 and A_2 = 0.77, and this MSS availabilities for two performance levels are A(1) = 0.95 and A(2) = 0.77.

III. Critical States of Multi-State System

The structure function represents all possible performance levels of MSS depending on system’s components states. But in point of view of reliability analysis, the boundary states are most interesting and information for MSS estimation. The conception of boundary states has been proposed for BSS in [22]. Two types of boundary states are used in reliability analysis of BSS dominantly.

The first of them is *critical system state* (CSS) that widely used in importance analysis [23, 24]. CSS for the i-th component is system state for the remaining n-1 components such that failure of component i causes the system to go from working state to a failed state. The second type of boundary states is minimal cat/path set [21, 25]. A cut of BSS is failure system state. A cut is minimal, if no component can be removed from it without losing its status as a cut. A path is operational state of system. A path is minimal, if no component can be removed from it without losing its status as a path.

The short review of the generalization of critical states conception for MSS is provided below.

The first type of critical states of MSS is minimal cat/path sets. The conception of minimal cat/path sets for MSS have been considered in [10, 21]. In difference of BSS, the boundary states of MSS (in particular minimal cats/paths) must be defined for every system performance level. So, cut set of MSS is defined for system performance level j based on state vector $\boldsymbol{x}$ as set of state vectors $\boldsymbol{x}$ for which $\phi(\boldsymbol{x}) < j$ and a cut vector $\boldsymbol{x}$ is minimal if $\phi(\boldsymbol{y}) \geq j$ for any $\boldsymbol{y} > \boldsymbol{x}$. The similar interpretation has minimal path set of MSS.

There other types of boundary states in MSS reliability analysis: Lower (Upper) Boundary Points for performance level j (j =0, …, m-1) introduced in [12, 13]. These points specify when a decrease (increase) in the state of any one of the n components forces a decrease (increase) in the system state and are defined by special component state vectors to the system performance level j. The state vector $\boldsymbol{x}$=(x_1, …, x_n) is Lower Boundary Points iff $\phi(\boldsymbol{x}) \geq j$ and $\boldsymbol{y} < \boldsymbol{x}$ implies that $\phi(\boldsymbol{y}) < j$. The state vector $\boldsymbol{x}$=(x_1, …, x_n) is interpreted as Upper Boundary Points iff $\phi(\boldsymbol{x}) \leq j$ and $\boldsymbol{y} > \boldsymbol{x}$ implies that $\phi(\boldsymbol{y}) > j$.

Next type of critical system states has been named as exact boundary states in [16]. These states are defined for fixed changes of system performance level j (j = 0, … m-1) and component i (i = 1, …, n). In paper [26] and [27] the correlations of these boundary states with minimal cut/path sets and Lower (Upper) Boundary Points have been investigated.

In paper [28] one more type of critical states of MSS has been introduced and named as CSS of MSS. The context of these critical states and exact boundary states are equal, but the calculations are based on different mathematical background. In this paper we will use the term of CSS for exact boundary states and define they according to [49] as states for which the change of the i-th component state from s to $\tilde{s}$ causes the system performance level change from j to $\tilde{j}$ ($s,\ \tilde{s} \in \{0,\ldots, m_i\text{-}1\},\ s \neq \tilde{s}$ and $j,\ \tilde{j} \in \{0,\ldots, M\text{-}1\},\ j \neq \tilde{j}$). CSS is defined by the *critical system state vector* (CSSV) unambiguously. Illustrate the correlation of CSS and CSSV by the example.

For example, determine the CSS of land on foamed runway for a twin-engine jet if the first engine been at half power before the failed (the structure function of this system is

in Table I). It means the change of the system performance level from state "2" to "1" depending on the change of the first component from state "1" to "0". According to the truth table (Table I), there are one CSSV for this condition: $\boldsymbol{x} = (x_1, x_2) = (1 \to 0, 2)$. This CSSV interprets the situation: land on foamed runway if the second engine is at full power and the first fails.

Therefore CSS and CSSV are useful approach in quantitative analysis of MSS and allow obtaining the sceneries of the system behaviour in critical situations. And methods for the formalization to indicate CSS/CSSv of MSS are necessary in reliability engineering. In MVL there are methods for the analysis and estimation of MVL-function behaviour depending on the changes of its variables values. These methods are developed and provided in Logic Differential Calculus that has been introduced for Boolean algebra in papers [29, 30] and developed and generalised for MVL in [20, 31]. One of possible mathematical approaches in Logic Differential Calculus for detail analysis of the influence of the variable value change to the MVL-function value is *Direct Partial Logical Derivatives* (DPLDs) [31, 32]. In papers [15, 16, 17] this mathematical approach has been used for analysis of MSS structure function.

The mathematical tool of DPLDs for calculation of CSS of MSS has been introduced in [15, 27]. In the paper [15, 17] the definition of DPLDs for MVL function has been adapted for a structure function (1). According to [17] DPLD with respect to variable x_i for the structure function (1) permits analyse the system performance level change from j to $\tilde{j}$ when the i-th component state changes from s to $\tilde{s}$:

$$\partial\phi(j \to \tilde{j})/\partial x_i(s \to \tilde{s}) = \begin{cases} 1, \text{ if } \phi(s_i, \boldsymbol{x}) = j \ \& \ \phi(\tilde{s}_i, \boldsymbol{x}) = \tilde{j} \\ 0, \text{ other} \end{cases} \quad (12)$$

where $\phi(s_i, \boldsymbol{x}) = \phi(x_1,\ldots, x_{i-1}, s, x_{i+1},\ldots, x_n)$; $\phi(\tilde{s}, \boldsymbol{x}) = \phi(x_1,\ldots, x_{i-1}, \tilde{s}, x_{i+1},\ldots, x_n)$; $s, \tilde{s} \in \{0,\ldots, m-1\}$, $s \neq \tilde{s}$ and $j, \tilde{j} \in \{0,\ldots, m-1\}, j \neq \tilde{j}$.

Therefore DPLD (12) allows indicating CSSV of the structure function of the system performance level j. These derivatives can be used for the calculation of CSSV and the investigation of influence of the i-th system component changes from s to $\tilde{s}$ to performance level j. The indication of different values for parameters $\tilde{j}$, s and $\tilde{s}$ permits to investigate all possible changes for the system performance level j that means the calculation of possible CSSVs.

IV. Qualitative and Quantitative Estimation of Critical System States of Multi-State System

A. Qualitative Analysis

Qualitative Analysis of CSS suppose the induction of scenario of changes of investigated system performance level depending of change of its component.

For example, let us define CSSV for a twin-engine jet [11]. The structure function of this system is shown in Table I. CSSV for the first component (it is represent by the first variable of structure function) can be calculated by DPLD (12) with respect to variable x_1. The CSSVs for the system performance level 1 and the first component are computed by DPLDs $\partial\phi(1\to0)/\partial x_1(2\to1)$, $\partial\phi(1\to0)/\partial x_1(1\to0)$ and $\partial\phi(1\to0)/\partial x_1(2\to0)$. These derivatives are in Table II.

TABLE II. CSSVs of Two-Engine Jet Crash

x_1	x_2	$\partial\phi(1\to0)/\partial x_1(2\to1)$	x_1	x_2	$\partial\phi(1\to0)/\partial x_1(1\to0)$	x_1	x_2	$\partial\phi(1\to0)/\partial x_1(2\to0)$
2→1	0	1	1→0	0	0	2→0	0	1
2→1	1	0	1→0	1	1	2→0	1	0
2→1	2	0	1→0	2	0	2→0	2	0

According to the Table II there is every of DPLDs $\partial\phi(1\to0)/\partial x_1(2\to1)$, $\partial\phi(1\to0)/\partial x_1(1\to0)$ and $\partial\phi(1\to0)/\partial x_1(2\to0)$ has non-zero values. These values reflect the system CSSV that are (2→1, 0), (1→0, 1) and (2→0, 0) These CSSVs allow us to define three scenarios of twin-engine jet crash depending of the first engine if troubles of engines been already. The first of them is jet crash depending on the first engine power reduction by half if the second engine is failed. The second scenario of jet crash is failure of the first engine if the second engine at half power only. The third scenario is similar to the first: the failure of the first engine causes jet crash if the second engine is failed. Therefore according to these scenarios the first engine power reduction by half or failure can cause jet crash if there are problem of the second engine.

The CSSVs for the system performance level 2 and the first component are computed by DPLDs $\partial\phi(2\to1)/\partial x_1(2\to1)$, $\partial\phi(2\to1)/\partial x_1(1\to0)$ and $\partial\phi(2\to1)/\partial x_1(2\to0)$, and $\partial\phi(2\to0)/\partial x_1(2\to1)$, $\partial\phi(2\to0)/\partial x_1(1\to0)$ and $\partial\phi(2\to0)/\partial x_1(2\to0)$. The CSSV calculated by derivatives $\partial\phi(2\to1)/\partial x_1(2\to1)$, $\partial\phi(2\to1)/\partial x_1(1\to0)$ and $\partial\phi(2\to1)/\partial x_1(2\to0)$ agree with land on foamed runway for a twin-engine jet (Table III). The scenarios of land of foamed runway of two-engine jet is defined based on these CSSVs. The first of these scenarios is problem in landing of jet depending on the first engine power reduction by half if the second engine is at half power. The second and third scenarios represent problem in landing depending of the first engine failure. It is possible if the second engine is at full power.

TABLE III. CSSVs of Land on Foamed Runway of Two-Engine Jet

x_1	x_2	$\partial\phi(2\to1)/\partial x_1(2\to1)$	x_1	x_2	$\partial\phi(2\to1)/\partial x_1(1\to0)$	x_1	x_2	$\partial\phi(2\to1)/\partial x_1(2\to0)$
2→1	0	0	1→0	0	0	2→0	0	0
2→1	1	1	1→0	1	0	2→0	1	0
2→1	2	0	1→0	2	1	2→0	2	1

The scenarios of jet crash in case of expected normally land is formed based on CSSV defined by DPLDs $\partial\phi(2\to1)/\partial x_1(2\to0)$, and $\partial\phi(2\to0)/\partial x_1(2\to1)$, $\partial\phi(2\to0)/\partial x_1(1\to0)$ and $\partial\phi(2\to0)/\partial x_1(2\to0)$. There is only one CSSV for these conditions (Table IV) that is (2→0, 1).

Therefore jet crash is possible if the second engine is at half power and the first engine is failing.

TABLE IV. CSSVs OF TWO-ENGINE JET CRASH

x_1	x_2	$\frac{\partial\phi(2\to0)}{\partial x_1(2\to1)}$	x_1	x_2	$\frac{\partial\phi(2\to0)}{\partial x_1(1\to0)}$	x_1	x_2	$\frac{\partial\phi(2\to0)}{\partial x_1(2\to0)}$
2→1	0	0	1→0	0	0	2→0	0	0
2→1	1	0	1→0	1	0	2→0	1	1
2→1	2	0	1→0	2	0	2→0	2	0

The map of scenarios of two-engine jet crash and degradation depending of the first engine state change is shown in Fig.1. This map combine all CSSs for this system for all possible degradation of the first component and any performance levels.

	The components		The system performance levels		
	x_1	x_2	2→1	2→0	1→0
Crash	2→1	0			x
	2→0	0			x
	1→0	1			x
	2→0	1		x	
Degradation	2→1	1	x		
	2→0	2	x		
	1→0	2	x		

Fig. 1. The map of two-engine jet behavior depending of the first engine state change

The specification of CSS and CSSV allows obtaining scenarios of the system behavior depending on its components states changes. Consider quantitative analysis of CSS that permits to estimate the probabilities of different CSS.

B. *Quantitative Analysis*

Qualitative analysis of CSS indicates probabilities of these states or provide the probabilities of changes of investigated system performance levels caused by change of one of components.

Indicate the probability of CSS $(a_1, \ldots, s_i, \ldots, a_n)$ for MSS performance level change from j to $\tilde{j}$ depending on the i-th system component change from s to $\tilde{s}$ as [19]:

$$p_{(a_1\ldots a_n)}\begin{pmatrix} j\to\tilde{j} \\ x_i \\ s\to\tilde{s}\end{pmatrix} = p_{1,a_1}\cdot\ldots\cdot p_{i-1,a_{i-1}}\cdot p_{i,s_i}\cdot p_{i+1,a_{i+1}}\cdot\ldots p_{n,a_n}, \quad (13)$$

where the symbol $\begin{pmatrix} j\to\tilde{j} \\ x_i \\ s\to\tilde{s}\end{pmatrix}$ represents CSS indicated by vector state $\boldsymbol{x} = (x_1,\ldots, x_i,\ldots, x_n) = (a_1,\ldots, s_i,\ldots, a_n)$ for which $\phi(a_1,\ldots, s_i,\ldots, a_n) = j$ and $\phi(a_1,\ldots, \tilde{s}_i,\ldots, a_n) = \tilde{j}$, and this CSS is calculated as non-zero value of DPLD (12).

For example, suppose that probabilities of states of two engines of the two engine jet are equal: $p_{1,0} = p_{2,0} = 0.1$, $p_{1,1} = p_{2,1} = 0.2$ and $p_{1,2} = p_{2,2} = 0.7$. CSSV of this system are indicated in Tables II – IV. Probabilities of CSS according to (13) are calculated in Table V.

TABLE V. PROBABILITIES OF CSS OF TWO-ENGINE JET DEPENDING ON THE 1ST COMPONENT STATE CHANGE

Crash		Land on Foamed Runway		Crash	
CSSV	$p_{(x_1a_2)}\begin{pmatrix}1\to0\\x_1\\s\to\tilde{s}\end{pmatrix}$	CSSV	$p_{(x_1a_2)}\begin{pmatrix}2\to1\\x_1\\s\to\tilde{s}\end{pmatrix}$	CSSV	$p_{(x_1a_2)}\begin{pmatrix}2\to0\\x_1\\s\to\tilde{s}\end{pmatrix}$
(2→1, 0)	0.07	(2→1, 1)	0.14	(2→0, 1)	0.14
(1→0, 1)	0.04	(1→0, 1)	0.04		
(2→0, 0)	0.07	(2→0, 2)	0.49		

Analysis of CSSV in Table V shows that there are some states for fixed changes of system performance level (for example, for the jet crash - 1→0 and land on foamed runway - 2→1). It is possible that there are some states for fixed changes of the i-th components from s to $\tilde{s}$ and MSS performance level from j to $\tilde{j}$ if DPLD $\partial\phi(j\to\tilde{j})/\partial x_i(s\to\tilde{s})$ has some non-zero values. The probability of CSSs for MSS performance level change from j to $\tilde{j}$ depending on the i-th system component change from s to $\tilde{s}$ is calculated as sum of all CSS for indicated conditions:

$$p\begin{pmatrix} j\to\tilde{j} \\ x_i \\ s\to\tilde{s}\end{pmatrix} = \sum p_{(a_1\ldots s_i\ldots a_n)}\begin{pmatrix} j\to\tilde{j} \\ x_i \\ s\to\tilde{s}\end{pmatrix} \quad (14)$$

The probability of CSS (13) can be used for the estimation of other properties of MSS, for example, as probability of the MSS performance level fixed change depending of all possible changes of the i-th system component or the probability of all MSS performance level change caused by fixed changes state of the i-th components from s to $\tilde{s}$. Such probabilities can be calculated as sum of suitable probabilities of CSS (13).

For example, compute the probability of the crash of two-engine jet (for the system performance changes 1→0) based on data in Table V:

$$p\begin{pmatrix}1\to0\\x_1\end{pmatrix} = p_{(20)}\begin{pmatrix}1\to0\\x_1\\2\to1\end{pmatrix} + p_{(11)}\begin{pmatrix}1\to0\\x_1\\1\to0\end{pmatrix} + p_{(20)}\begin{pmatrix}1\to0\\x_1\\2\to0\end{pmatrix} = 0.18 \quad (15)$$

The probabilities of two-engine jet complications depending on the failure of the first jet (the component state change from 2 to 0) is:

$$p\begin{pmatrix}x_1\\2\to0\end{pmatrix} = p_{(20)}\begin{pmatrix}1\to0\\x_1\\2\to0\end{pmatrix} + p_{(21)}\begin{pmatrix}2\to0\\x_1\\2\to0\end{pmatrix} + p_{(22)}\begin{pmatrix}2\to1\\x_1\\2\to0\end{pmatrix} = 0.70 \quad (16)$$

The proposed method for the analysis of CSS based on DPLD allows investigate all influence of the component state change to the any system performance level. Background for this investigation is the probability of CSS (13) that allows calculate other probabilities (for example, as (15) or (16)). Other probabilities of this system failure or deterioration of the performance levels are calculated similarly.

V. Conclusion

The theoretical background of the use of MVL mathematical methods in analysis of MSS is considered in this paper. In particular, in this paper the DPLD is used for the calculation of CSSV. In term of MSS reliability analysis these derivatives allow defining the state vector for which the change of fixed component state causes the change of the system performance level. Every DPLD has 4 parameters: the initial state of component s, the changed state of component $\widetilde{s}$, the initial performance level j and caused performance level $\widetilde{j}$. The combination of these parameters allows investigate all possible changes of the system performance levels depending on all possible changes of fixed component states.

DPLD can be used for qualitative and quantitative analysis. Under the qualitative analysis DPLD indicate CSSs for all possible changes of fixed component state and indicated system performance level. It is necessary information for the development of scenarios of system behavior. The probabilities of CSSs (13) permit provide the quantitative analysis of MSS.

References

[1] A. Birolini, Reliability Engineering. Theory and Practice, 7th ed. Springer, 2014, doi: 10.1007/978-3-642-39535-2.

[2] W. G. Schneeweiss, "A short Boolean derivation of mean failure frequency for any (also non-coherent) system," Reliability Engineering & System Safety, vol. 94, August 2009, pp. 1363–1367, doi: 10.1016/j.ress.2008.12.001.

[3] T. Aven, Risk Analysis, 2nd ed., Wiley, 2015, doi: 10.1002/9781119057819.

[4] W. Kuo and X. Zhu, Importance Measures in Reliability, Risk, and Optimization, Wiley, 2012.

[5] E. Zio, "Reliability engineering: Old problems and new challenges," Reliability Engineering System Safety, vol. 94, no. 2, pp. 125–141, 2009, doi: 10.1016/j.ress.2008.06.002.

[6] B.Natvig, Multistate Systems Reliability Theory with Applications. Wiley, 2011.

[7] T. Aven, "Risk assessment and risk management: Review of recent advances on their foundation", European Journal of Operational Research, vol 253, pp. 1–13, August 2016, doi: 10.1016/j.ejor.2015.12.023.

[8] J.D. Murchland, "Fundamental Concepts and Relations for Reliability Analysis of Multistate System," in: Reliability and Fault Tree Analysis, Theoretical and Applied Aspects of System Reliability. SIAM, 1975, pp.581–618.

[9] R.E. Barlow, A.S. Wu, "Coherent Systems with Multi-State Components," Mathematics of Operations Research, vol. 3, pp. 275–281, 1978, doi: 10.1287/moor.3.4.275.

[10] A. Lisnianski, G. Levitin, Multi-state System Reliability. Assessment, Optimization and Applications. World Scientific, 2003.

[11] A.P. Wood, "Multistate Block Diagrams and Fault Trees," IEEE Trans on Reliability, vol. R-34, pp.236-240, March 1985, doi: 10.1109/TR.1985.5222131.

[12] J.C. Hudson, K.C. Kapur, "Modules in Coherent Multistate Systems," IEEE Trans on Reliability, vol. 32, June 1983, pp. 183–185, doi: 10.1109/TR.1983.5221522.

[13] R.A. Boedigheimer, K.C. Kapur, "Customer-Driven Reliability Models for Multistate Coherent Systems," IEEE Trans on Reliability. Vol. 43, March 1994, pp.46–50, doi: 10.1109/24.285107.

[14] L.Caldarola, "Coherent System with Multi-State Components," Nuclear Engineering and Design, vol. 58, May 1980, pp.127–139 .

[15] E. Zaitseva, "Dynamic reliability indices for multi-state system," in Proc. IEEE the 33rd Int. Symposium on Multiple-Valued Logic (ISMVL) 2002, pp. 287-292.

[16] E. Zaitseva, V. Levashenko, "Investigation multi-state system reliability by structure function," in Proc. of Int. Conf. on Dependability of Computer Systems (DepCoS - RELCOMEX 2007), June 2007, pp. 81-90, doi: 10.1109/DEPCOS-RELCOMEX.2007.28.

[17] M. Kvassay, E. Zaitseva, V. Levashenko and J. Kostolny, "Minimal cut vectors and logical differential calculus," in Proc. IEEE 44th International Symposium on Multiple-Valued Logic (ISMVL) 2014, pp. 167–172, http://dx.doi.org/10.1109/ISMVL.2014.37.

[18] E. Zaitseva, V. Levashenko, "Multiple-Valued Logic Mathematical Approaches for Multi-State System Reliability Analysis," Journal of Applied Logic, vol.11, September 2013, pp. 350–362, doi: 10.1016/j.jal.2013.05.005.

[19] E. Zaitseva, V. Levashenko, "Reliability analysis of multi-state system with application of multiple-valued logic," International Journal of Quality and Reliability Management, vol.34 (6), pp.862-878, doi: 10.1108/IJQRM-06-2016-0081.

[20] M.A. Tapia, T.A. Guima, A. Katbab, "Calculus for a multivalued-logic algebraic system," Applied Mathematics and Computation, vol. 42, March 1991, pp.255-285, doi: 10.1016/0096-3003(91)90004-7.

[21] W.-C. Yeh, C. Bae, C.-L. Huang, "A new cut-based algorithm for the multi-state flow network reliability problem," Reliability Engineering & System Safety, vol 136, April 2015, pp. 1-7, doi: 10.1016/j.ress.2014.11.010.

[22] Z.W. Birnbaum, "On the importance of different components in a multicomponent system," in Multivariate Analysis, P. Krishnaiah, Ed, Academic Press, 1969.

[23] M.J.Armstrong, "Reliability-importance and dual failure-mode components, IEEE Trans Reliability," vol. 46, Februar 1997, pp.212-221, doi: 10.1007/s11269-013-0475-0.

[24] S. Beeson, J.D. Andrews, "Importance Measures for Non-Coherent-System Analysis," IEEE Trans on Reliability, vol.52, March, 2003, pp. 301-310, doi: 10.1109/TR.2003.816397.

[25] A. Rauzy, "Mathematical foundations of minimal cutsets," IEEE Trans on Reliability, vol. 50, December 2001, pp. 389–396, doi: 10.1109/24.983400 .

[26] M. Kvassay, E. Zaitseva, V. Levashenko, "Minimal Cut Sets and Direct Partial Logic Derivatives in Reliability Analysis," in Safety and Reliability: Methodology and Applications – Proc. of the European Safety and Reliability Conference, CRC Press, September 2014, pp. 241–248, doi: 10.1201/b17399-37.

[27] K.C. Kapur, E. Zaitseva, V.Levashenko, "New indices for measure of parallel and series multi-state system reliability," in Proc. of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (DETC2005), vol.4, 24 -28 September 2005, pp.653-662

[28] S. Wu, "Joint importance of multistate systems," Computers & Industrial Engineering," vol.49, 2005, pp.63-75, doi: 10.1016/j.cie.2005.02.001.

[29] S.B. Akers, "On a Theory of Boolean Functions," Journal of the Society for Industrial and Applied Mathematics," vol.7, 1959, pp.487-498, doi: 10.1137/0107041.

[30] A.D. Talantsev, "Analysis and Synthesis of Certain Electric Circuits by Means of Special Logical Operators," Automation and Remote Control, vol.20, September 1959, pp. 874-883.

[31] G.A.Kukharev, V.P. Shmerko, E.N. Zaitseva, E.N. Multiple-Valued Data Processing Algorithms and Systolic Processors, Minsk: Science and Engineering, 1990.

[32] M.D. Miller, M.A. Thornton, Multiple Valued Logic: Concepts and Representations. Synthesis Lectures on Digital Circuits and systems. Morgan and Claypool Publishers, 2008, doi: 10.2200/S00065ED1V01Y200709DCS012.

Temporal Data Group Management

Synchronization layer using attribute oriented approach

Michal Kvet, Karol Matiaško
Department of Informatics, Faculty of Management Science and Informatics
University of Žilina
Žilina, Slovakia
Michal.Kvet@fri.uniza.sk, Karol.Matiasko@fri.uniza.sk

***Abstract*—Database management requires fast and reliable access to the data stored in the database for complex evaluating in intelligent information systems. Significant data amount must be handled, stored and consequently retrieved for analysis. Each data tuple is delimited by the time of occurrence or reflected by the validity. Temporal database approach uses various architectures based on data structure and characteristics. Object oriented approach is on the one side, the second type is covered by the attribute oriented approach. In this paper, we define interlayer based on synchronization resources to reduce the cost of the server processing, which requires data group management over the time with regards on data group member loss. Thanks to that, performance is significantly improved resulting in temporal query performance optimization.**

Keywords—object level temporal approach; attribute oriented model; hybrid data architecture synchronization

I. INTRODUCTION

Data definition and characteristics have been influenced by the application and information system development during the whole history. In the first phase, only small amount of data was processed due to resource limitations reflected by the price as well as the performance, reliability and speed of the components. In 60ties of 20th century, relational database system has been proposed. It has been based on the entities and relationships between them mapping the real objects into database layer. Due to performance limitations, definition has been strictly bordered by actual object definition and management. Thus, only current valid data have been stored. Such paradigm is still valid, even nowadays, although the possibilities and technical resources are widespread offering high quality and robust solutions. Soon after the relational approach definition, also temporal model management has been defined to ensure complex data accessibility over the time [1]. However, it was rather concept than possibility for real usage. During long time, researchers were evaluating and improving temporal possibilities for database management in many spheres – logical, architectural or conceptual. It lasted almost 20 years till the acceptable solution has been proposed. It is partially used even today and is based on extension of the primary key using the validity interval called as "uni-temporal extension of the relational theory into temporal sphere". Till that definition, any change of the attribute value causes overwriting the existing values. Historical values were easily replaced with no possibilities to get those values later.

II. UNI-TEMPORAL MODEL

Uni-temporal model is the result of the significant evolution and strong stream and requirements for complex data management over the time. Primary key has been mostly extended by the time spectrum definition, by characterizing start and end point of the validity interval (interval model types are described in [5] [8] [9] or just by using one time point modelling begin point of the validity. In that case, each newer state borders the validity of direct previous one, which causes significant performance problems with query definition and location, therefore interval modelling was preferred. Later, another perspective has been introduced. Unlimited validity was originally modelled using *NULL* values, however such values cannot be part of the primary key definition. Moreover primary key should be unique and minimal. As the consequence, new data types have been defined and developed for complex temporal data management. The main role in this problem played just *period* data type modelling interval. In comparison with original interval definition characterized by two attributes, no *NULL* values can be placed from the definition, whereas begin position must be always defined. Moreover, whereas such attribute value was considered atomic, no discussions about the primary key effectivity definition could be done. Complex temporal protocol has been defined in subsequent years resulting in ISO standard certification attempts. Although several attempts have been done, no solution has been approved, which led to complete abolition and development possibilities in this area in 2001. Therefore, data must be managed explicitly ensuring no data object states overlay. In this part, it is useful to mention the Allen relationships covering all positional data characterizing time interval (fig. 1).

In principle, time interval can be characterized by the two attributes limiting the duration and validity. Solution can also be based on modelling only one attribute, for which each direct descendent limits the validity of the previous state. Naturally, individual solutions can be transformed each other.

Validity definition using two attributes can be modelled using four characteristic types, which influence consequent data retrieval principles. Concept of the temporal validity modelling suggests these four representations:

- closed-closed,
- closed-open
- open-closed,
- open-open.

The definition influences, whether the limiting date value is part of the interval (closed representation) or not (open representation). Characteristic representation is modelled for both borders. Fig. 1 shows the mapping solution for conversion one interval type to another. Notice, that the expression „$+1$“ expresses one point in time based on granularity.

Representation	Predicate
$[a_1, a_2]$ *equals* $[b_1, b_2]$	$a_1 = b_1$ AND $a_2 = b_2$
$[a_1, a_2]$ *equals* $[b_1, b_2)$	$a_1 = b_1$ AND $a_2 + 1 = b_2$
$[a_1, a_2]$ *equals* (b_1, b_i)	$a_1 = b_1 + 1$ AND $a_2 + 1 = b_1 + b_i$
$[a_1, a_2)$ *equals* $[b_1, b_2]$	$a_1 = b_1$ AND $a_2 = b_2 + 1$
$[a_1, a_2)$ *equals* $[b_1, b_2)$	$a_1 = b_1$ AND $a_2 = b_2$
$[a_1, a_2)$ *equals* (b_1, b_i)	$a_1 = b_1 + 1$ AND $a_2 = b_1 + b_i$
(a_1, a_i) *equals* $[a_2, b_1]$	$a_1 + 1 = a_2$ AND $a_1 + a_i = b_1 + 1$
(a_1, a_i) *equals* $[a_2, b_1)$	$a_1 + 1 = a_2$ AND $a_1 + a_i = b_1$
(a_1, a_{i_1}) *equals* (a_2, a_{i_2})	$a_1 = a_2$ AND $a_{i_1} = a_{i_2}$

Figure 1. Interval representation type transformation [9]

Allen relationships describe all possible positional relationships between two time periods along the common timeline. 13 types can be identified and used for time evaluation in temporal environment, to sort the states, but also to compose complex states, when multiple section structures are used [1] [2] [9]. Individual complex states based on the common object definition must be [*excluded*] modelled by either [*before*] relationship or by [*meets*] relationship – end point of the validity of the previous state is also begin point of the consequent state definition. On the other hand, complex environment composition and evaluation based on references is based on [*intersects*] relationship position. In that case, individual state definition can either [*overlaps*] or [*fills*] each other. If one interval fills (relationship [*fills*]) the second, two situations can occur (fig. 2). Either they are both identical - [*equals*], thus there is no time point belonging to one, which does not belong to second interval, as well. The second situation occurs when the first interval is a subset of the second interval. In this case we are speaking about the relationship [*occupies*]. Again, the opposite relationship does not apply. The general relationship [*occupies*] can be divided into the relationship [*aligns*] and [*during*]. The name of the relationship [*aligns*] defines its properties - two intervals have the common beginning or end time of the interval (exclusively – only one of them is true). In this case we deal with the relationship [*starts*], where intervals have a common beginning, otherwise we use the relationship [*finishes*]. However, if the intervals have a common beginning, and common end of the interval, this is, as we already mentioned the relationship [*equals*]) [13] [14].

If the relationship is defined as [*occupies*] and the intervals do not have a common beginning and end time of the interval, the relationship [*during*] is defined - beginning of the first period occurs later than the start of the second, end of the first period occurs before the end of the second interval. Next figure shows the positional relationships based on the relationship [*fills*].

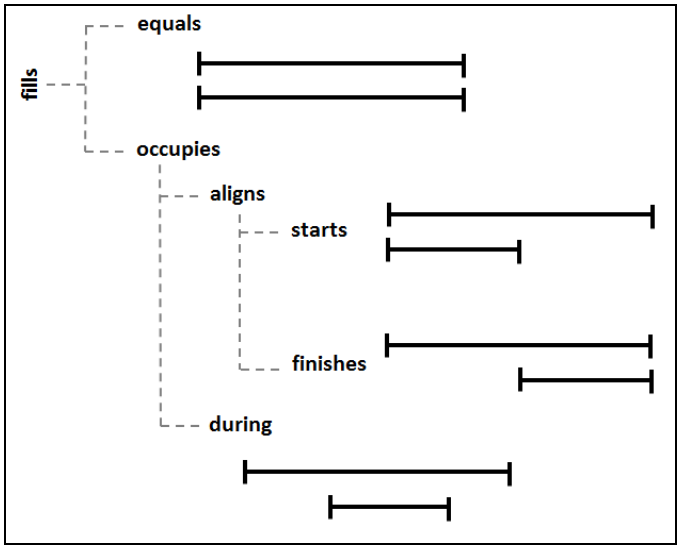


Figure 2. Interval definition [fills] covering [13]

Generalization of the uni-temporal model is just the bi-temporal system reflecting also transaction validity. In principle, we can define multi-temporal system by multiple time sphere definitions, like validity, transaction validity, locality, time perspective or time occurrence length [8]. Fig. 3 shows the models using various time spectrum definition.

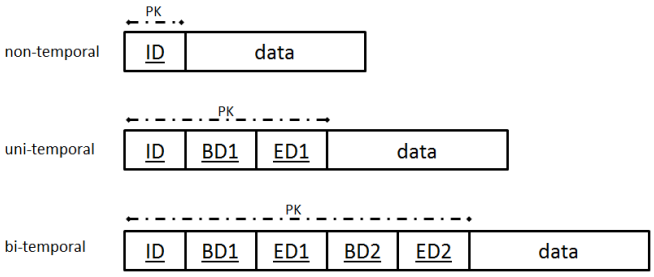


Figure 3. Temporal models (object level granularity) [9]

Previously defined evolution stream extends the characteristics using time borders. Therefore, primary key does not limit only the object itself, but defines validity of the particular object state during the time spectrum. It allows user to evaluate data and create complex future prognoses, statistics and reports.

Definition is associated with the whole object state. Thus, if only one attribute value is changed, the whole state is changed, which is correct, however, it is also physically reflected in the database by invoking new values forming the whole state. Therefore, existing values are copied into new states multiple times. *NULL* values do not solve the problems complexly. First of all, they reflect undefined values, however, for now, value is not undefined, but unchanged, which is significant difference [13] [15]. Moreover, *NULL* values have special denotation and meaning. Therefore, multiple duplicate tuples are stored causing significant performance degradation and demands for disc space and resource.

III. ATTRIBUTE ORIENTED APPROACH

Current temporal releases are mostly based on relational paradigm extended by the validity definition for the whole object state, which can bring many duplicates, if some attribute values do not evolve over the time. Solution based on data division possibility and frequency of changes is described in [18]. Universal solution provides transformation from the object granularity to the attribute definition itself. Thanks to that, any attribute temporal characteristics (conventional, temporal or even static attribute [13] [14] [18]) can be managed easily.

Performance and disc space demands force the developers to provide complex, robust and effective solution. Attribute value tuples are bordered by the validity and individual state based on object granularity is therefore created as composition of individual object parts – attributes. Moreover, in that case, undefined value of the attribute does not need to cause the whole object invalidity, which was particularity of the object access. Fig. 4 shows the developed and implemented architecture consisting of three layers. Current valid data are stored in accessible via the first layer – conventional access layer. Thus, existing applications can be easily shifted from conventional approach to temporal without necessity to rebuilding queries and the whole application itself. Second layer is temporal and forms the core part of the whole system. Individual object state or the attribute itself changes are managed using this layer. The last – third layer deals with the non-actual data (historical and future valid). Automatic transformation from future to current, current to historical and so on is provided by temporal management layer. Notice, that every communication is done by using temporal management. Actual valid states cannot manage non-actual directly at all to ensure security and consistency.

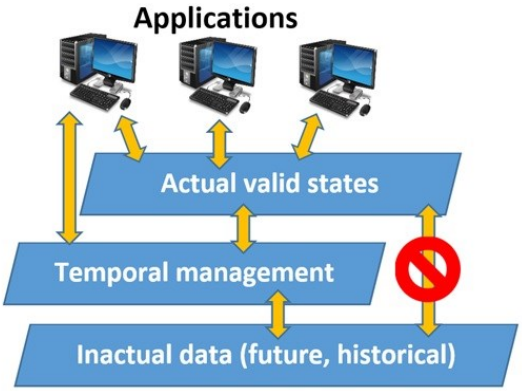


Figure 4. Attribute oriented temporal architecture

The core of the system is temporal table consisting of these attributes [9] [13] [14] [18]:

- *ID_change*
- *ID_previous_change* – references the last change of an object identified by *ID_orig*. This attribute can also have *NULL* value that means, data have not been updated yet, so the data were inserted for the first time in past and are still actual.
- *ID_tab* – references the table, record of which has been processed by DML statement (*INSERT, DELETE, UPDATE*).
- *ID_orig* - carries information about the identifier of the row that has been changed.
- *ID_column*, *ID_row* – hold the referential information to the old value of attribute (if the DML statement was *UPDATE*). Only update statement of temporal column sets *NOT NULL* value.
- *BD* – the begin date of the new state validity of an object.

Fig. 5 shows the data flow, when *Update* operation should be performed. In the first step *Update* statement is defined and temporal layer is contacted. Then, current data from the actual layer are moved into layer managing historical data, which can be done either directly or by using temporal layer. Then, in the step 4, new data are moved to actual layer. Last step confirms the operation by acknowledging data in all layers.

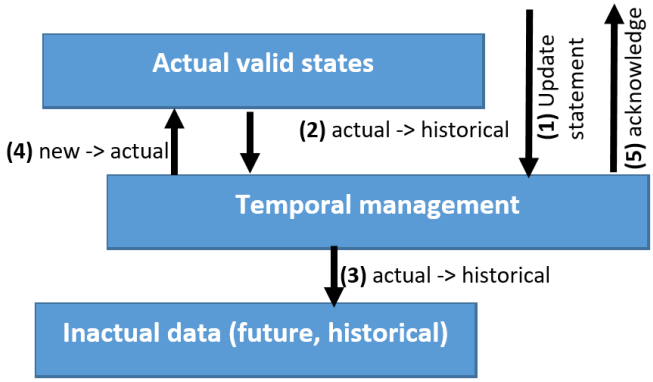


Figure 5. Data flow during the Update statement

In the recent past, temporal database approach has been extended by the various characteristics limiting locality position. Spatio-temporal database approach [3] [4] [10] is used for tracking moving objects determining single position at defined time point. Communication technology is mostly covered by the ad-hoc network architecture, where individual nodes are connected or disconnected based on communication possibilities and signal strength. In this field, reliability [11] of provided data must be highlighted. Moreover, to ensure communication possibilities and widespread signal, communication center positions must be optimized [7] [12]. Principles, definitions and limitations can be found in [3] [4] [6] [17]. In this sphere, attribute oriented approach shows its complexity and performance, however, some limitations should be also noticed. Therefore, we introduce our own solution based on the group management inside the attribute oriented approach. Thanks to that, hybrid solution is defined with aim to remove bottlenecks of the performance in communication and data protocol. Similar model used for spatial modelling is described in [16].

IV. Problem Statement

Temporal database characteristics are based on modelling object during the whole time spectrum with the possibility of data reconstruction during any timepoint or time interval. Object level temporal model uses extension of the primary key, thus each *Update* creates new state regardless the attributes, which have not been changed during the process. On the other hand, attribute oriented approach is based on attribute itself. These solutions can be considered as limiting factors (border solutions). In this paper, we combine existing solutions and create group management solution, if several attribute updates are synchronized. Group definition can be automatically recognized, evaluated, created and managed.

V. Own Solution

Our own proposed solution reflects the benefits from both object oriented architecture as well as attribute oriented model, but also highlight performance limitations. Sensor data management is suitable to be operated on attribute granularity, whereas individual values are updated with various granularity and frequency of changes. Some attribute values are updated very often, some of them even rarely. There may be even attributes, whose values are not changed and evolved over the time at all. Also attributes, which historical values can be stored only during strictly limited time period, should be mentioned. Therefore, the rest part of the attributes forming the state in the past will remain in the database, but some sensitive data should be removed. Notice, that by using object oriented approach, the whole state had to be removed, or unreliable value should be used to express removed value. Thus, another structure must be defined for such functionality management.

Data amount routed to the database interface is really significant and is enlarged over the time significantly. Although multiple listeners or even architecture of shared server session can be defined, by using dispatchers [2] [15], such solutions do not cover the complexity and do not solve the problems in total. In our proposed solution, we extend the architecture by group management. The aim is to lower the workload of the database system management. In principle, groups of data sensors are obviously defined, which are covered by the common board providing data synchronously. When attribute oriented architecture is used, data are provided in the batches. However, each data value reflects invoking separate *Insert* statement, which require significant system resources. Afterwards, similar batches are produced periodically, not only of the same group (board). Our architecture introduces the interlayers for the preprocessing. It identifies and evaluates the rate, frequency and quality of the input data flow, searches for the data synchronization and groups data together. Thanks to that, individual sensor management is disconnected from the database system itself and is routed to newly defined synchronization layer, which evaluates individual data and sends the processing based on Epsilon definition as well as reliability and correctness definition to the database system, which, however, does not deal with the attribute themselves, but with groups, which were created by synchronization layer. Moreover, although some data portions cannot be directly synchronized from the definition, if the evaluation, pre and post-processing of such values do not require to be time-strict, evaluated input values can be buffered using synchronization module and aggregated values are then stored. In this case, all attributes are automatically routed to the database system management processing. Running can be influenced only partially by dispatcher layer definition controlled by several listeners, however, it is not robust solution.

Architecture of the synchronization layer is proposed. In the first phase, automatic synchronization is searched for. If possible, such data are grouped. If not, subsequent step is based on evaluation in the distributed environment, where common parts and opportunities for synchronization are identified.

VI. Temporal Group Management Architecture

Our proposed solution extends the architecture of attribute oriented model by adding three new layers managing temporal groups. The first layer is based on automatic group detection. The second layer deals with the group definition, manages creations and dropping operations of the groups. The last group covers the synchronization processes (fig. 6).

Two parameters influence complex group management behavior – number of changes at the same time, after which group can be created. During the processing, possibility of group removal and destruction is monitored. Thus, the second parameter delimits, how many non-synchronous changes can occur until the group is removed. In principle, it is possible to drop the group immediately, but it should be noted, that such situation can be exceptional specific and immediate abolition would increase the costs to move association to the main database node, as well as the costs for the recreating group after defined number of changes. Therefore, it is necessary to find suitable solution with regards on group management, performance and significance of the change during the defined time point (cannot be processed later at all).

Fig. 6 shows the architecture of the solution. First layer is detector of the group. It has only *Select* privilege of the main database structure. The aim of such module is to detect and evaluate individual changes and reflecting the possibilities of group management (*create, alter, drop*) for the group manager layer (second layer of the architecture). This layer deals with group definition, reconstructs its definition, if some attribute should be removed or added, can form the whole group or even removes the definition completely. Group management and connection to the sensors is handled by the synchronization layer (third layer) to the database node or vice versa. Thus, group manager is the core of the system - master, the other layers are executors. Existing attribute oriented architecture is used for data management themselves.

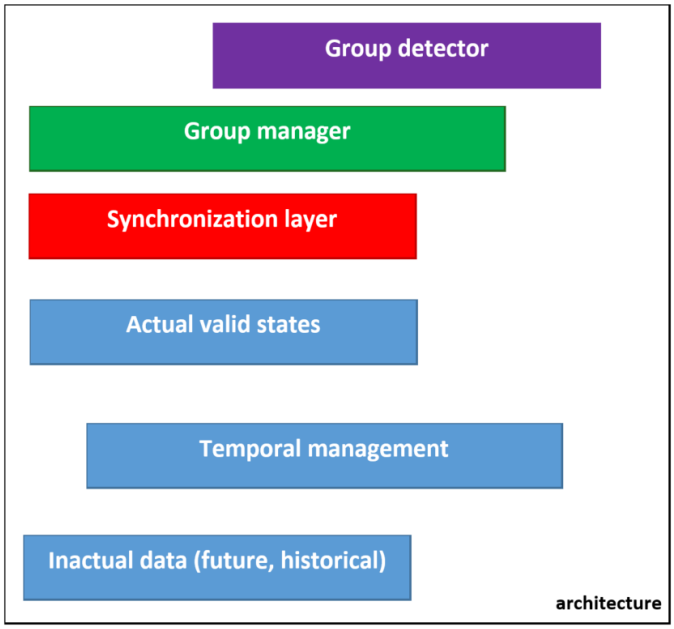


Figure 6. Temporal group management architecture

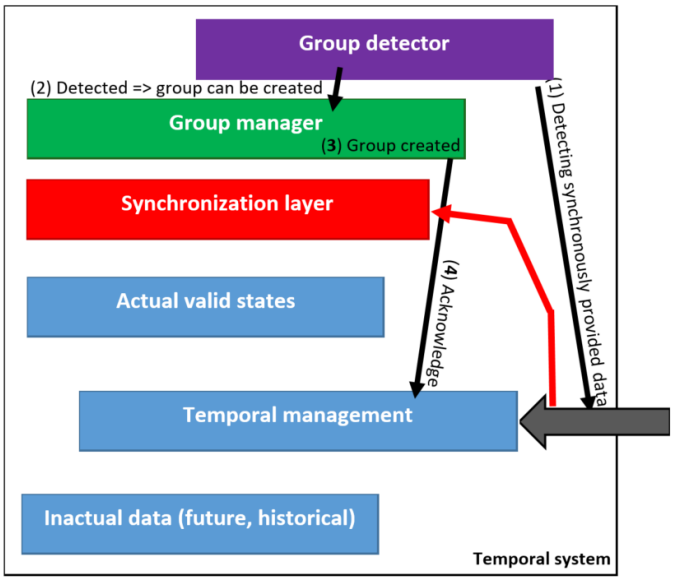


Figure 7. Group detection, management and synchronization of the values inside

Data management and message sending principle is shown in the fig. 7. If the group is detected (provided automatically), first layer (*detector*) sends the message to the second layer (*manager*) and asks for creating the group. Manager creates the group and notifies the *detector*. Afterwards, individual attributes are added to the group forming extension. Consequently, particular data are formed and rerouted from the database layer itself to the synchronization layer (red arrow in fig. 7), which has three tasks:

- Synchronize input data values and form the batch.
- Load the batch into the database.
- If data are coming asynchronously, it must notify group *manager* to evaluate the situation.

Some attributes cannot be grouped at all due to the data value structure or due to data character. Such information can be explicitly defined in *group manager layer*, which ensures such functionality by notifying *detectors* (particular data stream cannot be handled and evaluated by the *detector* layer) by status value - *Disable*. *Enable* value allows group definition using particular attribute. There is also another option – *Manual* – which ensures, that the user must confirm the possibility of creating group with a given attribute explicitly.

VII. PHYSICAL GROUP REPRESENTATION

At the beginning of the processing, each attribute is managed and evaluated separately, thus no group is defined. After the defined number of performed *Update* statements, individual groups can be formed. In principle, group can be created from individual attributes or can be considered as extension of the existing group. Therefore, from the database processing, same approach must be used for attribute itself as well as group definition. For these purposes, *ISA hierarchy* for the group management is proposed. Let consider data value (*data_val*) as the root (source) of the *ISA hierarchy*, which can be formed either by the attribute or the group (fig. 8).

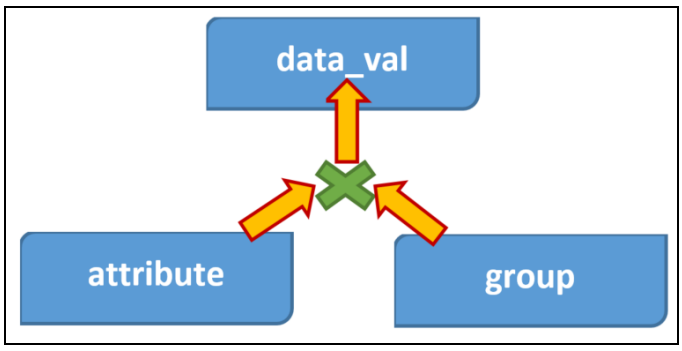


Figure 8. ISA hierarchy logical model

Individual group consists of several attributes, which can be encapsulated to existing group. Therefore, the complex model looks like following (fig. 9). Individual attribute association is delimited by the time, in our case, object level temporal model is used using time interval (*date_from, date_to – MaxValueTime* definition is used [9] [15]).

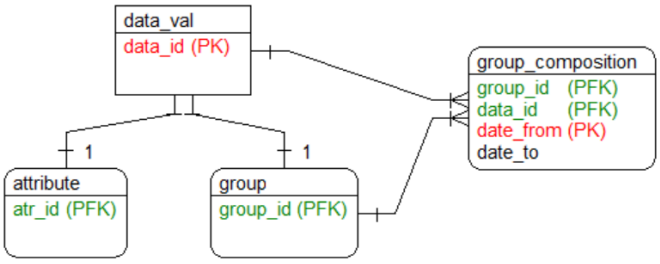


Figure 9. ISA hierarchy data model – physical representation

For the group management and evaluation, Allen relationships are used.

VIII. Data Security

Security of sensorial data is crucial part not only for temporal system, but also for each information system as well. Therefore, to ensure quality, consistency and security of the whole system, individual sensors are duplicated covering different communication techniques and protocols to provide data from the sensors themselves to the connection interface of the processing database and information system. Let have 3 sensors processing the same value. It is not necessary to store all the values, whereas they should be same (using the assumption, that all sensors as well as communication network show no errors). The evaluation can be done by the synchronization layer, which is securely connected to the database system. Notice, that the synchronization layer cannot be remote, it is mostly the subpart of the whole database system. The reason is based on possible data failures and complex data reliability.

For the purposes, technical parameters of the sensors should be also evaluated. If the data quality, reliability and frequency of data access is the same, no problem can occur. Many times, however, such control mechanisms are provided by several different measuring devices with various quality and characteristics. In this case, two situations can occur:

- All of the sensors measuring the same dimension have the same weight (fig. 10). In this case, principle two of three, respectively the majority is used. It causes the necessity for data provider synchronization – measurement and evaluation frequency is determined by the slowest member of the group. Such solution is inapplicable, if significant differences between individual sensors are detected. Obtained data cannot be significantly delayed, because at that changed time, such value does not need to reflect reality. Thus, to solve it, we proposed two possibilities – *Epsilon* value definition limiting the maximum displacement in time. Another solution is based on hierarchical architecture.

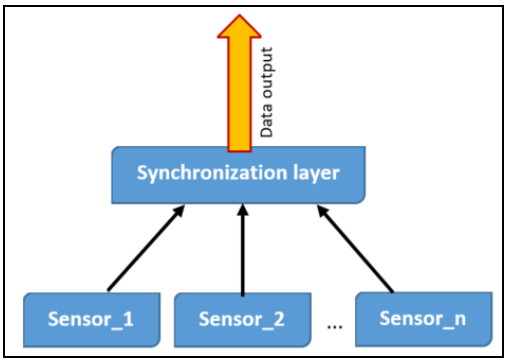


Figure 10. Data provider synchronization – same priority

- Hierarchical architecture (fig. 11) sorts the sensorial equipment based on quality reflected by the precision, reliability, error probability and so on. In this case, the most precise equipment is noted as main, other ones are supporting. Individual values are always sent to synchronization layer, however only value from the main (master) equipment is evaluated (such definition can evolve over the time e.g. by replacing individual sensors). Supporting sensor values are evaluated in time of accessing synchronization layer. Then difference between main and supporting values is evaluated. If it does not meet the defined precision range, original value is noted as non-reliable requesting user to investigate the situation manually. Otherwise, automatic sensor failure detector methods are launched to get rid of broken sensor. Naturally, user is noticed and particular data are no longer evaluated. Administrator must check the problem and test the provided data. Then, sensor must be replaced or fixed. Afterwards, user must close the issue in the administration tool, control and guide the system to solve the problem of inconsistency.

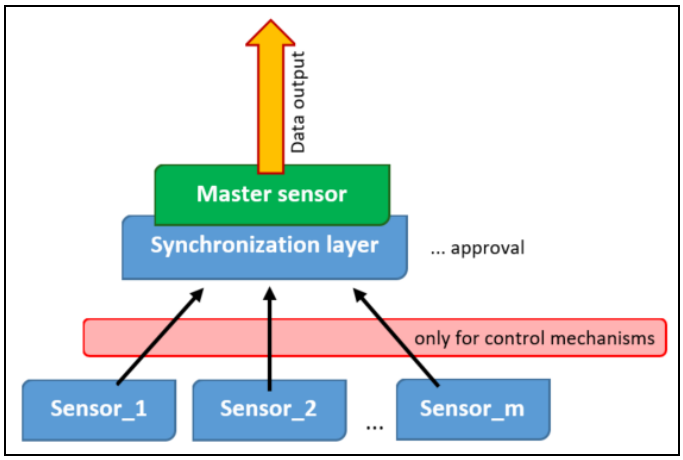


Figure 11. Data provider synchronization – master sensor architecture

IX. EXPLICIT GROUP DEFINITION

Previous data model and management principles assume automatic data group detection. However, as described in the previous chapter, such solution cannot be used generally. If you have multiple sensors of different quality, relevant group cannot be created automatically at all, whereas the time range of data provides is not the same. Another example forms the delays before the group formation and definition, since the group detector layer must evaluate multiple consecutive changes to decide the possibility to establish new drop or to force the system to remove existing group. Therefore, individual associations and definition can be done manually. How it works? During the definition or management, user can explicitly define and recreate defined groups to ensure, that individual attributes will be grouped based on his concept (or even not at all). Solution definition is interconnected with the data group manager, which ensures, that all user requirements and characteristics are met. Data group manager, moreover, continuously acknowledges rest part of the system (other layers). Nevertheless, if there is a situation, that some layer wants to change group definition, because information from data group manager has not been synchronized yet, defined change is suspended by data group manager layer, which is directly connected to the administration interface. It ensures, that it always has up to date information (fig. 12). Model representation is based on extension by another table - the group itself characterized by primary key – *exp_group_id* delimited by validity interval (*date_from, date_to*). Such group is composed from individual attributes using M:N relationships – association entity *group_part* is therefore added modelling bind of the attribute to the group. Notice, that triggers ensure, that attribute from the explicitly defined group cannot be removed automatically based on detector layer. Complete data model is shown in fig. 12.

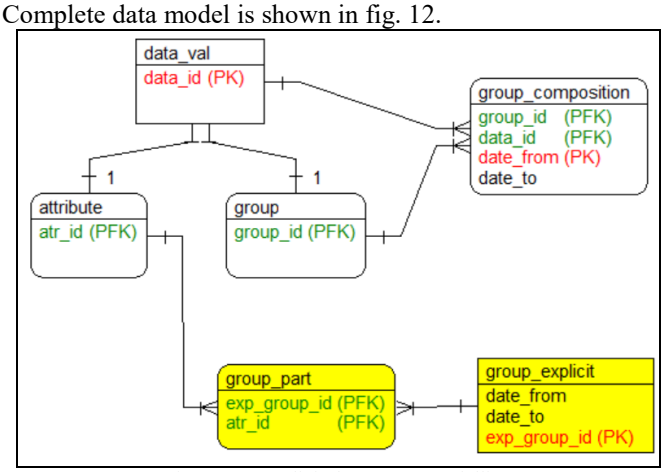


Figure 12. Explicit data group modelling

X. TRANSACTION MANAGEMENT

During the data management and processing, consistency of the database must be ensured as well as the complex reliability of calculations related to the execution of the query. Transaction definition is considered as the basic unit of the atomicity, consistency, reliability and provides data reflected to the database with regards on durability. Database is in consistent state, when all integrity constraints are met. When distributed database processing is used, transaction synchronization must be defined to provide the same transaction end in each node. Although the processing in our temporal architecture is distributed and parallel, it is necessary to highlight, that only temporal layer can change the data themselves stored in the database. The other layers can naturally query data, but cannot change them. Thanks to that, robust and safeguard solution is defined. As a consequence, two phase commit protocol can be used, however the logical journal consistency step is extended by the temporal definition of temporal access layer. So, *Ready To Commit* flag ensures data written in log files as well as temporal layer, as well. Moreover, for the security reasons, security flag can be set to ensure, that although transaction is roll backed, particular data are not removed from the temporal system, but are designated as unused due to transaction management. Fig. 13 shows the transaction protocol definition. *BOT* defines *Begin of Transaction*, *EOT* associates *End of Transaction*.

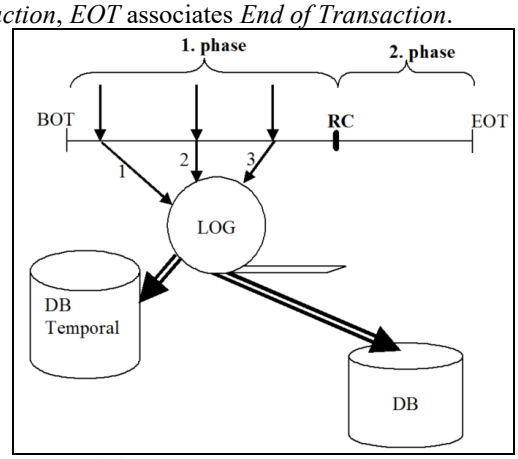


Figure 13. Extended two phase Commit protocol – temporal environment

XI. PERFORMANCE

Experiment results were provided using Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production; PL/SQL Release 11.2.0.1.0 – Production. Parameters of used computer are:

- Processor: Intel Xeon E5620; 2,4GHz (8 cores),
- Operation memory: 16GB,
- HDD: 500GB.

Experiment characteristics are based on real environment consisting of *1000* sensors producing data *ten times for one second*. If the difference between consecutive values is lower than *1%*, such values are not stored in the database and original value is considered as unchanged. Thus, based on our environment, average amount of new values is approximately *1000* per second. Amount of data after one hour is *3 600 000*.

The first part of the experiments consists of the comparison of the object and attribute oriented architecture. Processing time, costs and resource consumption (CPU percentage) are evaluated (tab. 1). For the consecutive evaluation, group is created after 10 consecutive value synchronization. On the other hand, group is reconstructed after two position desynchronization. As you can see, using

such defined environment, defined group management architecture propose 22,73% improvement in time processing. This is the consequence of parallelism possibilities based on synchronization layer and the group management. Processing costs are lowered using 13,12%. That is the consequence, that all the layers are using the same hardware resources. In the fig. 13, following abbreviations are used: *OLTA* – object level temporal architecture, *AOTA* – attribute oriented temporal approach and *TGMAOA* – temporal group management using attribute oriented approach.

TABLE I. RESULTS - OBJECT, COLUMN AND GROUP MODELLING

	OLTA	TGMAOA	AOTA
Costs	17 011	2 945	3 447
CPU (%)	52	9,6	11
Processing time (s)	257,5	40,4	52,3

When comparing our own proposed solution with object level temporal architecture, significant improvement is reflected – 82,69% for costs, 81,54% for CPU resources and 84,31% for processing time.

Graphical representation of the results is shown in fig. 14.

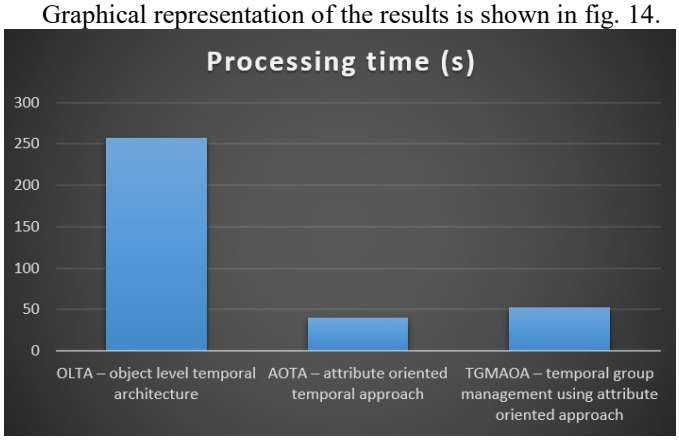


Figure 14. Processing time – object, column and group modelling

Second category of the experiments deals with the impact of group definition. 20% of the data can be grouped, probability of the desynchronization is 30%. The following table (tab. 2) shows the experiment results underlined by the graphical representation in fig. 15. For the easier evaluation, results are normalized and express the improvements in comparison with no data groups definition. With the increase of the number of groups, performance benefits, because of the no data shifting necessity.

TABLE II. RESULTS - GROUP DEFINITION PERFORMANCE BENEFITS

Number of groups	1	2	5	10	20	30	50
Performance improvements (%)	0,1	0,3	0,4	0,7	2,3	3,6	6,1

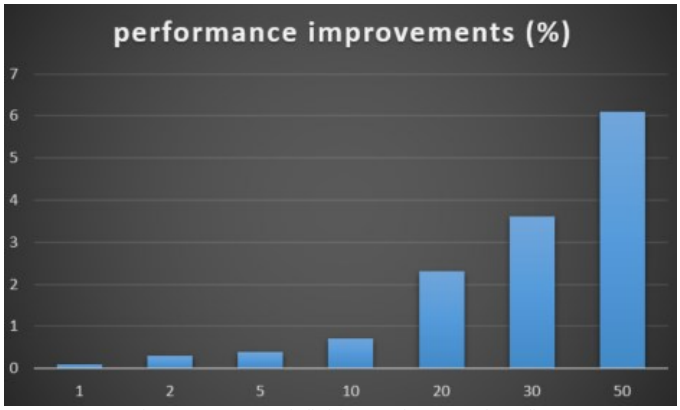


Figure 15. Group definition performance benefits

The last category deals with the principles of the group rebuilding and reconstruction. For the evaluation, we used the assumption, that the group can be created after double of the number of synchronized changes in comparison with the number delimiting group reconstruction or even removal. In principle, higher parameter values of the synchronization causes higher performance benefits. Fig. 16 shows the impact of border of desynchronization value characterizing requirement for group reconstruction. X-axis reflects the value of synchronization parameter, Y-axis models performance benefits in percentage.

TABLE III. RESULTS - GROUP REDEFINITION PERFORMANCE

Number of groups	1	2	5	10	20	30	50	100
Performance improvements (%)	0,10	0,13	0,22	5,82	8,23	9,26	9,32	9,31

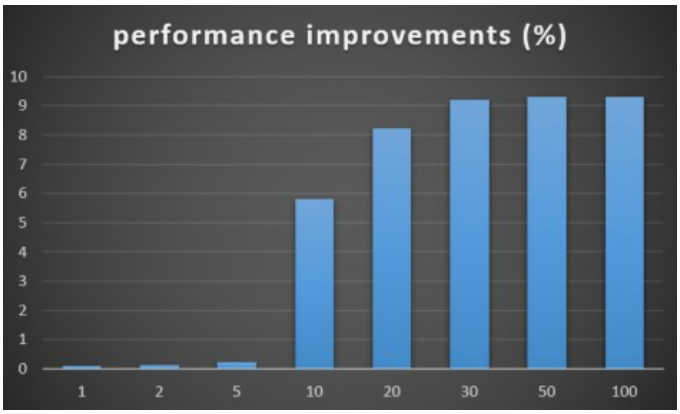


Figure 16. Group redefinition performance

XII. CONCLUSIONS

Current information systems deal with significant data amount, which must be evaluated, processed, stored and consequently obtained, if particular *Select* statements are defined. Important part is just the performance of the whole system. Individual states evolve over the time and must be stored in the database with regards on validity interval. It therefore reflects the transition from the conventional

approach, which processes only current valid data to the complex temporal structure dealing with the current, historical and also future valid states. For these purposes, several temporal definitions and characteristics have been proposed for particular area of deployment. Object level temporal architecture groups all attributes to the common block called object, thus object granularity is used, which can bring many duplicate values, if data updates are not synchronized. Opposite solution is just based on the attribute itself granularity, for which each attribute is associated with validity definition. Object state is composed dynamically based on actual states of each attribute definition. Thanks to that, temporal attribute definition can be managed dynamically, individual attributes can be added or removed from the system at any time. On the other hand, if some data portions are synchronized, effectivity and performance is violated. Therefore, in this paper, temporal group definition has been proposed by extending attribute oriented approach using three layers – *manager, detector* and *synchronizer*. The principle is based on automatic synchronized attribute changes detection reflected in automatic group definition and restructuralization, if necessary. Therefore, the database workload itself is lowered, however with no data loss. Temporal layer can then reflect not only attribute itself, but also the group during the defined time point of interval.

In the future, we will extend the group definition management into pure distributed environment with regards on data synchronization and rebalancing based on actual group definition and association. Group can be defined based on data on multiple nodes, however, after its creation, for the accessibility and easy manipulation, particular data must be shifted to the same node. Therefore, it is necessary to evaluate parameters and principles of the group definition as well as consequent data group reconstruction. Significant performance aspect is just rebalancing. Another spectrum, which will be researched, is based on automatic storage management in distributed environment.

ACKNOWLEDGMENT

This publication is the result of the project implementation: *Centre of excellence for systems and services of intelligent transport*, ITMS 26220120028 supported by the Research & Development Operational Programme funded by the ERDF and *Centre of excellence for systems and services of intelligent transport II.,* ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

This paper is also supported by the following project: "*Creating a new diagnostic algorithm for selected cancers*," ITMS project code: 26220220022 co-financed by the EU and the European Regional Development Fund.

"PODPORUJEME VÝSKUMNÉ AKTIVITY NA SLOVENSKU
PROJEKT JE SPOLUFINANCOVANÝ ZO ZDROJOV EÚ

REFERENCES

[1] K. Ahsan, P. Vijay. "Temporal Databases: Information Systems", Booktango, 2014.

[2] L. Ashdown. T. Kyte "Oracle database concepts", Oracle Press, 2015.

[3] G. Avilés et all. "Spatio-temporal modeling of financial maps from a joint multidimensional scaling-geostatistical perspective", 2016. In Expert Systems with Applications. Vol. 60, pp. 280-293.

[4] R. Behling et all., "Derivation of long-term spatiotemporal lanslide activity – a multisensor time species approach", 2016. In Remote Sensing of Environment, Vol. 136, pp. 88-104.

[5] C. J. Date, N. Lorentzos, H. Darwen. "Time and Relational Theory : Temporal Databases in the Relational Model and SQL", Morgan Kaufmann, 2015.

[6] M. Erlandsson et all., "Spatial and temporal variations of base cation release from chemical weathering a hisscope scale". 2016. In Chemical Geology, Vol. 441, pp. 1-13

[7] J. Janáček and M. Kvet, "Public service system design by radial formulation with dividing points". In Procedia computer science [elektronický zdroj], ISSN 1877-0509, Vol. 51 (2015), pp. 2277-2286

[8] T. Johnston. "Bi-temporal data – Theory and Practice", Morgan Kaufmann, 2014.

[9] T. Johnston and R. Weis, "Managing Time in Relational Databases", Morgan Kaufmann, 2010.

[10] A. Kadir and N. Adnan, "Temporal geospatial analysis of secondary school students´ examination performance", 2016. In IOP Conference Series: Earth and Environmental Science, Vol 37, No. 1.

[11] M. Kvassay, E. Zaitseva, J. Kostolny, and V. Levashenko, "Importance analysis of multi-state systems based on integrated direct partial logic derivatives", In 2015 International Conference on Information and Digital Technologies, 2015, pp. 183–195.

[12] M. Kvet and J. Janáček, "Relevant network distances for approximate approach to the p-median problem. In Operations Research Proceedings 2012: Selected Papers of the International Conference of the German operations research society (GOR)", Leibniz Univesität Hannover, Germany, Springer 2014, ISSN 0721-5924, ISBN 978-3-319-00794-6, pp. 123-128.

[13] M. Kvet, K. Matiaško, "Transaction Management in Temporal System", 2014. IEEE conference CISTI 2014, 18.6. – 21.6.2014, pp. 868-873

[14] M. Kvet and K. Matiaško, "Uni-temporal modelling extension at the object vs. attribute level", IEEE conference UKSim, 20.11 – 22. 11.2014, , pp. 6-11, 2013.

[15] D. Kuhn, S. Alapati, B. Padfield, "Expert Oracle Indexing Access Paths", Apress, 2016.

[16] S. Li, Z. Qin, H. Song. "A Temporal-Spatial Method for Group Detection, Locating and Tracking", In IEEE Access, volume 4, 2016.

[17] Y. Li et all., "Spatial and temporal distribution of novel species in China", 2016. In Chinese Journal of Ecology, Vol. 35, No. 7, pp. 1684-1690.

[18] A. Tuzhilin. "Using Temporal Logic and Datalog to Query Databases Evolving in Time", Forgotten Books, 2016.

Temporal database management

Temporal registration

Marek Kvet
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
Marek.Kvet@fri.uniza.sk

Michal Kvet
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
Michal.Kvet@fri.uniza.sk

***Abstract*—Temporal databases are characterized by significant data amount monitored over the time. Such values are stored with regards on validity during the whole time spectrum supporting decision making, reliability definition, analysis, progress monitoring and creating future prognoses. When dealing with various granulairty of changes, management must be shifted into attribute granularity level to remove duplicates. In this paper, we define our own proposed temporal architecture with regards on element registration in the system. Whereas data model, defined attributes as well as temporality definition evolve over the time, it is necessary to create complex environment and possibilities for dealing with these changes and reflections to the temporal management layer. I tis done by temporal registration concept.**

Keywords- element registration; temporal approach; attribute oriented approach; temporal evolution;

I. INTRODUCTION

Temporal databases are based on assumption of storing significant data amount with different types over the time. In such systems, many data granularity types are used with various precision and frequency of changes. Typically, data retrieval is composed as image of the whole system or its sub-part at some time point or monitors individual changes of the elements over the time. Therefore, storage effectivity is strongly important, when dealing with data characteristics and evolution over the time. Advantage of conventional approach storing only actual valid data is just hte fact, that each new data tuple replaces existing one with no reflection to the history. On the one hand, we are losing history and evolution, however great advantage is just the simplicity, if the structure must be changed, either by adding, removing individual attributes, changing their data types or if the structure should be changed more complexly. After executing *Alter,* respectively *Create* or *Drop* commands, each change is directly and immediately applicable. In temporal environment, situation is more complicated, because also evolution of the data model must be highlighted, defined and stored, to create possibility to reconstruct not only states, but also the structure (data model) itself. In case of changing precision – increasing accuracy, such problem is not so sharp, because it can be done automatically at the cost of ineffective data management in the past – values are stored with high precision, but the data themselves do not have such accuracy. In this paper, we define individual possibilities of changing data characteristics, definitions and model itself over the time with regards on the temporal managemennt, data registration and reflection over the time.

In the first part of the paper, brief history introdution is proposed focusing on current temporal technologies. Afterwards, temporal table definition and classification is proposed with regards on the pure attribute oriented architecture developed by us. Individual characteristics and definitions can be extended by various principles based on group management [17], transaction processing [15], decision making, tuning [19], indexing [12] [18], reliability [13] and distributed data modelling [1] [20]. Nowadays, strong stream of temporal definition extended by the positional, data forming spatio-temporal solution can be perceived. Principles are defined in [4] [5] [11] using spatio-temporal characteristics.

II. HISTORICAL EVOLUTION AND ELEMENTS

Requirement for temporal data definition has become strong in the earlier part of the database systems proposal themselves. In the first phase, security and high accessibility were provided by the backups and transaction log files. By accessing such structures, historical image of the data could be obtained. It was, however, too complicated process due to necessity of backup loading, which required user intervention. Moreover, such backup usually covered the whole database or its major part, therefore it lasted too much time, mostly if data were located externally in separate storage capacity. Operational data for decision making could not be obtained at all. If the log file was deleted, there was no evidence about that fact, because database system itself does not manage historical (obsolete) log files, whereas it was used only for security reasons, if some physical file was lost or corrupted. As the consequence, it could not be used complexly and could even provide insufficient data without any user warning. Fig. 1 shows the principles. Let have two backups managed and stored by *RMAN* (*Recovery manager).* Log files themselves, however, cannot be supervised by *RMAN*, thus any lost of its part cannot be detected. As the consequence, from the log file loss period until complete backup is obtained, particular states can be incomplete and therefore considered as unreliable.

However, user does not have any evidence about such problem, therefore he cannot not react any way [2] [3].

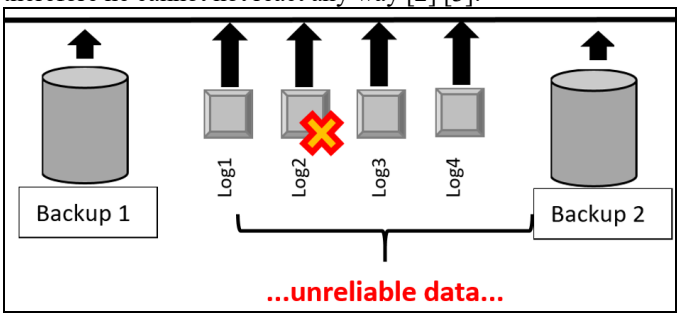


Fig. 1. Backup and log file data loss problem

Therefore, later, such functionalities and principles have been extended by the temporal transaction log definition and supported by the *Flashback technology* [3] [6] [9]. Concept has been proposed in Oracle 10g version and improved in Oracle 11g version. It allows you to define own methodology, own format and associate individual attributes to be managed temporally. It has brought the reduction of the data stored – only temporal attributes are monitored without any transaction information stored directly. Thus, data structure and evidence is optimized. It is automatically managed by *Flashback Repository process* based on defined attribute associations. Individual files are linked, referenced and managed automatically, so there is no possibility to delete referenced file. Thanks to that, provided solution is reliable, robust and powerful, provided data are correct. On the other hand, effectivity is ensured, only if the whole or partial data image should be obtained. In that case, it works like backup, but can be provided with regards on any timepoint in the past. Retrieving attribute or group evolution over the time is really complicated process resulting in inapplicability of defined solution, whereas it would require getting all partial images and changes reconstruction. Time processing of such data retrieval requirement is comparable with pure backup and log file management. Another limitation is just the system dependency – it can be used only in DBS Oracle. Other systems do not support it.

Fig. 2 shows the principles of *Flashback repository* management. If the data are updated and the change reflects temporal attribute, *Flashback Repository process* becomes active and stores particular historical data into *Flashback repository*. Thus, if the *Select* statement requires only actual data, they can be provided directly, historical data are accessed only by concacting *Flashback Repository process.* Data retrieval is then composition of the data provided by *PMON* (*Process Monitor*) – actual data, historical data are loaded from *Flashback repository* by particular process. Future valid data as core part of the temporality cannot be handled at all. Performance of such solution in comparison with our proposed solution is in section 6.

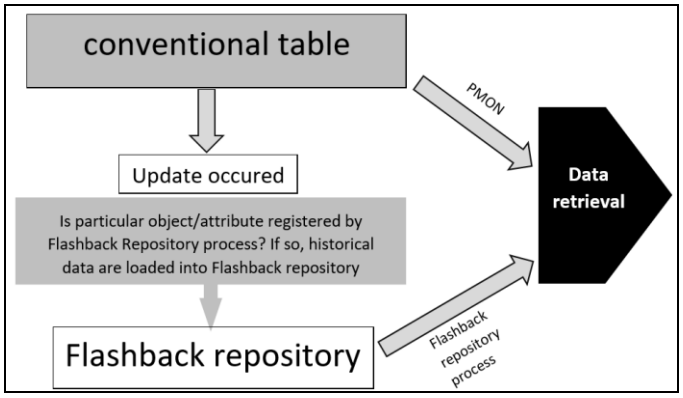


Fig. 2. Flashback repository management

The second research and development stream is based on temporal definition using the direct schema data. Thus, primary key of the table considered as temporal is extended by the time definition. In principle, it reflects at least validity, however, in real environment, such primary key extension can deal also with transaction definition (transaction validity), position, time occurrence, etc – fig. 3. However, there are several disadvantages to be mentioned. First of all, it requires data updates to be synchronized – if not all attributes are updated at the same time, several duplicate tuples would be stored. Moreover, such table must be fully temporal – each attribute is temporal regardless its definition. Last, but not least problem would cause table joining together – child validity intervals must be fully covered by master (parent) states (fig. 4). If it cannot be done by one tuple, composition must be executed to check correctness. In fig. 4 – there are 3 parent records associated with the same object. Each state is delimited by the time validity (*T1, T2, T3*). When joining with another table, particular child state must be covered by two parent records (*S1, S2*). If undefined object state would be defined between individual intervals, *join* would not be possible to execute at all, because of the covering [7] [8] [10].

As we can see, such developed solution is not user friendly, correct and robust management is too complicated. Therefore, the aim of the researchers was to create new temporal paradigm and management automatically covering illustrated problems. Validity intervals were transformed into values of the *period* data type proposed in temporal paradigm. Complete definition has been proposed even at the end of the 20th century with the aim to standardize it. Unfortunately, such definition has not been approved and was recommended to be improved. After several standardization attempts and consequent rejection of each one, defined concept has stopped to be developed and abolished completely in 2001. To conclude such management definition, it is available, but must be handled explicitly [2] [3].

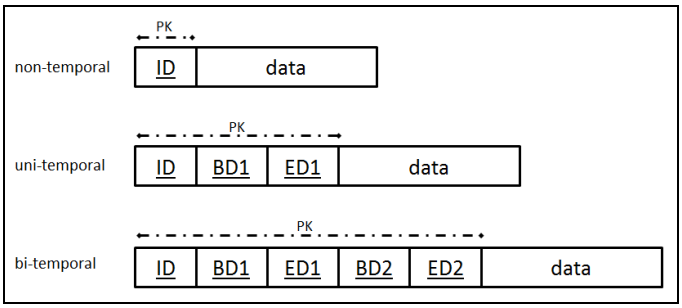


Fig. 3. Object level temporal definition [9]

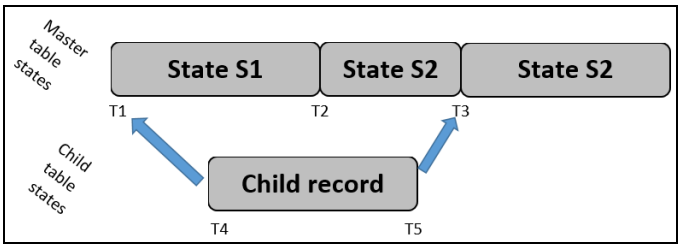


Fig. 4. States covering (referential integrity)

III. Own solution

In the previous chapter, research streams have been proposed with regards on the application usage and limitations. Our own proposed solution uses the benefits of the both stream systems with emphasis to eliminate boundaries. In comparison with validity temporal solution (it uses object granularity), our proposed solution uses attribute as the main element. Thanks to that, if granularity and frequency of the changes is not the same, no duplicate values are stored. It is based on three layers – the first layer deals with current valid states connected to the temporal layer, which forms the core of the system. Each change on temporal attribute is automatically referenced by such layer to ensure possibility of the state compositions. Non-actual states are stored in the third layer. It shows the perspective of the historical data as well as states valid in the future, which are automatically transformed to actual at defined timepoint. Architecture in shown in fig. 5.

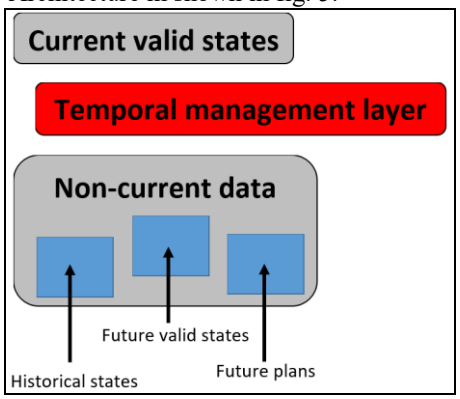


Fig. 5. Attribute oriented solution architecture

This concept has been developed by us and proposed in 2015. Data management, principles, properties can be found in [15] [16]. Later, the solution has been extended and optimized. In the pure way, each change of the temporal attribute directly invokes new *Insert* statement execution in the temporal layer. And it may be the source of inefficiency. Almost every temporal system is delimited by the sensor data processing (e.g. patient monitoring, industrial environment, transport systems). In principle, provided data have different granularity, precision and error handling. On the other hand, many times, to ensure data and measurement security, multiple sensors are used and consequently compared to create error immune measurement solution. In that case, partial data synchronization should be mentioned. Forming synchronized data into temporal groups provide effective solution and reflects performing only one statement into temporal layer. If the data are desynchronized, group is dropped or reconstructed either automatically or based on user requests. Principles are defined in [17].

Proposed attribute temporal approach with its improvements provide sufficient power, when dealing with data with various types, definitions, granularity and frequency of changes. It can robustly react to any change of the values.

In principle, each data attribute or tuple can be characterized by its temporal sphere. We can distinguish the following categories:

- ***Static*** – values stored inside cannot be changed later at all (onitialization parameters, code lists).
- ***Conventional*** - non-actual data are not necessary to be stored.
- ***Temporal*** – each attrribute definition is monitored over the time.
- ***Hybrid*** – combination of the previous options.

From this point of view, such defined attribute approach is hybrid.

So far, all temporal solutions have only dealt with the possibility of changing data, but the structure itself has remained the same. Temporal characteristics evolve over the time, provided data quality is still better and better, reliability and precision is higher. Therefore, also data model and complex structure must react to those situations and propose management of such changes with regards on structure defined in the past, but also states with impoved data model.

Therefore, in this paper, we deal with the possibilites to manage and solve such problems. Architecture of the proposed solution must be extended, specifically, another layer managing temporal model and structure is proposed. Processing is related, evaluated and secured by registration layer, which is managed by introduced background process – *Temporal Structure Manager (TSM).* It is the master process and if the number of changes is high, also executor processes can be defined (either automatically or manual) – *Temporal Registration Executor Processes (TRegExecn).*

IV. Logical view

Temporal *Data definition language* (*DDL*) statements are defined as the extension of the existing syntax by using kedwords definiting *temporality*. Access model can be

defined either using object or attribute granularity. If object granularity is used, each attribute is influenced by the definition. We propose the following temporality keyword options based on time management:

- ***Static*** (attribute value cannot be changed at all), applicable for table as well as attribute itself.
- ***Conventional*** (attribute can be changed at any time, historical value is not necessary to be stored later at all), applicable for table as well as attribute itself.
- ***Temporal*** (attribute is monitored over the time, each change is stored and accessible over the time), applicable for table as well as attribute itself,
- ***Hybrid*** (special type of the table, which allows you to store any type of the previously mentioned characteristics in the common table). Applicable only for the table. It cannot be used for attributes at all.

Fig. 6 shows the syntax of table definition command. In this case, keyword characterizing temporality is placed before the table structure and its name:

```
Create [{static | conventional | temporal | hybrid}] table table_name
 (atr1 data_type1,
  atr2 data_type2,
     ...
 );
```

Fig. 6. Table with temporality definition

Vice versa, *temporality* keyword can be defined for the attribute. The syntax of the solution would look like following (fig. 7). The first attribute (*atr1*) is *conventional*, the second attribute (*atr2*) is *temporal* and the last third attribute (*atr3*) is *static*.

```
Create table table_name
 (atr1 data_type1 conventional,
  atr2 data_type2 temporal,
  atr3 data_type3 static
 );
```

Fig. 7. Temporality of individual attribute definition

The previously defined commands are defined for table creation. Similar principle is also used for altering structure:

Fig. 8 reflects the transformation to the temporal, fig. 9 shows the syntax of individual attribute (*atr1*) transformation.

```
Alter table table_name convert to temporal;
```

Fig. 8. Changing existing table temporality

```
Alter table table_name convert attribute atr1 to temporal;
```

Fig. 9. Changing existing table attribute temporality

Object (table) as well as attribute dropping is a bit more complicated process. In that case, it is necessary to distinguish, whether defined object to be dropped is at least partially temporal (*temporal, hybrid*) or not (*conventional, static*). If data were not monitored over the time, particular object can be dropped directly. However, if the definition reflects temporality, two situations can occur:

1. particular referenced object can be dropped – in that case, keyword *purge* is used meaning that the whole reference is removed (from historical tables, planned calendar, future plans, actual states and also temporal table). Information about the existence in the past, however, remains in the temporal table for validation.
2. particular referenced object is not valid and will not be used for update in the future at all, however, it is necessary to store historical data continuosly along. In that case, archivation process is started to transfer data to the specific repository (*archive*). These activities are provided by the *Temporal Archive Master Process (TAM)* with its executors (*TAEn*) [17].

Principles are shown the fig. 10.

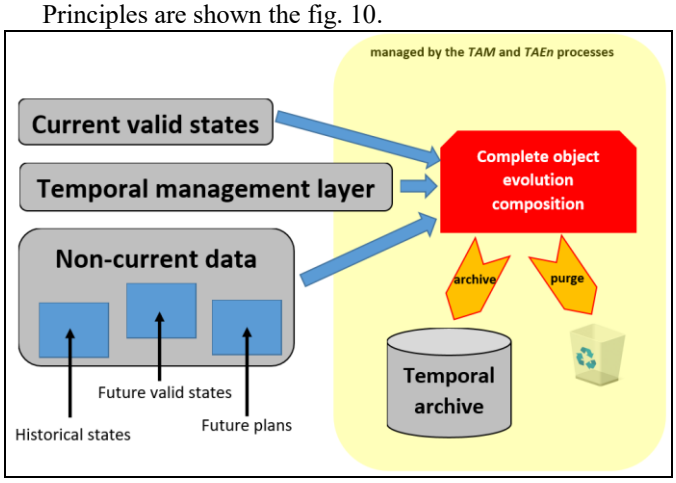


Fig. 10. Temporal Archive Master Process activities

The difference is based on using *purge* keyword. By default, archive process is used (fig. 11 – syntax).

```
Delete from table_name [{ archive | purge }];
```

Fig. 11. Removing object data syntax

From the logical perspective, script determines the temporal characteristics and processing environment. Before the *DDL* command execution, temporal declaration is identified and particular attributes or the whole tables are registered or de-registered using preprocessing layer, which identifies temporal extension of the commands and invokes *Temporal Registration Master Process (TRegMaster),* which is delimited by *Temporal Manager Process*, which is in charge of the entire temporal layer – fig. 12.

Fig. 12 shows also architectural model of the proposed solution – it is located in the temporal layer and interconnected to *Temporal Manager* process (yellow color box in fig. 12). As you can see, *Temporal manager* process invokes particular *Temporal Registration Master Process (TRegMaster)*, which provides required registration. It is done by the defined

package (*DBMS_TEMPORALITY*) with the following methods:

- ***Register*** (add new attribute/table)
- ***Change*** (reflects existing definition)
- ***Remove*** (purge/archive)

In physical way, *Register* method of the *DBMS_REGISTRATION* package is associated with altering table, which can be influenced by the following options:

- *Alter table => ADD*
- *Alter table => DROP*
- *Alter table => MODIFY*

The first two command definitions can be processed directly, however, if attribute definition is to be changed, it cannot be reflected as direct change of attribute data type in temporal layer, because existing data type would be lost. Therefore, physical denotation of the *MODIFY* option is replaced by the *DROP* of the exiting attribute (which is consequently moved into the archive repository) and *ADD* – adding new attribute definition.

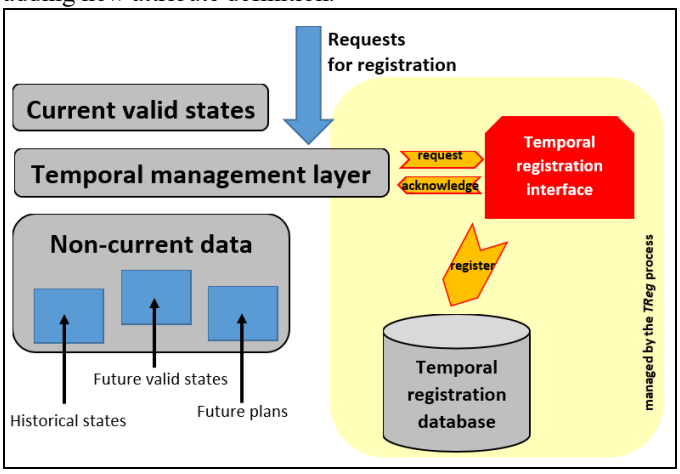


Fig. 12. Temporal registration

V. Physical Representation

For the physical representation, particular *Temporal Registration Master Process (TRegMaster)* must be launched. During the temporal database definition using our approach, it is started automatically immediatelly after *Temporal Manager Process* and becomes *inactive*. Status is changed to *active* only if the temporality definition is changed. Its aim is to to detect temporality from the *DDL* statements by the command decomposition. Afterwards, individual characteristics are registered. For these purposes, following table has been defined. As you can see, the structure is similar to attribute oriented architecture, the association is delimited by one attribute modelling interval – *BD*. *ED* can be calculated dynamically based on consecutive change. *Allen relationships* ensuring consistency are used for bordering previous definition [9] [10]. In principle, only relationship [*meets*] is used, whereas there can be no spaces (undefined states) inbetween. Each performed change of the model is triggered and enforced by current date delimiting previous valid structure definition. History and evolution can be obtained using linked list – *id_str_change* and *id_str_previous_change*. In theoretical way, also future structure redefinition operation can be used. In that case, structure would be automatically changed at the defined timepoint. However, it should be ensured, that no data states would be changed during the redefinition, otherwise such data would be lost.

Fig. 13 shows schema of the table used for storing structure and model over the time. Registration table itself consists of these attributes:

- *ID_str_change.*
- *ID_str_previous_change* – references the last change of the table/attribute identified by *ID_str_change*. It can also hold *NULL* value, which represents the process of adding (registering) new temporal table to the system.
- *Operation* – reflects the following options: *A* (add), *M* (modify), *D* (drop).
- *ID_tab* – references the table (each table has unique name (code) defined by the system and stored in data dictionary automatically).
- *ID_attribute* – carries information about added attribute – it is foreign key to the table, which holds the full attribute definition. The relationship type cardinality is 1:1.
- *BD* – begin data of the new state validity of the table specification.

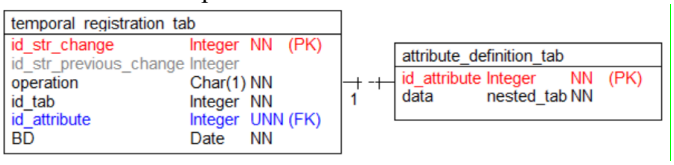


Fig. 13. Temporal registration data model

The fact that it was just a change of the type of attribute is expressed by a linked list of the changes (*id_str_previous_change*) in two layers. Main part (grey color in the fig. 14) represents the structural changes – adding or removing attribtues over the time. Changes on existing attributes are modelled for each attribute in separate deposit (blue color in fig. 14) – *NULL/NOT NULL, UNIQUE / DISTINCT*, *data types* characterizing *column* and *domain* integrity. Pointers are defined by the attribute *id_str_change* and *id_str_previous_change*.

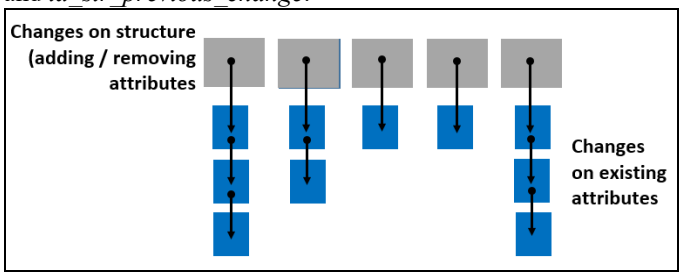


Fig. 14. Management of temporal structure

VI. Performance

Experiment results were provided using Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production; PL/SQL Release 11.2.0.1.0 – Production. Parameters of used computer are:

- Processor: Intel Xeon E5620; 2,4GHz (8 cores),
- Operation memory: 16GB,
- HDD: 500GB.

Experiment characteristics are based on real environment consisting of *1000* sensors producing data *ten times for one second*. If the difference between consecutive values is lower than *1%*, such values are not stored in the database and original value is considered as unchanged. Thus, based on our environment, average amount of new values is approximately *1000* per second. Amount of data after one hour is *3 600 000*.

The first part deals with the temporal structure definition using several approaches. First model (1) uses *object level temporal architecture* using *uni-temporality* defined by validity interval. Second model (2) uses pure solution using *Flashback backup and log files*, which is naturally the worst one, because of the necessity of data loading from the backup and activating stored log files. Notice therefore, that each backup as well as log file consists of the further set than data to be processed and retrieved, however, such data cannot be separated without full loading from binary files. Third model (3) is based on *Flashback repository* and *Flashback logs*. It provides robust solution, if image of the database should be provided at defined timepoint. However, it has significant weak side, if we need to get insights of the object changes over the time. Last – forth model is based on our *proposed temporal architecture*. In our solution, each temporal attribute data type is changed after 1 minute (totally performed 60 times).

Tab. 1 shows the provided results. Processing costs, CPU (%) and processing time (seconds) were measured using *autotrace* and *timing* functionality of the DBS Oracle.

For the evaluation, *object level temporal architecture* will be used as reference model – 100%. *Approach* (2) based on *backup* and *logs* is significantly worse, because of the loading necessity – slowdown more than 145% for costs, 155% for CPU and 148% for processing time. Therefore, such solution is totally inappropriate. On the other side, significant improvement provides *Flashback technology* – model (3) – improvement of 58% for costs, 59% for CPU and 59% for processing time. If the processing would deal with only database images and snapshots, performance would be much better – approximately 1% slowdown in comparison with model (4) performance (attribute oriented approach with registration functionality).

The best solution provides *attribute oriented model* (model 4), which can process and manage data model changes over the time. Global performance of the model (4) is following (reference is model (1)):

- Costs 80%
- CPU 76%
- Processing time 87%

TABLE I. Processing results

	Object level (1)	Backup & log files (2)	Flashback repository (3)	Attribute oriented approach + registration (4)
Costs	17 011	41 824	7 085	3 252
CPU (%)	52	133	21	12
Processing time (s)	257,5	640,8	103,3	32,8

Fig. 15 shows the graphical representation of the results – processing time.

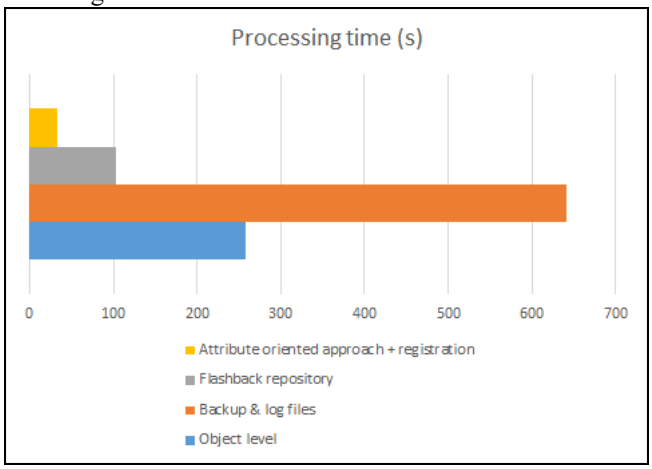


Fig. 15. Processing time (s)

Process of the registration based on our environment lasted totally 0,075s to the whole structure (in this case, only one master process has been used with no executors).

The second part of the experiments deals with the impact of the processing based on number of executor processes - *Temporal Registration Executor Process* (*TRegExecn*). Adding executor process significantly influence performance by reducing processing costs. However, adding too many executor processes does not provide sufficient power because of the synchronization necessity and resource competition. Based on experiments, limitation is 10 processes (tab. 2, fig. 16).

TABLE II. Processing costs based on number of executors

	Number of executor processes					
	1	2	5	10	50	100
Processing costs (%)	100	68	26	14	13	13

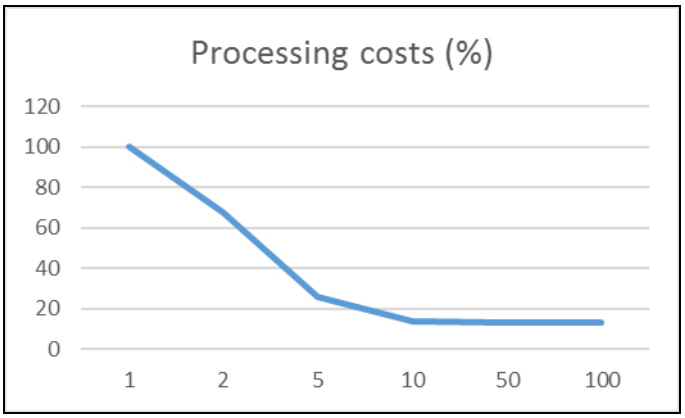


Fig. 16. Processing costs based on number of executors

VII. Conclusions

Each data tuple is delimited by the time validity from the definition. It´s on developers and users, whether such definition will be stored or not. Temporal database concept proposes management technology for data evaluations and storage possibilities. Performance of the data retrieval retrieval is a crucial part highlighting efficiency. Several solutions have been proposed for dealing with time bordered states depending on the granularity and frequency of changes. Conpect of using backups and log files for the temporal data evaluation has been significantly improved by the *Flashback repository,* which allows really efficient solution for deadling with images and snapshots of the database at defined timepoint, however does not provide sufficient power for monitoring changes over the time. Our proposed temporal solution core part is based on attribute granularity, which can be optionally formed into groups, if some data portions are synchronized.

Storing data over the long time period must deal and react to the possibilities of changing physical structure of the database – development of the data approach and model as well. It covers an option to add or remove the whole table, change the definition of the attribute or the group of attributes to become temporal. For these purposes, new processes have been introduced to reflect and manage such changes and reflect them into temporal access layer. Thanks to that, images over the time with regards on the structure at given time can be obtained. As we have shown, proposed solution is complex, secured by the new layer, which manages structure in the defined temporal table structure.

In the future, we will focus on the development of temporal registration management in the distributed environment with emphasis on the workload rebalancing on individual nodes in the temporal sphere.

Acknowledgment

This publication is the result of the project implementation:

Centre of excellence for systems and services of intelligent transport, ITMS 26220120028 supported by the Research & Development Operational Programme funded by the ERDF and *Centre of excellence for systems and services of intelligent transport II.,* ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

This paper is also supported by the following project: "*Creating a new diagnostic algorithm for selected cancers*," ITMS project code: 26220220022 co-financed by the EU and the European Regional Development Fund.

"Podporujeme výskumné aktivity na Slovensku
Projekt je spolufinancovaný zo zdrojov EÚ

References

[1] Q. Abbas, H. Shariq, I. Ahmad, S. Tharanidharan, "Concurrency control in distributed database system", 2016 International Conference on Computer Communication and Informatics (ICCCI)

[2] K. Ahsan, P. Vijay. "Temporal Databases: Information Systems", Booktango, 2014.

[3] L. Ashdown. T. Kyte "Oracle database concepts", Oracle Press, 2015.

[4] G. Avilés et all. "Spatio-temporal modeling of financial maps from a joint multidimensional scaling-geostatistical perspective", 2016. In Expert Systems with Applications. Vol. 60, pp. 280-293.

[5] R. Behling et all., "Derivation of long-term spatiotemporal lanslide activity – a multisensor time species approach", 2016. In Remote Sensing of Environment, Vol. 136, pp. 88-104.

[6] C. J. Date, N. Lorentzos, H. Darwen. "Time and Relational Theory : Temporal Databases in the Relational Model and SQL", Morgan Kaufmann, 2015.

[7] R. Heckman, "Data layer", Wiley-IEEE Press, 2016.

[8] J. Chomicki, J. Wihsen, "Consistent Query Answering for Atemporal Constraints over Temporal Databases", 2016

[9] T. Johnston. "Bi-temporal data – Theory and Practice", Morgan Kaufmann, 2014.

[10] T. Johnston and R. Weis, "Managing Time in Relational Databases", Morgan Kaufmann, 2010.

[11] A. Kadir and N. Adnan, "Temporal geospatial analysis of secondary school students´ examination performance", 2016. In IOP Conference Series: Earth and Environmental Science, Vol 37, No. 1.

[12] D. Kuhn, S. Alapati, B. Padfield, "Expert Oracle Indexing Access Paths", Apress, 2016

[13] M. Kvassay, E. Zaitseva, J. Kostolny, and V. Levashenko, "Importance analysis of multi-state systems based on integrated direct partial logic derivatives", In 2015 International Conference on Information and Digital Technologies, 2015, pp. 183–195.

[14] M. Kvet, K. Matiasko, "Improving performance of database system using architectural layer", CISTI 2016.

[15] M. Kvet, K. Matiasko, M. Kvet, Temporal index dispatcher layer service for intelligent transport system", Elektro 2016.

[16] M. Kvet, M. Vajsova, „Performance study of the index structures in audited environment", ICITST-2016, 2016.

[17] M. Kvet and K. Matiaško, "Temporal Data Group Management", *unpublished*.

[18] D. Kuhn, S. Alapati, B. Padfield, "Expert Oracle Indexing Access Paths", Apress, 2016.

[19] R. Niemiec, "Oracle Query Tuning", Oracle Press, 2014.

Context Based Visual Content Verification

Martin Lukac
Department of Computer Science
Nazarbayev University, Astana
Kazakhstan

Aigerim Bazarbayeva
Department of Computer Science
Nazarbayev University, Astana
Kazakhstan

Michitaka Kameyama
Department of Informatics
Ishinomaki Senshu University
Ishinomaki, Japan

***Abstract*—In this paper the intermediary visual content verification method based on multi-level co-occurrences is studied. The co-occurrence statistics are in general used to determine relational properties between objects based on information collected from data. As such these measures are heavily subject to relative number of occurrences and give only limited amount of accuracy when predicting objects in real world. In order to improve the accuracy of this method in the verification task, we include the context information such as location, type of environment etc. In order to train our model we provide new annotated dataset the Advanced Attribute VOC (AAVOC) that contains additional properties of the image. We show that the usage of context greatly improve the accuracy of verification with up to 16% improvement.**

I. Introduction

Image or scene understanding is a domain of computer vision aiming to provide a general and detailed description of an image content. The scene understanding goes beyond object detection, recognition, localization and segmentation, because it aims at providing also the description on a semantic level as well as explaining higher level relations between objects such as behavior or activity [1], [2], [3].

The scene understanding can be oriented towards some particular type of scenes or images. For instance Matsuyama [4] used logic reasoning with inference to improve the understanding of satellite imaging. But with the increasing computing power recently scene understanding has been focusing on general scenes segmentation, semantic segmentation or scene classification [5], [6], [7], [8].

With the advent of these more advanced methods and because many of the algorithms are based on machine learning, the quality of the processing heavily depends on the training data and as such there is a strong bias-variance trade-off. Consequently many algorithms resort to higher level content related heuristics such as co-occurrence statistics [9], [10] to provide a more reliable scene description.

Content verification [4], [11] is a problem related to the analysis of content obtained from sensors and is aimed to insure the quality of result of particular algorithm. It is most commonly used in areas that deal with real-world information such as computer vision, natural language processing or robotic application. Additionally with the increasing complexity of tasks intended to be performed by intelligent agents as well as required reliability, it is necessary to ensure that the perceived information and executed actions are performed correctly.

In this paper we extend a previously proposed method for content verification in [12], [13], [14] by enabling a context aware verification approach. While initially the verification method based on advanced co-occurrence statistics was developed for general set of images, we show that restricting context and algorithm can effectively increase the accuracy of verification process.

This paper is organized as follows. Section II describes the previous work, in Section III introduces the proposed method of verification and Section IV describes the experiments. Section V concludes this paper.

II. Previous Work

Content verification is required when the result of processing is not guaranteed to be accurate for all possible input scenarios. This is often the case in machine learning when the bias-variance is high due to training set data restriction. In such case the result of image processing for instance, must be verified for inconsistencies resulting from the imperfect processing, lack of information or due to complexity of the task. For instance, Figure 1 shows a set of examples of semantic segmentation where a high level semantic verification process can solve problems arising from the spurious results.

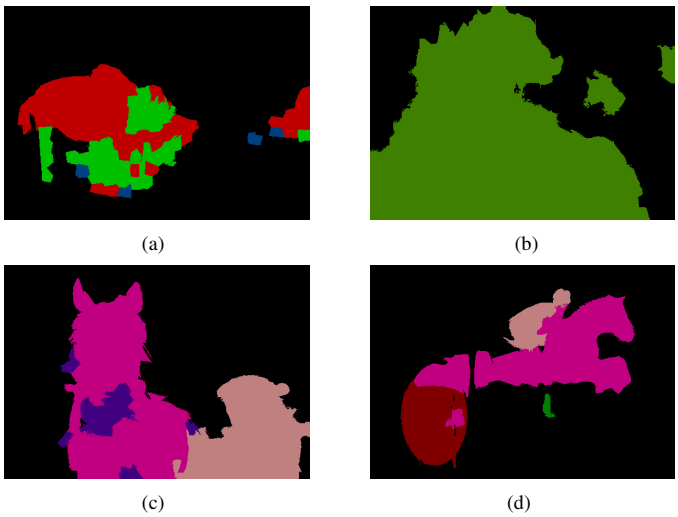


Fig. 1. Example of semantic segmentation results that can be improved by high-level semantic verification.

- Figure 1(a): parts of sofa (green) are mixed with parts of chair (red) and such result can be semantically detected.

- Figure 1(b): variations of sizes between several objects of the same class can be detected by relative size comparison
- Figure 1(c): the horse (pink) cannot be a dog (purple) at the same time
- Figure 1(d): a horse (pink) cannot be a cat (dark red) at a same time and a cat is also rarely under a horse.

Verification can be seen as a part of information restoration process in signal processing or circuit computing [15]. On a higher semantic level Natural Language Processing (NLP) often repairs the translated or generated content by co-occurrence metrics [16], [17], [18], [19]. NLP uses co-occurrence matrices to determine and to fill in correct words in a sentence or to determine relations between words and a context.

In computer vision reasoning is used to determine the correctness of the image content by analyzing higher-dimensional relations between the obtained content. For instance [20] uses co-occurrence statistics for object classification, in [21] similar approach is used to reason about the geometrical disposition to reason about missing elements using temporal and spatial co-occurrences. In [12], [13] the reasoning is used to remove or modify the existing elements of semantic segmentation in a meta-learning framework.

A common approach how to improve the accuracy of object recognition and segmentation is to use co-occurrence statistics [9]. Originally used to analyze language [22], the co-occurrence statistics can appear as a probabilistic generalization of the association analysis rule learning [23], [24].

Co-occurrence statistics represent information collected from sample data and represents probabilistic information about pairs of objects occurring together in one image. It was used on several occasions to improve the recognition accuracy in semantic segmentation and object recognition [9], [12]. In particular it was used in combination with Conditional Random Fields (CRF) [9], graph-cut [10] and with Algorithm Selection approach [13], [25].

In most cases the co-occurrence is used only as existential verifier, i.e. it represents only the statistical information of objects occurring at the same time. More advanced statistics have been proposed in [13] where the co-occurrence statistics were generalized to determine statistics of relative size, proximity, position and shape.

The main problem of co-occurrence statistics is the probabilistic bias that arises when calculating co-occurrence statistics from a data set; i.e. objects occurring more often will tend to bear stronger probabilistic presence than those that occur less often. This becomes a problem if the co-occurrence model or any statistics are used for generative purposes such as used in [13].

The advantage of the co-occurrence methods is however the relative easiness of using them and creating them from data. Consequently co-occurrence statistics are very useful and can be used both for verification as well as for model generation.

III. Proposed Method

In this paper we use the work from [12] as a starting point of our study. The platform introduced in [12] is based on meta-learning and uses algorithm selection to provide optimal output to a computer vision task. As a part of feedback to the algorithm selection mechanism, the content is verified using higher-level relational co-occurrence matrices.

The platform was used in [25] and outperformed by intelligent algorithm selection all of the used state-of-the-art semantic segmentation algorithms. Such result was obtained despite the fact that the used semantic verification had a accuracy of 60% for the 20 categories of objects available in the VOC2012 dataset.

The general scheme of the verification method used in [25] is shown in Figure 2.

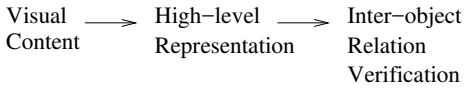


Fig. 2. The process of semantic verification using object relationships.

The verification uses the result of semantic segmentation such as shown in Figure 1, builds a relational model (called high level representation in Figure 2) and determines if these relations are semantically valid or lead to a contradiction.

For instance a human will never be in the air without the support unless the context is specific to sky diving or jumping in the air.

The main principle of this verification process is co-occurrence statistics of four different kinds. These co-occurrence statistics are used to determine the properties between any two objects obtained as a result of semantic segmentation. The collected information about objects were represented as a vector for each pair of objects in the image. Then from the set of all relationships a general conclusion on the object configuration is obtained by majority voting or by SVM. In the work used in this paper an SVM was trained to discriminate whether yes or no a contradiction exists in the semantic segmentation (Figure 3).

Unlike in simple co-occurrence statistics, the original work from [13] uses four distinct measures to estimate the contradiction. The schematic representation of the original method is shown in Figure 3. Unlike the work in [25] however in [13] the accuracy of semantic verification was up to 92% of accuracy but on a significantly smaller dataset and with specially prepared data. Consequently the semantic verification based on co-occurrence statistics should be studied in more depth to understand better the capacity to capture and represents real world information.

The five features obtained form semantic segmentation are described as follows:

1) The first is simple co-occurrence collected from the whole training data set. The result was a set of coefficient representing the probability that two objects occur at the same time in one image.

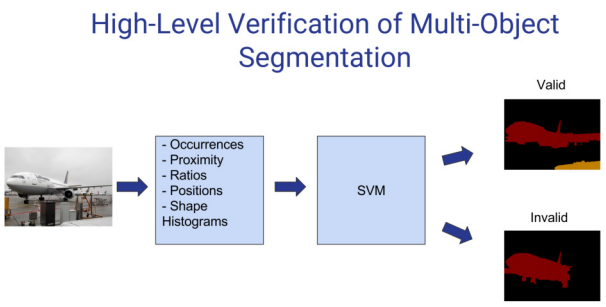


Fig. 3. Co-occurrences based Verification schematic of example

2) The second is the position statistics. For this to work each object in the image was segmented and center of gravity was calculated. For each two pairs of objects eight possible vector orientations were available. These are shown in Figure 4.
3) The third co-occurrence was calculated using proximity of objects. The coefficient indicating representing for any two objects the probability that the two objects are in visual contact. In fact in [25] the proximity was integrated and was given as additional four co-occurrence statistics: on, under, front and back.
4) The fourth co-occurrence is obtained from sizes of objects. Similarly to the previous statistics, from the segmented object the number of pixels is used as size. The normalized average of the size ratio indicates the probability of two objects relative sizes.

A shape characterization was added to increase the probability of recognizing the segmented object with higher accuracy. For this the shape histogram approach is used. The shape histogram is simply created by sampling the boundary of recognized object and clustering the distance of each sampled object from the center of gravity.

An example image and a high level representation is shown in Figure 4. Figure 4(a) shows three segmented objects in their relative position. The objects' gravity centers being the vertices of a triangle are joined by the edges representing pairwise relations between objects. The relations are represented by the four relative co-occurrences relative distance (RDist), relative size (RSize), relative proximity (RproX) and relative position (RPos) as shown in Figure 4(b). The relative position is ordered from left-to-right so that relative position L (left) indicates the position of the left object with respect to the right one and so on.

One of the main reasons of the low accuracy of the co-occurrence statistics used for semantic verification is due to statistical co-occurrence bias occurring because of lack of training data with same number of objects of different classes. For instance, because most of the images are human centric, human is present in most of the images and thus all the statistics are biased toward an over-presence of human.

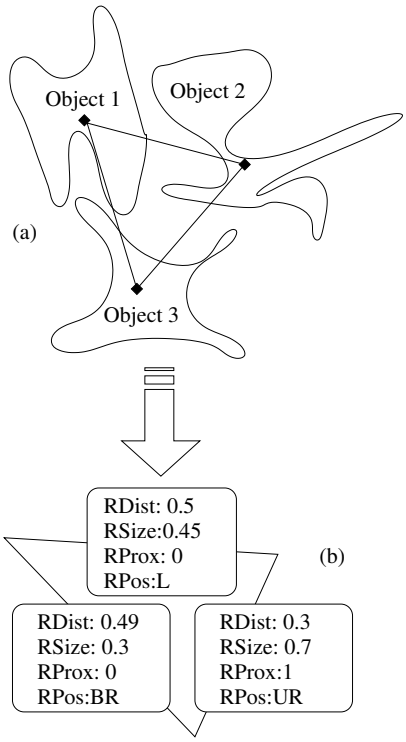


Fig. 4. Example of co-occurrence information extracted from a set of neighboring objects

In order to avoid this statistical co-occurrence based bias, the proposed approach formulates n models of contextually *tinted* models. The concept is simply explained as follows: a co-occurrence statistics will depend on the context that defines probability of certain objects occurring at all or in some particular configuration.

A. Context Aware Verification

Considering the context, the environment and the general specification of the application the verification process can be broken from a single general verification algorithm to a larger number of smaller verifiers. This idea is shown in Figure 5 where several context-aware verification models are used for each of the existing defined contexts.

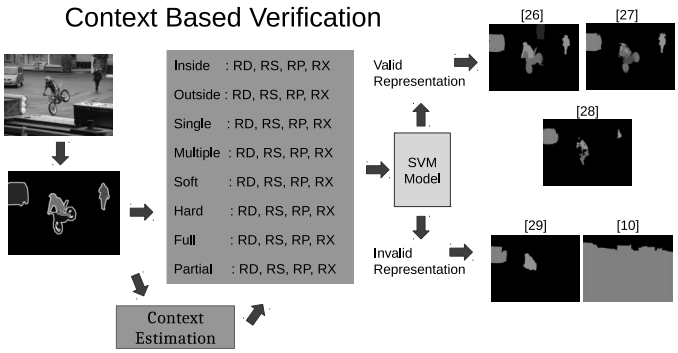


Fig. 5. Co-occurrences based Verification schematic of example

Figure 5 shows that for any input image, from the existing semantic segmentation or provided by user a context is estimated. The context is then used to select more accurate and smaller version of the general verifier.

The idea of the proposed improvement in this paper is to break the problem into categories or environments that allows to reduce efficiently the size of the co-occurrence matrices by eliminating objects that do not occur together or not at all.

For instance, from the VOC2012 data it was observed that if there is a person on a horse it is always outside, while the presence of cat is exclusively limited to inside environment.

Naturally such observation is highly depending on the dataset but it is reasonable to assume that the database images represent real objects in context proportionally to their natural existence in real world. For instance, domesticated cats are mostly inside while outside their occur much less often.

B. Selecting Contexts

Using the context characterization it is necessary to determine meaningful contexts that would allow improved verification and simplified reasoning. For this we generated a new set of visually relevant attributes for the VOC2012 data [26] set called the Advanced Attributes VOC (AAVOC). There exists already various augmented image sets by attribute however in most of the cases these attributes are object related or a particular view limited. Additionally the attributes are mostly and only focused on the image primary objects rather to the holistic properties of the environment. Consequently a new data set was created for this purpose solving the two above mentioned problems.

The used dataset provides thus the following improvements:

1) Object attributes labeling: only predefined attributes are considered as being or not being present
2) Scene attributes relating to global scene properties such as background or environment
3) Aperture attributes such as exposure quality or type of photo shot

A total of 44 attributes have been used. Out of these twelve are used to specify category of a given object and thus were not used in the candidates of global contexts. The remaining 32 attributes have together 172 different values. Consequently the provided labelling can be seen as having 172 binary attributes.

From each image in the AAVOC data set, each object was extracted sequentially and the attributes were provided. Thus according to the scopes of the attribute, for one image the value of particular attribute can remain the same or change.

For each of the available labels, first the empty values have been replaced by a placeholder. Then the entropy for each attribute and the labels was calculated using the Shannon's entropy the mutual information according to $I(L,A) = \sum_{l\in L}\sum_{a\in A} p(l,a) log\left(\frac{p(l,a)}{p(l)p(a)}\right)$ where $p(l)$ and $p(a)$ are prior probabilities of $l \in L$ and $a \in A$.

The attributes we are looking for are such context variables that best describes the data from a general point of view. The attribute of highest interest should have thus the following properties:

1) It has a defined value in all images
2) Values of the attribute are proportional to all images
3) Allows to estimate the classes of objects most reliably

For instance, environment location inside or outside is example of such attribute that satisfied most of the criteria. Specifically, any attribute specific to any object is in most cases unsuitable because it is specific to either a type of object, a group or a description.

The main rule for selecting such attribute is also one that has defined values for all images and one that allows to best predict and separate between classes of objects. For instance, an attribute separating objects in two groups of non overlapping sets is desired. Attribute *inside* will indeed put all sky in one group (being not *inside*) and all TV monitors in another group 8being *inside*).

IV. Experimentation

To evaluate the proposed method the VOC2012 challenge dataset was used. Two subsets were used. The *train* subset was used to build our models and the *val* dataset was used to evaluate the proposed approach.

The verification framework is the one used in [13]. In the proposed approach verification is required in order to verify the semantic content of each semantic segmentation. The method evaluation and accuracy of predicting semantic contradiction from a segmentation results are shown in Table I on the VOC 2012 validation dataset.

TABLE I
VERIFICATION OF SEMANTIC SEGMENTATION ACCURACY RESULT USING THE METHOD FROM [13]

Algorithm	Verification Accuracy
[27]	79%
[28]	75%
[29]	74%
[10]	67%
[30]	39%
Average	66.8%

The accuracy of semantic segmentation represents a pixel-wise difference between a ground truth semantic segmentation and the algorithm's result. An accuracy of 50% can represent several different results: (a) exactly 50% of objects have been perfectly segmented only and all others have not been even detected, (b) the average accuracy of detecting and segmenting images is 50% and (c) all objects were properly detected and have been segmented with 50% of accuracy.

The original algorithm tested in [13], [25] had an accuracy of detecting a contradiction of 60%. The algorithm was trained and tested according to the description given in [13].

The experiments using the context based verification were separated in two sets. First, an algorithm and context dependent verification algorithm is implemented and was evaluated on the validation data set of the VOC2012 dataset. The idea behind this approach to semantic verification is to apply fine

grain verification able to trained for each available algorithm individually. Such approach, could precisely detect semantic segmentations contradictions resulting from the particular structure and architecture of each algorithm.

In the second set of experiments, a more general approach is used. Verification algorithms have been constructed only for each of the available image contexts.

In order to evaluate the method on the simplest available cases of scene configuration only images that contained single instances of class were used for both training and testing. Example and counter example of such images are shown in Figure 6. Following the approach in [13] the used data to

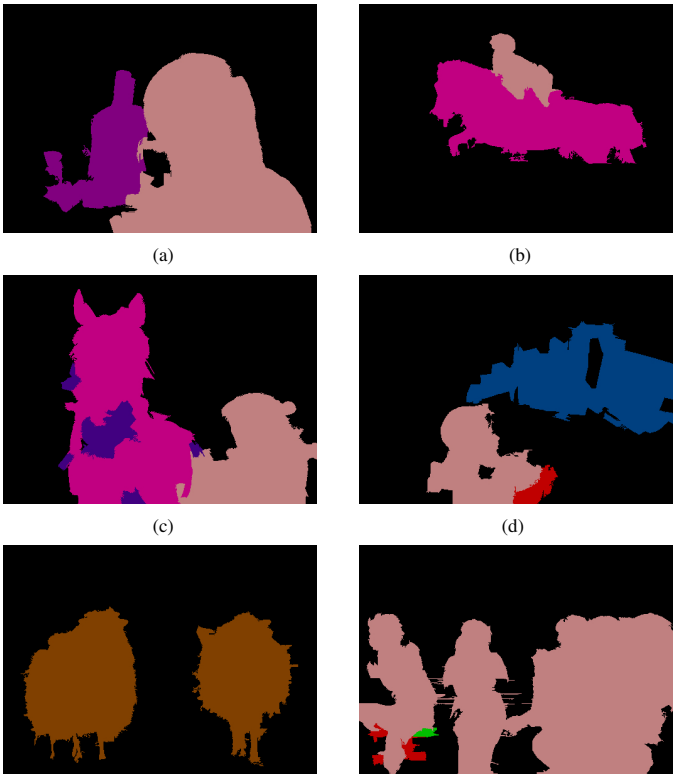


Fig. 6. Example of (a-b) valid images containing no contradiction, (c-d) valid images containing contradiction and (e-f) invalid images for the verification algorithms

train the verification algorithms were obtained from the tested algorithms output. The valid images in Figure 6 represent the negative examples for the training of the contradiction detector. Positive examples are generated by removing one of the available objects.

Table II shows the results and accuracy of a context and algorithm dependent verification. Each row represents one of the selected contexts and each column represents one particular algorithm.

The results in the Table II shows that certain algorithms are more consistent and thus more predictable than others. In particular, three out of the five tested algorithm were convolutional neural networks (CNN) [29], [28], [27] while the two remaining ones are older algorithms based on Conditional Random Fields (CRF) [10] and on machine learning of region segmentation [30].

TABLE II
CONTENT VERIFICATION ALGORITHMS FOR EACH ALGORITHM AND FOR EACH CONTEXT.

Context / Dataset	[28]	[30]	[10]	[29]	[27]	AVG
outside	80 %	58 %	62 %	77 %	81 %	71.6%
inside	69 %	76 %	72 %	76 %	76 %	73.8%
single	75 %	63 %	66 %	81 %	78 %	72.6%
multiple	81 %	50 %	75 %	75 %	50 %	66.2%
soft	72 %	57 %	62 %	82 %	79 %	70.4%
hard	73 %	67 %	62 %	85 %	76 %	72.6%
full	75 %	63 %	67 %	79 %	80 %	72.8%
partial	74 %	55 %	62 %	76 %	76 %	68.6%
Average	75 %	61 %	66 %	79 %	75 %	71.2%/71.07%

Interestingly, the accuracy of determining if an algorithm's semantic segmentation contains a semantic contradiction is proportional to the average accuracy of the algorithm itself. This indicates that the used contradiction detection method principles are not capturing accurately and efficiently the algorithm-dependent error structure.

To verify this claim another set of experiments was conducted. This set of experiments was only context dependent and the contradiction detection method was applied on the semantic segmentation of all used algorithms. The results are shown in Table III. Column one indicates the context, column 2, 3 and 4 indicates the number of images used to train the verifier, column five indicates the obtained accuracy of predicting a contradiction, column six shows the results reported in Table II and column seven shows the relative improvement. The results from this second set of experiments

TABLE III
CONTENT VERIFICATION FOR EACH CONTENT BUT ALGORITHM INDEPENDENT.

Context	Valid	Invalid	Total	Result of verification on all datasets	Average of previous results	Improvement
inside_all	1732	739	2471	74.00%	73.80%	0.20%
outside_all	3301	793	4094	76.00%	71.60%	4.40%
multiple_all	1537	474	2011	68.00%	66.20%	1.80%
single_all	4241	1300	5541	75.00%	72.60%	2.40%
partial_all	2872	913	3785	72.00%	68.60%	3.40%
full_all	3448	1023	4471	76.30%	72.80%	3.50%
soft_all	3168	1110	4278	73.00%	70.40%	2.60%
hard_all	2806	683	3489	73.00%	72.60%	0.40%
Average	–	–	–	73.41%	71.07%	–

are shown in Table III. Interestingly the results show that indeed he verification is same or even performs better when performed in a algorithm independent manner rather than when done on an algorithm by algorithm basis. This implies that if there is an underlying structure to each algorithm's error then either it is negligible with respect to the context or it cannot be properly captured by the model used here.

The general remarks on the results include the observation that while the improvement of the verification accuracy compared to the initial result from [25] is considerable the

accuracy claimed in [13] was not achievable in any of the here performed experiments. The conclusion from this observation is that while in [13] the train and test data was the problem, in [25] the problem was clearly the fact that single verifier for all algorithms and all contexts is too complex and inaccurate.

Surprisingly splitting the verifier to a set of smaller ones w.r.t. context and algorithms perform worse than only train verifiers one for each available contexts. This is very interesting because this indicates that the type of error obtained as a result of semantic segmentation is strongly dependent on the context of the image rather than on the type of the used algorithms.

V. Conclusion

In this paper, we showed a context based approach to semantic segmentation content verification. We experimentally demonstrated that the context is an important variable allowing to strongly increase the accuracy of the verification while a finer granularity of verification (algorithm dependent verification) did not improve the accuracy.

Additionally we have experimentally shown that the most important factor in increasing the accuracy of content verification is indeed the context. Thus in the future work more development in the area of automatic context extraction and identification is to be studied and developed.

References

[1] A. Chella, M. Frixione, and S. Gaglio, "Understanding dynamic scenes," *Artificial intelligence*, vol. 123, no. 1-2, pp. 89–132, 2000.

[2] J. Luo, A. E. Savakis, and A. Singhal, "A bayesian network-based framework for semantic image understanding," *Pattern recognition*, vol. 38, no. 6, pp. 919–934, 2005.

[3] G. Sagerer and H. Niemann, *Semantic networks for understanding scenes*. Springer Science & Business Media, 2013.

[4] T. Matsuyama, "Knowledge-based aerial image understanding systems and expert systems for image processing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-25, no. 3, p. 305, 1987.

[5] N. Serrano, A. E. Savakis, and J. Luo, "Improved scene classification using efficient low-level features and semantic cues," *Pattern Recognition*, vol. 37, no. 9, pp. 1773–1784, 2004.

[6] L. Cao and L. Fei-Fei, "Spatial coherent latent topic model for concurrent object segmentation and classification," in *Proceedings of the ICCV*, 2007.

[7] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," vol. 29, no. 2, pp. 300–312, 2007.

[8] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2015.

[9] L. Ladicky, C. Russell, P. Kohli, and P. Torr, "Graph cut based inference with co-occurrence statistics," in *Proceedings of the 11th European conference on Computer vision*, 2010, pp. 239–253.

[10] ——, "Inference methods for crfs with co-occurrence statistics," *International Journal of Computer Vision*, vol. 103, no. 2, pp. 213–225, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0583-y

[11] T. J. Palmeri and I. Gauthier, "Visual object understanding," *Nature Reviews Neuroscience*, vol. 5, no. 4, pp. 291–303, 2004.

[12] M. Lukac and M. Kameyama, "An algorithm selection based platform for image understanding using high-level symbolic feedback and machine learning," *International Journal of Machine Learning and Cybernetics*, vol. 6, pp. 417–434, 2015.

[13] ——, "Bayesian-network-based algorithm selection with high level representation feedback for real-world information processing," *IT in Industry*, vol. 3, no. 1, pp. 10–15, 2015.

[14] M. Lukac, A. Zhurtanov, and A. Ospanova, "High-level verification of multi-object segmentation," in *2016 International Conference on Information and Digital Technologies (IDT)*, July 2016, pp. 173–179.

[15] T. K. Moon, *Error Correction Coding*. New Jersey: John Wiley & Sons, 2005.

[16] S. Bordag, "A comparison of co-occurrence and similarity measures as simulations of context," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2008, pp. 52–63.

[17] M. Baroni and R. Zamparelli, "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1183–1193.

[18] Z. Hai, K. Chang, and J.-J. Kim, "Implicit feature identification via co-occurrence association rule mining," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2011, pp. 393–404.

[19] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[20] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[21] R. Fablet and P. Bouthemy, "Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1619–1624, 2003.

[22] P. Kroeger, *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005. [Online]. Available: https://books.google.kz/books?id=rSglHbBaNyAC

[23] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991, pp. 229–248. [Online]. Available: http://dblp.uni-trier.de/db/books/collections/PiatetskyF91.html#Piatetsky91

[24] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993. [Online]. Available: http://doi.acm.org/10.1145/170036.170072

[25] M. Lukac, K. Abdiyeva, and M. Kameyama, "Reasoning and algorithm selection augmented symbolic segmentation," *CoRR*, vol. abs/1608.03667, 2016. [Online]. Available: http://arxiv.org/abs/1608.03667

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the ICLR*, 2015.

[28] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014. [Online]. Available: http://arxiv.org/abs/1412.7062

[29] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*, 2014, pp. 297–312.

[30] A. Ion, J. Carreira, and C. Sminchisescu, "Probabilistic joint image segmentation and labeling," in *25th Conference on Neural Information Processing Systems*, 2011.

An Analytic Sifting Approach to Optimization of LNN Reversible Circuits

Martin Lukac
*Department of Computer Science
Nazrabayev University, Aatana, Kazakhstan
e-mail:martin.lukac@nu.edu.kz

Pawel Kerntopf
†Faculty of Physics and Applied Informatics
University of Lodz, Lodz, Poland
Email: pawel.kerntopf@gazeta.pl

Michitaka Kameyama
†Ishinomaki Senshu University
Ishinomaki, Japan
Email: kameyama@ecei.tohoku.ac.jp

***Index Terms*—Reversible Circuits, Gate reordering, Qubit sifting, LNN Model**

***Abstract*—In this paper we propose an analytic approach to the variable sifting based on weighting of qubits and gates. The proposed scheme allows to optimally sift gates (multi-control single-target reversible gates) within a linear number of steps of computation and provides in general smaller amount of SWAP gates required to transform a reversible circuit into an Linear Nearest Neighbor (LNN) model than other competing approaches. The method is analyzed for two different models of implementations, is verified on experimental data and results are compared to the state of the art algorithms for the design of LNN circuits.**

I. Introduction

The design of reversible and quantum circuits has been originally motivated by several factors. Landauer [1] has shown that every bit erasure results in heat dissipation which in turn leads, with the current level of VLSI, to very serious problem of heat dissipation. This problem was also pointed out by Moore [2] who indicated the doubling of the number of transistors per square unit every 18th month.

Consequently and because reversible computing is the model that quantum computer directly implements a considerable effort was redirected to the study of reversible computers as one of the possible alternative to the currently most popular CMOS irreversible computing technology.

The quantum computer still being relatively far from a fully implemented device the design of reversible circuits requires more and more to take into account several factors directly related to the underlying quantum realization. One of the initial considerations was the decomposition of arbitrary C^nNOT gates into at maximum C^2NOT (also known as Toffoli gates) due to the fact that it is not possible to directly design quantum gates with more than two control qubits. Similarly, while in most of initial designs of reversible and quantum circuits the $CV/CV^\dagger/CNOT$ quantum gates have been used, recently the so called Clifford-T set of quantum gates is used due to higher fault tolerance capability [3], [4], [5]. Finally, the realization of many quantum technologies requires that all qubits interacting with each other be direct neighbors (this is also called the Linear Nearest Neighbor (LNN) model). Thus a relatively large body of works either proposed methods for directly implementing quantum circuits in an LNN architecture [6], [7], [8] or proposed algorithms to efficiently transform already realized quantum circuit into LNN compatible quantum circuits [9], [6], [10], [11]

In this paper we examine the LNN model of reversible computers: we propose a simple but effective method to calculate the minimal number of SWAP gates required to make a circuit LNN model compatible. Additionally, the insertion of SWAP gates is required in general quantum circuit model because multi-qubit quantum gates can be realized from at maximum two-qubit quantum gates. Th minimizing the number of SWAP gates is required to make the realization cost smaller and more compact. Our method can be applied as pre or post processing and requires in practice only fast enumerating and counting algorithms. Two variants are proposed that address directly circuits with different level of complexity. The first one is an algorithm for circuits realized directly from Programmable Logic Array (PLA) specification [12] as a set of C^2NOT reversible gates. In the second one the proposed method is evaluated on a set of reversible circuits and is compared with the current state-of-the-art algorithms. The verification is performed over the data from set of optimally realized reversible circuits posted at RevLib [13].

This paper is organized as follows. Section II provides the necessary background to understand the problem and Section III shows the impact of qubit ordering and gate ordering on the problem of making a circuit LNN. Section IV introduces the algorithm for circuits realized directly from the PLA specifications and Section V describes the algorithm for arbitrary reversible circuits. Section VI shows the experiments and the results while the Section VII concludes the paper.

II. Background

Definition 1 (Reversible Logic Function)**.** *A function $A \rightarrow B$ is reversible if it is bijective. In other words, $f(X_a) = f(X_b) \implies X_a = X_b$.*

Table I shows examples of reversible and irreversible functions.

A reversible circuit is built from a universal reversible gate set. The simplest gate set includes the Toffoli gate and negation. Toffoli gate is also known as CCNOT or C^2NOT gate and negation gate is also known as NOT. Pictorial representation of both gates are shown in Figure 1. Other gates that can be used for reversible design is the Fredkin gate (also

TABLE I: Example of (a) reversible CNOT function and (b) irreversible function

AB	A'B'
00	00
01	01
10	11
11	10

(a)

AB	A'B'
00	01
01	11
11	00
10	11

(b)

known as CSWAP) shown in Figure 1(d). The CSWAP gate has also been used for building reversible circuits [14], [15], [16]; in addition of being reversible the CSWAP gate is also conservative.

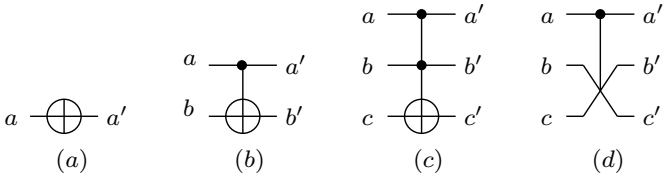


Fig. 1: Basic reversible gates (a) NOT, (b) CNOT, (c) CCNOT, (d) CSWAP

In many technologies for the implementation of reversible and quantum circuits it is required that the circuit is built in the Linear Nearest Neighbor (LNN) configuration. That is, any multi-qubit quantum gate can be applied only on directly neighboring qubits. An example of LNN and of not LNN configured circuits is shown in Figure 2.

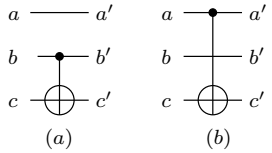


Fig. 2: Example of (a) LNN configuration and (b) non-LNN configuration of reversible gate

One of the possible approaches to achieve the LNN architecture is the insertion of SWAP gates. Insertion of SWAP gates will change the position of two directly adjacent qubits by switching one with the other. Example of the SWAP gate effect is shown in Figure 3(b). SWAP gate can be used to inverse the order of input qubits to a circuit as shown in Figure 3(b).

It is also used to re-arrange the qubits so that every desired gate is applied in an LNN fashion. Example of moving qubits around is shown in Figure 3(c).

III. Gate Ordering and Variable Sifting in Reversible Circuits

A. Order of Variables

The order of variables being important and required to allow for LNN realization of quantum and reversible circuits, the initial order of qubits in the circuit have effect as well on how many SWAP gates need to be inserted in order to create the circuit in the LNN manner. This can be seen in Figure 4.

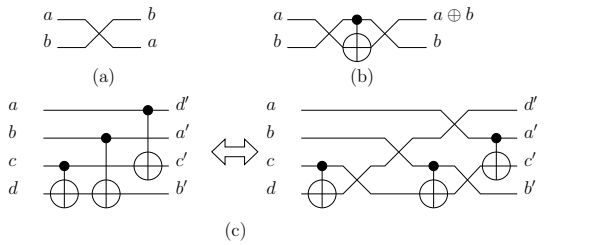


Fig. 3: Example of usage of SWAP gate, (a) a SWAP gate, (b) inverting CNOT using SWAP gate, (c) bit rearrangement using SWAP gates.

Figure 4(a) shows the circuit in its original configuration using C^nNOT gates. In order to achieve the LNN configuration Figure 4(c) shows how many SWAP gates are required to be inserted at every indexed position in both circuits Figure 4(c) and Figure 4(d). For instance, at position 1 in Figure 4(c), six SWAP gates are required in order to make the C^2NOT gate on a, e and f LNN compatible. Similarly, for all other indexed positions. Proceeding in this manner for both circuits in Figure 4(a) and (b) the required amount of SWAP gates is 28 and 20, respectively. Thus the initial variable ordering has a significant impact on the required number of SWAP gates to realize the circuit in the LNN architecture. Note that if one

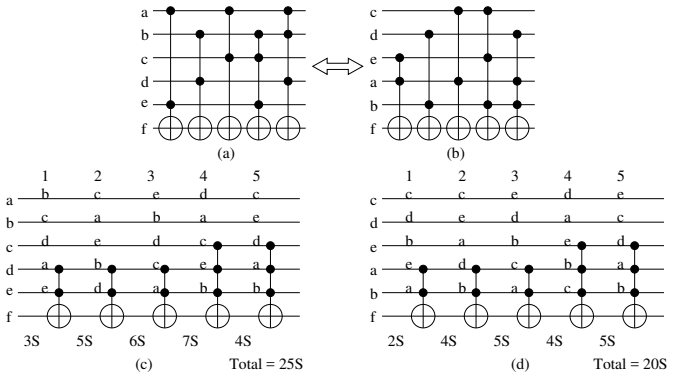


Fig. 4: Impact of the initial ordering of variables on the number of SWAP gates required: (a) total swap gates = 28 , (b) total SWAP gates = 20.

would like to keep the initial order of input variables the same as the order of the output variables, the circuit from Figure 4(d) would require 22 SWAP gates.

B. Order of Gates

The order of variables being crucial for the realization of LNN quantum circuits, the insertion of SWAP gates allows to reorder the variables so that every gate can be applied on only neighbouring qubits. The possibilities how to apply the SWAP gates are however not unique and different gate arrangements will result in more or less SWAP gates required for the LNN realization. Figure 5 shows the same circuit realized by three C^2NOT gates: depending on the gate ordering used to build the circuit different numbers of SWAP gates have to be

inserted. Similarly to Figure 4, if one would like to keep the

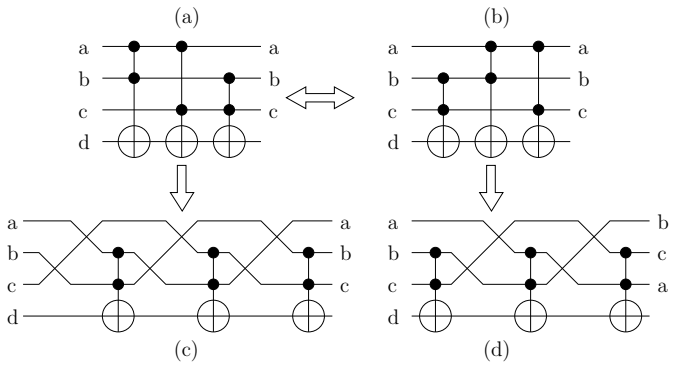


Fig. 5: Different amount of swap gates have to be inserted for different orders of reversible gates: (a) 6, (b) 4

initial order of input variables the same as the order of the output variables, the circuit from Figure 5(d) would require 6 SWAP gates.

The reason for this is that the order of the gates matters and does not commute because insertion of each SWAP gates modifies the order of variables locally. Thus to move from an LNN gate $G1$ to another LNN gate $G2$ might change the number of SWAP gates.

IV. Sifting Algorithm

Before introducing the proposed method some definitions are required.

Definition 2 (Gate Weight (GW))**.** *Gate Weight (GW) is the number of control qubits the gate is using plus the number of SWAP gates that are needed to bring this gate to an LNN realizable form.*

Let G be a particular reversible or quantum gate and $b_G = \{c_0, \ldots, c_{k-2}, t\}$ be an ordered set of the indexes of control and target qubits of this gate. Each c_i in b_G takes value from $\{1, \ldots, n\}$ representing what wire is on the i^{th} position in the wire array. For instance for the initial order of variables $\{a, b, c, d, e\}$, $c_0 = a$, $c_1 = b$, etc, thus swapping a and d variables can be represented as $\{c_3, c_1, c_2, c_0, c_4\}$.

Additionally let there be n qubits available in a given circuit and let

$$d(i) = \begin{cases} |c_i - c_{i+1}| - 1 & \text{if } i > |b_G| - 1 \\ c_i - t - 1 & \text{otherwise} \end{cases} \quad (1)$$

.

Then the calculation of $GW(G)$ can be written as follows

$$GW(G) = \sum_{i=0}^{|b_g|-1} d_i + d_{i+1} \quad (2)$$

For instance, taking the leftmost gate from the circuit shown in Figure 4(a), the gate weight is computed as $GW = |1 - 5| - 1 + |5 - 6| - 1 = 3$.

Lemma 1. *The GW is the exact indicator of the number of SWAP gates required for a gate G be implemented in a LNN manner.*

Proof. The proof is a direct consequence of Definition 2. The GW is given by a recursive additive cost between each two consecutive control bits and between a single control and a target bit. Because the cost of each difference d_i is exact (eq. 2) the sum of all d_i is exact by transitivity of addition as well. □

Definition 3 (Qubit Weight (QW))**.** *is the number of gates in which this particular qubit is a control variable. It applies to all qubits except to target ones.*

For instance, in the circuit in Figure 4(a) the qubit weight of qubits a, b, c, d, e is $3, 3, 2, 2, 2$, respectively. Note that for simplicity we will denote $\overline{GW}$ and $\overline{QW}$ to represent the gate weight and the quantum weight of the whole circuit respectively.

Integrating all these observations we propose a new algorithm for optimal input variables sifting. The algorithm has several parts. The first components are calculations of $QubitWeight$ and $GateWeight$ and they are performed by the pseudocode shown in Algorithm 1.

Algorithm 1 Weight Calculating Algorithm

```
1: function GateWeight(GateWeight[])
2:     j ← 0
3:     i ← n − 1
4:     for j < k − 1; j++ do
5:         for i >= 0; i−− do
6:             if c_i ∈ G_j then
7:                 GC ← GC + (c_i − T)
8:                 T ← T + 1
9:             end if
10:        end for
11:        GateWeight[j] ← GC
12:    end for
13:    return GateWeight
14: end function
15: function QubitWeight(GateWeight[])
16:    j ← 0
17:    i ← n − 1
18:    for j < k − 1; j++ do
19:        for i >= 0; i−− do
20:            if c_i ∈ G_j then
21:                QubitWeight[c_i] ← QubitWeight[c_i] + 1
22:            end if
23:        end for
24:    end for
25:    return QubitWeight
26: end function
```

The algorithm processes as follows. Initially a circuit is given in the ESOP or in any canonical EXOR forms such as FPRM, PPRM or GRM. This form is extracted from the PLA specification file. An example of such a circuit is shown

in Figure 4(a) or 4(b). The algorithm initializes to 0 several variables in particular a $QubitWeight$ representing the weight of a qubit and $GateWeight$ the weight of a gate. Then for each gate in the array the $GateWeight$ is calculated in line 9 and for each control qubit the $QubitWeight$ is updated at line 11. The algorithm finishes and the resulting weights are shown in Figure 6(a).

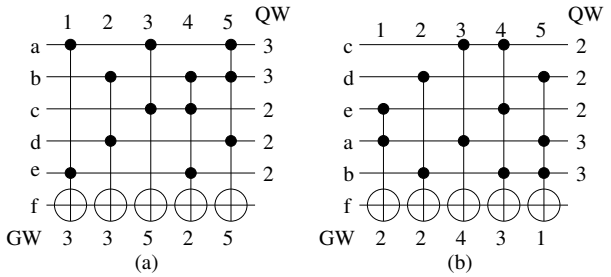


Fig. 6: $QubitWeight$ (QW) and $GateWeight$ (GW) for the circuit from (a) Figure 4(a), (b) Figure 4(b)

Once Algorithm 1 is finished, Algorithm 2 reorders the variables in increasing $QubitWeight$ by iteratively testing each move of a variable down. Algorithm 2 results in a circuit

Algorithm 2 Sifting Algorithm

```
1: function SIFT(QubitWeight)
2:     while True do
3:         for i = 1, i < n do
4:             if QubitWeight[i − 1] > QubitWeight[i] then
5:                 temp ← QubitWeight[i − 1]
6:                 QubitWeight[i − 1] ← QubitWeight[i]
7:                 QubitWeight[i] ← temp
8:             end if
9:         end for
10:    end while
11:    return QubitWeight
12: end function
```

with new order of variables shown in Figure 6(b). As can be seen the sum of all $GateWeights$ have been considerably reduced; from original circuit (Figure 6(a) with $\overline{GW} = 18$) to the circuit with sifted variables (Figure 6(b) with $\overline{GW} = 12$).

Definition 4. *Cross-Gate Weight (CGW) is the minimal number of SWAP gates required to bring the ordered set of variables* $v_a = \{c_0, \ldots, c_i, c_j, \ldots, c_{k-2}, t\}$ *to* $v_b = \{c_0, \ldots, c_j, \ldots, c_i, \ldots, c_{k-2}, t\}$.

For instance the CGW between gates 1 and 2 in the circuit from Figure 6 is the number of gates required to get from $\{b, c, d, a, e, f\}$ to $\{c, a, e, b, d, f\}$. This weight can be calculated as the sum of required SWAP gates of each of the moved wire. An example of CGW is shown in Figure 7(a). Figure 7(b) shows a more complex example of CGW and again the minimal cost of SWAP gates is equal to the number of times the arrows indicating the variable move do cross each

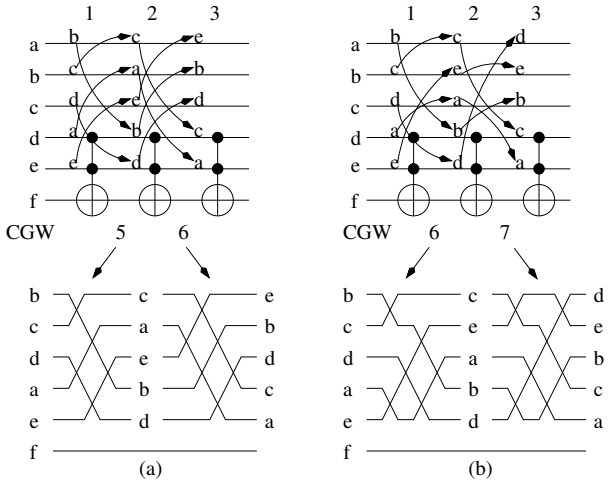


Fig. 7: (a) CGW between gates 1 and 2, and between 2 and 3 from Figure 6 and (b) more complicated CGW between two adjacent gates.

other. The exact algorithm to calculate CGW is very simple and can be inferred from example shown in Figure 7. It can be described as follows:

1) Draw arrows between the corresponding states from the initial order to the final order
2) Count the number of intersections upon each line
3) The number of intersections is the minimal required number of SWAP gates.

In order for the sifting algorithm to obtain even lower SWAP gate requirements the next step is the gate reordering. A simple heuristic such as the one used in [17] can be used. In fact it is possible to prove that there exists an optimal gate ordering in the FPRM/PPRM/GRM forms (such as considered here) that minimizes the CGW.

Let $s = \{G_1, \ldots G_k\}$ be a sequence of C^*NOT gates forming an ESOP expression with n being the number of variables. Also let $j = \sum_{i=0}^{|b_{G_j}|-2} 2^{c_i}$ be the integral representation of the control qubits. For instance, for the gates in Figure 6(a) and starting from qubit a with index 0, the integers assigned to gates 1 to 5 are $17, 10, 5, 24, 11$.

Lemma 2. *The order of gates* $s_O = \{G_\alpha \ldots G_\gamma\}$ *is optimal if it minimizes the CGW cost for every pair of gates* $\{G_\gamma, G_\delta\}$ *for every two consecutive gates from the set* s_O.

Proof. *The CGW cost is the number of SWAP gates required to change qubits in an ordered set* v_i *to* v_j. *Minimizing CGW for each pair of consecutive gates results in a minimal* $\overline{CGW} = \sum_{j=0}^{|s_O|-2} CGW_j$, *where* CGW_j *is the CGW between* v_j *and* v_{j+1}.

Variable Ordering that minimizes the CGW is natural integer order [17] or Gray code.

Conjecture 1. *The ordering of the gates by either integer encoding or by Gray code requires the same number of SWAP gates.*

Finally by combining these different procedures and functions we obtain the final algorithm for sifting variables and gates shown in Algorithm 3.

Algorithm 3 Weight Calculating Algorithm

1: $GateWeight[] \leftarrow GateWeight(GateWeight[])$
2: $QubitWeight[] \leftarrow QubitWeight(QubitWeigh[])$
3: $QubitWeight \leftarrow$Sift($QubitWeight$)
4: $GateWeight \leftarrow$Reorder($GateWeight$)

Lemma 3. *Algorithm 3 guarantees the minimal number of SWAP gates required to bring circuit C into an LNN configuration*

Proof. *The proof for this property follows directly from Lemma 2. Reordering the variables according to the weight enforces the smallest amount of SWAP gates by placing lines with the highest CGW close to the target qubit. Moreover the gate ordering minimizes the CGW joint results leading to the smallest cost of the required SWAP gates.*

V. Circuits with Different Target Qubits

In Section III the proposed method assumed that all gates have the same target qubit, i.e. the minimization of the number of required SWAP gates can be done with respect to a selected qubit in the entire circuit. This naive method however is usable only in very particular realizations of quantum circuits such as in [17].

This is problematic in the sense that many methods are optimized for various and different quantities rarely resulting in such well ordered circuits. In many cases however efficients methods for realizations result in circuits where target qubits may occur on any lines in the circuit. An example of such circuit is shown in Figure 8.

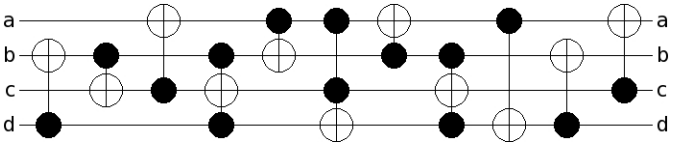


Fig. 8: Example of a circuit realization with target qubits being placed on different qubits (Courtesy of Revlib [13])

For such circuits the method introduced in Section III does not work because:

1) The calculation of CGW is not appropriate as the gates might not have any variables in common. Thus CGW must be recomputed in a different manner
2) Selection of a single target qubit also does not work because many gates might not even be using that qubit
3) Reordering the variables with global constraints affecting all gates is not optimal and can negatively affect the original circuit (increase to an increase in the cost)

Such circuits cannot be efficiently minimized using the simple approach introduced in Section III but require a decomposition of the circuit into subcircuits and recursive application of the algorithm introduced. Using such an approach, to an arbitrary circuit the problem can be called a CSP.

For instance, the circuit from Figure 8 can be reordered as follows into groups that either contain single gates or multiple gates with targets on similar qubits. Then for each such a group we can formulate the problem of minimal amount of SWAP as:

$$CSP(C) = \sum_{c_i \in C} min\left(\sum_{j=0}^{k} QW_j\right) \quad (3)$$

Simply apply the initial algorithm ordering the lines by their weights now became a grouping of high weight lines for each group c_i. Thus instead of calculating QW in a single step after each qubit is being reordered, a minimization for sub QWs is calculated and the new ordering is accepted if the sum is equal or reduced.

VI. Experimental Evaluation

To evaluate our approach we compared the algorithm with RevLib [13] realizations of circuits. The experiments were chosen so that most of the evaluated circuits have been previously used by other authors. This allows us to provide qualitative analysis.

In particular, we compared the proposed approach on previously synthesized quantum circuits obtained from RevLib [13] and compared it to the results provided in [6] and [10]. The results from the experiments are shown in Table II. The first column shows the function name, the second column shows the total number of variables in the circuits (including input, output and garbage) and the third column indicates the number of terms (gates) that are used to build the function in question. The last three columns show the result in the number of required SWAP gates to make the given function realization in LNN.

Table II shows that the proposed method for inserting SWAP gates is more efficient than the two other methods. On average our method improves over the method from [6] by $\approx 85\%$ while compared to method from [10] our method improves the previous results on average by $\approx 45\%$. Some examples of reordered circuits are shown in Figure 9.

Additionally the algorithm is very fast. The whole Revlib database of 319 circuit realizations was run on a Intel Pentium I7 (4 cores) in a sequential mode (no parallelism) with 8G memory. The results were generated in less than two minutes. The largest realization was *alu_319* and took 12 seconds while the smallest circuit realization of *ex-1_166* was instantaneous.

VII. Conclusion

In this paper we presented an algorithmic analytical approach to transform circuits to the LNN model. We showed that our approach is faster than any previously proposed method and provides on average better results for smaller amount of computation. While the scaling of this approach is only $O(n^2)$ for circuits as described in Section V the scaling depends on several factors such as circuit structure. Moreover, for arbitrary realization the true optimality cannot be guaranteed. However, due to the running cost being pseudo-linear in the number of variables, this approach could be

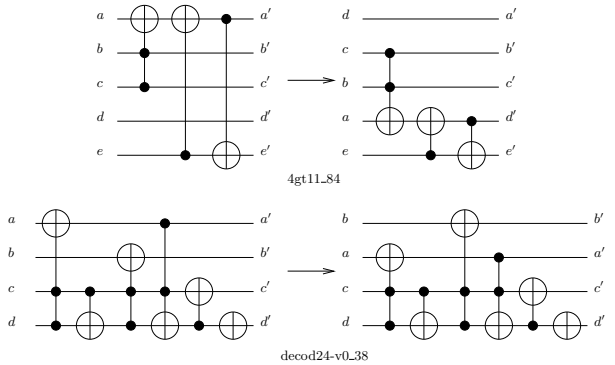


Fig. 9: Example of two circuits before (left) and after (right) reordering of two realizations for the *4gt11_84* and *decod24-v0_38* functions

transformed and applied to similar problems in another areas which is the future work plan.

REFERENCES

[1] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

[2] G. Moore, "Cramming more components onto integrated circuits," in *Electronics, April 19*, 1965.

[3] A. Ahlbrecht, L. S. Georgiev, and R. F. Werner, "Implementation of clifford gates in the ising-anyon topological quantum computer," *Phys. Rev. A*, vol. 79, p. 032311, Mar 2009. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevA.79.032311

[4] P. Selinger. (2012, December) Efficient clifford+t approximation of single-qubit operators. Preprint arXiv:1212.6253.

[5] B. Giles and P. Selinger, "Exact synthesis of multiqubit clifford+t circuits," *Phys. Rev. A,*, vol. 87, p. 032332, 2013. [Online]. Available: arXiv:1212.0506

[6] A. Mohammad and A. Imtiaz, "A greedy algorithm for low cost lnn reversible circuit realization," in *Proceedings of the 3rd International Conference on Advances in Computing and Emerging Learning Technologies (ICAC2ET 2013)*, 2013.

[7] M. Saeedi, R. Wille, and R. Drechsler. (2012) Synthesis of quantum circuits for linear nearest neighbor architecture. [Online]. Available: arXiv:1110.6412v2

[8] M. M. Rahman, G. Dueck, A. Chattopadhay, and R. Wille, "Integrated synthesis of linear nearest neighbor ancilla-free mct circuits," in *ISMVL*, 2016.

[9] M. AlFailakawi, L. AlTerkawi, I. Ahmad, and S. Hamdan, "Line ordering of reversible circuits for linear nearest neighbor realization," no. 12, p. 33193339, 2013.

[10] R. Wille, A. Lye, and R. Drechsler, "Optimal swap gate insertion for nearest neighbor quantum circuits," in *ASP-DAC*, 2014.

[11] A. Lye, R. Wille, and R. Drechsler, "Determining the minimal number of swap gates for multi-dimensional nearest neighbor quantum circuits," in *ASP-DAC*, 2015.

[12] A. Kent, "A texas instruments application report: Mos programmable logic arrays," 1970.

[13] R. Wille, D. Große, L. Teuber, G. W. Dueck, and R. Drechsler, "RevLib: An online resource for reversible functions and reversible circuits," in *Int'l Symp. on Multi-Valued Logic*, 2008, pp. 220–225, RevLib is available at http://www.revlib.org.

[14] E. Fredkin and T. Toffoli, "Conservative logic," *International Journal of Theoretical Physics*, vol. 21, pp. 219–253, 1982.

[15] D. Maslov, G. W. Dueck, and D. M. Miller, "Synthesis of Fredkin-Toffoli reversible networks," *IEEE Transactions on VLSI*, vol. 13, no. 6, pp. 765–769, 2005.

[16] J. Donald and N. K. Jha, "Reversible logic synthesis with Fredkin and Peres gates," *ACM J. on Emerging Technologies in Computing Systems*, vol. 4, no. 1, pp. 1–19, 2008.

[17] M. Lukac, M. Kameyama, M. Perkowski, and P. Kerntopf, "Decomposition of reversible logic function based on cube-reordering," *Facta Universitatis*, vol. 24, no. 3, pp. 403–422, 2011.

TABLE II: Comparative results on selected circuits realization.

Function Name	Vars	Terms	Result [6]	Results [10]	analytic sifting
0410184_169	14	46	-	-	69
0410184_170	14	74	-	-	83
3_17_13	3	7	3	2	2
3_17_14	3	7	-	-	1
3_17_15	3	7	-	-	4
4_49_16	4	18	-	-	11
4_49_17	4	13	-	-	9
4gt10-v1_81	5	6	16	-	9
4gt11-v1_85	5	4	-	-	0
4gt11_82	5	12	-	-	14
4gt11_83	5	8	-	6	5
4gt11_84	5	3	1	1	0
4gt12-v0_86	5	14	-	-	16
4gt12-v0_87	5	10	-	-	5
4gt12-v0_88	5	5	-	-	0
4gt12-v1_89	5	5	26	-	0
4gt13-v1_93	5	4	-	3	4
4gt13_90	5	14	-	-	16
4gt13_91	5	10	-	-	4
4gt13_92	5	3	4	-	0
4gt4-v0_72	5	6	-	-	5
4gt4-v0_73	5	17	-	-	13
4gt4-v0_78	5	13	-	-	11
4gt4-v0_79	5	9	-	-	3
4gt4-v0_80	5	5	-	-	0
4gt4-v1_74	5	5	-	-	4
4gt5_75	5	5	1	-	6
4gt5_76	5	13	-	-	8
4gt5_77	5	4	-	-	5
4mod5-bdd_287	7	8	-	-	12
4mod5-v0_18	5	10	-	-	7
4mod5-v0_19	5	6	-	3	4
4mod5-v0_20	5	6	-	2	2
4mod5-v0_21	5	7	-	-	7
4mod5-v1_22	5	6	-	3	4
4mod5-v1_23	5	9	15	-	12
4mod5-v1_24	5	6	-	10	4
4mod5-v1_25	5	5	-	1	4
4mod7-v0_94	5	5	-	-	7
4mod7-v0_95	5	6	22	-	7
4mod7-v1_96	5	7	-	-	7
4_49_17	4	13	16	-	9
5xp1_194	17	88	-	-	446
9symml_195	10	133	-	-	252
C17_204	7	12	-		18
C7552_205	21	83	-	-	788
alu-v0_26	5	6	-	0	3
alu-v0_27	5	7	-	18	3
alu-v1_29	5	8	-	15	4
alu-v4_37	5	8	-	14	4
alu-v4_37	5	7	17	-	10
decod24-v0_38	4	6	-	4	2
decod24-v0_39	4	5	-	5	2
decod24-v0_40	4	9	-	3	4
decod24-v1_42	4	9	-	2	5
decod24-v2_43	4	6	-	3	0
decod24-v3_46	4	9	-	3	3
graycode6_48	6	6	-	0	0
mod5d1_63	5	8	-	10	6
mod5d2_70	5	8	-	18	3
mod5mils_71	5	5	-	7	4
rd32-v0_66	4	5	-	4	2
rd32-v0_67	4	3	-	2	2
rd32-v1_68	4	6	-	4	2
rd32-v1_69	4	3	-	2	1
hwb4_52	4	12	9	-	8
rd73_140	10	20	75	-	42
rd53_135	7	16	68	-	28
rd84_142	15	28	142	-	75
cnt3-5_180	16	25	170	-	55
hwb5_55	5	25	60	-	18
hwb6_58	6	43	143	-	62
hwb7_62	7	332	3762	-	754
cycle10_2_110	12	19	2052	-	41
sym9_148	10	210	5492	-	711
Average	-	-	12094	145	1849/80

Building Detection on Aerial and Space Images

P.Lukashevich, B.Zalessky, A.Belotserkovsky
Department of Intelligent Information Systems
United Institute of Informatics Problems, National Academy of Sciences
Minsk, Belarus
pavel.lukashevich@newman.bas-net.by

***Abstract* — An approach for detection and segmentation of individual buildings on space images and aerial photos is proposed. The approach allows intuitively constructing the system of rules to select objects without prior training, using only simple geometric characteristics of their form.**

Keywords — building detection, clustering, form analysis, satellite image

I. INTRODUCTION

A task of automatic detection of buildings is one of the most important stages of the analysis of aerial photographs and satellite images. Although it is actively developed from the 1980s, the challenge now is far from the final solution, which would able to build a software systems for automatic building detection without human interaction.

A lot of work has been devoted to the subject of building detection on aerial photographs, and a large number of algorithms using various techniques have been developed, so to describe the current state of research is not an easy task [1]. There are a lot of approaches used to extract buildings or building features: the Hough transform; active contours; different variations of oriented gradient and geometric primitives (such as lines, corners and rectangles); image color analysis; analysis of shadows; roof photometric properties analysis; model based approaches, neural networks, etc.

Over the last decade, algorithms became more sophisticated, new approaches appeared to detect buildings. Modern algorithms are multicriterial due to variety of forms and colors of buildings. It based on non-trivial analysis of a set of characteristics of objects to be detected [2, 3, 4].

There need be no doubt that last trend in the field of image recognition and classification is convolutional neural network (CNN), so it also has been applied to satellite images [3, 4]. It is also worth noting that major space survey data providers and high performance computing (HPC) equipment manufacturers show considerable interest to CNN data processing challenges [5].

We do not deny the considerable successes in image processing and understanding tasks caused by the use of neural networks. However, the algorithm presented here has valuable advantage – absolutely clear, understandable and interpretable rules for buildings allocation, unlike the "black box" on CNN.

To solve a task of automatic detection of buildings we use RGB satellite image with a resolution of 3-4 m/pixel (and better), filmed at the nadir or at a slight angle to the surface. We can extract buildings of polygonal shape for one of two cases:

1) a priori information about the color of the roofs of buildings is unknown;

2) the color of the roof is known.

Note that the second case about a color of a roof makes the task easier, providing better results. Although it is almost never turning to trivial image segmentation by color, as in most cases the image includes a sufficiently large number of objects with similar color.

Specifics of approach proposed here lies in using of clustered images. Clusters and their contour are distinguished and labeled. For building (or building part) detection we mainly use information about cluster contour geometric characteristics. Also we use information about the presence of shadows, detected corners with predetermined values, color characteristics of cluster candidates.

To solve the problem several criteria for the buildings detection are used, where each of them allows selecting only part of buildings, but with a small probability of error of the second kind: almost does not detect false constructions where they don't exist.

Extraction of buildings and their elements is based on the analysis of clusters and contour by each criterion separately avoiding increasing the error of the second kind. Each criterion assigns binary deterministic tag to clusters, checking them for given feature. Then a composite decision making rule is applied (it can be represented as acyclic graph). The rule analyzes all assigned tags and makes the final decision.

Such an approach allows gaining a close to 90% result of extracting of buildings on test images.

II. ALGORITHM

As mentioned above, there are two cases for the problem of detection of buildings which have polygonal shapes: a complete lack of information about the dominant color of the roof (case1) and its presence (case 2). It is assumed that we know minimum and maximum dimensions of rectangles bounding buildings, as pairs of numbers h, w ($h < w$) and H, W ($H < W$), respectively.

Detection algorithm consists of extensible set of independent parts (blocks), and obligatory parts are:

1. preprocessing – clustering the original color image, identifies the contour of clusters, finds a polygon that approximates cluster (to simplify feature extraction and

improve the performance) and calculates a simple cluster features (shape, shadow, color and other characteristics);

2. analysis – performs a joint multi-criteria analysis of clustered color image based on all features, selected on the previous stages.

A. Image clustering, border segmentation & vectorization

Let's describe briefly the first part of the algorithm. Denote, $P = \{0,..., m-1\} \times \{0,..., n-1\}$ is the set of pixels $p = (i, j)$ of RGB-image I with sizes of $m \times n$ with intensities I_p, $p \in P$.

By the cluster of image C we understand the maximum sized pixel set, which is connected at 8-point neighborhood system, so that pixels have with the same color, i.e., $I_{p1} = I_{p_2}$ for any $p_1, p_2 \in C$. Clustered image is denoted by $\widetilde{I}$, the set of its clusters denoted by C, and the color of pixels $\widetilde{I}$ on a cluster of C denoted by $I(C)$. Cluster size is the number of pixels $|C|$ in it.

To perform color clustering one can apply different approaches, based on known clustering algorithms (smart blur, median with adaptive window, posterization) bilaterally filtering, combinatorial clustering algorithms and other. For mentioned task, we have used modification of graph cut algorithm described in [6, 7].

Contour $\Gamma(C) = (\gamma_0(C), \gamma_1(C),..., \gamma_{M-1}(C))$ of the cluster C – it is an ordered set of cluster boundary pixels belonging to C and forming a closed raster curve such that $\gamma_{i-1}(C), \gamma_i(C), \gamma_{i+1}(C)$ – are neighbors in an 8-point neighborhood system. Length of the boundary $\Gamma(C)$ will be understood as a number of pixels in the set, denoting it $|\Gamma(C)|$.

It is used an approximation of boundary by polygon at sup-norm to improve performance of the algorithm. Such an approximation is well established analyzing objects' shapes [8]. We apply simple and fast one which is based on iterative assignment of β_i at nodes of the boundary $\Gamma(C)$ on polygon, so the most distant from the current approximation of the cluster's boundary.

Denote the desired accuracy of the approximation by ε. Then algorithm of approximation of polygon's boundary $\Gamma(C)$ looks as described below.

We select pixels, which are the ends of the contour $\Gamma(\mathrm{C})$ diameter, i.e. for the Euclidean norm $\|\cdot\|$ couple

$$(\mathrm{d}_0, \mathrm{d}_1) = \mathrm{argmax}_{(\gamma_0,\gamma_1)\in\Gamma(\mathrm{C})} \|\gamma_0 - \gamma_1\| \qquad (1)$$

It is assumed that d_0 is met at $\Gamma(\mathrm{C})$ earlier than d_1. We believe that at the first step the cluster's boundary is approximated by diameter $\mathrm{d}_0, \mathrm{d}_1$ – so by polyline $(\mathrm{d}_0, \mathrm{d}_1, \mathrm{d}_0)$.

We scan the part boundary pixels $\Gamma(\mathrm{C})$ and between d_0 and d_1, find the pixel farthest from the line segment with endpoints $\mathrm{d}_0, \mathrm{d}_1$. If its distance greater than ε, add it to the approximating polyline. Repeat a similar calculation for the second part of pixels $\Gamma(\mathrm{C})$, lying between d_1 and d_0.

Then we repeat the same calculations for chain of the current iterative border approximation $\mathrm{B_i(C)}$. Calculations are finished when there are no pixels $\Gamma(\mathrm{C})$ at the current step, with which have more than ε from current polyline distant.

Nodes of obtained polylines that approximate boundary $\Gamma(\mathrm{C})$ of clusters $\mathrm{C} \in \mathrm{C}$ with a given accuracy ε, are denoted as

$$\mathrm{B(C)} = \big(\beta_0(\mathrm{C}), \beta_1(\mathrm{C}), \dots, \beta_{\mathrm{K}-1}(\mathrm{C})\big), \qquad (2)$$

where $\mathrm{K} = \mathrm{K}(\varepsilon, \mathrm{C})$ depends on threshold ε, shape and size of cluster C, limited from above as $\mathrm{K} \le \mathrm{M}$.

B. Basic cluster characteristics

Now, on the basis of cluster C pixels, cluster boundary $\Gamma(C)$ and its polygonal approximation B(C) we can calculate several characteristics and binary variables for multicriterial building detection. Note that when solving the task there were calculated over 30 different characteristics of buildings, that are obtained on the basis of cluster representations of images. The most valuable of them, that we have used for described in paper approach, are shown below.

We start with the simplest feature $\rho_{\mathrm{size}}(\mathrm{C}) = |\mathrm{C}|$, which characterizes the area of the cluster, and the associated binary variable $\phi_{\mathrm{size}}(\mathrm{C})$, which is equal to one if $\mathrm{h} \cdot \mathrm{w} \le \rho_{\mathrm{size}}(\mathrm{C}) \le \mathrm{H} \cdot \mathrm{W}$, otherwise it is equal to zero.

One of the most basic features, allowing to avoid cluster group which is not part of a roof is the isoperimetric ratio (C), given by the fraction $\rho_{\mathrm{iso}}(\mathrm{C}) = |\Gamma(\mathrm{C})|/\sqrt{|\mathrm{C}|}$. Value of associated binary variable $\phi_{\mathrm{iso}}(\mathrm{C}) = \phi_{\mathrm{iso}}\big(\rho_{\mathrm{iso}}(\mathrm{C})\big)$ is defined by the threshold τ_{iso}: $\phi_{\mathrm{iso}}(\mathrm{C}) = 1$, if $\rho_{\mathrm{iso}}(\mathrm{C}) \ge \tau_{\mathrm{iso}}$, and $\phi_{\mathrm{iso}}(\mathrm{C}) = 0$ otherwise. For results presented here we used threshold $\tau_{\mathrm{iso}} = 10$.

Along referred simplest features it is possible to calculate a complex characteristic $\rho_{\mathrm{par\&ort}}(\mathrm{C})$, which analyzes approximating polyline B(C) and finds long parallel and perpendicular segments with maximum total length $\rho_{\mathrm{par\&ort}}(\mathrm{C})$, and compares it with the length of the boundary $\Gamma(\mathrm{C})$. If the given number $0 < \tau_{\mathrm{par\&ort}} < 1$ the following is true

$$\rho_{\mathrm{par\&ort}}(\mathrm{C}) \ge \tau_{\mathrm{par\&ort}} |\Gamma(\mathrm{C})|, \qquad (3)$$

the value of connected binary variable relies $\phi_{\mathrm{par\&ort}}(\mathrm{C}) = 1$, otherwise $\phi_{\mathrm{par\&ort}}(\mathrm{C}) = 0$ ($\tau_{\mathrm{par\&ort}} = 0.3$ for presented here results).

The last feature is well suited for the extracting object of urban buildings or its parts, usually represented by rectangular area.

C. Form approximation features

Beyond simple roofs of a multistorey buildings (which are well approximated by quadrilaterals) there are building with another configurations that may also be presented on many aerospace images. So they are: flat roof, gable roof, hipped roof (tent), hipped roof and roofs with two trapezoidal ramps and two triangular end, as well as combinations thereof within the same building of complex shape. Such roofs are clearly visible

on satellite and aerial images, and some of it can be approximated by simple geometric figures – triangles, parallelograms, trapezoids, etc.

This approach greatly enhances the automatic extraction of buildings' roofs. For a quick approximation of the cluster form by simple geometric figure it is used approach based on the Hough transform [9].

The basis of used Hough transform is a voting procedure, which is applied to the space of parameters (collecting space). Classic conversion algorithm operates with the identification of lines in the image. In raster case of Hough transform each point of an image "votes" for a set of lines $h(\alpha,\tau)$ containing it, where τ is the distance from the origin to the closest point on the straight line, and α is the angle between the x axis and the line. Building Hough space for each pixel of entire image is computationally complicated operation. However, when use a pre-found boundary $\Gamma(C)$ of cluster C we can achieve a substantial acceleration. Furthermore, it is possible to optimize it by using an approximating polyline B(C) instead of boundary itself. In this case, to build $H(\alpha,\tau)$ line segments (forming the boundary of a cluster) are used, which are formed already. Each of the segments forms a spot in Hough space and its size depends on the length of the polyline and program settings.

The next stage is to determine the most suitable geometric figure from the set to approximate the cluster boundary. To do this it is used a histogram of distribution of angle α in $h(\alpha,\tau)$:

$$\mathrm{Hist}(\alpha) = \sum_{\tau=0}^{\tau_{max}} h(\alpha,\tau) \qquad (4)$$

Based on this histogram one can select the most suitable angle from a set of possible classes of geometric shapes $F = (F_1,\dots,F_k)$, given by a sequence of angles of its sides $F_i = (\alpha_{i,1},\dots,\alpha_{i,l})$. That angle should give the highest correlation with the histogram of angles. Furthermore, the shift angle in the histogram providing the highest correlation, will also set the orientation of the shape α_{opt}:

$$(F_{opt},\alpha_{opt}) = \mathrm{argmax}_{F_j,\alpha}\left(\sum_{i=0}^{|F_j|}\mathrm{Hist}\left(\alpha_{F_j,i}+\alpha\right)\right) \qquad (5)$$

Next, when the type F_{opt} of shape is set and its orientation is the angle α_{opt}. It is only remains to determine the position and size of its sides. So, the maximum values for the rotation angle of optimum geometrical figures are determined:

$$\tau_i = \mathrm{argmax}_{\tau}\left(\sum_{\tau=0}^{\tau_{max}} h\left(\alpha_{F_{opt},i}+\alpha_{opt},\tau\right)\right), \quad i = 0,\dots,|F_{opt}|-1 \qquad (6)$$

Now we have a parametric description of the geometric shape from the list of given shapes, which approximates better than others the boundary of the cluster (Fig.1).

As a measure of proximity (quality criterion) of found approximation the normalized Jaccard index of the polygon and the cluster is used:

$$\rho_{form}(C) = 1 - \frac{|F_{opt}\ominus\beta_i(C)|}{|F_{opt}\cup\beta_i(C)|}, \qquad (7)$$

where $\ominus$ – is operation of symmetric difference ("XOR") for a set's elements), $|\ \ |$ – area of sets.

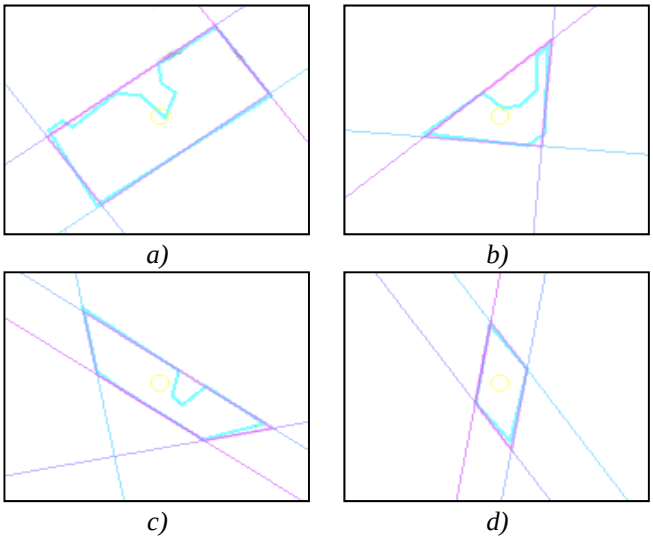


Figure 1. Examples of automated apporximation of a cluster by simple poligons: a) by a rectangle; b) by a triangle; c) by a trapezoid; d) by a parallelogram

When calculating (7) it is added a normalization (unit), which supports the condition $0 \le \rho_{form} \le 1$. Thus, to determine a binary variable $\varphi_{form}(C)$ one can use some pre-determined threshold $0 \le \tau_{form} \le 1$. For mentioned image set the threshold are chosen from the interval $0.75 \le \tau_{form} \le 0.95$

The resulting approximation feature $\rho_{form}(C)$ carries significantly greater information about the cluster. Furthermore, due to its high informativity, the feature of such type is used as the main or even as the only criterion to extract objects with a simple shape.

D. Shadow based features

As practice shows, most satellite and aerial images of urban buildings contain contrasting shadows, suitable for automatic selection. Therefore, if the image has well distinguishable shadows to improve the quality of automatic extraction of buildings it is calculated additional shadow criteria $\rho_{shad}^{not}(C)$ and $\rho_{shad}^{make}(C)$.

To calculate mentioned shadow criteria firstly it is needed to extract shadows from initial aerospace image. Specially developed algorithm [10] is used here which is based on the form analysis of histogram of image brightness in the sliding window. The result of this block is a binary image, where shadows correspond to the black areas, and unshaded areas – to white ones. Note that frequently there are aerospace images, which do not have shadows. So it is no sense to use the current block when working with such images.

After extracting shadows we can proceed with calculation of shadows features. It is proposed to use two criteria, based on a binary mask of shadows.

First one $\rho^{not}_{shad}(C)$ characterizes the degree of filling of the cluster with shadow. It is needed to avoid shadow regions which looks like a roof of a building.

Second one $\rho^{make}_{shad}(C)$ refers to the probability if the cluster C is a part of a roof of a building which gives a shadow.

Calculating a first feature $\rho^{not}_{shad}(C)$ is to determine the part of the cluster, which is not filled with shadow:

$$\rho^{not}_{shad}(C) = 1 - \frac{|C_{shad}|}{|C|}, \quad (8)$$

where C_{shad} – a subset of the cluster C, containing parts of binarized shadows. Thus, the feature is in the range $0 \le \rho^{not}_{shad}(C) \le 1$. Associated with $\rho^{not}_{shad}(C)$ binary variable $\phi^{not}_{shad}(C)$ is determined by the threshold $\tau^{not}_{shad}(C)$ and set equal to one if the cluster has a small intersection with shadows $\rho^{not}_{shad}(C) \ge \tau^{not}_{shad}(C)$, and zero if otherwise (Fig. 2). In this paper we use $\tau^{not}_{shad}(C) = 0.7$.

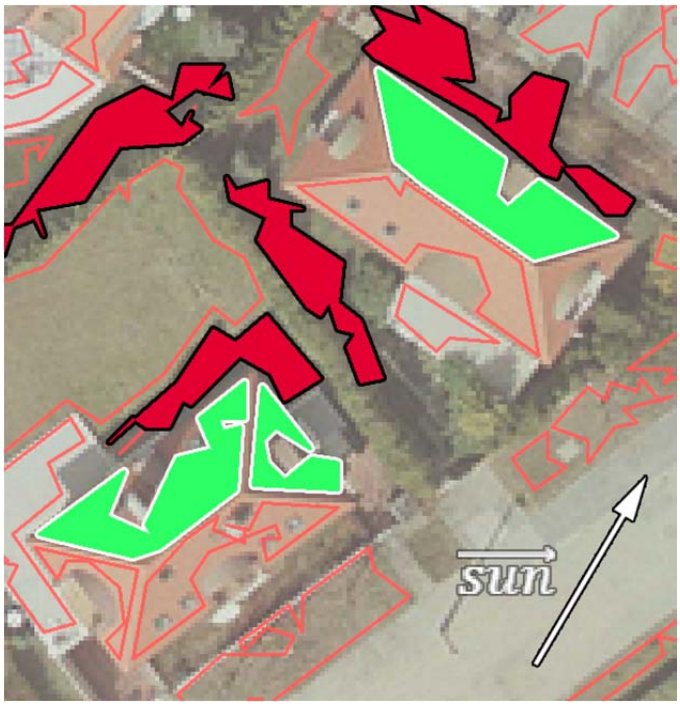


Figure 2. Visualisation of clusters with high $\rho^{not}_{shad}(C)$ shadow criteria (dark red cluster with a black outline) and ρ^{make}_{shad} shadow criteria (light green cluster with a white outline); unfilled contours indicate all remaining clusters

The second criterion $\rho^{make}_{shad}(C)$ is calculated on the basis of the relative position of the shadows and the current cluster. It determines the likelihood that the cluster belongs to the roof and creates a shadow nearby in the direction of the sun's rays $\overrightarrow{sun}$.

To determine the $\rho^{make}_{shad}(C)$ criterion it is proposed to calculate the increase rate ΔS of shadow pixels count in the cluster when it shifted in the sunlight $\overrightarrow{sun}$ direction:

$$\Delta S = \left|C^{\overrightarrow{x_1}}_{shad}\right| - \left|C^{\overrightarrow{x_0}}_{shad}\right|; \ |\overrightarrow{x_1}| > |\overrightarrow{x_0}|; \ \overrightarrow{x_0} || \overrightarrow{x_1} || \overrightarrow{sun} \quad (9)$$

where $C^{\vec{x}}_{shad}$ – is a subset of shadow pixels belonging to a cluster C, shifted by $\vec{x}$ along the sunlight rays $\overrightarrow{sun}$.

As in general the growth rate of ΔS depends on shape, size and cluster C orientation, so to normalize a criterion $\rho^{make}_{shad}(C)$ we use a coefficient $\Delta W^{ort}_{\overrightarrow{sun}}(C)$, equal to the width of the projection of cluster on the axis, which is perpendicular to the $\overrightarrow{sun}$:

$$\rho^{make}_{shad}(C) = \frac{\left|C^{\overrightarrow{x_1}}_{shad}\right| - \left|C^{\overrightarrow{x_0}}_{shad}\right|}{\Delta W^{ort}_{\overrightarrow{sun}}(C) \cdot (|\overrightarrow{x_1}| - |\overrightarrow{x_0}|)} \quad (10)$$

When calculating the feature $\rho^{make}_{shad}(C)$, it is necessary to select a value $\overrightarrow{x_0} > 0$ as it will avoids errors related to the accuracy of the extraction of cluster's border and shadow. Thus, the feature lies in the interval $0 \le \rho^{make}_{shad}(C) \le 1$.

Associated with $\rho^{make}_{shad}(C)$ binary variable $\phi^{make}_{shad}(C)$ is determined by the threshold $\tau^{make}_{shad}(C)$ and set equal to one if $\rho^{make}_{shad}(C) \ge \tau^{make}_{shad}(C)$ and zero if otherwise. Value $\tau^{make}_{shad}(C)$ chosen for the experiments is in the range [0.75,0.95]. Example of feature $\rho^{make}_{shad}(C)$ extraction is shown on Fig.2.

E. Color based features

In most cases it is possible to extract most (but not all) of buildings with high accuracy on the basis of described features $\rho_{size}(C)$, $\rho_{iso}(C)$, $\rho_{par\&ort}(C)$, $\rho_{form}(C)$, $\rho^{not}_{shad}(C)$, $\rho^{make}_{shad}(C)$.

However, image may contain building of non-standard form. Thus it makes sense to use a color of *already found buildings* using the above features so to extract the remaining buildings regardless of a priori information about the color of the roofs. Doing so at the final stage we increase the overall quality of recognition significantly and make analysis more uniform and reliable.

Thus, denote $\rho_{col}(C)$ as the average RGB-color of the cluster. Value of the binary variable $\phi_{col}(C)$ is set to one if the value $\rho_{col}(C)$ is close to color value $\rho_{col}(C')$ of one of any cluster C' which are recognized as part of a building. Otherwise, $\phi_{col}(C) = 0$.

As a measure of similarity it were chosen a criteria based on the comparison of colors in HSV color space (Hue, Saturation, Value), as well as criteria based on correlation of RGB color components.

The formula of comparison of colors X and Y in HSV color space has the following expression (if we normalize color component from 0 to 1):

$$h_{XY} = (1 - |X_H - Y_H|)^{W_H} \cdot (1 - |X_S - Y_S|)^{W_S} \cdot (1 - |X_V - Y_V|)^{W_V} \quad (11)$$

Used weights W_H, W_S, W_V in the formula are determined empirically from a series of experiments based on a fixed set of data. Good results have been achieved when $(W_H, W_S, W_V) = (1.4, 0.5, 0.7)$.

Another promising approach is to compare three-(or more) component color vectors using the Pearson correlation coefficient. Prerequisites for choosing this approach are basic

correlation properties: simple normalization of similarity degree from -1 to +1; sustainability linear changings of brightness (incremental); sustainability linear changings of contrast (multiplication); ability to apply without pseudo-colors (more than three spectral components for satellite images).

The formula to compare color X and Y under correlation for RGB-color space is following:

$$\boldsymbol{r}_{XY} = \frac{\boldsymbol{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=R,G,B}(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \sigma_Y} \qquad (12)$$

Both criteria give comparable results when it testing (with small variations at different samplings).

Analogically, the value $\phi_{\text{col}}(C)$ is set by threshold τ_{col}, and $\phi_{\text{col}}(C) = 1$ if $\rho_{\text{col}}(C) \geq \tau_{\text{col}}$, $0 < \tau_{\text{col}} < 1$, and $\phi_{\text{col}}(C) = 0$ if otherwise (value τ_{col} differs for different types of color comparison criteria).

F. Construction of generalizing feature based classifier for building detection

Generalizing classifier D is equal to one if the cluster belongs to building, and to zero if otherwise. As mentioned above, the features ρ and the corresponding binary variables φ are constructed so that each feature may detect a part of buildings with the least possible error of the second kind (false alarm).

Meanwhile, variables $\phi_{\text{size}}(C)$, $\phi_{\text{iso}}(C)$, $\phi_{\text{par\&ort}}(C)$, $\phi^{\text{not}}_{\text{shad}}(C)$, $\phi^{\text{make}}_{\text{shad}}(C)$ – are basic in sense that each of them is calculated independently. A variable $\phi_{\text{col}}(C)$ – is a derivative from basic features, since its value is defined on the basis of color of already classified clusters.

Thus, values classifier $D(C)$ can be represented as: value of auxiliary classifier $D_0(C)$, which supports independent filtering of clusters by basic features; and values of $\phi_{\text{col}}(C)$, constructed using auxiliary classifier $D_0(C)$.

Classifier $D_0(C)$ can be represented as following conjugation of values of basic variables:

$$D_0(C) = \phi_{\text{size}}(C) \wedge \phi_{\text{iso}}(C) \wedge \phi^{\text{not}}_{\text{shad}}(C) \wedge [\phi_{\text{form}}(C) \vee \phi_{\text{par\&ort}}(C) \vee \phi^{\text{make}}_{\text{shad}}(C)], \qquad (13)$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Components of form criterion - $\phi_{\text{par\&ort}}(C)$, $\phi_{\text{form}}(C)$ – gives mostly similar (but do not coincided) results. Thus, for a task of detection buildings with simple shape when the boundary is clearly visible it might be sufficient to use the simplest criterion $\phi_{\text{par\&ort}}(C)$. For more complicated cases, it is reasonable to use the feature $\phi_{\text{form}}(C)$.

It should be noted that a criterion of belonging of a cluster to the roof $\phi^{\text{make}}_{\text{shad}}(C)$ does not depend on the shape of the cluster, that it significantly expands the clustering capabilities, making it more sensitive to the buildings of irregular shape. However, improper set of threshold $\tau^{\text{make}}_{\text{shad}}$ can cause false clusters.

Another criterion of belonging to the roof of the building is the criterion of color $\phi_{\text{col}}(C)$, calculated on the basis of the auxiliary classifier $D_0(C)$ and color image information. The criterion is independent of the shape of the cluster. After inclusion in the general decision rule, the final classifier $D(C)$ can be written as:

$$D(C) = D_0(C) \vee \{\phi_{\text{size}}(C) \wedge \phi_{\text{iso}}(C) \wedge \phi^{\text{not}}_{\text{shad}}(C) \wedge \phi_{\text{col}}(C)\}, \qquad (14)$$

It is easy to see that the classifier $D_0(C)$, which selects a part of buildings without color analysis, can be implemented as a tree, and $D(C)$ – as an acyclic graph.

In general, we can construct more complex decision making rule (for example, based on voting of features, thresholding of their weighted sum or learning with a teacher). But, such an approach did not give significant benefits in our experiments, while losing the ease of construction of the final rule as a conjunction of binary criteria.

Solution of a task under 2nd case makes it sense to use the same classifier D with the only difference: it is in need to use the set of available colors of buildings when calculating the value of a variable $\phi_{\text{col}}(C)$. However, if to make the classification of a given set of colors only, it may results to falsely detected buildings with higher probability.

III. EXPERIMENTS

For experiments it have been used a sets of color satellite images of urban landscape with a resolution of about 1 m/pixel. Binary images of shadows to highlight features have been calculated in advance by algorithms [10] and used additionally to satellite (aerial) image, when necessary.

Detecting buildings on color test image of 1920×1080 pixels (Fig. 3) under 1 core of Intel® Core™2 Duo E6400 (2MB Cache, 2.13 GHz, 1066 MHz FSB) takes approximately 4 seconds (without taking into account of analysis of shadows).

For example on the test image it were found 90 building parts, missed 2 buildings and 10 clusters classified wrongly (Fig. 3).

The results of tests on a set of four aerospace images have showed that the average number of missed buildings is less than 5 % of their total number. Average percentage of falsely detected clusters (which are recognized as belonging to the buildings) does not exceed 15% of the total number of clusters.

To investigate the sustainability of the algorithm to the parametric choice it has been tested satellite images with for parameters values taken for other images of the same type. For example, the values found for detection of buildings on the satellite image of part of a city, has been used to solve the same problem on the other satellite image of another part of the same city. Classification accuracy in such cases remained high and achieved to be 90 %.

IV. CONCLUSION

In this paper, a new approach to the problem of automatic detection of buildings on grayscale and color aerial and satellite images is proposed. Its specificity lies in usage of cluster representations of images of the urban landscape, making it

Figure 3. Results of automatic detection of parts of buildings on the image. The red contour highlighted the identified parts of buildings (90); Blue – missed buildings (2); The wrong classified clusters (10) are highlighted in yellow.

possible to extract contour of clusters, and then to build their approximating polylines. Using of cluster contour and their approximating polylines it increases accuracy of the result and reduces the cost of the task. In addition, the cluster representation facilitates interactive error correction and interactive selection of buildings.

The second feature of the approach is a set of understandable criteria; each of them allows us to extract a part of buildings with a very low error probability of false detection (in contrast with the convolutional neural network approach). Based on the proposed set of criteria it is easy to construct an optimized decision making rule, thus intending into account an individual characteristics of each task.

The experimental results demonstrate the possibility of buildings detection based on the proposed classifier, which is constructed with logical conjunction of a set of simple binary features.

REFERENCES

[1] Mayer H. Automatic Object Extraction from Aerial Imagery – A Survey Focusing on Buildings. Computer Vision and Image Understanding. 1999, Vol. 74, № 2, P. 138–149.

[2] Yang X. Urban Remote Sensing: Monitoring, Synthesis and Modeling in the Urban Environment. Blackwell : John Wiley & Sons, 2011. 408 p.

[3] Saito S, Aoki Y. Building and road detection from large aerial imagery. Proceedings of SPIE - The International Society for Optical Engineering. 2015. Vol. 9405.

[4] Marmanis D. [et al.] Semantic segmentation of aerial images with an ensemble of CNNs. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. III-3. 12–19 July 2016. Prague, Czech Republic.

[5] Exploring the SpaceNet Dataset Using DIGITS. https://devblogs.nvidia.com/parallelforall/exploring-spacenet-dataset-using-digits/

[6] Zalesky B., Kravchonok A., Lukashevich P. Methods of object recognition that use cluster representation of images. 9-th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA`2008). Nizhny Novgorod, RF. P. 339–342. September, 14-20, 2008.

[7] Zalessky B., Lukashevich P. Bayesian classifier. Theory of Probability and Mathematical Statistics. No. 78. 2009. P. 23–35.

[8] Bai X., Yang X., Latecki L.J. Detection and Recognition of Contour Parts Based on Shape Similarity. Pattern Recognition. 2008. Vol. 41(7). P. 2189–2199.

[9] Duda R.O., Hart P.E. Use of the Hough transformation to detect lines and curves in picture. Communications of the ACM. 1972. Vol. 15. Iss. 1.
P. 11–15.

[10] Lukashevich, P., Zalesky B. Shadow Extraction on Images Using Histogram Analysis . Informatika., 2011 № 1 (29). P. 118–128.

Preparation of Medical Students for Cadaveric Anatomy using Multimedia Education Tools

Jaroslav Majerník, Lenka Szerdiová
Department of Medical Informatics
Pavol Jozef Šafárik University in Košice, Faculty of Medicine
Košice, Slovakia
jaroslav.majernik@upjs.sk, lenka.szerdiova@upjs.sk

***Abstract*—Teaching of human anatomy at Faculty of Medicine of Pavol Jozef Šafárik University in Košice, Slovakia is based on combination of lectures and practical lessons. Modernization of curricula changed traditional teaching methods and brought new technologies into the education process during past decade. Because of positive improvements including visualization of individual anatomical structures, these changes were accepted quite well by both the teachers as well as by the students. However, the range of lectures and practical lessons was reduced to 126 teaching hours respectively, taught during first and second study year. Also, the lack of cadavers and increased number of students was considered as a problem that may affect not only the students' competence but also their confidence in later clinical praxis. These, from our point of view, negatives were a big challenge for us to keep cadaveric dissections in practical lessons and to support preparation phase of learning by imaging and multimedia resources.**

Keywords—anatomy; cadaver; dissection; education; multimedia

I. INTRODUCTION

Anatomy, as one of the core subjects in undergraduate medical education, was taught using solely traditional teaching methods during long time history [1]. This oldest subject of medical education resisted various changes regarding teaching methods and approaches. However, the latest advances in modern information and communication technologies and their utilization in pedagogical processes forced also the needs to change this historical and teacher-centered approach.

Considering the nature of human anatomy, the emphasis was placed on cadaveric dissections to enhance competence of medical students. However, changes of anatomical teaching minimized the role of cadaveric dissections or, in some cases, it was completely replaced by plastic models, multimedia learning packages, simulation tools or virtual reality animations [2]. Many of the changes have been caused by limited financial resources, deficiency or absence of cadaver donation, optimization of curricula and related reduction of teaching hours, increasing number of students or insufficient capacity of dissecting rooms [3]. Despite of long historical background, the anatomists discuss about how to best teach human anatomy and the debate still continues [4] as the trends to move away from cadaveric dissections brought no significant improvements [5].

The teaching methods used in human anatomy education cannot be easily assessed as there is a lack of standardisation between medical universities across the countries [6]. One way is to compare results and performance of the students during examinations, however this will also require some type of cooperation between anatomists working at different institutions and/or in different countries. Also, the teaching materials have to be evaluated and improved to be well accepted and understood by the students [7].

Clinical professionals consider the anatomy as an important medical subject with very high and direct relevancy to clinical praxis [8], [9]. It is because the doctors, clinicians and physicians have to know how the combination of all human body subsystems works, what is the normal and healthy harmony between these subsystems and they have to recognize the pathology that disturbs this harmony in terms of proper diagnosis specification. It was also demonstrated that cadaver oriented courses are essential for example in general surgery procedures [10]. Improvements of competences were reported in many other clinical professions [11] and implementation of clinical references, including images of typical diagnostic, intraoperative and clinical findings in undergraduate medical education was recognised as inseparable part of adapted learning objectives [12].

Technologies that were already applied in teaching brought various benefits including motivation of students [13]. Realistic and highly detailed 3D structures were created for virtual dissecting rooms [14]. However, applied in medical education, some of the new and innovative environments already failed in production of medical graduates with sufficient anatomy competence and capability [15]. Nevertheless, the students feel the anatomical knowledge is essential in understanding of individual clinical cases [16]. They often realise this once they enter the residency programs and clerkships, where the limited anatomy experiences become obvious.

The main aim of our work is to optimize teaching of anatomy respecting decreased dedicated time, space limitations of dissecting rooms and shortage of cadavers while keeping the importance of teaching based on cadaveric dissections. Prior to the dissection, the individual topics are introduced to the students by didactic lectures. These lectures build up knowledge base, needed to understand terminology, structures, functions and interaction of particular anatomical systems.

II. MATERIALS AND METHODS

In our teaching approach we combined traditional teaching methods with innovative multimedia tools to find and reach the most effective way in increasing the anatomical knowledge of our students. Respecting the allocated amount of teaching hours (Table 1) online available environment was used to fulfil this aim as well as to allow medical students to be prepared prior to the lecture or practical lesson.

TABLE I. COMPULSORY ANATOMY SUBJECTS FOR MEDICAL STUDENTS

Subject	Allocated Range for General Medicine				
	Lectures	*Lessons*	*Year*	*Term*	*Credits*
Anatomy 1	42	42	1	W	7
Anatomy 2	42	42	1	S	9
Anatomy 3	42	42	2	W	9

W – winter terma, S – summer term

As it was already confirmed in our previous works [17], the students are more active and more prepared to discuss significant topics and to solve any potential disagreements. Our teaching approach follows four phases as it is shown in Fig. 1.

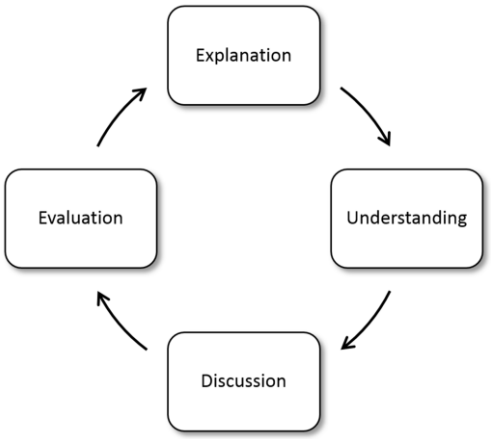


Figure 1. Teaching approach used to improve anatomical knowledge and competence of medical students.

New topics are explained to the students in the form of online available multimedia education materials, lectures and practical lessons. Students reactions as well as their feedback are used to explore problematic parts in detail and to focus on understanding during face-to-face teaching. Discussions are used in practical lessons to resolve any significant students' disagreements. Finally, the students' knowledge is evaluated to reveal either encompassment of particular topic and to be able to move to the explanation of new topics.

Individual phases are supported by face-to-face lectures (33,3%) and practical lessons (33,3%) as well as by online learning environment (33,3%). Practical lessons consist of instructed prosections (11,1%) followed by cadaver dissections (17,5%). Some of the anatomical structures are explained using plastination (4,7%). Multimedia outputs, developed mostly for our conditions at our faculty are shared using faculty learning management system and include outputs from prosections and dissections (28,6%) and partially also the imaginary from medical imaging techniques (RTG, CT, MRI, USG etc.) and living anatomy (4,7%). Representation of individual teaching strategies integrated in our curricula is shown in Fig. 2.

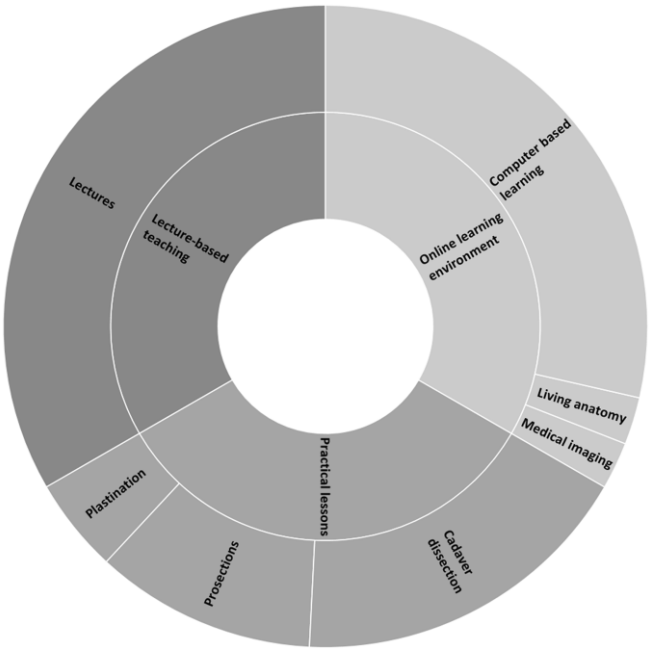


Figure 2. Teaching strategies applied in education of human anatomy.

From our perspective, the better preparation of our students for practical lessons and for cadaveric dissections was obtained by developing and integration of multimedia materials into the curricula. The students can use all such materials anytime and everywhere, no matter what is the range of allocated time for anatomy in presence teaching form.

III. RESULTS

Our teaching scheme of anatomy subjects integrate verified traditional lecturing with explanations, exercises in dissecting rooms, study of anatomical structures in computer laboratory, lecturing supported by 2D and 3D virtual projection as well as online access to education multimedia outputs intended for distance and self-learning and self-repetition.

During last year, we developed a technological background that supports systematic access to the lectures, images, animations and education movies, all prepared and commented by our teachers to harmonize curriculum needs and requirements. The most illustrative video recordings of cadaver dissections realized in our conditions were utilized in all above mentioned teaching forms. Everything was done in the way that allows us to prepare our students to practical lessons with cadaver dissections from both professional and psychological aspects. Of course, multimedia outputs are suitable for and can be used as a substitute in the case of cadaver deficiency, if this happens during teaching periods with large groups of students.

All dissections intended for educational purposes were recorded at our Department of Anatomy, Faculty of Medicine

in Košice, Slovakia. Interventions were captured in HD using both 2D and 3D video cameras. Final education material was processed afterwards embedding texts and audio comments of anatomists.

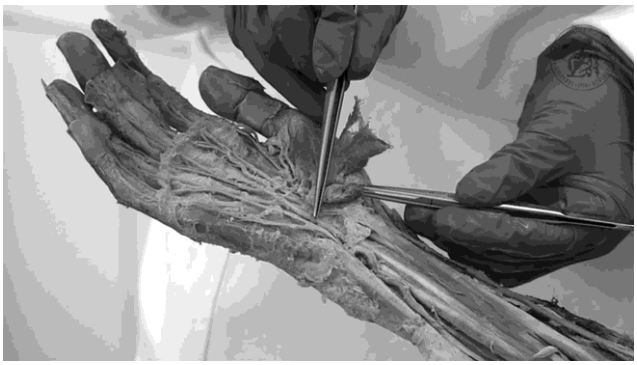


Figure 3. All cadaver dissections were captured in 2D and 3D.

The anatomical structures of upper and lower limbs were processed as first. Final set of upper limb dissections consists of 20 video clips with the length of 63 minutes. The set of video clips for lower limb has 18 video clips with total time of 82 minutes. Recently, the structures of thorax, abdomen, pelvis, head and neck are recorded and processed.

The video-clips are available to all our students in full HD quality via computers in department's computer laboratory located close to the dissection rooms. This allows students to study particular details through multimedia prior to practical dissections. Furthermore, the compressed video sequences are used in online education resources.

The faculty's portal of multimedia support in the education of clinical and health care disciplines (portal.lf.upjs.sk) was used as the online learning environment and as a platform to share already finished multimedia outputs. Here, based on the video gallery principles, the individual multimedia materials are published in the form of articles with embedded movies. An example of the video gallery published at the faculty' portal equipped with 3D animations is shown in Figure 4.

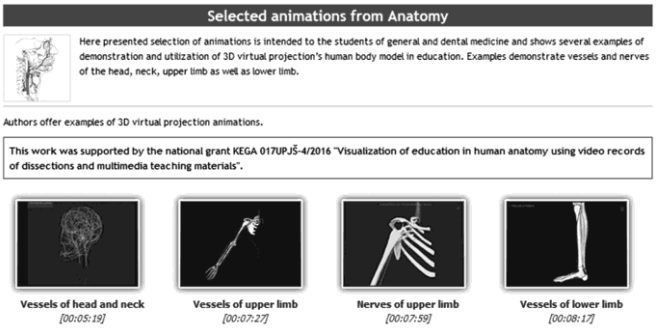


Figure 4. Videogalery of 3D animations shared at faculty's portal.

Video gallery, located in the body of the article uses Video LightBox technology to share movies in well-arranged and attractive way. Individual video files can be stored in external servers or directly in the portal's repository. We did not specify any restrictions to access individual lectures. However, respecting principles of the portal, the content can be available for various groups of users, including nonregistered anonymous users; registered anonymous users who accept the terms of use within his/her registration; users of MEFANET network, i.e., students or teachers from any Czech or Slovak medical faculty; users of local university or faculty, whose affiliation to that university/faculty has been verified at the portal or via the local information system of that university/faculty; or users to whom attachments are made available only on the author's explicit consent. Interconnection of another medical faculties' portals allows to share these materials at international level across all faculties integrated in MEFANET network.

To make didactic lectures more demonstrative and attractive, the 3D virtual projection system was implemented into to the lecturing process. However, its main aim is not to replace, but to support traditional form of face-to-face education. The system is installed in the lecture room with the capacity of 200 students. Wearing anaglyph glasses, the students feel an existence of 3D space like it is in cinemas and they are allowed to study human body systems in much bigger proportions than it is on small computer screens. Scenes of 2D and 3D content recorded from dissections are presented without the need to change glasses the users wear while watching education materials.

To monitor effects on students caused by changes in teaching approach and related inclusion of multimedia education outputs, we compared study results two groups of students that completed all three terms of anatomy.

This preliminary evaluation compared only the final grades from final examination as the evidence of students' skills obtained from both the theoretical and the practical tasks. First group of here included students consisted of 193 participants that passed the final examination from anatomy in academic year 2015/16 and completed only the face-to-face lectures and practical lessons in dissection laboratories. The second group consisted of 241 students that completed anatomy in 2016/17. This group had access to online learning environment and our first set of dissection video records and multimedia content additional to traditional lectures and practical lessons.

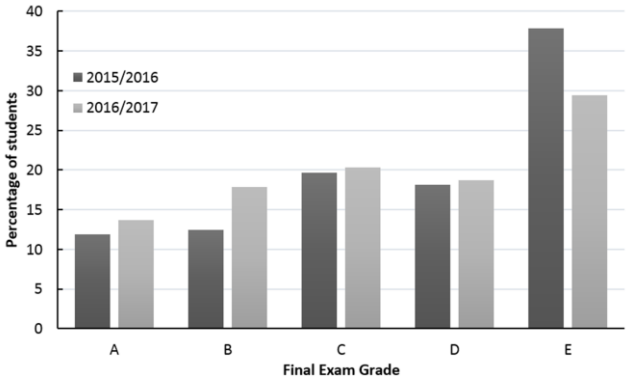


Figure 5. Anatomy final exam results in 2015/2016 and 2016/2017.

As it is shown in Fig. 1, the students having access to supporting multimedia materials reached better results in final examination. The best students (grades A and B) moved from 24,4% (47 students) in 2015/16 to 31,5% (76 students) in 2016/17. The middle grades C and D remains almost the same, i.e. 37,8% (73 students) in 2015/16 and 39% (94 students) in 2016/17, while the worst grade (E) significantly decreased from 37,8% (73 students) in 2015/16 to 29,5% (71 students) in 2016/17.

However, these results have to be carefully interpreted as we have to take into account also another factors like abilities to employ anatomical knowledge in another preclinical and clinical courses, performance of the students etc. Furthermore, the comprehensive evaluation will be able to be done after we will prepare multimedia outputs from cadaver dissections for all anatomical systems. Therefore, deep analysis will be required after all the multimedia materials will be integrated in curricula of atomy and shared to our students via online learning environment.

IV. Conclusions

There are various teaching methods used to teach human anatomy all around the world. Some of them already moved away from dissections, however, we suggest this is an essential part of education in medical students to increase their skills and competence. Opponents, may claim it can be stressful part of the study, but we have to bear in mind that, as physicians and doctors they will work with real patients having real problems that are rarely solved without the stress and emotions.

Therefore, we still continue in teaching of human anatomy based on cadaveric dissection. One of the most resonant reasons is to enhance students' manual skills and to foster their understanding of relationships between symptoms and pathology. Also, the variations in anatomical structures can be easily presented and understood.

On the other hand, we suggest to utilize the potential of modern and innovative teaching methods to compensate reduced time dedicated to the anatomy as well as the lack of cadavers we register during last decade. Using multimedia outputs, adopted to our curricula we obtained positive feedback of our students and we also noticed they are better prepared for both the lectures and practical lessons. We also registered better results of our students in the sense of better performance during final examinations comparing results of previous year where the students were not equipped by latest multimedia materials we created in dissection laboratories. However, we do not want to replace cadaveric dissections and we disagree with those doing so without quantification of benefits and consequences.

Our following work will be aimed to evaluate differences between group of students having cadaveric dissections and group of students without cadaveric dissections. We also plan to quantify differences between the groups where the students will have access to multimedia outputs and students without this access based either on study results and abilities to reach individual learning objectives.

Acknowledgment

Results presented in this work were obtained with the support of the national agency's grant KEGA 017UPJS-4/2016 "Visualization of education in human anatomy using video records of dissections and multimedia teaching materials".

References

[1] E. Kurt, S.E. Yurdakul, Adnan Atac, "An Overview Of The Technologies Used For Anatomy Education In Terms Of Medical History", Procedia - Social and Behavioral Sciences 103, 2013, pp. 109–115.

[2] S.J. Chapman, A.R. Hakeem, G. Marangoni, K.R. Prasad, "Anatomy in medical education: Perceptions of undergraduate medical students", Annals of Anatomy, 2013, 195, pp. 409–414.

[3] M. Estai, and S. Bunt, "Best teaching practices in anatomy education: A critical review", Annals of Anatomy 208, 2016, pp. 151-157.

[4] G.Gradl-Dietscha, T. Kordena, A. Modabberb, T.T. Sönmezb, J.P. Strompsc, B. Gansea, H.Ch. Papea, M. Knobe, "Multidimensional approach to teaching anatomy - Do gender andlearning style matter?", Annals of Anatomy 208, 2016, pp. 158-164.

[5] Ch. Schulz, "The Value of Clinical Practice in Cadaveric Dissection: Lessons Learned From a Course in Eye and Orbital Anatomy", Journal of Surgical Education, Volume 74, Number 2, 2017, pp. 333-340.

[6] S.N.H. Hadie, et al., "Developing constructs of anatomy education environment, measurement: A Delphi study", Procedia - Social and Behavioral Sciences, 2014, 116, pp. 4219–4223.

[7] D.M. Coombs, and S.J. Peitzman, "Medical Students' Assessment of Eduard Pernkopf's Atlas: Topographical Anatomy of Man", Annals of Anatomy 212, 2017, pp. 11–16.

[8] J. Majerník, J. Živčák, I. Staško, "Creation and sharing of human anatomy multimedia education outputs acros medical and biomedical studies", Acta Mechanica Slovaca, 2016, vol. 20, no. 4, pp. 20 – 25.

[9] R. Hudák, V. Rajťúková, J. Živčák, "Automatization of contact pressure measurement between trunk orthosis and patient's body using a matrix tactile sensor", Acta Mechanica et Automatica,9 (1), 2015, pp. 38-43.

[10] G. Sharma, M.A. Aycart, P.A. Najjar, T. van Houten, D.S. Smink, R. Askari, J.D. Gates "A cadaveric procedural anatomy course enhances operative competence", J. of surgical research 201, 2016, pp. 22 -28.

[11] J.J. Meyer, M.M. Obmann, M. Gießler, D. Schuldis, A.K. Brückner, P.C. Strohm, F. Sandeck, B. Spittau, "Interprofessional approach for teaching functional knee joint anatomy", Annals of Anatomy 210, 2017, pp. 155–159.

[12] A. Kranz, I. Bechmann, Ch. Feja, K.R. Kohlhaw, T. Bürkigt, L. Lippross, N. Dietze, S. Löffler, "Implementation of clinical references for undergraduates in anatomy", Annals of Anatomy 210, 2017, pp. 164–169.

[13] Ch.M. Hammer, F. Paulsen, P.H.M. Burger, Michael Scholz, "Integration of the musculature in the course "functional anatomy ofthe locomotor system" - Preparing medical students for the dissection course", Annals of Anatomy 208, 2016, pp. 234-240.

[14] M. Zilverschoon, K.L. Vincken, R.L.A.W. Bleys, "The virtual dissecting room: Creating highly detailed anatomy models for educational purposes", Journal of Biomedical Informatics 65, 2017, pp. 58–75.

[15] R. Pabst, A. Schmiedl, S. Schrieber, T. Tschernig, V.C. Pabste, "Ceremonies of gratitude following the dissection course: A report on procedures in departments of anatomy in German speaking countries", Annals of Anatomy 210, 2017, pp. 18–24.

[16] J. Kubicek, T. Rehacek, M. Penhaker, M., I. Bryjova, "Software simulation of CT reconstructions and artifacts", Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, Volume 165, 2016, pp. 428-437.

[17] J. Vecanová, et al.: Virtuálna anatomická pitva - výučbové videá, Košický morfologický deň: zborník vedeckých prác, Košice, Slovenská lekárska spoločnosť, 2016, ISBN 9788081290527, pp. 103-105.

NAO robot as experimenter: social cues emitter and neutralizer to bring new results in experimental psychology

O. Masson [*,1], J. Baratgin [1,2], F. Jamet [1,3]

[1] *CHArt (P-A-R-I-S), Université de Paris 8 and EPHE, France*
[2] *Institut J. Nicod, ENS, Paris, France*
[3] *Université de Cergy-Pontoise, France*
[*] *Corresponding author, olivier.masson@univ-paris8.fr*

***Abstract*—The aim of this study is to highlight the impact of social factors on the outcome of well-known experimental situations, as the exchange paradigm of Knetsch, used to measure the endowment effect. This bias in decision making brings individuals to assign a widely greater value to an object when they own it. Using a humanoid robot NAO taking the place of the experimenter and left alone with the participant permits to standardize the social factors and the pragmatics engaged within any participant-experimenter interaction. This feature allowed us to validate the hypothesis that the endowment effect could be produced by pragmatical factors, as politeness rules, inherent to all exchange situations between humans. As a continuation of our previous work presented for IDT 2016, we refined our choice of programming methods, keeping a "Wizard of Oz" method for any time the robot must appear as a social entity and applying a completely preprogrammed behavior when all pragmatic factors related to the presence of an experimenter should be annihilated. These methods were applying to the exchange paradigm. We based here on the same paradigm to measure the endowment effect, the exchange paradigm of Knetsch, and using a NAO robot to conduct the experiments. But in contrast with our previous works, a dissociation of programming technics was introduced: a wizard of Oz and fully pre-programmed behavior. After finding that results obtained do not differ regarding the programming technic chosen from previous experiments, we set up a new pilot study to find out if a specific nonverbal clue emitted by the experimenter and varying within a single fixed method (fully preprogrammed behavior) would be sufficient to produce an endowment effect. For this purpose, the first results obtained by only varying the voice intonations of the robot show that vocal features can constitute an important factor for a human individual to activate the application of social standards towards a robot, bringing him to produce an endowment effect.**

Keywords— humanoid robots, reactive systems, embodied cognition, HRI, NAO, social norms, social context, pragmatics, relevance theory, politeness rules, cognitive bias, decision-making, class inclusion, endowment effect

I. INTRODUCTION

Could different approaches of programming a robot as an experimenter for research in psychology could impact the pertinence of results?

These present works aim to refine the programming methods used in our previous study presented in the IDT 2016 [17] to answer this new question. In these previous works, a unique method of using a NAO robot as an experimenter (a derivation from the classical "wizard of Oz" method) was exposed for conducting the exchange paradigm of Knetsch [14]. This experimental paradigm has been widely used to measure the "endowment effect", constituting a bias in decision making, and following a specific set of steps:

Two objects of an identical economic value A and B are chosen through a first control task within a group of participants, ensuring that there is now preference between the two objects the participants want to keep.

Three experimental steps are then engaged: the experimenter gives the participant a first object A; then the participant is asked to perform a masking step (for example some reasoning task) during around fifteen minutes; and finally the experimenter propose to the participant to exchange A against B. These steps are reproduced in a second group by permuting the presentation order of A and B.

The endowment effect consists in attributing a greater economic value to an object when it is possessed and a much lower value to that object when it is not [32]. So, in the exchange paradigm, this bias is revealed by a widely higher refusal rate than acceptance within the two groups of participants [14].

Humans' reasoning and decision making can be impacted by some pragmatical factors, leading them to make biases and reasoning differently than what is expected [2], [3], [24], [29], [30]. This phenomenon could bring to make irrational decisions in a large set of fields, as economics for example. In our previous works, to test the hypothesis according to an endowment effect could be produced by pragmatical factors and the activation of social norms [4], [10], [11] - that is to say the application of politeness rules in the exchange paradigm situation -, the use of a NAO robot allowed us to control the non-verbal cues emitted by the experimenter, influencing the social norms activated in the context of a participatory-experimental relationship [17]. As nonverbal cues emitted in interaction brings each interlocutor to consider the other as a social entity [7], [16-18], these cues could call the application of all social norms attached to a specific situation, and so

would have an important impact on the result obtained in all exchange paradigm sessions with a human experimenter.

As a continuation of this work, we bring a variation in the programming methods used for this same experimental paradigm, using two distinct methods: a fully preprogrammed behavior and the variation of the "magician method of Oz" we used, in order to determine whether the programming method of the robot could influence the results obtained.

In order to provide a complete answer to this main issue, we propose a new experimental protocol, which consists in varying a type of nonverbal cues emitted by the experimenter in isolation, while using only one programming method: Fully preprogrammed robot to conduct in an entire autonomous way the experiment with the participant.

Although robots have been often used in HRIs to observe the impact of their appearance or their behavior on the outcome and several dimensions of interactions, studies reporting their impact on bias production in decision-making or in reasoning fields are not so numerous [4], [16-18].

The exchange paradigm could involve contextual and pragmatics effects that could not be entirely controlled by a human experimenter. As this one cannot standardize every social cue emission, behaviors commonly observed from participants could be induced by social norms applications towards the experimenter, as this one is still considered as a social dialog partner. These pragmatical factors induce the use of social norms, especially politeness rules applied to all exchange situations: for example in our western countries, this is "forbidden" to refuse or redeem a gift [9-11].

To control these pragmatic effects, robots can be helpful because they can be perceived by humans as social entities or not, according their behavior [5], [17]. Thus robots bring exclusive benefits in studying the impacts of social factors on decision making in HRIs [16-18].

A. Why robots should be used to study social factors activated in within humans' interactions

Humans have a natural tendency to anthropomorphize objects, that is, to attribute to them human capacities such as thinking, feeling emotions and having their own intentions. This phenomenon is magnified when the "non-human" object is in motion and when a situation of dialogue is encouraged [19].

This phenomenon is thus amplified when a human interacts with a robot, and even more so when a humanoid robot behaves, in addition to its appearance, as a human. It has been shown in many studies that human-like behavior has far more impact on anthropomorphization than the appearance of the object itself [5], [13], [19], [26], [28]. This phenomenon would be useful for humans to regain some feelings of familiarity, especially in unknown situations and would not be related to the level of knowledge on technics and in computer sciences [1].

Between humans, the act of influencing other individuals is a necessary ability in all of social interactions [8]. This act is possible when both agents are considered as social beings and so a robot could influence the behavior of an individual when this one anthropomorphizes it.

Research in robotics has identified some well-known factors according to a robot could appear as a social entity, making it able to influence a human behavior:

- Appearance and behavior of the robot: the level of presence of the artificial agent – physical or virtual; the agent gender – male or female. These factors impact the level of confidence felt towards the robot and the degree of responsibility the individual can yield to it [33], [35].
- Several modalities of interactions with the robot also produce some different behaviors in humans, as cooperation in action with the artificial agent: robots must be able to interpret and to reproduce non-verbal cues of humans, as body language and visual contact [7], [16], [34].

Some recent studies use robots to analyze human cognition by manipulating an all set of visual and behavioral factors, showing that to study human cognition by this way is far from being an aberrant idea [7]. Some studies, for example, underline the strong effect of anthropomorphism with traditional computers, as everyday tools, showing that individuals could take them as real social partners when technologies enhance this phenomenon by including some subtle cues [20].

This tendency was called the "Media Equation" and was induced by introducing in computers some expressions linked to a gender or by calling some cooperative behavior from humans [27]. In latter studies, this precept has been applied to HRIs [32]. For example, the manner humans physically respond to robots is exactly the same that the one they use to interact with other humans. Especially, the proxemical distance taking place between humans and robots is identical as within humans' interactions [8].

Since the youngest age when children play with toys, humans train to build interactions with everything. These human skills make individuals able to represent all artificial agents as social entities [8], [27]. By analyzing factors triggering social norms activation and applications within humans, it is possible to recreate these behavioral factors or to control them by applying them to robots. Moreover, a robot could be presented to an individual so as to lend it a specific social role, changing the context of interactions with it.

This last point will be very major in the fact that the perceived social role of the robot can impact the contextual situation of every experiment in psychology research. Thus, by the use of a humanoid robot, it is possible to study all impacts of behavioral factors on HRIs social quality, by isolation or by combination. So, a robot brings the inestimable possibility of control all non-verbal and automatic behavior susceptible to induce any social skills in all interactions, including in within humans relations.

Based on Bateson's « double-bind » theory suggesting that inconsistency between information's content and presentation

could produce mental disorders [6], some authors use the same methods to show impacts produced by robots when they express some inconsistency different politeness between postures and statements. This inconsistency would impact reasoning and memorization, especially with a rude posture and polite statements. These studies underline the importance of some different factors effects combined in HRIs on the social quality of interactions [20].

The studies mentioned above focus on comparison between HRIs and within human's interactions, in terms of impacts on interactions quality and social perception of agents. These works often use new experimental methods [1], [7], [31]. The study presented here addresses to use well-known experimental paradigms and to derive them according a new paradigm: to follow traditional experimental procedures but by replacing the human experimenter by a humanoid robot.

B. Outcomes from our previous works

For our previous works presented for IDT 2016, our choice went toward two important topics in cognitive psychology: the Piaget's "inclusion task" and the "endowment effect" studied through the exchange paradigm of Knetsch. We assume here that both a human experimenter and the social context where the experiments take place could bring participants to make additional implicatures and to reinterpret what they are asked to do by an experimenter. Thus, results obtained in these well-known experimental paradigms would be induced by some implicit social and contextual factors, activating social norms in the experimenter-participant interaction [4], [10], [11], [16-18], [23]. These factors would impact for a major part on results.

Some previous researches showed the importance of social norms impacts on outcomes in the exchange paradigm: in western countries, refusing or redeeming a gift is perceived as a violation of politeness rules. The endowment effect produced in any exchange situation would result from some kind of social norms activation applied to this exchange situation [10], [11], [18].

Focusing on the exchange paradigm of Knetsch and basing on a practice with children within the inclusion task of Piaget to improve our wizard of Oz method [17], we will then present two different methods to use NAO: a totally "preprogrammed" behavior and a derivation of the "wizard of Oz" method.

I. ISSUE AND HYPOTHESES

Regarding the main question we asked in our present work: "*How to demonstrate and to measure the impact of social norms on the outcomes of these experiments when even the experimenter cannot control the social cues he emits and his social position inferred by the participant?"*, it was shown that the only serious way to answer this question is to use a robot to entirely standardize nonverbal cues emitted by the experimenter, using a robot to conduct the experiments.

The *new technical issue* arising here is to investigate whether different programming approaches could impact the results we obtained. To answer this question, we focus our present study on the endowment effect, still conducting the exchange paradigm of Knetsch, as an experimental model we use for this purpose.

The endowment effect, as bias in decision making, could arise in every exchange situation. Humans have always traded objects. So exchange appears as a conduct which has been socialized – so as to say including social rules. As every socialized conduct meets social norms, social behavioral factors as non-verbal cues for example or the perceived role of an individual could have an important impact on outcomes of all trades. This impact would so apply to the exchange paradigm of Knetsch too, as the experimenter leading experiments is human. The exchange paradigm regroups three experimental main steps as following, in individual sessions, with two objects A and B by an identical economic value:

- The experimenter gives to the participant a first object A.
- The experimenter gives the participant some cognitive task to treat (as a questionnaire containing reasoning items). This task has to take around fifteen minutes and is called the "masking step" as it is considered by the participant as the main task of the experiment.
- The experimenter asks the participant if he accepts to exchange the first given object A against the second object B.

These steps are led separately with two independent groups where the presentation order of objects is permuted. According to this paradigm, an endowment effect is highlighted by significantly higher rates of refusals in the both two groups. In psychological and economical literatures, the endowment effect can be explained by two main classes of explanations: the "loss aversion" [12] and the "ownership" [25] phenomenon.

In latter studies in which the experimenter endorsed different roles or behaviors, the endowment effect could be minimized or annihilated, these studies varying the social context of the experimental situation. These results show that politeness rules, activated by the social context framing the experiments, constitute a major factor to produce an endowment effect [10], [11], [16-18].

The following hypotheses were investigated, with a NAO robot replacing the human experimenter [10], [17]:

- The endowment effect will be annihilated when NAO will behave neutrally regarding a social point of view, considered as a simple tool to collect data.
- The endowment effect will reappear when NAO will behave socially, applying social norms in interaction.

Results we obtained in our previous works presented in IDT 2016 supported these hypotheses. Experiments were there led by a NAO robot, programmed with a wizard of Oz method.

In what way changing the programming methods applied to the artificial experimenter could impact these results?

II. BUILDING SEVERAL APPROACHES IN USING NAO

We suggest two methods to program NAO to make him able to lead every experiment without any physical presence of the experimenter.

1) Programming a fully "autonomous" NAO

In this configuration NAO, all NAO's actions are launched by itself according to certain stimuli for which he has been programmed to be receptive. This has the advantage to standardize to perfection the robot behavior: robot will just wait for some answer from participant's behavior to continue on the experiment. After the main program is launched, no behavior resulting from the human experimenter comes into play in the course of the experiment. Expecting from the participant, there is no other human actor to determine the experiment rhythm in terms of timing of actions.

Nevertheless, in this case there is also a disadvantage brought by many technical constraints, able to make the experiment turning off, as for example:

- Visual recognition may take some time to recognize a specific object in space and requires precise and lengthy adjustments to be made.
- Depending on the room acoustics and the participant's vocal features, NAO's speech recognition module may not understand certain sentences. This acoustic configuration is never exactly the same from one experiment to another.
- NAO can only catch sentences it has been programmed for.

However, with a well programmed robot coupled to some explicit experimental steps that the robot can recognize properly, this method is suitable to let NAO to lead the exchange paradigm, as this only contain three simple main steps. So we use this method to review the exchange paradigm of Knetsch.

For this method, Harel's statecharts are very useful to build a complete reactive program because only sensors and its reactivity to environment will be used to carry out the experiment. Indeed, Harel's statecharts are useful to model every system behavior based on events, as called as a "reactive system". Statecharts, as we will see can be used as a starting framework to build many programs.

2) Source of a "semi-wizard of Oz" method

This method, widely used in experiments in robotics and psychology, consisting of controlling a remote robot without the participant's knowledge and making him believe that the robot is perfectly autonomous, allows avoiding lost results where a totally preprogrammed autonomous behavior could totally stop during experiments, due to some technical malfunction.

Nevertheless, this technique does not allow complete standardization of all actions led by the experimenter. In the Wizard of Oz method, rhythms of gestures are never exactly the same as the experimenter launches himself every step. An extreme would reside in the fact that the robot would reproduce all gestures expressed by a distant human experimenter. In this case, uncontrolled non-verbal cues could also be retransmitted, parasitizing the social perception of the participant towards the robot.

For this purpose, a traditional wizard of Oz could be criticized in the fact that the robot is totally operated by human experimenters and then, pragmatical factors and social behaviors could arise, even if experimenters are not in the same room. To avoid this, we suggest a "hybrid" wizard of Oz method, used to simultaneously preserve a maximized standardization all by avoiding malfunctioning. This method was improved basing on previous experiments led with young children.

Children do not behave as adults in front of a robot and this method allows the experimenter to deal with any behavior that only an autonomous robot could not manage. This constituted a source for our idea to set up a technic that takes place between Wizard of Oz methods and a fully preprogrammed behavior in the robot.

This technic revealed to be really useful to conduct an entire experiment to the child's rhythm, in the fact that every following step does not start while the child does not proceed or understands the current task.

Then, this new programming approach was applied to the inclusion task of Piaget. According to Piaget's theory, children do not master the concept of inclusion of objects classes until when they meet a specific development step, at the age of around eight years old [21]. Before achieving this step, children would confound a whole set of objects (superclass – for example "fruits") with some of its parts (subclasses – for example "apples"), preventing to make vertical comparisons.

Basing on the same experimental steps from the traditional inclusion task of Piaget, using two subcategories of objects (5 apples and 3 pears) included in a supercategory "fruits" placed on a table in front of the child, we proceed the experiments with NAO replacing the experimenter:

- NAO tells the child "he" wants to learn everything about the world, remaining he does not know anything and that the child, as a professor, would teach him.
- Showing the fruits set on the table in front of the child, NAO asks him to show the apples, and pears in a second time. Then NAO asks him to give a name for both apples and pears. This preparing step assures that the child knows the objects used.
- NAO asks the child: "Are there more apples, or more fruits?"

In the traditional task conducted by a human – in most case a teacher -, this question is ambiguous. Indeed, asking to compare a subcategory as its supercategory is an unusual request. Furthermore, the social context framing the task is far from neutral: as the adult present an authority figure as a teacher, the child can make a different of the task than what is exactly asked. A child could reinterpret the task by assigning some expectations to the experimenter, understanding for

example he has to show he is able to make comparison between objects and to count them. Thus, this different implicit interpretation could impact the results found in traditional experiments [23].

By using a robot NAO, we could entirely control the social framework of experiments and pragmatics factors. Moreover, robot presented itself as a stupid being, wanting to learn everything of the world and positioning the child in the role of its "teacher". Thus, these works built an entirely new frame for the inclusion task, completely annihilating ambiguity: a "stupid" robot could only ask for questions in order to get an answer for this exact question. Or the child, the robot does not have any expectation of him. In these experiments, to perform each step, we have to ensure that the child understands the current task before going further. The semi-wizard of Oz method allowed us to achieve it. We then applied it to the exchange paradigm with adults, showing that method was efficient [17]. This set of experiments permitted to improve our programming methods and its application in live sessions.

III. MATERIALS AND PROCEDURE

A. Materials

The two objects chosen for the exchange paradigm are a black "BIC" pen and a little candy box (smarties). The masking step consists in filling a questionnaire containing reasoning tasks.

The robot used in all experiments is a humanoid robot NAO (version 5 – "Evolution" by Softbank Robotics). Our choice oriented towards NAO for advantages residing both in its appearance and its body abilities. NAO can hold little and soft objects with its hands and embeds useful modules needed as a vocal synthesis, a voice recognition module, the ability to store audio files, and holds two cameras for processing visual information in a three-dimensional space, allowing face tracking [17].

NAO's behavior is programmed with the software Choregraphe in version 2 built by Aldebaran, and stored on a laptop connected to the robot via Wi-Fi from another room. The software "Choregraphe", proposing a visual interface used to program NAO's behavior is also very suitable for this kind of experiments as this will be explained in the next section.

B. Procedures using NAO

The same experimental steps are used in the exchange paradigm of Knetsch, with the same kinds of experimenter we used in our previous work [17]:

- The donation step: the experimenter gives a first object to the participant
- The "masking" step: the experimenter asks the participant to perform a reasoning task
- The exchange proposal: the experimenter asked the participant if he would exchange the first given object against a second object.

In contrast with our previous works, we use both of programming approaches in these experiments: a fully preprogrammed behavior in the "neutral robot experimenter condition" and the semi-wizard of Oz we derived in the "sociable robot experimenter" condition. These two new conditions were added to the traditional "human experimenter" condition. In the "neutral robot experimenter" condition, the robot was presented as a simple tool to collect data, and all non-verbal cues were annihilated. In the "sociable robot experimenter" condition, main non-verbal cues were reintroduced, considered by research as to have strong impacts in every HRIs: gazes, face tracking, vocal intonations.

IV. DESIGN TOOLS

The experimental design has to be built around one main constraint: NAO, being operated by distance – Wizard of Oz method - or not, must be able to lead a whole experiment with a participant without any human experimenter in the same room. This is the main technical goal of this work: testing if both methods work to make it possible to let a robot lead experiments alone, avoiding all malfunction.

A. Harel's statecharts

Statecharts (SC) of Harel bring a useful to program an autonomous robot and also to build a model of actions. As some systems exist and allow programming them directly from SC, this is not the case for NAO. Nevertheless, the visual interface is much closed to it. SC present a good tool to start the design of NAO's behavior, as a framework. Starting directly from Choregraphe is also possible but visual contents can become quickly confused. As Choregraphe is more a programming tool, SC are in this case used as a modeling tool for NAO's behavior and this step has to be achieved before to start programming. In that way, we could also directly draw SC on a paper with a pen. However some software can help to make this step very easy. The software used here to do so is "SmartDraw". Our choice oriented to this software because it is not necessary to bind the SC to complex programmed functions, making it possible to use it only in a visual modeling step.

In the exchange paradigm situation, waiting for the answer of the participant when the robot asks him if he would accept the trade is a crucial step because the measurement of endowment effect is based on this answer. Nevertheless, the voice recognition here is used at the end of the experiment and even if NAO does not understand this answer, the human experimenter can save the real answer and whereas NAO would not hear it, the experiment does not need to continue from what it could understand.

Then, as we did in our previous work, we started designing the experiments by the help of a modelization from SC [17], for both methods (Wizard of Oz or fully programmed behavior). Actually, the modelization for both has not to be different. On the other hand, the application of the model in terms of action will differ: in the wizard of Oz method, the linking action between to states of the robot implies a manual activation of the next state of the robot, from distance by the human experimenter. In contrasts, bounds between states

drawn in SC in the case of a fully preprogrammed behavior induce that NAO's sensors are programmed to wait for some stimuli of the environment, to automatically activate the appropriate behavior. This reaction does not require any added action from human experimenters from the other room.

B. From StateCharts to the programmed behavior

SC take in account of different sates of a system regarding stimuli of the environment (signal from a sensor or a manual call for action by a human experimenter). This is possible to derive it in Choregraphe as it takes in account all events caught by NAO's sensors.

The difference here between the two methods is that when NAO is fully autonomous, the next actions are launched automatically after some stimuli has been captured by NAO's sensors, in contrast with the semi-wizard of Oz method, where following actions have to be launched manually by the human experimenter. Especially, this method allows avoiding using the vocal recognition module.

In the wizard of Oz method, all boxes do not have to be linked to let the experimenter launch the following actions of the robot himself. Thus the experimenter can launch each block of actions separately as many times this is necessary, making NAO's behavior fully adaptive to the participant's actions (see Fig. 1).

Another benefit of the Choregraphe software is that it is possible to edit any program code manually, each box being written in Python. It is also possible to edit the whole behavior, as order of some action for example, directly from source code of the Choregraphe project. This one contains a file in ".xar" extension, reporting all boxes used in the visual interface and all libraries loaded for it. Including blocks containing all Pyhton codes can also be found here in one file. This file, containing XML nodes appears as a main controller by calling all used functions. So this makes it easy to everyone knowing some XML basics to edit some features of the whole NAO behavior from there.

C. Choices towards methods

Our choices regarding methods used were oriented according the kind of response quality needed from NAO. When NAO must endorse a social identity, as in "sociable "experimenter robot" condition, we assumed that the possibility to launch all action manually would be more human likeness. Indeed, the human experimenter can immediately catch and understand every subtle behavior from the participant and then we argued that the rhythm would be more natural. So we oriented towards the wizard of Oz method in this case. The only condition using an entirely preprogrammed behavior was the "neutral robot experimenter". Indeed, this method only uses NAO's sensors to change of state and launch next actions. At this time, sensors cannot catch any stimuli from the environment and NAO's reactions in terms of time response can be slower or break a natural rhythm of actions. In this condition, NAO must not be perceived as a social entity so these technical limits come at the right time here.

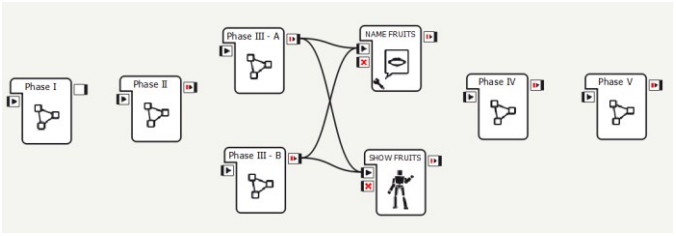


Figure 1. A capture from Choregraphe used to design our "semi-wizard of Oz" method. Each "phase" contains several subbehavior. The simplified view allows the human experimenter to quickly launch the appropriate program in the good time.

V. EXPERIMENTAL TRIALS

A. Configuration

In each session for both experiments, NAO sits in front of the participant on a table. Also, human experimenters always oversee all experiments in a separate room. A Wi-Fi connection between NAO and the computer makes it possible to follow all the sessions in live thanks to NAO's cameras focusing on the participant's actions. One of the two small objects (a bic pen or a little candy box) is positioned on the table near the robot. NAO holds the second object in one arm placed behind its leg as to keep this object hidden from participant.

B. Experimental conditions

The main experimental varying factor is the robot behavior, impacting the social context. We expose here the added conditions using NAO, in comparison with the traditional situations with a human experimenter:

In addition to a traditional "human experimenter" condition, two conditions are applied to the robot. In the former ("neutral robot experimenter") NAO is presented as a simple tool used to collect data. It does not emit any non-verbal cues and its discourse is courteous without containing other verbal information than needed. In that condition the whole performance of NAO has been entirely programmed. In the latter ("sociable robot experimenter"), several non-verbal cues have been reintroduced to NAO's behavior: face tracking and head moves, voice intonations, gaze expressions and eye blinking. Added to this, NAO's discourse is more user-friendly, presents itself as a colleague of human experimenters and repeats the participant's name during the introduction of the experiment. In that last condition NAO's behavior has been programmed ion submodules manually launched separately by the human experimenters.

VI. RESULTS AND DISCUSSION

Differing programming technics from our previous works had not any significant impact on results. Furthermore, a larger set of data show that the use of a robot to make experiment could change the social context, activating different representation of the role endorsed by the artificial experimenter. The absence of any significant difference on results from previous ones shows that the programming method is not a main factor impacting the endowment effect

production. Nevertheless, the fully preprogrammed behavior technic answers to all the standardization main issues, allowing letting NAO only by itself complete an entire experimental session.

With a randomized sample of 60 adults distributed according three conditions (20 adults per group):

- A standard human experimenter condition has been replicated and according to literature, refusals rate was of 90% against 10% of acceptations for the exchange proposition. So a strong endowment effect was measured.
- In the “neutral robot” condition, the endowment was totally annihilated and we did not find any significant difference between acceptances and refusals rates.
- In the “sociable robot” condition, refusal rates were about 63% against 37% of acceptances. An endowment effect was revealed, but not as strong as with a human experimenter.

This experiment reproduces that presented in our previous work for IDT 2016, although here two different programming methods are used (a fully preprogrammed and autonomous behavior for the "neutral experimental robot" condition and the use of a method Of the Oz magician type for the "sociable experimental robot" condition). The results obtained confirm those of the pilots that were carried out and are similar.

Nevertheless, the use of different programming methods did not show any impact on results, compared to those obtained in our previous study [17].

Moreover, this confirms that the only factors susceptible to make difference in results on the strongest of the endowment effect are the social nonverbal cues emitted by the experimenter himself.

VII. New experimental design to isolate pragmatical factors

We suggested in the above experiments that a wizard of Oz method would enhance the sociable quality of interactions between the participant and the robot. Nevertheless, no significant difference was found in this changing from our previous results.

All prospective programming effect discarded – the main objective of our present work -, this now allows to turn toward a major issue: which nonverbal cue would have a maximum impact on the strength of endowment effect produced?

In preparation for our next works and basing on the exchange paradigm used to measure the endowment effect, we suggest a new experimental design applied to other pilot experimental sessions, using the “fully preprogrammed behavior” programming technic, only varying one factor: NAO’s vocal intonations. For this purpose we set up two conditions: “no vocal intonations” and “natural vocal intonations”. By fixing on the method of a fully preprogrammed NAO’s behavior for both conditions, we ensure that only the impact of a specific non-verbal factor (vocal intonation here) is measured. For the “natural vocal intonations” condition, NAO’s statements were the same than in other condition but spelled by prerecorded human vocals, treated by a vocoder to reproduce NAO’s voice.

With two independent groups composed by 20 adults, first results show a notable difference between the two conditions: in the “no vocal intonations”, 8 participants (40%) refused the exchange proposal and in “natural vocal intonations”, 13 participants refused (65%). Thus, as a major part refuse to exchange in “natural vocal intonation condition”, in contrast with the “no vocal intonation” condition, these first results tend to show that only varying this specific vocal intonations factor would be sufficient to produce a bias.

These results are totally new from our previous works. Thus, future investigations in that way could show that the vocal intonation factor, applying to a robot, could bring people to produce some biases in decision-making.

VIII. Conclusion

In the continuation of our work presented at IDT 2016, dissociation in programming technic was introduced to let the NAO robot conduct experiments with the participants.

We did not obtain any significant difference in results caused by this variation, showing that once the robot is able to stay alone with the participant to conduct the experiment, the programming technics cannot be considered as main experimental factors for the production of results.

Nevertheless, this gives a guideline to choose the best programming technics to apply, according the type of experiments to conduct and the type of participants. For example, it seems that in every study using a robot with children, a wizard of Oz method will be more appropriate, as children, in front of a robot, do not react as adults.

Thus, results obtained in this present study confirmed the tendency we obtained to validate our starting hypothesis: pragmatical and contextual factors have a strong impact on the endowment effect production. Indeed, by taking NAO as a tool to standardize the social context impacting each experiment, we ensure that only the NAO’s behaviors and the way it is introduced to the participant are sufficient to produce large differences, especially from outcomes widely observed in literature.

So, this study is a good starting block to guide a choice between two different methods of using NAO: a fully preprogrammed behavior method and an efficient variation of the traditional wizard of Oz method. Both methods are easily accessible and efficient, though the first one would be more delicate to set up with children for some technical constraints reasons. Mainly, the voice recognition module will demand more adjustments.

These two different approaches used here could be resumed as some kind of program templates and we are working to make it the easiest to reuse, even for neophyte people in informatics. Thus, other programming methods, compatible with Choregraphe, could be developed as for example friendly

web interfaces, adapted to each experiment template allowing to directly edit main common variables and then to apply these variation to a custom downloadable version of the program, usable by the robot. Some enterprises in robotics work on this dimension to allow designing program from web interfaces but this would be interesting to work on a fully customizable process to do it from scratch, as Choregraphe projects structure is entirely editable as XML nodes.

As we did not find any significant outcome caused by a variation in programming technics, we finally present last pilot experimental sessions in which only vocal features of the robot varied within the exchange paradigm situation.

First results, tending to show a strongest endowment effect in the "natural vocal intonation" condition, highlight the fact that vocal features would constitute a major factor modulating the application of social norms towards a robot or a virtual agent.

This tendency will have to be validated with a larger number of participants and the same paradigm could be derived with other factors taken in isolation, as for example the two other main non-verbal cues: movements and gazes clues expressed by the robot.

APPENDIX

Picture of an experimental session in inclusion task with young children

ACKNOWLEDGMENT

Financial support for this work was provided, by a grant from the ANR Chorus 2011 (project BTAFDOC), and by a grant of Institut des Sciences Complexes (2014-ISC-PIF petits et moyens équipements).

REFERENCES

[1] R. Baddoura and G. Venture, "Social vs. useful HRI: experiencing the familiar, perceiving the robot as a sociable partner and responding to its actions," Int. J. Soc. Robotics *5*, pp. 529–547, 2013.

[2] J. Baratgin, Le raisonnement humain: Une approche finettienne [Human Reasoning: A Finettian Approach]. Hermann Publishers, Paris, France, in press.

[3] J. Baratgin and G. Politzer, "The psychology of dynamic probability judgment: order effect, normative theories, and experimental methodology," Mind & Society, vol. 6, no. 1, pp. 53–66, 2007.

[4] J. Baratgin, F. Jamet, F. Ruggieri, and O. Masson, "Stupid NAO and Piaget's class inclusion question: a new argument for the relevance-theoretic explanation," The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics, 19–22 septembre 2016, IEEE ICDL-EPIROB2016. Etis, Université de Cergy-Pontoise, Cergy-Pontoise: France.

[5] C. Bartneck and J. Forlizzi, "A design centred framework for social human-robot interaction," in Proc. of the 13th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN 2004), pp. 591-594, 2004.

[6] G. Bateson, Steps to an ecology of mind, New York: Ballantine, 1972.

[7] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd and D. Mulanda, "Humanoid robots as cooperative partners for people," IJHR, vol. 1, no. 2, pp. 1–34, 2004.

[8] L. J. Byom and B. Mutlu, "Theory of mind: mechanisms, methods, and new directions," Front Hum Neurosci, vol. 7, 2013.

[9] J. Elster, "Social Norms and Economic Theory," JEP, vol. 3, no. 4, pp. 99–117, 1989.

[10] F. Jamet, J. Baratgin, and P. Godin, "Don, droit, coutume, cultures. Études expérimentales sur « l'effet de dotation »" [Gift, law, custom, cultures. Experimental studies on "the endowment effect"], in C. Puigelier, C. Tijus, and F. Jouen (Eds), Droit, Décision et prise de décision [Law, decision and decision-making]. Paris, France: Mare et Martin (collection Science et Droit), in press.

[11] F. Jamet, J. Baratgin and C. K. Bearune, "Effet de dotation approche développementale chez des enfants Kanak", [Endowment effect, developmental approach in a population of kanak children]. In (S. Minvielle, Ed.), L'école calédonienne du destin commun [The Caledonian school of the common destinity], Nouméa, france: Presses Universitaires de Nouvelle-Calédonie and Centre de Documentation Pédagogique de la Nouvelle-Calédonie, in press.

[12] D. Kahneman et al., "Anomalies: the endowment effect, loss aversion, and Status Quo Bias," JEP, Vol. 5, no. 1, pp. 193-206, 1991.

[13] S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey, "Anthropomorphic interactions with a robot and robot–like agent," Social Cognition, vol. 26, no. 2, pp. 169–181, Apr. 2008.

[14] J. L. Knetsch, "The endowment effect and evidence of nonreversible indifference curves," AER, Vol. 79, no. 5, pp. 1277-1284, 1989.

[15] K. F. MacDorman and H. Ishiguro, "The uncanny advantage of using androids in cognitive and social science research," Interaction Studies, vol. 7, no. 3, pp. 297–337, 2006.

[16] O. Masson, J. Baratgin, J., and F. Jamet, "NAO robot, a social clues transmitter: what impacts? The example with endowment effect," in 30th International Conference on Industrial, Engineering, Other Applications of Applied Intelligent Systems (IEA/AIE) 2017. Université d'Artois, Arras, France, in press.

[17] O. Masson, J. Baratgin, F. Jamet, F. Ruggieri, and D. Filatova, "Use a robot to serve experimental psychology: Some examples of methods with children and adults," in Proceeding of the International Conference on Information and Digital Technologies 17, pp. 190-197, Poland: IEEE, Rzeszow, 2016.

[18] O. Masson, J. Baratgin, and F. Jamet, "NAO robot and the endowment effect," in Proceeding of the International Workshop on Advanced Robotics and its Social Impacts (pp. 1–6). Lyon, France: IEEE Robotics & Automation Society, 2015.

[19] C. Nass and Y. Moon, "Machines and mindlessness: social responses to computers," JSI, vol. 56, no 1, pp. 81‑103, janv. 2000.

[20] K. N. T. Nomura, "Influences of inconsistency between phrases and postures of robots: A psychological experiment in Japan," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4438-4443, 2010.

[21] J. Piaget and B. Inhelder, "La genèse des structures logiques élémentaires" ["The Early Growth of Logic in the Child: Classification and seriation"], Neuchâtel: Delachaux et Niestle, 1959.

[22] G. Politzer, “Reasoning, Judgement and Pragmatics,” in Experimental Pragmatics, I. A. Noveck and D. Sperber, Eds. Palgrave Macmillan UK, pp. 94–115, 2004.
[23] G. Politzer, “The class inclusion question: a case study in applying pragmatics to the experimental study of cognition,” SpringerPlus. 5(1):1133, 2016.
[24] G. Politzer, and I. Noveck, “Are conjunction rule violations the result of conversational rule violations?,” JPR, vol. 20, pp. 83–103, 1991.
[25] J. Reb and T. Connolly, “Possession, feelings of ownership and the endowment effect,” JDM, vol. 2, pp. 107-114, 2007.
[26] B. Reeves and C. Nass, “The media equation: how people treat computers, television, and new media like real people and places,” Cambridge, MA, Cambridge University Press, 1996.
[27] N. Schwarz, Cognition and Communication, Mahwah, NJ: Lawrence Erlbaum, 1996.
[28] M. Siegel et al., “Persuasive robotics: the influence of robot gender on human behavior,” in IROS 2009, IEEE/RSJ International Conf., pp. 2563-2568, 2009.
[29] D. Sperber, F. Cara, and V. Girotto, “Relevance theory explains the selection task,” Cognition, vol. 57, no. 1, pp. 31–95, Oct. 1995.
[30] D. Sperber and D. Wilson, Relevance: Communication and Cognition, Cambridge, MA, USA: Harvard University Press, 1986.
[31] D. S. Syrdal, “Exploring human mental models of robots through explicitation interviews”, 2010.
[32] R. Thaler, “Toward a positive theory of consumer Choice,” JEBO, vol.1, no. 1, pp. 39–60, 1980.
[33] L. Takayama, C. Pantofaru, “Influences on proxemic behaviors in human-robot interaction,” in Proceedings of Intelligent Robotic Systems: IROS 2009, St. Louis, MO, USA, pp. 5495-5502, 2009.
[34] G. Wilcock and K. Jokinen, “Speech, gaze and gesturing: multimodal conversational interaction with Nao robot”, in enterface’12 summer workshop, final rep. project p1, 2012.
[35] T. N. Yasser and F. O. Mohammad, “Toward combining autonomy and interactivity for social robots” AI Soc., vol. 24, no. 1, pp. 35–49, 2009.

On the Reed-Muller Spectrum of Symmetric Boolean Functions

Claudio Moraga

Faculty of Computer Science, TU Dortmund University, 44221 Dortmund, Germany
claudio.moraga@tu-dortmund.de

Dedicated to Prof. Chen Xiexiong, Hangzhou University, China

Abstract.

The paper considers a class of symmetric Boolean functions called Reed-Muller type Fundamental Symmetric Functions, it reviews some of their properties and presents some new ones. The main contribution of the paper is a proof that the Reed-Muller transform of a symmetric Boolean function is also symmetric and that of a rotation symmetric Boolean function is also rotation symmetric. Since symmetric *n*-place Boolean functions may be given a compact representation with a value vector of *n*+1 elements and this holds also for its Reed-Muller spectrum, some methods are reviewed, to calculate the Reed-Muller spectrum of a symmetric Boolean function based on its compact value vector. Furthermore a method is presented to calculate the Reed-Muller spectrum of a rotation symmetric Boolean function from the compact value vector representation of the function.

I Introduction

Symmetric Boolean functions started to be studied since the early times of Switching Theory, mostly in terms of Boolean algebra leading to AND-OR-NOT realizations. (It should be recalled that in the early times, an EXOR was an "expensive circuit", requiring two ANDs, two NOTs and one OR gates, since x EXOR $y = \bar{x}y \vee x\bar{y}$.) Hardware close interest lead also to *theoretical* results like "The number of NOT elements required for a symmetric function in n variables does not exceed $n/2$" [7], and to use symmetric functions as benchmarks for the implementation complexity of circuits. Symmetric functions are the only functions that have known implementations with a gate count which is linear in the number of arguments [33]. Methods were developed to realize arbitrary Boolean functions as symmetric functions with repeated arguments (see e.g. [1]) and to detect the symmetry of given Boolean functions [16]. On the other hand, studies on cryptographic properties of symmetric functions started at the end of the past century [27], [4]. including new kinds of symmetry, like "rotational symmetry" (also called "circular symmetry") [22], [6], [29].

Although there are only 2^{n+1} (straight) symmetric functions out of 2^{2^n} Boolean functions of n variables, it should not be surprising if symmetric functions continue to be the source of new challenging research subjects.

II Preliminaries

Definition 1: A Boolean function $f: \{0,1\}^n \rightarrow \{0,1\}$ is (totally) symmetric if it takes the same value for all value assignments to its arguments, which have the same Hamming weight. The Hamming weight of a binary structure –(finite set, n-tuple, vector)– is obtained as the arithmetic sum of the value of its elements.

Definition 2: A Boolean function $f: \{0,1\}^n \rightarrow \{0,1\}$ is rotation symmetric if it takes the same value for all value assignments to its arguments, which represent cyclic shifts of an assignment pattern.

The interest in symmetric Boolean functions may be traced back to the work of Claude Shannon [28] in 1938. Later, the study of symmetric functions in the context of developing circuits to build a computer received strong support with the work of Yasuo Komamiya [13], [26] in the 1950 decade.

A relevant contribution of Shannon [28] was to show that a symmetric n-place function may be given an unambiguous compact representation as a value vector of n+1 elements ordered by increasing Hamming weight of the value assignment to its arguments, instead of using a list or map of 2^n minterms.

Rotation symmetric Boolean functions were introduced by Pieprzyk J. and Qu C.X. in 1999 [22] in the context of hashing. Their cryptographic properties were presented later by Cusick T.W. and Stănică P. [6] and by Stănică P. and Maitzra S. [29].

It is simple to see that cyclic shifting a value assignment preserves the Hamming weight, but there may exist assignment patterns with the same Hamming weight, however emerging from different "seed patterns". This means that a set of value assignments with a common Hamming weight may be partitioned and each block of the partition will have a different seed pattern. This has three immediate consequences. The first one is that there are no "real" rotation symmetric Boolean functions for $n < 4$, since no partition of the value assignments with Hamming weight 0, 1, 2, or 3 by shifting a seed assignment pattern is possible. The

second one is that for $n > 3$, symmetric Boolean functions (Definition 1) are a subset of rotation symmetric Boolean functions (Definition 2), since blocks with the same Hamming weight may exhibit the same function value. The third one is that (real) rotational symmetric Boolean functions may also be represented by a compact value vector with a length equal to the number of blocks, which will be larger than $n + 1$ but much smaller than 2^n.

Example 1: (Part 1).

Let the symmetric function f: $\{0,1\}^4 \rightarrow \{0,1\}$ be specified as the Marquand map [15] shown in Figure 1(a) and by a compact value vector $\mathbf{F}_4$ of length 5, shown in Figure 1(b). The value assignment to the variables x_1, x_2, x_3 and x_4 is given by the vector $[v_1,v_2,v_3,v_4]$. Marquand maps order the value of the arguments of a function lexicographically and this allows a simple presentation of the symmetry of a Boolean function, as may be seen in Figure 1(a). (It will later be shown, that a Marquand map also has a representation structure that coincides with that of the vec^{-1} operation applied to the value vector of a Boolean function to generate a representation as a matrix [10].)

Karnaugh maps, most well known in Switching Theory, constitute a (partial) 2-D representation of a torus upon which a Boolean functions is projected. Karnaugh maps order the value of the arguments of a function following a Grey code. This has the desired effect that adjacent minterms have a Hamming distance of 1, and this is used to minimize Boolean expressions, not to represent symmetries.

(a)

$f(\boldsymbol{x})$		v_1v_2			
		00	01	10	11
v_3v_4	00	0	1	1	0
	01	1	0	0	0
	10	1	0	0	0
	11	0	0	0	1

(b)

	Hamming weight of $[v_1, v_2, v_3, v_4]$				
	0	1	2	3	4
$\mathbf{F}_4 =$	0	1	0	0	1

Fig. 1: A symmetric Boolean function.
(a) Specification as a Marquand map
(b) Specification as a compact value vector

Example 2 (Part 1)**:**

Let the rotation symmetric function f: $\{0,1\}^4 \rightarrow \{0,1\}$ be specified as the Marquand map [15] shown in Figure 2(a) and by a compact value vector $\mathbf{F}_4$ of length 6, shown in Figure 2(b). The value assignment to the variables x_1, x_2, x_3 and x_4 is given by the vector $[v_1,v_2,v_3,v_4]$. The entries in bold illustrate the partition of the set of value assignments to the arguments with Hamming weight 2. In the compact vector specification (b) the columns headed by 2_a and 2_b correspond to the partition of the set of value assignments with Hamming weight 2. The block 2_a is determined by cyclic shifting the value assignment [0,0,1,1] while 2_b, by cyclic shifting the value assignment [0,1,0,1]. It should be noticed that blocks do not necessarily have the same cardinality.

(a)

$f(\boldsymbol{x})$		v_1v_2			
		00	01	10	11
v_3v_4	00	0	1	1	**1**
	01	1	**0**	**1**	0
	10	1	**1**	**0**	0
	11	**1**	0	0	1

(b)

	Hamming weight of $[v_1, v_2, v_3, v_4]$					
	0	1	2_a	2_b	3	4
$\mathbf{F}_4 =$	0	1	**1**	**0**	0	1

Fig. 2: A rotation symmetric Boolean function.
(a) Specification as a Marquand map
(b) Specification as a compact value vector

III FUNDAMENTAL SYMMETRIC FUNCTIONS

Chen Xixiong [5] introduced the concept of Reed-Muller type *Fundamental Symmetric Functions* in 1994 and studied several of their basic properties. (Komamiya used similar basic symmetric functions, but without assigning a particular name to them).

Definition 3: [5] Reed-Muller type Fundamental Symmetric Functions (FSFs) are n-place functions specified in GF(2) as follows:

$R_0 = 1$

$R_1 = x_1 \oplus x_2 \oplus x_3 \oplus \ldots \oplus x_n$

$R_2 = x_1x_2 \oplus x_1x_3 \oplus \ldots \oplus x_1x_n \oplus x_2x_3 \oplus \ldots \oplus x_{n-1}x_n$

...

$R_k = x_1x_2\ldots x_k \oplus \ldots \oplus x_{n-k-1}\ldots x_{n-1}x_n$

...

$R_n = x_1x_2 \ldots x_{n-1}x_n$

(The symbol $\oplus$ denotes the sum modulo 2 or EXOR)

Notice that in the product terms, the variables appear ordered after their respective indices. Products with this structure will be called ordered products. Unless otherwise specified, in the rest of this paper, all products in formal expressions will be ordered.

If the number of arguments is not specified, or if work with different number of arguments is done, the notation $R_k^{(n)}$ will be used.

Remark:

It should be mentioned the name *Fundamental Symmetric Functions* and the notation of Def. 3 was introduced by Xiexiong Chen in 1994 [5]. Early this century, in [25], [30], the notation SB(n,k) was used. Recently, in [8], 2016, the name *Elementary Polynomial-Unate Symmetric Boolean Functions* for the FSFs and the notation E_n^k is used. The notation E_n^k had been introduced in [32], but without giving a name to the functions. In [3], [7] and [21] the name FSF was used, but with a meaning different to the one given by Chen and considered in the present work.

The main properties of Fundamental Symmetric Functions may be summarized in a series of Lemmas. Some of them will be presented without proof for space reasons. Proofs may be found in [17], available from the author upon request.

Lemma 1: Fundamental Symmetric Functions are idempotent.

Proof: In GF(2), $x^2 = x$, $(x \oplus y)^2 = x^2 \oplus y^2 = x \oplus y$.

Assume that for some u,

$$(x_1 \oplus x_2 \oplus \ldots \oplus x_u)^2 = x_1 \oplus x_2 \oplus \ldots \oplus x_u.$$

Increasing u by 1,

$$(x_1 \oplus x_2 \oplus \ldots \oplus x_u \oplus x_{u+1})^2 \equiv$$
$$\equiv (x_1 \oplus x_2 \oplus \ldots \oplus x_u)^2 \oplus (x_{u+1})^2,$$

which with the induction hypothesis and the idempotence of x_{u+1} leads to

$$(x_1 \oplus x_2 \oplus \ldots \oplus x_u \oplus x_{u+1})^2 \equiv$$
$$\equiv (x_1 \oplus x_2 \oplus \ldots \oplus x_u \oplus x_{u+1})$$

Therefore this is valid for all n, proving that

$$(R_1^{(n)})^2 = R_1^{(n)}$$

Since in GF(2) both addition and product are closed, then in $R_k^{(n)}$ every one of the $\binom{n}{k}$ product terms is an element of GF(2). Therefore $R_k^{(n)}$ is isomorph with $R_1^{\binom{n}{k}}$. Since $R_1^{\binom{n}{k}}$ is idempotent, then $R_k^{(n)}$ is also idempotent.

□

Lemma 2: For $q \geq k \geq 2$,

$$R_k^{(q+1)} = R_k^{(q)} \oplus R_{k-1}^{(q)} x_{(q+1)}$$

Proof: See [17]

Lemma 3: For all $j \in \mathbb{N}$ such that $2^j \leq n$, holds:

$$R_{2^j-1}^{(n)} = \prod_{i=1}^{j} R_{2^j-i}^{(n)} = R_{2^j-1}^{(n)} \cdot R_{2^j-2}^{(n)} \cdot \ldots \cdot R_1^{(n)}$$

This Lemma was introduced and proven by Yasuo Komamiya [13] in 1959. The Lemma has later been reformulated by different authors. (See e.g. [5], [25], [30]).

A recursive application of Lemma 3 leads to

Corollary 3.1:

For all $j \in \mathbb{N}$ such that $2^j \leq n$, holds:

$$R_{2^j-1}^{(n)} = R_{2^j-1}^{(n)} \cdot R_{2^{j-1}-1}^{(n)}.$$

Lemma 4:

For all $k \in \mathbb{N}$, $k \geq 2$, such that $2^{k+1} - 2 \leq n$, holds:

$$R_{2^k-2}^{(n)} = R_{2^{k-1}}^{(n)} \cdot R_{2^{k-1}-2}^{(n)}$$

Proof: See [17]

Lemma 5: (Variety of Corollary 3.1)

For all $j \in \mathbb{N}$, $j \geq 2$, such that $2^j < n$, (also) holds:

$$R_{2^j-1}^{(n)} = R_{2^j-1}^{(n-1)} \cdot R_{2^{j-1}-1}^{(n)}.$$

Proof:

Applying Lemma 2 to the second term,

$$R_{2^j-1}^{(n)} = R_{2^j-1}^{(n-1)} \cdot \left(R_{2^{j-1}-1}^{(n-1)} \oplus R_{2^{j-1}-2}^{(n-1)} x_n\right) =$$
$$= \left(R_{2^j-1}^{(n-1)} \cdot R_{2^{j-1}-1}^{(n-1)}\right) \oplus \left(R_{2^j-1}^{(n-1)} \cdot R_{2^{j-1}-2}^{(n-1)} x_n\right).$$

Applying Corollary 3.1 to the first product and Lemma 4 to the second, leads to:

$$R_{2^j-1}^{(n)} = R_{2^j-1}^{(n-1)} \oplus R_{2^j-2}^{(n-1)} x_n.$$

Since this expression satisfies Lemma 2, the assertion follows.

□

Lemma 6:

Let $j \in \mathbb{N}$. For all $n > 2^j$ holds:

$$R_{2^j+1}^{(n)} = R_{2^j}^{(n)} \cdot R_1^{(n)}$$

Proof:

Induction hypothesis: Assume that there exists $u > 2^j$ such that $R_{2^j+1}^{(u)} = R_{2^j}^{(u)} \cdot R_1^{(u)}$.

Induction step: increase u by 1.

$$R_{2^j+1}^{(u+1)} = R_{2^j}^{(u+1)} \cdot R_1^{(u+1)}.$$

Applying Lemma 2 and the definition of $R_1^{(u+1)}$,

$$R_{2^j+1}^{(u+1)} = \left(R_{2^j}^{(u)} \oplus R_{2^{j-1}}^{(u)} \cdot x_{u+1}\right) \cdot \left(R_1^{(u)} \oplus x_{u+1}\right)$$

$$= R_{2^j}^{(u)} \cdot R_1^{(u)} \oplus R_{2^{j-1}}^{(u)} \cdot R_1^{(u)} \cdot x_{u+1} \oplus$$

$$\oplus R_{2^j}^{(u)} \cdot x_{u+1} \oplus R_{2^{j-1}}^{(u)} \cdot x_{u+1}.$$

Applying the hypothesis and reordering:

$$R_{2^j+1}^{(u+1)} = R_{2^j+1}^{(u)} \oplus R_{2^j}^{(u)} \cdot x_{u+1} \oplus$$

$$\oplus \left(R_1^{(u)} \oplus 1\right) \cdot R_{2^{j-1}}^{(u)} \cdot x_{u+1}.$$

Expanding $R_{2^{j-1}}^{(u)}$ with Lemma 3:

$$R_{2^j+1}^{(u+1)} = R_{2^j+1}^{(u)} \oplus R_{2^j}^{(u)} \cdot x_{u+1} \oplus$$

$$\oplus R_{2^{j-1}}^{(u)} \cdot R_{2^{j-2}}^{(u)} \cdot \ldots \cdot R_1^{(u)} \cdot \left(R_1^{(u)} \oplus 1\right) \cdot x_{u+1}.$$

Notice that $R_1^{(u)} \cdot \left(R_1^{(u)} \oplus 1\right) \equiv 0 \bmod 2$. Therefore

$$R_{2^j+1}^{(u+1)} = R_{2^j+1}^{(u)} \oplus R_{2^j}^{(u)} \cdot x_{u+1}.$$

Since this expression satisfies Lemma 2, the assertion follows.

□

Lemma 7: (Variety of Lemma 6)

Let $j \in \mathbb{N}$. For all $n > (2^j + 1)$ also holds:

$$R_{2^j+1}^{(n)} = R_{2^j}^{(n-1)} \cdot R_1^{(n)}$$

Proof:

Since $R_1^{(n)} = R_1^{(n-1)} \oplus x_n$, with Lemma 6 and with Lemma 2, follows that

$$R_{2^j}^{(n-1)} \cdot R_1^{(n)} = R_{2^j}^{(n-1)} \cdot \left(R_1^{(n-1)} \oplus x_n\right) =$$

$$= R_{2^j}^{(n-1)} \cdot R_1^{(n-1)} \oplus R_{2^j}^{(n-1)} \cdot x_n =$$

$$= R_{2^j+1}^{(n-1)} \oplus R_{2^j}^{(n-1)} \cdot x_n = R_{2^j+1}^{(n)}.$$

□

Example 3:

Calculation of $R_3^{(4)}$ as $R_2^{(4)} \cdot R_1^{(4)}$ and as $R_2^{(3)} \cdot R_1^{(4)}$. See Figure 2.

Analysis:

Products of two variables appear twice in every row (grey shaded) and will cancel in the modulo 2 sum. Similarly, products of three variables appearing twice are shaded in light blue or marked with a blue diagonal. In both product tables, the remaining four product terms of three variables (in bold) are the product terms corresponding to $R_3^{(4)}$.

$R_3^{(4)}$		$R_1^{(4)}$			
		x_1	x_2	x_3	x_4
$R_2^{(4)}$	x_1x_2	x_1x_2	x_1x_2	$x_1x_2x_3$	$x_1x_2x_4$
	x_1x_3	x_1x_3	$x_1x_2x_3$	x_1x_3	$x_1x_3x_4$
	x_1x_4	x_1x_4	$x_1x_2x_4$	$x_1x_3x_4$	x_1x_4
	x_2x_3	$\boldsymbol{x_1x_2x_3}$	x_2x_3	x_2x_3	$x_2x_3x_4$
	x_2x_4	$\boldsymbol{x_1x_2x_4}$	x_2x_4	$x_2x_3x_4$	x_2x_4
	x_3x_4	$\boldsymbol{x_1x_3x_4}$	$\boldsymbol{x_2x_3x_4}$	x_3x_4	x_3x_4

$R_3^{(4)}$		$R_1^{(4)}$			
		x_1	x_2	x_3	x_4
$R_2^{(3)}$	x_1x_2	x_1x_2	x_1x_2	$x_1x_2x_3$	$\boldsymbol{x_1x_2x_4}$
	x_1x_3	x_1x_3	$x_1x_2x_3$	x_1x_3	$\boldsymbol{x_1x_3x_4}$
	x_2x_3	$\boldsymbol{x_1x_2x_3}$	x_2x_3	x_2x_3	$\boldsymbol{x_2x_3x_4}$

Fig. 2: Product tables of $R_2^{(4)} \cdot R_1^{(4)}$ and $R_2^{(3)} \cdot R_1^{(4)}$

Example 1 (Part 2).

Since the FSFs $R_1^{(4)}, R_2^{(4)}, R_3^{(4)}$ and $R_4^{(4)}$ are symmetric, they may also be given a compact specification, as shown in Figure 3. Whenever the Hamming weight of the value assignments is odd, $R_1^{(4)}$ will take the value 1, since in GF(2) the sum of an odd number of 1s equals 1. When the Hamming weight is 2 there is a single pair of 1s, and when the Hamming weight is 3, there are 3 pairs of 1s. Hence in both cases $R_2^{(4)}$ will take the value 1. When the Hamming weight is 3, $R_3^{(4)}$ will take the value 1, because there will be a single group of 3 1s and 4 groups of 3 1s when the Hamming weight is 4, however these last 4 groups add up to 0. Finally when the Hamming weight is 4, there will be a single group of 4 1s. Figure 3 shows the summary. Notice that to calculate the entry of $R_k^{(n)}$ when the Hamming weight of the value assignment to the arguments is *w*, it is needed to deduce whether the number $\binom{w}{k}$ of products that will take the value 1 is even or odd. If it is even, the entry will be 0 and if it is odd, the entry will be 1. For large values of *n*, and consequently, large Hamming weights, as suggested in [4], an application of Lucas Theorem [14] strongly simplifies the problem. Given two non-negative integers *a* and *b*, binary coded as $a_{m-1}\, a_{m-2} \ldots a_1\, a_0$ and $b_{m-1}b_{m-2} \ldots b_1b_0$, respectively, then

$$\binom{a}{b} \equiv \prod_{i=0}^{m-1} \binom{a_i}{b_i} \bmod 2$$

Moreover if for $0 \le i \le m-1, a_i \ge b_i$ then $\binom{a}{b} \equiv 1 \bmod 2$. Otherwise $\binom{a}{b} \equiv 0 \bmod 2$.

Notice that Figure 3 shows that

F = $R_1^{(4)} \oplus R_3^{(4)} \oplus R_4^{(4)}$.

Furthermore if only the FSFs $R_{2^j}^{(4)}$, for $0 \le j \le 2$ are considered, then $\sum_{j=0}^{2} 2^j R_{2^j}^{(4)}$ returns the vector of Hamming weights.

	Hamming weight of $[v_1, v_2, v_3, v_4]$				
	0	1	2	3	4
$\mathbf{F}_4$ =	0	1	0	0	1
$R_1^{(4)}$	0	1	0	1	0
$R_2^{(4)}$	0	0	1	1	0
$R_3^{(4)}$	0	0	0	1	0
$R_4^{(4)}$	0	0	0	0	1

Fig. 3: Compact representation of the symmetric function and the FSFs on the same arguments

IV Reed-Muller Transform

The mathematical fundaments of what today is known as the Reed-Muller transform may be found in the work of the Russian mathematician I.I. Zhegalkin [34], [35]. The work however remained unknown in the "non-understanding-Russian" world. Later on, Irving S. Reed [24] and David E. Muller [20], starting from work on coding theory, provided the basis for further development of the transform, which received their names. In the literature, the coefficients of the Reed-Muller transform of a Boolean function are frequently called Reed-Muller spectrum.

An important characterizing property of the Reed-Muller transform is that it constitutes a bijection in the set of Boolean functions: The Reed-Muller transform of an *n*-place Boolean Function is another –(eventually the same)– *n*-place Boolean function. This is not the case with the Walsh or the Haar transforms (see e.g. [12]). Another important characterizing property of the Reed Muller transform is its Kronecker product structure [9], starting with a basic transform (See [12]). Let $\mathfrak{R}_n$ denote a $2^n \times 2^n$ matrix representation of the Reed-Muller transform on *n* arguments. The following holds:

$$\mathfrak{R}_n = \mathfrak{R}_1 \otimes \mathfrak{R}_{n-1} = \mathfrak{R}_{n-1} \otimes \mathfrak{R}_1,$$

where the symbol $\otimes$ denotes the Kronecker product.

Symbolically, $\mathfrak{R}_1 = [\ 1 \ \ x\ ]$ and in matrix form,

$$\mathfrak{R}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Then, for example,

$$\mathfrak{R}_2 = \mathfrak{R}_1 \otimes \mathfrak{R}_1 = [\ 1 \ \ x_1\ ] \otimes [\ 1 \ \ x_2\ ] =$$

$$= [\ 1 \ \ x_2 \ \ x_1 \ \ x_1x_2\ ]$$

and in matrix form,

$$\mathfrak{R}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

With respect to the Reed-Muller spectrum of symmetric Boolean functions, the literature is not quite clear. It seems that Valery Suprun [31] proved that the Reed-Muller spectrum of a symmetric Boolean function is also symmetric, but his paper was published in Russian, thus being difficultly accessible to the non-Russian speaking community. Before addressing the next contribution it should be mentioned that people of the Cryptography community speak of the "Algebraic normal form" of a Boolean function, where people in Switching Theory speak of the Reed-Muller expansion. In [4] a Proposition is stated saying that a Boolean function of *n* variables is symmetric iff the coefficients of its algebraic normal form may be represented by a vector of length (*n*+1) (however without proof). On the other hand, a proven method is given, to calculate the compact Reed-Muller spectrum of a symmetric function, from the compact value vector of the function. Calculation of the compact Reed-Muller spectrum of a symmetric Boolean function based on a Reed-Muller sub-matrix is presented in [23], *assuming* that the Reed-Muller spectrum of a symmetric Boolean function is symmetric. In [32] the *transeunt triangle* method was introduced, to effectively calculate the compact Reed-Muller spectrum of different fixed polarities. This method was further elaborated in [2] and applied to a set of benchmarks showing its effectivity.

Lemma 8: Let $n = 2$ and let $\mathfrak{R}_2 = [r_{i,j}]$, $i, j \in \{0, 1, 2, 3\}$, where *i* and *j* are the integers corresponding to (i_1, i_2) and (j_1, j_2), which are binary, respectively. Then $\mathfrak{R}_2$ is invariant with respect to the transposition of j_1 and j_2 –(which corresponds to transposition of the arguments x_1 and x_2)– or the transposition of i_1 and i_2.

Proof: (Adapted from [19])

Recall that symbolically $\mathfrak{R}_2 = [\ 1 \ \ x_1 \ \ x_2 \ \ x_1x_2\ ]$, from where if the arguments are exchanged, the expression becomes $[\ 1 \ \ x_2 \ \ x_1 \ \ x_1x_2\ ]$. It is simple to see that this is equivalent to exchange j_1 and j_2 in the matrix representation.

Consider first $\mathfrak{R}_1 = [r_{i,j}] = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $i, j \in \{0, 1\}$. It is simple to show that $r_{ij} = 1 \oplus j \oplus ij$.

Then, for $\mathfrak{R}_2$ holds

$$\mathfrak{R}_2 = [r_{i,j}] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, i, j \in \{0, 1, 2, 3\},$$

$$r_{i,j} = (1 \oplus j_1 \oplus i_1 j_1)(1 \oplus j_2 \oplus i_2 j_2).$$

If j_1 and j_2 are transposed, let the corresponding index be denoted as (j). (Similarly, when transposing i_1 and i_2, the index (i) will be used). Then

$$r_{i(j)} = (1 \oplus j_2 \oplus i_1 j_2)(1 \oplus j_1 \oplus i_2 j_1),$$

$$r_{(i)j} = (1 \oplus j_1 \oplus i_2 j_1)(1 \oplus j_2 \oplus i_1 j_2).$$

Since the product is commutative, $r_{i(j)} = r_{(i)j}$.

□

Corollary 8.1:

Let $\mathbf{P}_2$ be a permutation matrix which when applied to $\mathfrak{R}_2$ (from the right) induces the transposition of the middle columns. Lemma 8 then proved that

$$\mathfrak{R}_2 \cdot \mathbf{P}_2 = \mathbf{P}_2 \cdot \mathfrak{R}_2$$

Lemma 9:

Consider $f: \{0,1\}^n \rightarrow \{0,1\}$ to be a Boolean function with value vector $\mathbf{F}_{n.}$. Let $\mathbf{P}_n$ be a permutation matrix, which when applied to $\mathbf{F}_n$ has the same effect as exchanging (only) the arguments x_k and x_{k+1}, ($k < n$), in f and preserving the position of all other arguments. The following holds:

$$\mathfrak{R}_n \cdot \mathbf{P}_n \cdot \mathbf{F}_{n.} = \mathbf{P}_n \cdot \mathfrak{R}_n \cdot \mathbf{F}_{n.}.$$

Proof:

$\mathbf{P}_n$ may be decomposed as $\mathbf{I}_{k-1} \otimes \mathbf{P}_2 \otimes \mathbf{I}_{n-k-1}$ to produce the transposition of the selected neighbour arguments.

Accordingly, let $\mathfrak{R}_n = \mathfrak{R}_{k-1} \otimes \mathfrak{R}_2 \otimes \mathfrak{R}_{n-k-1}$.

Then, (see [10]),

$$\mathbf{P}_n \cdot \mathfrak{R}_n = (\mathbf{I}_{k-1} \otimes \mathbf{P}_2 \otimes \mathbf{I}_{n-k-1})(\mathfrak{R}_{k-1} \otimes \mathfrak{R}_2 \otimes \mathfrak{R}_{n-k-1})$$
$$= (\mathfrak{R}_{k-1}) \otimes (\mathbf{P}_2 \cdot \mathfrak{R}_2) \otimes (\mathfrak{R}_{n-k-1})$$

and

$$\mathfrak{R}_n \cdot \mathbf{P}_n = (\mathfrak{R}_{k-1} \otimes \mathfrak{R}_2 \otimes \mathfrak{R}_{n-k-1})(\mathbf{I}_{k-1} \otimes \mathbf{P}_2 \otimes \mathbf{I}_{n-k-1})$$
$$= (\mathfrak{R}_{k-1}) \otimes (\mathfrak{R}_2 \cdot \mathbf{P}_2) \otimes (\mathfrak{R}_{n-k-1}).$$

With Corollary 8.1, the assertion follows.

□

Since any permutation and particularly, cyclic shifting the element of an n-tuple may be obtained with a sequence of transpositions and any transposition may be obtained with a cascade of elementary transpositions of neighbour elements, then:

Corollary 9.1:

Consider $f: \{0,1\}^n \rightarrow \{0,1\}$ to be a Boolean function with value vector $\mathbf{F}_{n.}$. Let $\mathbf{P}_n$ be a permutation matrix, which when applied to $\mathbf{F}_n$ has the same effect as permuting the arguments of f. The following holds

$$\mathfrak{R}_n \cdot \mathbf{P}_n \cdot \mathbf{F}_{n.} = \mathbf{P}_n \cdot \mathfrak{R}_n \cdot \mathbf{F}_{n.}.$$

Main Theorem :

The Reed-Muller spectrum of a symmetric Boolean function is symmetric.

The Reed-Muller spectrum of a rotation symmetric Boolean function is rotation symmetric.

Proof:

i) Let $f: \{0,1\}^n \rightarrow \{0,1\}$ be a symmetric Boolean function with value vector $\mathbf{F}_{n.}$.

Then, $f(x_1, x_2, \ldots, x_n) = f(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$, where π represents a permutation of the elements of $\{1, 2, \ldots, n\}$. The equality holds for all $n!$ permutations π.

Let $\mathbf{P}_n$ denote a permutation matrix, which when applied to $\mathbf{F}_n$ has the same effect as applying π to the indices of the arguments of f. Then for all such $\mathbf{P}_n$ holds that

$$\mathbf{P}_n \cdot \mathbf{F}_n = \mathbf{F}_n.$$

Since from Corollary 9.1

$$\mathfrak{R}_n \cdot \mathbf{P}_n \cdot \mathbf{F}_{n.} = \mathbf{P}_n \cdot \mathfrak{R}_n \cdot \mathbf{F}_{n.},$$

then $$\mathfrak{R}_n \cdot \mathbf{F}_{n.} = \mathbf{P}_n \cdot (\mathfrak{R}_n \cdot \mathbf{F}_n).$$

The first assertion follows.

ii) Let $f: \{0,1\}^n \rightarrow \{0,1\}$ to be a rotation symmetric Boolean function with value vector $\mathbf{F}_{n.}$.

Then, $f(x_1, x_2, \ldots, x_n) = f(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$, where π represents a cyclic shifting of the elements of $\{1, 2, \ldots, n\}$. The equality holds for all n such permutations π.

Let $\mathbf{P}_n$ denote a permutation matrix, which when applied to $\mathbf{F}_n$ has the same effect as applying π to the indices of the arguments of f. Then for all such $\mathbf{P}_n$ holds as above,

$$\mathfrak{R}_n \cdot \mathbf{F}_{n.} = \mathbf{P}_n \cdot (\mathfrak{R}_n \cdot \mathbf{F}_n).$$

The second assertion follows.

□

Corollary :

The Reed-Muller spectrum of a symmetric n-place Boolean function has a compact representation as a vector of length n+1 with entries ordered according to the Hamming weight of the value assigned to the arguments.

Following [5] any symmetric Boolean function may be expressed as

$$f(x_1, x_2, \ldots, x_n) = \oplus_{j=0}^{n} c_j \cdot R_j^{(n)}(x_1, x_2, \ldots, x_n). \quad (1)$$

For instance, the function of example 1 may be expressed as follows (recalling that $R_0^{(4)} = 1$):

$f(x_1, \ldots, x_4) = c_0 \oplus c_1 R_1^{(4)} \oplus c_2 R_2^{(4)} \oplus c_3 R_3^{(4)} \oplus c_4 R_4^{(4)}$,

which since (from Lemma 3) $R_3^{(4)} = R_2^{(4)} R_1^{(4)}$, may be rewritten as

$$c_0 \oplus c_1 R_1^{(4)} \oplus c_2 R_2^{(4)} \oplus c_3 R_2^{(4)} R_1^{(4)} \oplus c_4 R_4^{(4)}.$$

It may be seen that this expression corresponds to a Reed-Muller expansion of f in terms of the "variables" R_0, R_1, R_2 and R_4. This may well be the reason why Chen called "Reed-Muller type" his Fundamental Symmetric Functions. The coefficients c_i, $0 \le i \le 4$, correspond to the spectral coefficients of the Reed-Muller spectrum of the function when the Hamming weight of the values assigned to the arguments equals i.

Example 1 (Part 3):

The Reed-Muller spectrum of the function f with value vector $\mathbf{F}_4$, may be calculated "in the classical way", e.g. taking advantage of the space efficient Lemma 4.2.1. of [10], by using the entries of the Marquand map as $vec^{-1}\mathbf{F}_4$ to be two-sided Reed-Muller transformed.

$$\mathbf{S}_f = \mathfrak{R}_4 \cdot \mathbf{F}_4 = (\mathfrak{R}_2 \otimes \mathfrak{R}_2) \cdot \mathbf{F}_4 =$$

$$= vec\,((\mathfrak{R}_2) \cdot vec^{-1}\mathbf{F}_4 \cdot (\mathfrak{R}_2)^{\mathrm{T}}) \mod 2$$

$$= vec\left(\begin{bmatrix}1&0&0&0\\1&1&0&0\\1&0&1&0\\1&1&0&0\end{bmatrix} \cdot \begin{bmatrix}0&1&1&0\\1&0&0&0\\1&0&0&0\\0&0&0&1\end{bmatrix} \cdot \begin{bmatrix}1&1&1&1\\0&1&0&1\\0&0&1&1\\0&0&0&1\end{bmatrix}\right)$$

$$= vec\begin{bmatrix}0&1&1&0\\1&0&0&1\\1&0&0&1\\0&1&1&1\end{bmatrix},$$

where *vec* denotes the vectorizing operation that concatenates the columns of a matrix to build a column vector and vec^{-1} transforms a vector of length 2^n into a $2^{n/2} \times 2^{n/2}$ square matrix.

It is easy to see –(particularly after the Theorem)– that the Reed-Muller spectrum of f is symmetric. Therefore it has a compact vector representation with entries ordered after the Hamming weight of the values assigned to the arguments, as follows:

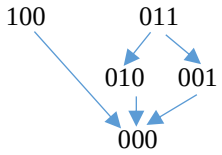
	Hamming weight of $[v_1, v_2, v_3, v_4]$				
	0	1	2	3	4
$\mathbf{S}_f =$	0	1	0	1	1

Accordingly, $c_0 = 0$, $c_1 = 1$, $c_2 = 0$, $c_3 = 1$, and $c_4 = 1$, leading to $f(x_1, \ldots, x_4) = R_1^{(4)} \oplus R_2^{(4)} \cdot R_1^{(4)} \oplus R_4^{(4)}$, as noticed at the end of Part 2 of the Example.

Using the equation of [4] and Lucas Theorem [14],

$$S_f(i) = \oplus_{k=0}^{i} \binom{i}{k} \mathrm{F}(k).$$

The required relation "$i \ge k$" for the Hamming weights may be represented in binary as follows:



Therefore:

$S_f(0) = F(0) = 0$

$S_f(1) = F(1) \oplus F(0) = 1 \oplus 0 = 1$

$S_f(2) = F(2) \oplus F(0) = 0 \oplus 0 = 0$

$S_f(3) = F(3) \oplus F(2) \oplus F(1) \oplus F(0) =$

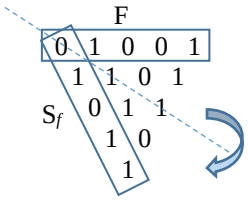
$= 0 \oplus 0 \oplus 1 \oplus 0 = 1$

$S_f(4) = F(4) \oplus F(0) = 1 \oplus 0 = 1.$

Using the sub-matrix method of [23], which is equivalent to the method of [4],

$$S_f = \begin{bmatrix}1&0&0&0&0\\1&1&0&0&0\\1&0&1&0&0\\1&1&1&1&0\\1&0&0&0&1\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\\0\\1\end{bmatrix} = \begin{bmatrix}0\\1\\0\\1\\1\end{bmatrix}.$$

The *transeunt triangle* [32], which is built starting with the compact value vector of the function and reduces this vector operating as in the Pascal triangle, but adding modulo 2 provides a simple method to obtain the compact vector of the Reed-Muller spectrum of f. Notice that, since the Reed-Muller transform matrix is self-inverse, by rotating the triangle around the middle axis –(dash line)– the inverse spectrum is obtained.



It has been shown that the calculation using the transeunt triangle has complexity $O(n^2)$ [32], but is very "transparent" when working by hand with a reasonable number of variables, as shown in the example. Furthermore, the triangle also supports the calculation of Fixed Polarity Reed-Muller spectra [32], [2].

In [8], a very efficient algorithm is presented, which uses Lucas Theorem and has complexity $O(n)$, which however does not exhibit the "transparency" of the above transeunt triangle method. Nevertheless it allows working possibly for the first time, with symmetric functions with a large number of arguments.

Recall that $\mathfrak{R}_n$ is self-inverse, and $\mathbf{S}_f = \mathfrak{R}_n \cdot \mathbf{F}_n$. Therefore, Eq. (1) may also be applied "backwards", leading to:

$$S_f(x_1, x_2, \ldots, x_n) = \bigoplus_{j=0}^{n} f(j) \cdot R_j^{(n)}(x_1, x_2, \ldots, x_n). \quad (2)$$

Since from Fig. 3 the compact vector of $\mathbf{F}_4$ is [0,1,0,0,1] the compact spectrum follows from $R_1^{(4)} \oplus R_4^{(4)}$ giving $\mathbf{S}_f$ = [0,1, 0, 1, 1].

Notice that if the FSFs are calculated *once* for several values of *n* and saved, this method of calculating the Reed-Muller spectrum of a symmetric function from the compact value vector is the simplest.

Example 2 (Part 2):

The Reed-Muller spectrum of the rotation symmetric function *f* with value vector $\mathbf{F}_4$, may be calculated taking advantage of the space efficient Lemma 4.2.1. of [10], by using the entries of the Marquand map as $vec^{-1}\mathbf{F}_4$ to be two-sided Reed-Muller transformed.

$$\mathbf{S}_f = \mathfrak{R}_4 \cdot F_4 = (\mathfrak{R}_2 \otimes \mathfrak{R}_2) \cdot \mathbf{F}_4 =$$

$$= vec\,((\mathfrak{R}_2) \cdot vec^{-1}\mathbf{F}_4 \cdot (\mathfrak{R}_2)^T) \mod 2$$

$$= vec\left(\begin{bmatrix}1&0&0&0\\1&1&0&0\\1&0&1&0\\1&1&0&0\end{bmatrix} \cdot \begin{bmatrix}0&1&1&1\\1&0&1&0\\1&1&0&0\\1&0&0&1\end{bmatrix} \cdot \begin{bmatrix}1&1&1&1\\0&1&0&1\\0&0&1&1\\0&0&0&1\end{bmatrix}\right)$$

$$= vec\begin{bmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&1\end{bmatrix},$$

It may be seen that when the Hamming weight of the value assignments is 2, just by coincidence in this example, the Reed-Muller spectrum has the same values as the function, and for the places with other Hamming weights, the symmetry conditions are preserved. It clearly follows that the Reed-Muller spectrum of *f* is indeed rotation symmetric.

This leads to the following compact specification:

	Hamming weight of $[v_1, v_2, v_3, v_4]$					
	0	1	2_a	2_b	3	4
$\mathbf{S}_f =$	0	1	1	0	1	1

Notice that the block "2_a" comprises the places where the value assignments are [0,0,1,1], [0,1,1,0], [1,1,0,0], and [1,0,0,1], while the block "2_b", the places [0,1,0,1] and [1,0,1,0]. If $R_2^{(4)}$ is partitioned accordingly, one obtains the following "extended" FSFs:

$$R_{2a}^{(4)} = x_1x_2 \oplus x_2x_3 \oplus x_3x_4 \oplus x_1x_4$$

$$R_{2b}^{(4)} = x_1x_3 \oplus x_2x_4$$

In analogy to the case of symmetric functions, from $\mathbf{S}_f$ the following coefficients are obtained: $c_0 = 0$, $c_1 = 1$, $c_{2a} = 1$, $c_{2b} = 0$, $c_3 = 1$, and $c_4 = 1$.

In Figure 4, the representation of the function as a compact value vector, the representation of the Spectrum as a compact vector and the corresponding representation of the FSFs is shown. It may be seen that as in the case of the symmetric functions, the following holds:

$$\mathbf{F}_4 = R_1 \oplus R_{2a} \oplus R_3 \oplus R_4.$$

	Hamming weight of $[v_1, v_2, v_3, v_4]$					
	0	1	2_a	2_b	3	4
$\mathbf{F}_4 =$	0	1	1	0	0	1
$R_1^{(4)}$	0	1	0	0	1	0
$R_{2a}^{(4)}$	0	0	1	0	0	0
$R_{2b}^{(4)}$	0	0	0	1	1	0
$R_3^{(4)}$	0	0	0	0	1	0
$R_4^{(4)}$	0	0	0	0	0	1
$\mathbf{S}_f =$	0	1	1	0	1	1

Fig. 4: Compact representation of the rotation symmetric function and the (extended) FSFs on the same arguments

From Fig. 4 may be seen that Eq. (2) may be extended to rotation symmetric functions if the blocks are preserved:

$$\mathbf{S}_f = \mathbf{F}_4(1)R_1^{(4)} \oplus \mathbf{F}_4(2a)R_{2a}^{(4)} \oplus \mathbf{F}_4(4)R_4^{(4)} =$$

$$= R_1^{(4)} \oplus R_{2a}^{(4)} \oplus R_4^{(4)}$$

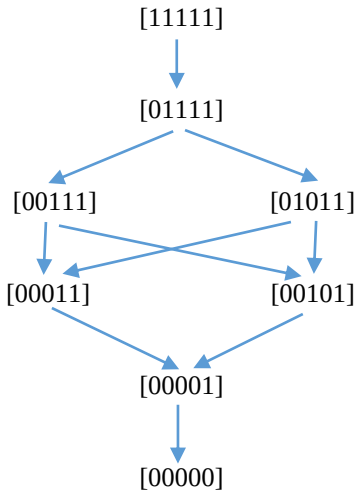
Example 3: Let the structural space for a 5-place rotation symmetric function be shown in Fig 5, where the cells –("blocks")- indicate the Hamming weight of the value assignments and, if a partition is possible, the different blocks with the letters a and b.

The following blocks are considered:
1 : {00001, 00010, 00100, 01000, 10000}
2a: {00011, 00110, 01100, 11000, 10001}
2b: {00101, 01010, 10100, 01001, 10010}
3a: {00111, 01110, 11100, 11001, 10011}
3b: {01011, 10110, 01101, 11010, 10101}
4 : {01111, 11110, 11101, 11011, 10111}
0 : {00000} ; 5: {11111}

		$x_1x_2x_3$							
		000	001	010	011	100	101	110	111
x_4 x_5	00	0	1	1	2a	1	2b	2a	3a
	01	1	2b	2b	3b	2a	3b	3a	4
	10	1	2a	2b	3a	2b	3b	3b	4
	11	2a	3a	3b	4	3a	4	4	5

Fig. 5: Blocks for a 5-place rotation symmetric Boolean function.

Let the first element of each block written with square brakkets represent the whole block. Then the following Hasse diagram summarizes the space structure. The levels correspond to the Hamming weight of the value assignments. At the levels 2 and 3, the blocks of the corresponding partitions are shown. (The fact that [00101] appears following [01011] should not lead to confusion: notice that 10101 is an element of [00101]. Then 00101 follows from 10101. The correspondence for the other rotations follows similarly.)



Let a rotation symmetric function f: $\{0,1\}^5 \rightarrow \{0,1\}$ be specified as the Marquand map [15] and as a compact (block-oriented) vector in Figure 6.

As shown in Example 2 Part 2, the Reed-Muller spectrum of f may be calculated as:

$$\mathbf{S}_f = \mathfrak{R}_5 \cdot F_5 = (\mathfrak{R}_3 \otimes \mathfrak{R}_2) \cdot \mathbf{F}_5 =$$

$$= vec\,((\mathfrak{R}_2) \cdot vec^{-1}\mathbf{F}_5 \cdot (\mathfrak{R}_3)^T) \mod 2$$

The result of the calculation is shown in Fig. 7

$f(x)$		$x_1x_2x_3$							
		000	001	010	011	100	101	110	111
x_4 x_5	00	0	1	1	0	1	1	0	0
	01	1	1	1	1	0	1	0	0
	10	1	0	1	0	1	1	1	0
	11	0	0	1	0	0	0	0	1
Block		0	1	2a	2b	3a	3b	4	5
$\mathbf{F}_5$		0	1	0	1	0	1	0	1

Fig. 6: Full and compact specification of a 5-place rotation symmetric Boolean function.

S_f		$x_1x_2x_3$							
		000	001	010	011	100	101	110	111
x_4 x_5	00	0	1	1	0	1	1	0	0
	01	1	1	1	0	0	0	0	1
	10	1	0	1	0	1	0	0	1
	11	0	0	0	1	0	1	1	0

Block	0	1	2a	2b	3a	3b	4	5
$\mathbf{S}_f$	0	1	0	1	0	0	1	0

Fig. 7: Full and compact specification of S_f, the Reed-Muller spectrum of f.

In order to relate the function and the spectrum to the FSFs, they have first to be specified and their vector representation must be calculated.

$$R_1^{(5)} = x_1 \oplus x_2 \oplus x_3 \oplus x_4 \oplus x_5$$

$$R_{2a}^{(5)} = x_1x_2 \oplus x_2x_3 \oplus x_3x_4 \oplus x_4x_5 \oplus x_1x_5$$

$$R_{2b}^{(5)} = x_1x_3 \oplus x_2x_4 \oplus x_3x_5 \oplus x_2x_5 \oplus x_1x_4$$

$$R_{3a}^{(5)} = x_1x_2x_3 \oplus x_2x_3x_4 \oplus x_3x_4x_5 \oplus x_1x_4x_5 \oplus \oplus x_1x_2x_5$$

$$R_{3b}^{(5)} = x_1x_3x_4 \oplus x_2x_4x_5 \oplus x_2x_3x_5 \oplus x_1x_3x_5 \oplus \oplus x_1x_2x_4$$

$$R_4^{(5)} = x_1x_2x_3x_4 \oplus x_1x_2x_3x_5 \oplus x_1x_3x_4x_5 \oplus \oplus x_2x_3x_4x_5 \oplus x_1x_2x_4x_5$$

$$R_5^{(5)} = x_1x_2x_3x_4x_5$$

Fig. 8 shows the complete relationship among the compact value vector of the function, the FSFs per block and the Reed-Muller spectrum of the function.

It may be seen that Eq. (2) holds showing the interrelation between the compact representation of the value vector of the function and its Reed-Muller spectrum.

Block	0	1	2a	2b	3a	3b	4	5
$\mathbf{F}_5$	0	1	0	1	0	1	0	1
$R_1^{(5)}$	0	1	0	0	1	1	0	1
$R_{2a}^{(5)}$	0	0	1	0	0	1	1	1
$R_{2b}^{(5)}$	0	0	0	1	1	0	1	1
$R_{3a}^{(5)}$	0	0	0	0	1	0	0	1
$R_{3b}^{(5)}$	0	0	0	0	0	1	0	1
$R_4^{(5)}$	0	0	0	0	0	0	1	1
$R_5^{(5)}$	0	0	0	0	0	0	0	1
S_f	0	1	0	1	0	0	1	0

Fig. 8: Relationship between the compact representation of the functions, the FSFs, and the compact representation of the Reed-Muller spectrum of the function.

$$\mathbf{S}_f = \mathbf{F}_5(1)R_1^{(5)} \oplus \mathbf{F}_5(2b)R_{2b}^{(5)} \oplus \mathbf{F}_5(3b)R_{3b}^{(5)} \oplus \oplus \mathbf{F}_5(5)R_5^{(5)}$$

$$= R_1^{(5)} \oplus R_{2b}^{(5)} \oplus R_{3b}^{(5)} \oplus R_5^{(5)}$$

$$= [0\ 1\ 0\ 0\ 1\ 1\ 0\ 1] \oplus$$
$$\oplus [0\ 0\ 0\ 1\ 1\ 0\ 1\ 1] \oplus$$
$$\oplus [0\ 0\ 0\ 0\ 0\ 1\ 0\ 1] \oplus$$
$$\oplus [0\ 0\ 0\ 0\ 0\ 0\ 0\ 1]$$

$$= [0\ 1\ 0\ 1\ 0\ 0\ 1\ 0] = \mathbf{S}_f$$

V Other symmetries

In Chapter 5 of [11] the concepts of "Equivalence Symmetry" and "Non-equivalence Symmetry" were introduced and characterized in the spectral domain. They require that particular pairs of subsets of neighbour minterms –(Hamming distance 1)– have the same value. A neighbourhood requirement is clearly not compatible with a common Hamming weight. Similarly, in [18] the concepts of "Co-Symmetries" were introduced, which are close related to those of [11]. These symmetries do not satisfy Definition 1. Therefore, the main theorem of the present paper does not apply for these special symmetries.

VI Conclusions

A proof is given to the Proposition that the Reed-Muller spectrum of a (rotation) symmetric Boolean function is also (rotation) symmetric. Reed-Muller type Fundamental Symmetric Boolean functions are presented and some of their properties reviewed and adapted to blocks. The relationship between Fundamental Symmetric Boolean functions and the Reed-Muller spectrum of symmetric and rotation symmetric Boolean functions is shown. Methods are recalled to work with compact representations of symmetric Boolean functions and their respective (compact) Reed-Muller spectra. A method is presented to calculate the Reed-Muller spectrum of a (rotation) symmetric Boolean function from its compact value vector representation and to recover a (rotation) symmetric Boolean functions from the compact representation of its Reed-Muller spectrum, as weighted sums of Fundamental Symmetric Functions.

References

[1] Born R., Scidmore A.K.: Transformation of switching functions to completely symmetric switching functions. *IEEE Trans. Comput.* C-**17**, 596-599, 1968

[2] Butler J.T., Dueck G.W., Shmerko V., Yanuskevich S.: Comments on "Sympathy": Fast exact minimization of fixed polarity Reed-Muller Expansion for Symmetric functions. *IEEE Trans. CAD*, **19** (11), 1386-1388, 2000

[3] Butler J.T., Schueller K. A.: Worst Case Number of Terms In Symmetric Multiple-valued Functions. In: *Proc. 21st. Int. Symposium on Multiple-valued Logic.* IEEE Press, 1991.

[4] Canteaut A., Videau M.: Symmetric Boolean functions. *IEEE Trans. Information Theory*, **51** (8), 2791-2811, 2005

[5] Chen X.: Research of symmetric functions based on the AND – Exclusive OR algebraic system. *Journal of the Hangzhou University*, China, **21** (3), 291-297, 1994

[6] Cusick T.W., Stănică P.: Fast evaluation weights and nonlinearity of rotation-symmetric functions *Discrete Math*., 258 (1–3), 289–301, 2002

[7] Epstein G.: Synthesis of Electronic Circuits for Symmetric Functions. *IRE Transactions on Electronic Computers*, EC-**7** (1), 57-60. DOI: 10.1109/TEC.1958.5222097, 1958

[8] Gorodecky D. A.: A novel approach of polynomial expansion of symmetric functions. In: *Problems and New Solutions in the Boolean Domain*. B. Steinbach (Ed.). Newcastle upon Tyne, UK. Cambridge Scholars Publishing, 96-115. 2016

[9] Graham, A.*: Kronecker products and matrix calculus with applications.* Ellis Horwood Ltd., Chichester UK, 1981.

[10] Horn, R.A., Johnson, Ch.R.: *Topics in matrix analysis.* Cambridge University Press, New York, 1991

[11] Hurst S., Miller D.M., Muzio J.: *Spectral Techniques in Digital Logic*. Academic Press, New York, 1985

[12] Karpovsky, M.G., Stanković, R.S., Astola, J.T.: *Spectral Logic and its Applications for the Design of Digital Devices.* John Wiley, N.J. 2008

[13] Komamiya Y.: Theory of computing networks. Research Report of the Applied Mathematics Section, Electrotechnic Laboratory of the Japanese Government, July 10, 1959

[14] Lucas E.: Théorie des fonctions numériques simplement périodiques. (French). *Amer. J. Math.* **1** (2), 184–196, 1878

[15] Marquand A.: On logical diagrams for *n* terms, *Philosophical Magazine*, vol. **12**, pp. 266-270, 1881.

[16] McCluskey Jr., E.J.: Detection of group invariance or total symmetry of Boolean functions. *Bell Sys. Tech. J.*, **35**, 1445-1453, 1956

[17] Moraga C.: Fundamental Symmetric Boolean Functions Revisited. *Research Report* 876, Faculty of Computer Science, TU Dortmund University, 2017

[18] Moraga C.: Spectral Analysis of Co-Symmetries. *Proceedings 3rd. International Workshop on Spectral Techniques*, 127-139, T.U. Dortmund. ISSN 0933-6192, 1988

[19] Moraga C.: Permutations under Spectral Transforms. In *Proc. 38th Int. Symposium on* Multiple-valued Logic, 76-81, IEEE Press, 2008

[20] Muller D.E.: Application of Boolean algebra to switching circuit design and to error correction. *IRE Trans. on Elec. Computers* EC-**3** (3), 6-12, 1954

[21] Muzio J.C.: Concerning the maximum size of the terms in the realization of symmetric functions. In *Proc. 20th Int. Symposium on Multiple-valued Logic*, 292-299, 1990

[22] Pieprzyk J., Qu C.X.: Fast hashing and rotation-symmetric functions *J. Universal Comput. Sci*., 5 (1), 20–31, 1999

[23] Pogossova E., Egiazarian K.: Reed-Muller representation of symmetric functions. *J. Multiple-valued Logic and Soft Computing*, **10** (1), 51-72, 2004

[24] Reed I.S.: A class of multiple-error-correcting codes and the decoding scheme. *IRE Trans. on Information Theory* PGIT-**4,** 38-49, 1954

[25] Sasao T.: A new expansion of symmetric functions and their application to non-disjoint functional decompositions for LUT type FPGAs, In *Proc. IEEE Inter. Workshop on Logic Synthesis, IWLS-2000*, IEEE Press, 2000

[26] Sasao T., Stanković R.S., Astola J.T. (Eds.): Contributions of Yasuo Komamiya to Switching Theory. In *Reprints from the Early Days of Information Sciences.* TICSP Series # 65, Tampere International Center for Signal Processing, Tampere. ISBN 978-952-15-3453, 2015

[27] Savicky P.: On the bent Boolean functions that are symmetric. *Europ. J. Combin.* **15**, 407-410, 1994

[28] Shannon C.E.: A svmbolic analysis of relay and switching circuits. *Transactions of the AIEE*, 57, 713-723, 1938

[29] Stănică P., Maitzra S.: Rotation symmetric Boolean functions—Count and cryptographic properties. *Discrete Applied Math*., 156 (10), 1567-1580, 2008

[30] Stanković R.S., Astola J.T., Egiazarian K.: Remarks on symmetric binary and multiple-valued functions. In: *Proceedings of the 6th International Workshop Boolean Problems*. B. Steinbach (Ed.). 83-87. Press TU Bergakademie Freiberg, ISBN: 3-86012-233-9. 2004

[31] Suprun V.P.: Polynomial expression of symmetrical Boolean functions. *Izvestija AN USSR. Techn. Kibernetika*, **4**, 123-127, in Russian, 1985

[32] Suprun V.P.: Fixed Polarity Reed-Muller expressions of symmetric Boolean functions. In *Proc. IFIP WG 10.5 Workshop on Applications of the Reed-Muller Expansions in Circuit Design*, 246-249, 1995

[33] Wegener I.: *The Complexity of Boolean Functions.* Wiley, New York, 1987

[34] Zhegalkin I.I.: O tekhnyke vychyslenyi predlozhenyi v symbolytscheskoi logykye, *Math. Sb.*, Vol. **34**, 9-28, in Russian, 1927

[35] Zhegalkin I.I.: Aritmetizatiya symbolytscheskoi logyky, *Math. Sb.*, Vol. **35**, 311-377, in Russian, 1928

Detection of dynamical properties of flow in an eye vessels by video sequences analysis

A.Nedzved
Belorussian State University,
United Institute of Informatics Problems
Minsk, Belarus
NedzvedA@tut.by

O.Nedzved, A.Glinsky, G.Karapetian
Belorussian State Medical University
Minsk, Belarus
Olga_Nedzved@tut.by

I.Gurevich, V.Yashina
Federal Research Center of the RAS "Computer Science and Control"
Moscow, Russia
igourevi@ccas.ru, werayashina@gmail.com

Abstract—**In this paper the method of dynamical properties of flow in some eye vessels is described. It is based on algorithm of segmentation on base CNN, morphological processing and optical flow. It allows to define new properties of changing blood flow in vessels that is depends from structure of vessels net.**

Keywords— vessel segmentation; dynamical properties; optical flow; CNN; morphological properties

I. INTRODUCTION

The rate of blood flow is one of the most important parameters, which characterizes the functionality of the circulatory system of the body. This indicator depends on the frequency of contractions of the heart muscle, the amount and quality of blood, the geometrical characteristics of blood vessels, blood pressure, human age, genetic characteristics and etc.

This indicator plays a huge role not only in diagnosis of heart and vascular diseases, but also in monitoring of the state of the whole body. Today there are many portable devices (fitness bracelets, holders) for monitoring of blood flow. However, the interpretation of the results is not always adequate. This is due to the fact that most methods of analysis do not take into account the features of the changes in processes that occur in the nodes and complex fragments of the vasculature. Methods of analyzing and processing images allow to trace such changes.

The analysis of images of vascular systems is referred to the task of pattern recognition and has a long history. This is the process of detecting the set of images and their interrelation for vessels description, which is associated with a number of operations or steps. Typically, these are methods using the processing of low-quality images, which allow to move to more complex analysis. Digital images of vessels are processed with a sequence of algorithms that can consist of many pre-processing steps preceding the stages of image segmentation, characterization and classification. Pre-processing can be used to equalize the brightness of the image, correct for irregularities, suppress noise or eliminate distortion. Segmentation breaks the image into fragments of blood vessels. In the process of calculating characteristics, numerical information on individual objects obtained by segmentation is evaluated. The calculated characteristics can be used to classify objects according to predetermined criteria, such as size, structure or color.

Today, there are many methods for image analysis, for which several reviews are done. In 2003, Sharp in his report for the NHS (Health Technology Assessment) published an overview of digital imaging technologies in the field of diabetic retinopathy, which was completed in 1998. According to the authors, their initial goal was to conduct a numerical analysis of various technologies for working with digital diagnostic images. They failed to do this, since digital technologies were only at an early stage of their development in this field [1]. The first reviews of algorithms for the definition of vascular structures on medical images can be seen in [2, 3]. Kirbas and Quek [4] presented a comparative overview of methods and algorithms for isolating vessels and elongated objects on both two-dimensional and three-dimensional medical images used in various tasks. A review of the algorithms for segmentation and registration of the retina is presented by Mabrouk et al. [5], which limited the discussion to the tasks of isolating the boundaries and central lines of the vessels. In recent study [6] algorithms for the automatic diagnosis of diabetic retinopathy in retinal images are discussed. The review [1] is unique as it presented an analysis and categorization of literature related to digital imaging technologies in the field of diabetic retinopathy published between 1998 and 2008 and focused on algorithms and methods of segmentation on two-dimensional color images of the retina received with the help of fundus cameras. This review is still the best known for diabetic retinopathy. The article focuses only on diabetic retinopathy and includes an analysis of such signs of diabetic retinopathy as microaneurysms, small spot hemorrhages, blood spots, lipid exudates and cotton-like spots

that are essentially micro-nerve fibers. In [8] it provides an overview of algorithms primarily focused on the isolation of vessels on two-dimensional, color images of the retina obtained with fundus cameras or fluorescent angiography from 1995 to 2010, and the focus is only on studies related to segmentation of blood vessels of retina.

The analysis of the reviews made it possible to group the articles according to methods and determine existing trends, problems in the field of digital vascular processing and future directions. In most cases, existing methods have some drawbacks and many of the articles are aimed at developing an improved approach based on already existing ones. It is very difficult for researchers to determine the optimal algorithm for use at each stage in order to ensure the most efficient processing of vessel images. This point of view is confirmed in his paper by Abramoff et al. [9]. He claims that at present the use of an automated system for the detection of vascular pathology based on existing algorithms is not recommended for clinical use.

Thus, a large number of tasks in the field of analysis of vascular systems have been solved as particle solution. The task of determining the optimal set of algorithms for creating an automated system for monitoring and diagnosing vascular pathologies remains urgent.

Most methods are focused on static images. There are very few methods these were developed for determining dynamic properties. And almost all of them are aimed at the general flow and do not take into account the morphology of the vasculature.

In this paper, we propose a new method that define dynamical properties of flow in every points of vessel net.

II. PROPERTIES OF VIDEOSEQUENCE OF VESSEL NET

Today, the most convenient object for research is the flow in the blood vessels of the fundus. The position of the vessels in this area of eye allows to avoid many distortions. Therefore we used video sequence of fundus for the debugging our algorithms. Nevertheless, we have experience working with images of the vessels themselves of different organs. It can confidently state that researchers have the same set of troubles in the observation of dynamic properties of any vessels.

The first trouble appears at the stage of obtaining of a video sequence. This is instability of frame position in video. The acquisition is performed in real time for live objects that have in uncontrolled motion. It is not possible to provide a solution to this problem by fixing the object. The shifting of position between frames is random value and it can change in big diapason at any time. It is inherent property of micro-scale video of live objects. Therefore, at this stage the most important task is the stabilization of the video sequence.

The second serious trouble is definition of pattern of vessel net. There are many methods and algorithms for vessel segmentation. But vessels are very complicated objects with complex geometry. Usually a vessel can be studied in three-dimensional space. As a result, fragments of their two-dimensional images can lose sharpness and brightness. Therefore, the problem of segmentation is very important at this stage.

The third trouble is related to determining of the instantaneous changes in the vessel. Some of these changes happen instantly. In the vessel the blood flow can slow down or accelerate, the blood concentration may change in any time. In addition, the vessel is an elastic structure, as result geometric characteristics can change. Therefore, it is very important to determine the instantaneous characteristics of the vessel, including the flow velocity in the vessel.

A large number of changes in the vessel leads to the fourth trouble associated with constructing an adequate description of events in the vessel.

III. VIDEO STABILISATION AND PREPROCESSING

Today there are many algorithms for stabilization of the video sequence. Most of them are designed to work with the usual dynamic scene of the macro image. For detection of changes they use the image features and based on algorithms of SIFT and SURF. This allows to change images through projective transformation and combine features of two images. These algorithms are built into hardware and software packages. Researchers try to use them in many cases; unfortunately, these methods have a good result only in short time segments in the video sequence data. Therefore, a simple stabilization of the video stream in the vessels is much better. It corresponds to the vessel properties. The features of the images of the vasculature can be repeated on all frames, and in this case the geometric binding of the image fails. The simple algorithm is based on image correlation. The shifting coordinates of frame are calculated from maximum at correlation image. This algorithm is showed on Fig. 1.

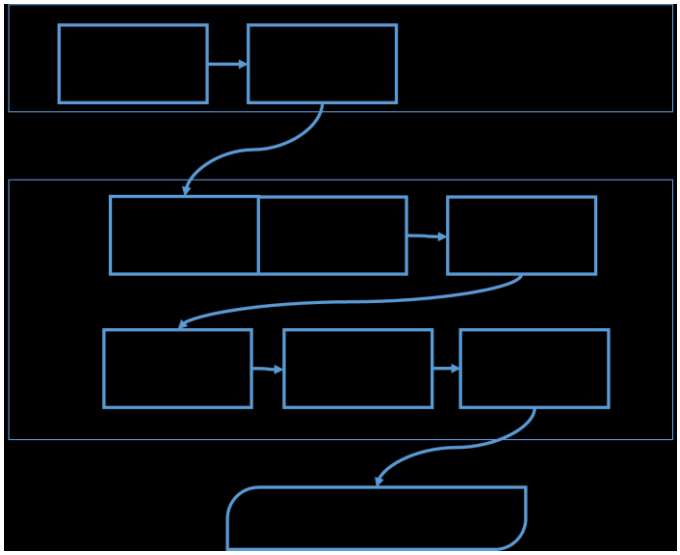


Figure 1. Sheme of stabilisation of video sequences of vessel net

The processing of the first frame is very important. The sharpness of the vessels is not constant across the image in the frame. Therefore, it is necessary to define a sharp fragment for its subsequent search in next frames. This fragment with sharp vessels is determined on base of the maximum values with help of the first derivative of the image. It corresponds to any raster gradient filter, for example, Sobel. The detection of such maximum is started from center of image, because next frame

can be shift to any direction. We define region of interest (ROI) as sharpest fragment of the image near from center. A new template image is cut out from this ROI.

This template image is used as kernel of correlation for definition of shifting coordinate. They are used for stability of video.

IV. SEGMENTATIN OF VESSELS

The segmentation of vessel is a very complex process. It was described above. For improving the quality of segmentation, the pattern of the vessel arrangement is determined. It is realized by image accumulation that is represented on Fig. 2.

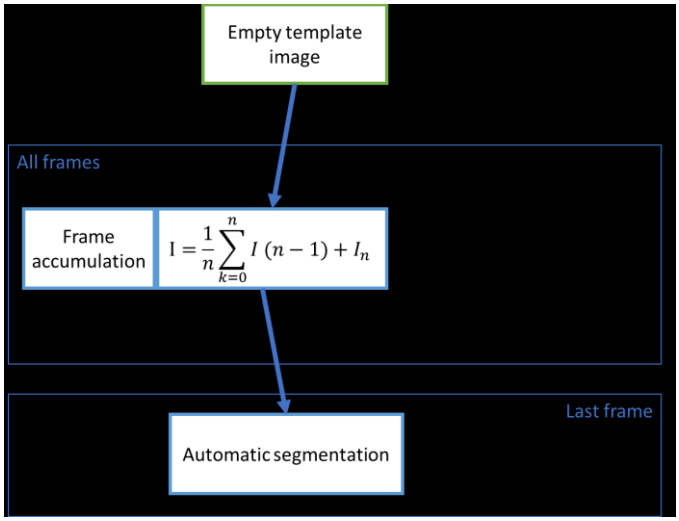


Figure 2. Sheme of vessel sharping at the image

This accumulated image is used to enhance the image of the vessel before the segmentation. This is realized by averaging this image and the current frame.

The solution of the problem of segmentation of the vascular net is performed as a classification problem. The classifier is implemented as a method of learning the convolutional neuron network (CNN) with a sliding window. It allows to predict the label of a class for each pixel, based on the selection of the pattern around it as in article [10]. A small area around the pixel is used as a source data (Fig. 3).

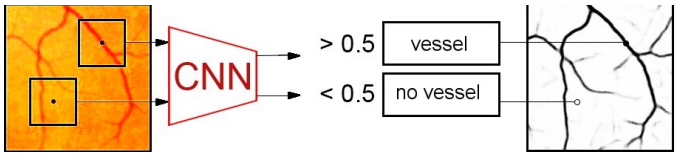


Figure 3. Example image for learning

This approach has a number of important advantages:

– Binding a local area with a pixel allow to the network increase ability of localization of this area;

– The proposed method of formation of the initial data dramatically increases the amount of data. Every pattern represents, in fact, a training image;

– The orientation of each pattern, its dimensions, brightness and color characteristics can be arbitrarily changed, that increases the amount of training images for CNN training.

The network architecture is shown in Fig. 4. It is constructed from two practically symmetrical branches: tapering (left side) and expanding (right side). The converging part corresponds to the typical architecture of the SNS. It consists of two consecutive 3x3 bundles (without indentation), followed by a ReLU layer and a 2x2 max pooling operation in 2 steps to reduce the dimension. After each reduction in dimension, the number of features doubles. Each step of the expanding branch consists of the development layers and a set of attributes. Layers consist of 2x2 convolution that increase the resolution. Attributes are formed on the base of these branches, which reduces the number of features by half. Then concatenates are gone with the corresponding set of signs from the convergent part. Two convolutions 3x3 are processed after ReLU.

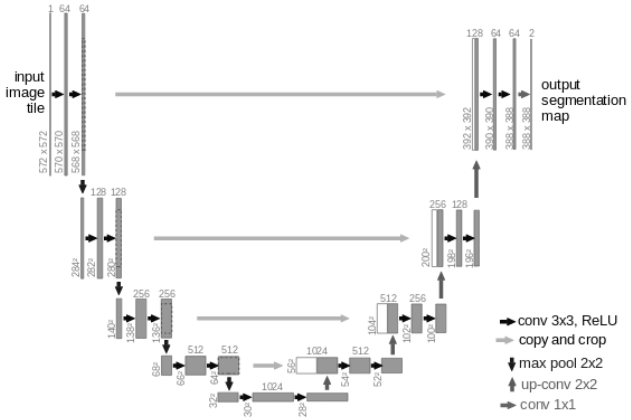


Figure 4. U-net architecture of CNN

Cutting of edges is caused by the loss of border pixels after each convolution. On the last layer, convolutions of 1x1 size are used to bring each 64-component vector to the required number of classes. As a result, the CNN has 23 convolutional layers.

For seamless coverage of the output markup image, it is necessary to select the size of the input image. All 2x2 max pooling operations are applied to the layer with even height and width values.

The training of the network for the segmentation of ophthalmic images was carried out on a publicly accessible DRIVE basis, which allowed to compare with the classical architecture (Table 1) and demonstrate the advantage of using the U-net network.

The quality of segmentation is shown in Fig. 5. Examples are taken of a picture of the fundus of a healthy person and an image that obtained by the laparoscopic method.

TABLE I. RESULTS OF COMPARISON U-NET WITH CLASSICAL ARCHITECTURE

Method	AUC ROC on DRIVE
Melinscak et al. [7]	0.9466
Our segmentation	.9712

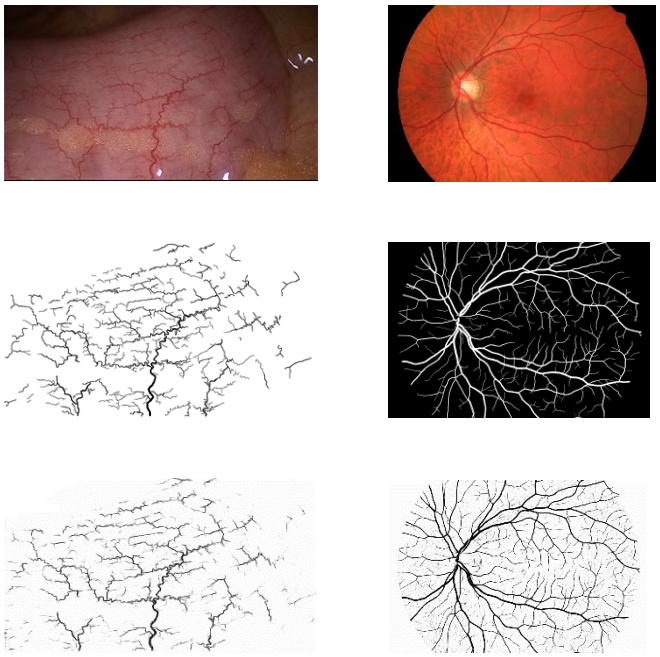


Figure 5. Results of vessels segmentation by CNN

The architecture of U-net demonstrates high quality in solving problem of segmentation of the vascular network on ophthalmic and laparoscopic images. The addition of data with elastic deformations and the use of ZCA whitening transforms allowed to abstract from the brightness and geometric distortions in the original images. This method allows to segment the vessels with high qualitatively.

This method has long training time (about three days). It is compensated by the high velocity of segmentation by the trained network. For a full segmentation of the FullHD image resolution (1920 x 1080) on the NVIDIA GTX 950 GPU, it takes less than 10 seconds. We believe that this architecture and the solutions are the most optimal for implementation in medical institutions.

V. DETECTION OF DINAMICAL PROPERTIES OF VESSEL

A. *Velocity of flow in point of vessel*

The instantaneous velocity at the point of the vessel can be detected by an optical flow. The optical flow corresponds to the apparent movement of objects. The optical flow contains information on the direction and velocity of movement of visual characteristics [11]. It can be determined as the change at position on the image. The optical flow is usually represented as vector map.

The use of optical flow throughout the image does not allow to determine the rate of blood flow in the vessel, because in this case changes of brightness around the vessel are taken into account. To solve this problem, the thinning of the vessel image is performed before calculation of the optical flow.

For vessel analysis, the optical flow mast be determined only for vessel regions. There are a few types of blood flow velocity in these regions. Their determination is very complex process due to low hemodynamics in such vessels. We use simplification and calculate optical flow only for middle line of vessel.

This line is extracted by thinning operation at binary result of vessel segmentation (Fig. 6). Then, the conjunction of the resulting skeleton is performed for each frame of the video sequence. The resulting video sequence contains only changes that happen in the vessel.

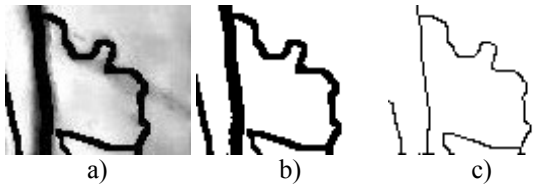


Figure 6. Vessel prepocessing: a) source image, b) result of segmentation, c) vessel skeleton after thinning.

The optical flow calculate into skeleton of vessel. We use block algorithm to find the dense optical flow that computes the optical flow for all the pixels at the frames. This realisation is based on Gunner Farneback's algorithm [12]. As result a 2-channel array with optical flow vectors (u,v) is calculated. These vectors are translated to magnitude and direction by polar transformation. At a new image a magnitude corresponds to intensisty and a direction corresponds to hue of color. The obteined color code is used visualization. Only magnitude is used for velocity definition. It is possible to trace skeleton line and to construct profile intensity for it (Fig. 7).

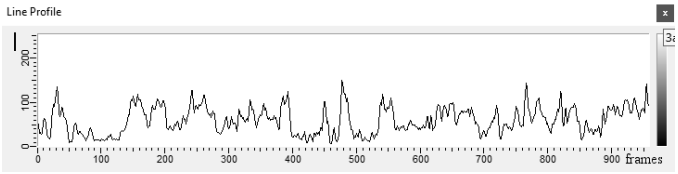


Figure 7. Intensity profile for skeleton line that shows changing of velocity for blood flow

This profile of skeleton line represents the changing of instantaneous velocity of blood flow for any point of vessel.

B. *Width of vessel for any point*

The calculation of width is a complex procedure of analysis of vessels in net. It is connected with difficulties of definition of shape. There are not algorithms of quality detection of such distribution. The algorithm of calculation of width distribution based on distance map (Fig. 8). The ride analysis of these maps allow to determine the width distribution and characteristics of its changing.

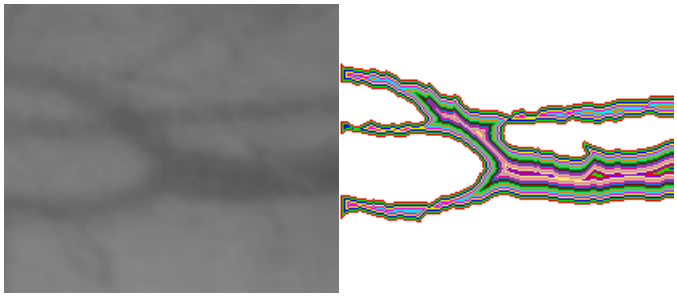


Figure 8. Distance map construction: a) source image, b) distance map

The determination of width is not fully clear for every point of vessel. The width of vessel at the point is dependent from morphological and topological shape complexity of objects. We construct a distribution of widths of vessel that corresponds to collection of sizes for every pixels of medial line of vessel determined on the distance map.

The width distribution corresponds to distribution of intensity points at skeleton of distance map without cross-section and ending regions.

The gray thinning of distance map allows to extract line that have intensity with width value in every point (Fig. 9). It is complex characteristics but it includes important practical information. In this case, it is possible to describe interactions between different fragments of vessel net. Usually these interactions are cause of shape creations or condition of existing such structures and have influence to blood flow properties.

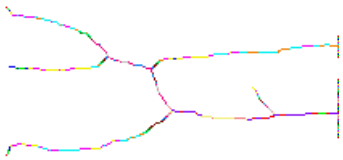


Figure 9. Skeleton with width value at every point.

Therefore, this characteristic can be used for investigations of processes in human organism.

VI. Determination of Instantaneous Changes

The determination of instantaneous changes that happen in the vessel is based on characteristics like velocity of blood flow and width of vessel at every point.

Unfortunately, the problem of discretization of time and space do not allow to get real values. Therefore, we used the amplitude of the optical flow to determine the instantaneous velocity. In this case, the velocity was determined in relative units.

The flow rate usually depends from the width of the vessel. This dependence gives possibility to determine different properties of blood vessels and quality of blood flow. Therefore, for each point of the vessel, the volume velocity is determined as relation of magnitude of optical flow (V) to vessel width (d):

$$v = \frac{\mathrm{V}}{\mathrm{d}}$$

Definition of velocity as relation of optical flow and vessel width can be going in parallel branch of calculation process. As result, the pipeline of such analysis is described by scheme on Fig. 10.

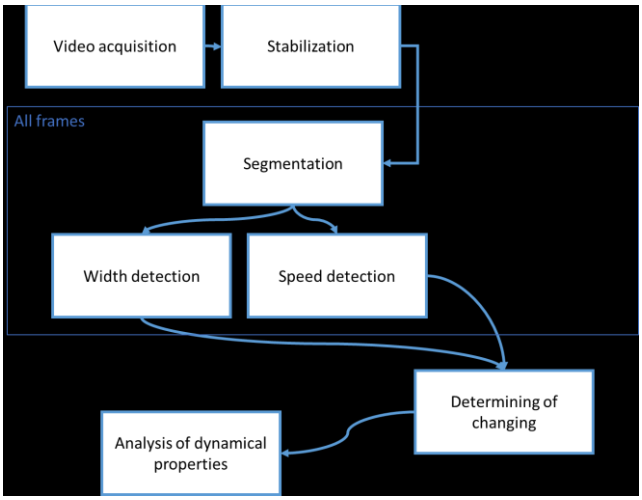


Figure 10. The pipeline of blood flow analysis in vessels

VII. Application of Method of Dinanical Propeties Analysis at Vessels

Testing of this method was carried out on the video sequence of the blood vessels of the bulbar conjunctiva of the eye. The change in the rate of blood flow in them reflects the change in blood flow in the microcirculatory bed. According to the flow rate characteristics in the vessels of bulbar conjunctiva, one can investigate the state of blood flow in other tissues and organs or vascular pathologies. The research was carried out using a high-resolution monochrome digital video camera Imperx Bobcat IGV-B1410M with a microscope objective having a focal length of 40 mm.

As a result of the test, the initial value of the velocity for a vessel with a diameter of 1.91 microns will be 0.50379 relative units, that correspond to $5 * 10^{-8}$ m/s. When using a mixture with a high content of carbon dioxide (5%), the velocity has value about 0.260568 relative units, and using a carbogen (30% CO_2 and 70% O_2), the velocity equal 0.783431 relative units. Thus, the blood flow velocity will vary by 51% and 55%, respectively.

Result of this research correspond real values that were obtained by Doppler methods.

VIII. Conclusion

The proposed method is designed to study the specific features of the vascular network. It is based on the determination of instantaneous and volume velocity at each point of the vessel.

The developed method allows to spend a quantitative estimation of diameter, cross-sectional area, linear and volume velocity in vessels of healthy people under different conditions.

This allows to perform an assessment of the processes that occur in the vessels and perform a prognosis of the human health condition.

ACKNOWLEDGMENT

The research was supported by Belorussian Foundation for Basic Research (project no. F16R-180) and Russian Foundation for Basic Research (project no. 16-57-00231)

REFERENCES

[1] R.J. Winder, P.J. Morrow, I.N. McRitchie, J.R. Bailie, P.M. Hart "Algorithms for digital image processing in diabetic retinopathy" Computerized Medical Imaging and Graphics. – 2009. – Vol. 33(8). – pp. 608-622.

[2] P. Felkel, R. Wegenkittl, A. Kanitsar "Vessel tracking in peripheral CTA datasets – an overview" Computer Graphics (Spring Conference on), 2001. pp. 232-239.

[3] K. Buhler, P. Felkel, A.L. Cruz "Geometric methods for vessel visualization and quantification – a Survay" Geometric Modelling for Scientific Visualization. 2003. pp. 399-421.

[4] C. Kirbas, F. Quek "A review of vessel extraction techniques and algorithms" ACM Computing. Vol. 36(2). 2004. pp. 81-121.

[5] M.S. Mabrouk, N.H. Solouma, Y.M. Kadah "Survey of retinal image segmentation and registration" International Journal on Graphics, Vision and Image Processing GVIP Journal. Vol. 6(2). 2006. pp. 1-11.

[6] O.R.A.U. Faust, E.Y.K. Ng, K.-H. Ng, J.S. Suri "Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review" Journal of Medical Systems. Vol. 36(1). 2012. pp. 145-57

[7] / M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara,

[8] A.R. Rudnicka, C.G. Owen, S.A. Barman "Blood vessel segmentation methodologies in retinal images" Comput Methods Programs Biomed. Vol. 108(1). 2012. pp.407-433.

[9] M.D. Abramoff, M. Niemeijer, M.S. Suttorp-Schulten, M.A. Viergever, S.R. Russell, B. van Ginneken "Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes" Diabetes Care. Vol. 31(2). 2008.– pp.193-198.

[10] Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural net-works segment neuronal membranes in electron microscopy images. In: NIPS. pp.2852–2860 (2012)

[11] J.L. Barron; D.J. Fleet, S. Beauchemin "Performance of optical flow techniques" International Journal of Computer Vision. Springer. 12, 1994. pp.43–77

[12] G.Farnebäck "Two-Frame Motion Estimation Based on Polynomial Expansion" Proceedings of the 13th Scandinavian Conference on Image Analysis, 2003. pp. {363--370

Performance analysis of data security algorithms used in the railway traffic control systems

Waldemar Nowakowski
Faculty of Transport and Electrical Engineering
K. Pulaski University of Technology and Humanities
26-600 Radom, Poland
w.nowakowski@uthrad.pl

Piotr Bojarczak
Faculty of Transport and Electrical Engineering
K. Pulaski University of Technology and Humanities
26-600 Radom, Poland
p.bojarczak@uthrad.pl

Zbigniew Łukasik
Faculty of Transport and Electrical Engineering
K. Pulaski University of Technology and Humanities
26-600 Radom, Poland
z.lukasik@uthrad.pl

***Abstract*—First of all, the railway traffic control process should ensure the safety. One of the current research areas is to ensure the security of data in the distributed rail traffic control systems using wireless networks. Emerging security threats are the result of, among others, an unknown number of users who may want to access the network, and an unknown number and type of equipment that can be connected to the network. It can cause potential threats resulting from unknown format of data and hacker attacks. In order to counteract these threats, it is necessary to apply safety functions. These functions include the use of data integrity code and encryption methods. Additionally, due to character of railway traffic control systems, it is necessary to keep time determinism while sending telegrams. Exceeding the maximum execution time of a cryptographic algorithm and creating too large blocks of data constitute two critical factors that should be taken into account while developing the system for data transmission. This could result in the inability to transmit data at a given throughput of the transmission channel (bandwidth) at a certain time. The paper presents analysis of delays resulting from the realization of safety functions: such as to prepare the data for transfer and their later decoding. Following block encryption algorithms have been analyzed: Blowfish, Twofish, DES, 3DES, AES-128, AES-192 and AES-256 for modes: ECB, CBC, PCBC, CFB, OFB, CTR and data integrity codes: MD-5, SHA-1, SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/224 and SHA-512/256. The obtained results can be very helpful in the development of new rail traffic control systems in which wireless data transmission is planned.**

Keywords— railway; wireless; cryptography; security; performance

I. INTRODUCTION

New technologies for data transmission are frequently used in modern railway traffic control systems [1, 2]. One of the current research areas is to ensure the security of data in the distributed rail traffic control systems using wireless networks [3, 4]. The main security threat is the failure by the recipient in obtaining valid and authentic telegram. This condition can be caused by repeating telegram (repetition), delete telegram (deletion), the creation of a telegram by an unauthorized sender (insertion), changing the order of telegrams (resequence), damage telegram (corruption), the delay in receiving the telegram (delay), masquerade (masquerade) [5, 6, 7]. In such case, according to the recommendations of CENELEC (European Committee for Electrotechnical Standardization), it is necessary to apply methods based on security functions. These methods include, among the others, use of data integrity codes and data encryption. At the same time, modern railway traffic control systems can be treated as a distributed real-time systems (DRTS) [8]. In most cases we are dealing with scattering data, which must take into account the time constraints (timing constraints). Since the exchange of information in such systems is made via the data communications network, it is important to keep time determinism for sending telegrams [9, 10]. Exceeding the maximum execution time of a cryptographic algorithm and creating too large blocks of data constitute two critical factors that should be taken into account while developing the system for data transmission [11, 12]. This could result in the inability to transmit data at a given throughput of the transmission channel (bandwidth) at a certain time. It is therefore necessary

to analyze the delays arising from the implementation of safety functions, such as: to prepare the data for transfer and their later decoding [13, 14, 15, 16].

II. A HIERARCHICAL STRUCTURE OF RAILWAY TRAFFIC CONTROL SYSTEMS

Railway traffic control systems serve for a safe running of the railway traffic. Computerization and the use of modern information and communication technologies influence not only an improvement of functionality of these systems, but also their structure. Modern railway traffic control systems are distributed systems in which one can distinguish three layers. The biggest layer is the executive layer, which consists of:

- electronic interlocking systems,
- level-crossing protection systems,
- automatic block signalling system,
- track-vehicle transmission systems,
- additional systems (e.g. communication, failure detection).

Next layer is created by a group of systems serving as a remote control. These systems allow for the remote control of many objects of the executive layer. The last layer is the traffic management layer. An example system of this layer is the European Rail Traffic Management System (ERTMS), which includes the European Train Control System (ETCS) and the GSM-Railways (GSM-R). The ETCS allows for cab signalling and a continuous control over the engine driver's work. The GSM-R, on the other hand, is a railway version of the GSM system, which enables users to transmit voice and data and to send short text messages. Realization of the remote control and the traffic management function is connected to the exchange of information between the systems of the executive layer and the systems of higher layers. Because the traffic management systems belong to the safety-related systems, it is obligatory to ensure data transmission safety. Transmission systems of these systems should be treated as unreliable. That is why it is recommended to use safety ensuring methods defined as security functions, including methods ensuring data integrity and confidentiality.

III. METHODS ENSURING DATA INTEGRITY AND CONFIDENTIALITY

Data protection in computer systems and networks is an area of expertise handling methods ensuring [3, 9]:

- data integrity – protecting data against a deliberate or an accidental modification through the usage of integrity codes,
- data confidentiality – protecting data against illegal access and modification through encryption.

Data integrity can be ensured through e.g. the usage of hash functions.

A cryptographic hash function assigns to an arbitrarily large message M a short message, which usually has a constant size $h = H(M)$ (the so called hash message - a checksum). The hash functions have the following properties:

- for given M, it is easy to calculate h,
- for given h, it is hard to calculate $H(M) = h$,
- for given M, it is hard to find other message M' such that $H(M') = h$.

The hash functions are public – they do not hide any secrets. Security of the hash functions consists in their one-way direction. In no visible way does the output data string depend on the input one. A change of the value of any input bit causes a change of approximately half of the bits of the output string. For a given hash value it is practically impossible to find the input string which, in result, can give the same hash value. The simplest checksum calculation method is the Cyclic Redundancy Check. For example, the CRC32 means a 32 bits hash. Another algorithm is the MD-5 (Message-Digest algorithm 5), which for data of any length creates a 128-bits hash. Another example is the family of SHA algorithms (Secure Hash Algorithms), including the SHA-1, which generates 160-bits hash and six algorithms of the type SHA-2, which generate respectively 224, 256, 384 or 512 bits hashes: SHA-224, SHA-256, SHA-384, SHA-512 [17].

Data confidentiality can be achieved with the help of data encryption. Encryption are methods of enciphering data in the process of transforming them with mathematical methods; cryptography, on the other hand, is a field investigating encryption. A cryptographic algorithm, a cipher, is a mathematical function which, with the help of a key, serves for encoding and decoding data. One can distinguish two groups of cryptographic algorithms: symmetric algorithms and public key algorithms. One of the symmetric algorithms are block ciphers. The next part of the article includes the following block encryption algorithms: Blowfish, Twofish, DES, 3DES and AES. The Blowfish cipher is a block cipher created by Bruce Schneier in 1993; it was created as a fast and freeware alternative for existing, in that times, algorithms. It operates on 64-bits input data block written as a plaintext, which is processed in 16 rounds to ciphertext by using a key from 32 to 448 bits. Bruce Schneier is the co-author of the next algorithm, called Twofish. In this algorithm one deals with 16 rounds operating on blocks of 128 bits by using keys from 128 to 256 bits (usually 128, 192, and 256). Another group of algorithms are the DES and the 3DES (Data Encryption Standard). The DES is a symmetric algorithm which encrypts consecutive 64-bits blocks using a 56-bits key in 16 rounds. In the years 1976-2001, the DES was a Federal Standard in the USA, and from 1981 it has been an ANSI standard for the private sector. Due to a short key length and a failure in providing adequate security, the 3DES algorithm was introduced. This algorithm was created by repeating the DES algorithm 3 times, but using 3 different keys which, as a result, increased the encryption time. In 2000, the AES algorithm (Advanced Encryption Standard), also known as Rijndael, became a new standard. The AES algorithm operates on 128-bits blocks, using keys of 128, 192, and 256-bits. A round number is variable and depends on the key length: 10 (for 128-bits key), 12 (for 192-bits), and 14 (for 256-bits key) [17].

IV. SIMULATION SOFTWARE

In order to verify methods to counter threats, that can be used in the wireless exchange of data by the railway traffic control systems, specialized simulation software in the form of two modules (Client and Server) has been developed. The software includes selected methods of defence against threats, comprising block encryptions: Blowfish, Twofish, DES, 3DES, AES-128, AES-192 and AES-256 modes: ECB, CBC, PCBC, CFB, OFB and CTR and integrity codes: MD-5, SHA-1, SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/224 and SHA-512/256 [3, 9, 18, 19]. Because the DES and Blowfish encryption algorithms, as well as the ECB mode and the MD5 and SHA1 hash functions are not fully secure, their usage in the research will serve for comparison with newer cryptographic algorithms. The software allows us to create telegrams type B0, such as: an encrypted message and data integrity code [20, 21]. While developing this software, authors used cryptographic library (native-code open source cryptographic library) *"TurboPower Lockbox 3"*. The research was carried out on a computer with an Intel Core i3 2.4 GHz under Windows 8 operating system. Measuring the execution time of each function was carried out using a high-resolution timer, the applications were prioritized in real-time.

A. Client Module

The "Client" is responsible for generating, encrypting and sending messages to the module "Server". User can choose the algorithm, encryption mode and methods of determining the data integrity code. The size of the created data block is defined by the user. Additionally, the "Client" module can also measure the time that is needed to encrypt telegram and generate data integrity code. This allows for the assessment of the influence of the message size on the delay introduced by the encryption process and data integrity code. An example of the main screen of the "Client" is shown in Figure 1.

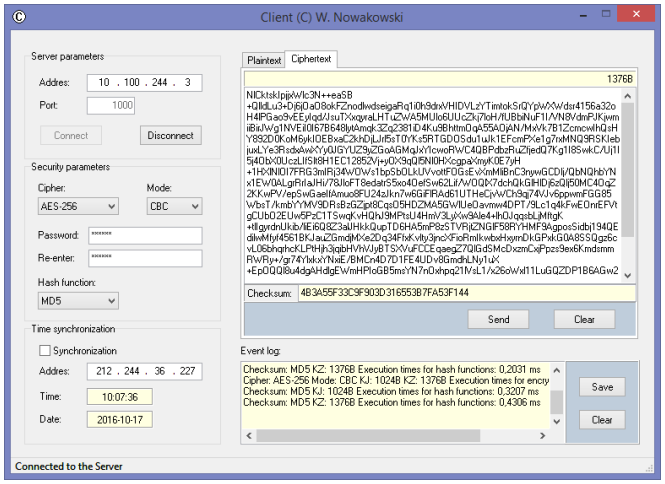


Fig. 1. The main screen of the "Client"

B. ServerModule

The "Server" is responsible for receiving and decrypting messages sent by the module "Client". User can choose the algorithm and encryption mode and methods of determining the data integrity code as well as the size of generating message. Another functionality of the "Server" is a time measurement of decryption and data integrity code generation. This allows the user to evaluate the influence of the size of the message on the delay introduced by the process of decryption and data integrity codes. Figure 2 presents an example of the main screen of the "Server"

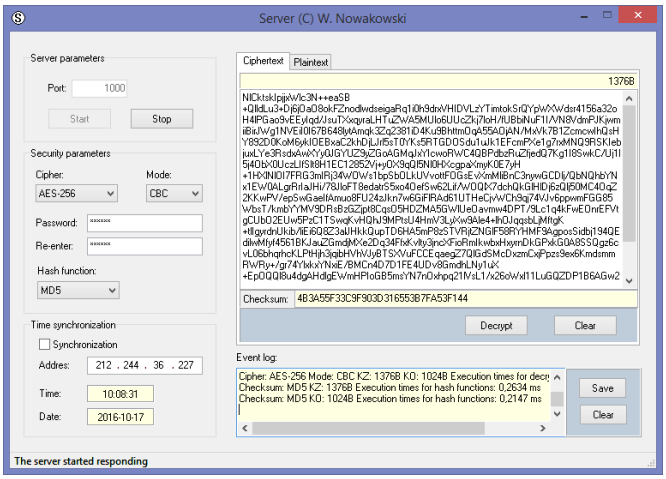


Fig. 2. The main screen of the "Server"

V. ANALYSIS OF CRYPTOGRAPHIC HASH FUNCTIONS

The test was made for the selected hash functions, such as: MD-5, SHA-1, SHA-256, SHA-512, taking into account the different size of the data blocks: 1024, 2048, 4096, 8192, 16384, 32768 bytes [3, 9]. The results of the delay introduced by the hash functions are presented in Table 1 and graphically in Figure 3.

TABLE I. EXECUTION TIMES FOR CRYPTOGRAPHIC HASH FUNCTIONS AND DATA BLOCK SIZES

	1024 (byte)	**2048** (byte)	**4096** (byte)
MD5	0.27 ms	0.32 ms	0.41 ms
SHA-1	0.22 ms	0.29 ms	0.36 ms
SHA-256	0.25 ms	0.32 ms	0.42 ms
SHA-512	0.34 ms	0.46 ms	0.66 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
MD5	0.61 ms	0.96 ms	1.65 ms
SHA-1	0.50 ms	0.79 ms	1.31 ms
SHA-256	0.63 ms	1.03 ms	1.81 ms
SHA-512	1.06 ms	1.91 ms	3.47 ms

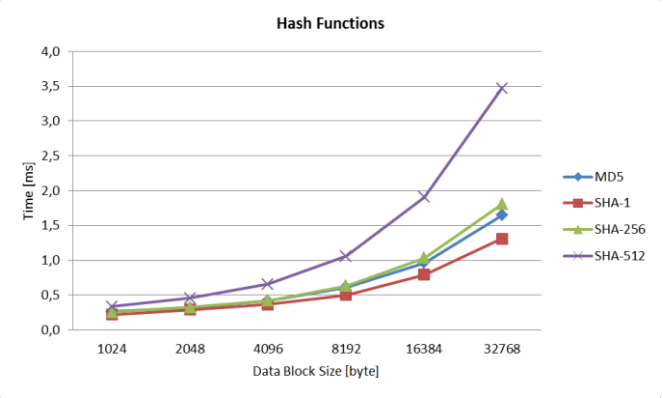


Fig. 3. Comparison performance of cryptographic hash functions

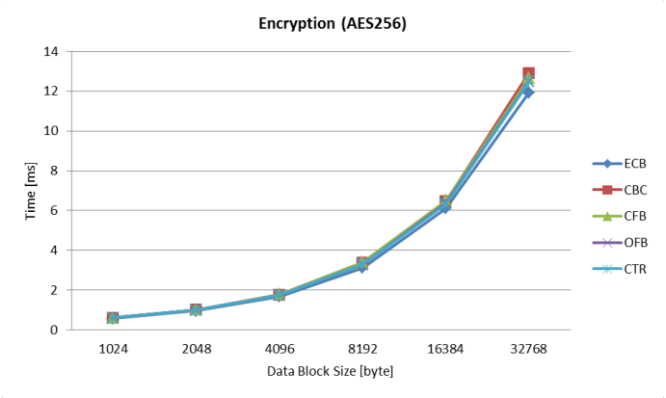


Fig. 4. Comparison performance of AES256 encryption

The largest delays occur for the SHA-512 algorithm. Other methods also generate similar delay.

VI. Analysis of Data Encryption Algorithms

Following block encryption algorithms have been analyzed: Blowfish, Twofish, DES, AES128, AES192, AES256, taking into account the different encryption modes: ECB, CBC, CFB, OFB and CTR [3, 9]. The test shows that the choice of encryption mode does not affect the amount of delay. Therefore, safer modes such as CBC or CFB should be used instead of ECB. Test results for the encryption and decryption with AES256 algorithm for the various modes are shown in Table 2 and 3 and in Figure 4 and 5.

TABLE II. Execution times for AES256 encryption and data block sizes

	1024 (byte)	**2048** (byte)	**4096** (byte)
ECB	0.58 ms	0.96 ms	1.68 ms
CBC	0.61 ms	1.01 ms	1.77 ms
CFB	0.60 ms	1.01 ms	1.79 ms
OFB	0.61 ms	1.00 ms	1.76 ms
CTR	0.61 ms	1.00 ms	1.75 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
ECB	3.12 ms	6.10 ms	11.93 ms
CBC	3.40 ms	6.46 ms	12.91 ms
CFB	3.41 ms	6.49 ms	12.66 ms
OFB	3.28 ms	6.39 ms	12.49 ms
CTR	3.31 ms	6.30 ms	12.48 ms

TABLE III. Execution times for AES256 decryption and data block sizes

	1024 (byte)	**2048** (byte)	**4096** (byte)
ECB	0.60 ms	1.03 ms	1.75 ms
CBC	0.64 ms	1.06 ms	1.87 ms
CFB	0.63 ms	1.06 ms	1.86 ms
OFB	0.64 ms	1.05 ms	1.85 ms
CTR	0.63 ms	1.06 ms	1.85 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
ECB	3.28 ms	6.37 ms	12.89 ms
CBC	3.50 ms	6.80 ms	13.44 ms
CFB	3.55 ms	6.82 ms	13.57 ms
OFB	3.48 ms	6.75 ms	13.16 ms
CTR	3.45 ms	6.67 ms	13.15 ms

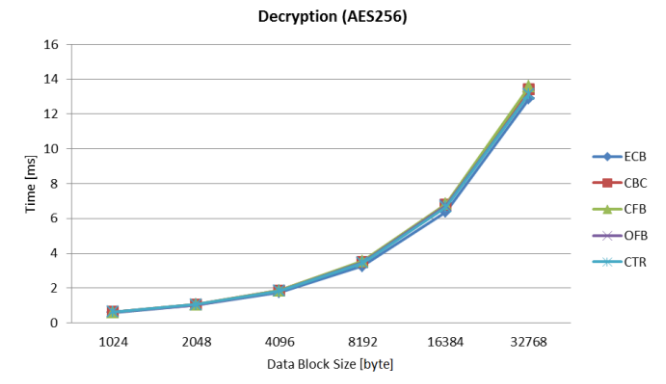


Fig. 5. Comparison performance of AES256 decryption

Since, as already mentioned, the choice of encryption mode does not affect significantly the amount of delay, and therefore further study (Tables 4, 5, 6, 7 and Figures 6, 7, 8, 9) was

carried out for two selected modes of encryption such as: CBC and CFB and various block encryption algorithms.

TABLE IV. EXECUTION TIMES FOR ENCRYPTION IN CBC MODE AND DATA BLOCK SIZES

	1024 (byte)	**2048** (byte)	**4096** (byte)
Blowfish	0.49 ms	0.61 ms	0.79 ms
Twofish	0.29 ms	0.38 ms	0.51 ms
DES	0.63 ms	1.04 ms	1.88 ms
AES128	0.49 ms	0.79 ms	1.37 ms
AES192	0.55 ms	0.91 ms	1.58 ms
AES256	0.61 ms	1.01 ms	1.77 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
Blowfish	1.17 ms	1.89 ms	3.41 ms
Twofish	0.77 ms	1.27 ms	2.36 ms
DES	3.46 ms	6.66 ms	13.21 ms
AES128	2.53 ms	4.89 ms	9.35 ms
AES192	2.92 ms	5.66 ms	11.05 ms
AES256	3.40 ms	6.46 ms	12.91 ms

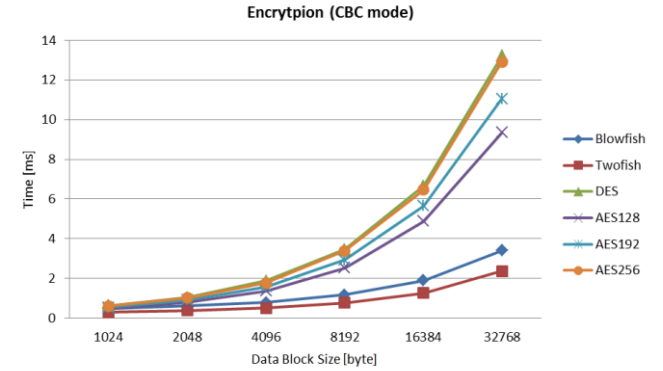


Fig. 6. Comparison performance of encryption in CBC mode

TABLE V. EXECUTION TIMES FOR AES256 DECRYPTION AND DATA BLOCK SIZES

	1024 (byte)	**2048** (byte)	**4096** (byte)
Blowfish	0.54 ms	0.72 ms	0.95 ms
Twofish	0.33 ms	0.45 ms	0.64 ms
DES	0.67 ms	1.14 ms	2.03 ms
AES128	0.54 ms	0.85 ms	1.46 ms
AES192	0.59 ms	0.96 ms	1.66 ms
AES256	0.64 ms	1.06 ms	1.87 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
Blowfish	1.47 ms	2.53 ms	4.68 ms
Twofish	1.05 ms	1.84 ms	3.43 ms
DES	3.91 ms	7.52 ms	14.71 ms
AES128	2.68 ms	5.18 ms	10.36 ms
AES192	3.10 ms	6.14 ms	11.82 ms
AES256	3.50 ms	6.80 ms	13.44 ms

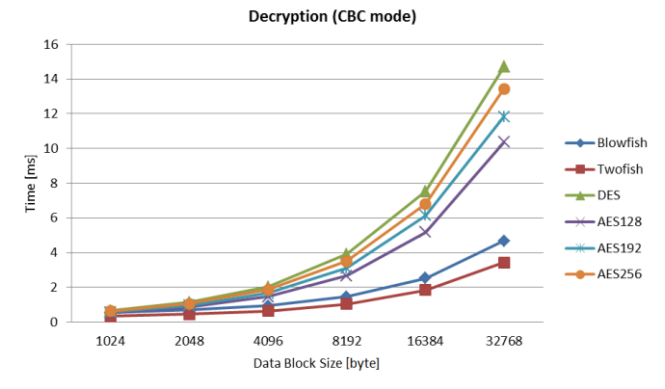


Fig. 7. Comparison performance of decryption in CBC mode

TABLE VI. EXECUTION TIMES FOR ENCRYPTION IN CFB MODE AND DATA BLOCK SIZES

	1024 (byte)	**2048** (byte)	**4096** (byte)
Blowfish	0.50 ms	0.60 ms	0.79 ms
Twofish	0.29 ms	0.37 ms	0.50 ms
DES	0.62 ms	1.04 ms	1.83 ms
AES128	0.50 ms	0.80 ms	1.36 ms
AES192	0.55 ms	0.91 ms	1.57 ms
AES256	0.60 ms	1.01 ms	1.79 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
Blowfish	1.15 ms	1.91 ms	3.45 ms
Twofish	0.76 ms	1.28 ms	2.33 ms
DES	3.48 ms	6.69 ms	13.13 ms
AES128	2.54 ms	4.84 ms	9.39 ms
AES192	2.95 ms	5.62 ms	11.35 ms
AES256	3.41 ms	6.49 ms	12.66 ms

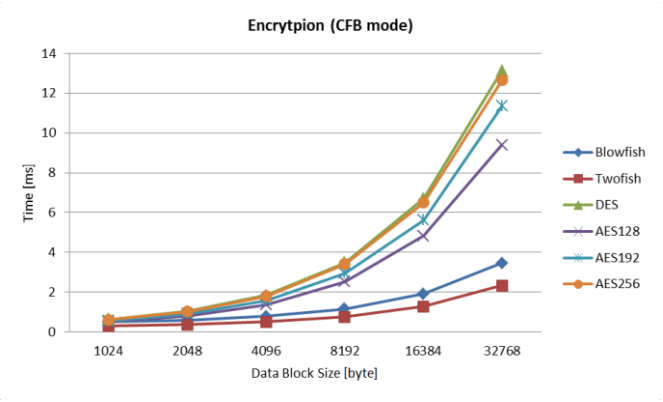


Fig. 8. Comparison performance of encryption in CFB mode

TABLE VII. EXECUTION TIMES FOR AES256 DECRYPTION AND DATA BLOCK SIZES

	1024 (byte)	**2048** (byte)	**4096** (byte)
Blowfish	0.52 ms	0.66 ms	0.89 ms
Twofish	0.32 ms	0.43 ms	0.60 ms
DES	0.65 ms	1.09 ms	1.94 ms
AES128	0.52 ms	0.84 ms	1.44 ms
AES192	0.58 ms	0.95 ms	1.66 ms
AES256	0.63 ms	1.06 ms	1.86 ms

	8192 (byte)	**16384** (byte)	**32768** (byte)
Blowfish	1.37 ms	2.34 ms	4.25 ms
Twofish	0.97 ms	1.70 ms	3.14 ms
DES	3.69 ms	7.14 ms	13.97 ms
AES128	2.65 ms	5.14 ms	10.19 ms
AES192	3.16 ms	6.00 ms	11.71 ms
AES256	3.55 ms	6.82 ms	13.57 ms

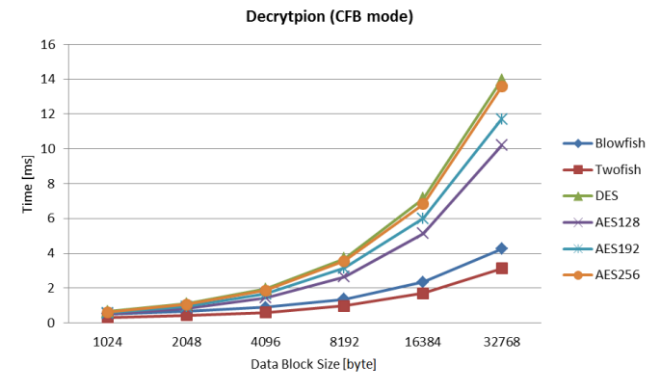


Fig. 9. Comparison performance of decryption in CFB mode

In the case of block encryption algorithms, the smallest delay are for algorithms: Blowfish, Twofish, which use 128-bit keys. The longest encryption and decryption time occurs for AES-256 and DES algorithms. Due to the level of safety, AES algorithms using 256-bit key should be used instead of DES algorithm using 56-bit key.

VII. CONCLUSIONS

Wireless transmission systems used to control railway traffic, in terms of safety should be treated as untrusted systems. Emerging security threats are the result of, among others, an unknown number of users who may want to access the network, and an unknown number and type of equipment that can be connected to the network. This creates a potential threat to the safety of rail traffic control systems, mainly from the possible emergence of data with unknown formats, as well as unknown quantities, as well as the possibility of network attacks from unauthorized users. In order to ensure safety, according to the recommendation presented in the standards, it is necessary to use cryptographic methods. At the same time, modern railway traffic control systems are distributed real-time systems.

Therefore during design process, the response time to changes in the controlled objects should be taken into account. It is the sum of response times for individual traffic control devices. We say then that the system keep time determinism, which takes into account the response times of individual components of a distributed system. It becomes essential in such cases the choice of methods to ensure the required speed of data exchange while maintaining their integrity and confidentiality. The study allowed for the experimental estimate of the time required for determining the data integrity and data encryption, taking into account the different cryptographic algorithms. The obtained results can be very helpful in the development of new rail traffic control systems in which wireless data transmission is planned.

ACKNOWLEDGEMENT

This material is based upon work supported by National Centre for Research and Development under Grant No. PBS3/A6/29/2015.

REFERENCES

[1] Z. Łukasik, W. Nowakowski, "Bezprzewodowe systemy sterowania ruchem kolejowym", Infrastruktura Transportu, nr 4/2013, str. 22-25, 2013.

[2] W. Nowakowski, M. Szczygielski, "Analiza bezpieczeństwa transmisji w systemie zabezpieczenia przejazdów SZP-1", Technika Transportu Szynowego nr 9/2012.

[3] N. Ferguson, B. Schneier and T. Kohno, "Cryptography Engineering: Design Principles and Practical Applications", John Wiley and Sons, 2010.

[4] Z. Łukasik, W. Nowakowski, T. Ciszewski, "Bezpieczeństwo danych w diagnostyce systemów sterowania ruchem kolejowym", Autobusy: technika, eksploatacja, systemy transportowe, R.17 nr 6, str. 264-267, 2016.

[5] E. Earle Aaron, "Wireless Security Handbook", Auerbach Publications, 2005.

[6] P. Patil, P. Narayankar, D. G. Narayan, et al., "A Comprehensive Evaluation of Cryptographic Algorithms: DES, 3DES, AES, RSA and Blowfish", 1st International Conference on Information Security and Privacy, Nagpur, India, 2015, Procedia Computer Science, Vol.:78, pp. 617-624, 2016.

[7] M. J. Wang and Y. Z. Li, "Hash Function with Variable Output Length", 2015 International Conference on Network and Information Systems for Computers (ICNISC), pp. 190-193, 2015.

[8] C. Y. Chuang, Y. C. Chen and C. W. Hsueh, "Scheduling Low-Utilized Real-Time Systems with End-to-End Timing Constraints", 22nd IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, RTCSA 2016, Daegu, South Korea, 2016, IEEE Computer Society, 2016.

[9] W. Stallings, "Cryptography and Network Security: Principles and Practice", 6th Edition, Prentice Hall, 2014.

[10] M. Sumila and A. Miszkiewicz, "Analysis of the Problem of Interference of the Public Network Operators to GSM-R", Mikulski J. (eds) Tools of Transport Telematics, Communications in Computer and Information Science, Springer, Volume 531, pp. 253-263, 2015.

[11] B. Ndibanje, Y. J. Kang, M. Sain, et al., "An Approach to Designing a Network Security-based Application for Communications Safety", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 2015.

[12] M. Saadatmand, A. Cicchetti and M. Sjodin, "Design of Adaptive Security Mechanisms for Real-Time Embedded Systems", 4th International Symposium on Engineering Secure Software and Systems, Eindhoven, Netherlands, 2012, Engineering Secure Software and Systems, Vol. 7159 pp. 121-134, 2012.

[13] A. Odeh, S. R. Masadeh and A. Azzazi, "A Performance Evaluation of Common Encryption Techniques with Secure Watermark System (SWS)", International Journal of Network Security & Its Applications (IJNSA) Vol.7, No.3, pp. 31-38, 2015.

[14] A. Nadeem and M. Younous Javed, "A Performance Comparison of Data Encryption Algorithms", 1st International Conference on Information and Computer Technologies, pp. 84-89, 2005.

[15] L. J. Chen, Z. Y. Shan, T. Tang, et al., "Performance analysis and verification of safety communication protocol in train control system", Computer Standards & Interfaces, Vol. 33 Issue 5 pp. 505-518, 2011.

[16] L. Dai and K. Cooper, "Modeling and performance analysis for security aspects", 4th International Workshop on Systems/Software Architectures, Las Vegas, USA, 2005, Science of Computer Programming, Vol. 61 Issue 1, pp. 58-71, 2006.

[17] W. Nowakowski, "Information security and privacy protection in emergency management software systems", Logistyka 4/2015, str. 8072-8077, 2015.

[18] O. B. Montoya Alber, A. G. Munoz Mario and T. Kofuji Sergio, "Performance Analysis of Encryption Algorithms on Mobile Devices", 47th International Carnahan Conference on Security Technology (ICCST), Medellin, Colombia, 2013.

[19] S. Ji, T. Chen and S. Zhong, "Wormhole Attack Detection Algorithms in Wireless Network Coding Systems", IEEE Transactions on Mobile Computing, Vol. 14 Issue 3, pp. 660-674, 2015.

[20] EN 50159, "Railway applications - Communication, signalling and processing systems - Safety -related communication in transmission systems", 2010.

[21] Z. Łukasik, W. Nowakowski W, "Wymiana informacji w systemach związanych z bezpieczeństwem", Logistyka 6/2008.

Dynamical properties of red blood cell model in shear flow

Mariana Ondrušová[1], Ivan Cimrák[2]

Cell-in-fluid Research Group, http://cell-in-fluid.fri.uniza.sk
Faculty of Management Science and Informatics, University of Zilina, Slovakia
Corresponding author: mariana.ondrusova@fri.uniza.sk

Abstract — **Cells in a shear flow exhibit tumbling and tank-treading. This rotational movement is characterized by rotational frequency. In this work, we analyze and test a computational model of red blood cell by comparing simulated movement of a cell in a shear flow to the experimental data. We set up a computational experiment that recasts dynamical behavior of cells in a shear flow. We analyze the dependence of the rotational frequency of cells on the shear rate. Results show that the model including the stretching, bending, area and volume preservation moduli does not recover frequencies from the biological data. After adding the visco-elastic modulus, the simulations show compliance with the data.**

Keywords— computational modelling; red blood cell; shear flow; simulations

I. Introduction and Motivation

Understanding the mechanical behavior of individual red blood cells (RBC) is crucial in the dynamics of the human circulatory system. The flow of blood is mostly determined by the elastic properties of cell membranes and their interaction with each other. Gaining insights into the dynamics and morphologies of RBCs was addressed in several experimental studies [1,2].

Recently, computational models proved to be useful in computer-added discovery in biological sciences. In [3] for example, the authors provided evidence that the shear induced red blood cell tumbling-to-tank-treading transition also occurs at quite high volume fractions. Computational models also belong to powerful tools for design of microfluidic devices [4].

Proper validation of a computational model is provided by direct comparison of simulations with experimental results. For assessing static properties of red blood cells, the data from stretching experiment is widely used [2]. Here, RBC is stretched on opposed sides with known forces using optical tweezers. The prolongation is measured and the dependence of deformation index on stretching force is used to determine elastic coefficients of the model.

Further model validation concerns dynamic properties of RBC in a shear flow. Because of their elastic nature and shape, RBCs exhibit various behaviors in shear flows. First, an RBC elongates and aligns itself at a constant angle to a flow when embedded in a shear flow. Second, it may tumble, exhibit a tank-treading motion of the membrane, or both, depending on the shear rate [5]. The biological laboratory experiments concerning the tumbling and tank-treading frequency have been reported [6, 7].

A spring-network based model have been introduced in [8, 9]. Here, the cell is modeled by a triangulation of the membrane that defines the spring network. The static validation of elastic properties has been done [10] and the software implementation of the model was described in [11].

In this article, we aim at further validation of this model. We will investigate the model behavior under the shear flow. In Section II, we describe the computational model and ist main components: fluid and immersed object. Next section is devoted to description of the simulation setup. We provide the geometry of the computational domain and present how the shear flow is generated. In Section IV we present the results concerning the rotation frequency of the cell in a shear flow. We show that the model does not fit experimental data. Further in Section V we point out the visco-elasticity of biological cell's membrane and we suggest to test the model with additional visco-elastic modulus. We show the results of the adjusted model. Here, we demonstrate that inclusion of the visco-elasticity enables the model to fit the biological data. Finally, in Section VI we summarize the findings and draw conclusions.

II. Model of Red Blood Cell

Detailed description of the model has been presented elsewhere [8,9]. Briefly, to describe the mechanical processes of the cells flow, we need to take into account two basic phenomena: the fluid dynamics and cell deformation. Important part is also the coupling between those two components.

The fluid dynamics. Evolution of fluid is governed by the well-known and documented lattice–Boltzmann method. This

[1] The work of M.Ondrušová was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0751
[2] The work of I.Cimrák was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the contract No. VEGA 1/0643/17 and by the Slovak Research and Development Agency under the contract No. APVV-15-0751

method is based on fictive particles which propagate and collide over a fixed three-dimensional discrete lattice. The unknown variable is the particle density function defined for each lattice point. We use the D3Q19 version of the LB method (three dimensions with 19 discrete directions along the edges and diagonals of the lattice). The bulk properties such as fluid density and fluid velocity can be computed directly from the particle density function. More details can be found in [12].

Cell deformation. Cells are described by their membranes and these are represented by a triangular mesh containing points on the surface of the object. Mesh points are moving under the influence of fluid–cell interaction forces, as well as elasto-mechanic forces generated by the elasticity of the membrane. The elasticity is modeled using five elastic moduli: stretching, bending, local and global area conservation and volume conservation. Each module has its own stiffness coefficient, similar to the stretching coefficient of a linear spring.

The simplest module is the stretching module. It generates a force between each pair of mesh points connected by the edge in the triangulation. This force models essentially a non-linear spring [13], which pulls the points together when they are further than in the relaxed state and pushes them apart, when they are closer than in the relaxed state. The explicit formula for this force between two mesh points A and B is given by

$$F_s = k_s \kappa(L)(L - L_0). \quad (1)$$

Here, k_s is the stretching coefficient, L is the current length of the edge between A and B, L_0 is the length of the edge between A and B in the relaxed state without any external forces acting on the cell, and $\kappa(L)$ represents the neo-Hookean nonlinearity of the stretching force. From this expression, it is evident that if the edge is stretched than $L > L_0$ and the force directs towards the shortening of the edge. If the edge is squeezed than $L < L_0$ and the resulting force has opposite direction and thus supports prolongation of the edge.

Expressions for other moduli like bending, surface preservation and volume preservation can be found e.g. in [9,11].

Coupling between fluid and cells. The fluid and immersed cells interact with each other. This interaction is implemented by introducing a drag force F_d between the fluid and the mesh points

$$F_d = \xi(u - v), \quad (2)$$

where $\xi$is a phenomenological coefficient, v is the velocity of the mesh point and u is the velocity of the fluid at the position of the mesh point. This approach penalizes the difference $u - v$ and thus mimics the natural no-slip condition at the boundary of the flowing cell.

III. Simulation Setup

The computational domain comprises a cubic box with dimensions $20\ x\ 20\ x\ 20\mu m$. The shear flow in our simulations

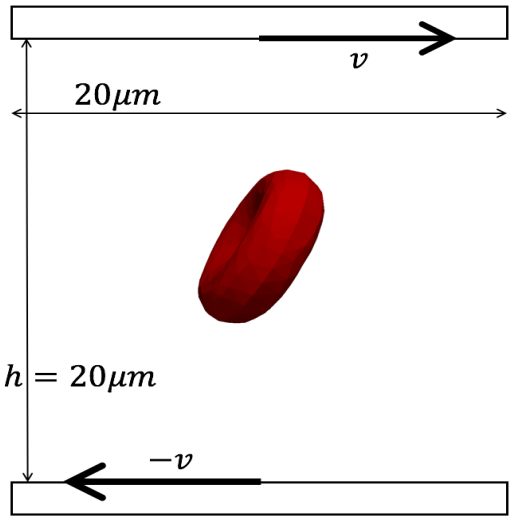


Figure 1. Two-dimensional view of 3D channel with horizontal x-axis and vertical z-axis. Simulation setting with indicated velocities of the fluid at the boundaries generating the uniform shear flow.

is generated by setting the constant velocity v and $-v$ at the top and bottom boundaries of the channel, see Figure 1. In this setting for an empty channel, the velocity field has zero y and z components and the horizontal x component linearly decreases from the value v at the top boundary to value $-v$ at the bottom boundary. This means that the shear rate is constant over the whole channel and equals to

$$\dot{\gamma} = \frac{2v}{h} \quad (3)$$

In our simulations, we work with one cell in the middle of the channel with shear flow. Simulation showed, that red blood cell in a shear flow at a certain speed exhibits tumbling motion. At higher speeds this motion becomes tank-treading, which corresponds to previously presented results [14]. During tumbling motion, the cell rotates as a whole, while during tank-treading motion, the cell leans slightly and its membrane begins to rotate around the interior of the cell.

The speed of the rotation during the tumbling and tank-treading depends on the shear rate. This dependence was measured on live cells and the data is available in [6,7,14].

To reproduce these experiments, we use the density of fluid $1050\ kg.m^{-3}$and viscosity $5\ Pa.s$. These values correspond to biological solutions of dextran that is typically used in experiments as in [6,7]. In recent study [15], the authors derived the expression for the friction coefficient and using this expression, we compute the value $\xi = 5.0$. For elastic coefficients of the model we use the following values

$$k_s = 0.008, k_b = 0.0003,\ k_{al} = 0.006,$$
$$k_{ag} = 0.9, k_v = 0.5. \quad (4)$$

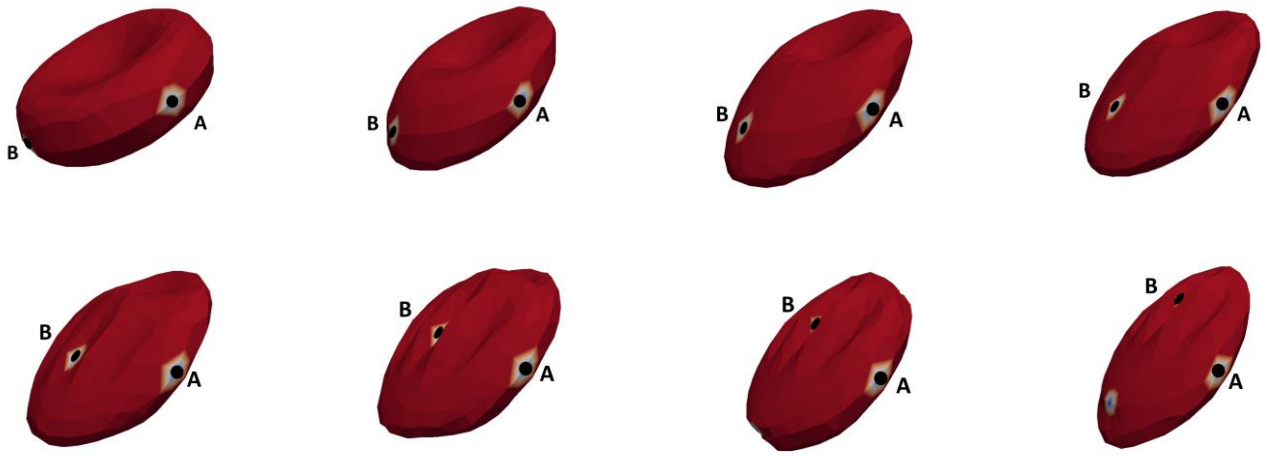


Figure 2. Deformation of the cell membrane during tank-treading motion of a cell under shear rate 200 s^{-1}. Two specific mesh points are highlighted. Point A is located on the y-axes of rotation. Point B rotates around the inside of the cell.

These values have been determined from stretching experiments mentioned in [2]. For discretization of time we use the time step equaling to $0.1\mu s$. Snapshots of tank-treading cell is depicted in Figure 2. From these snapshots we can clearly see that the membrane rotates around the inner part of the cell and the shape changes from the relaxed bi-concave shape into ellipsoid-like shape that does not change much.

IV. Results

During the simulation, we record coordinates of the mesh points. In Figure 3, x-coordinate of one specific point is depicted. The curve shows characteristic periodic evolution

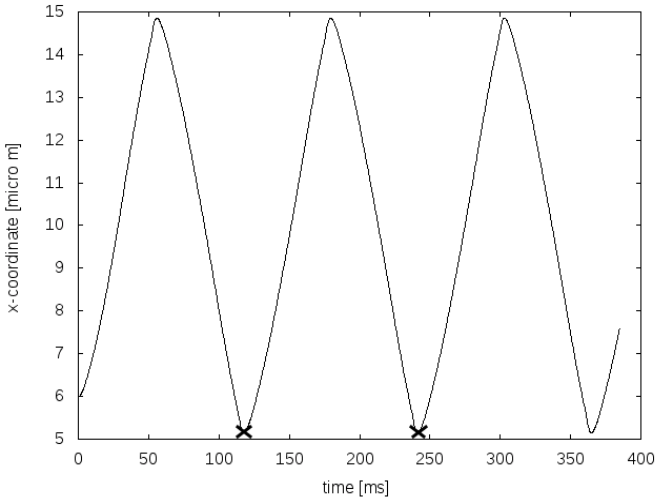


Figure 3. Evolution of x-coordinate of one specific point during the tank-treading motion. The difference of times at horizontal axis between two marked minimum points represents duration of one cell rotation.

from which we extract period of one rotation T. Then we calculate the rotation frequency of the cell f as

$$f = \frac{1}{T} \tag{5}$$

In [14] they work with shear rate up to $200s^{-1}$. To have the same values in our experiment, we use (3) to compute the corresponding value of velocity resulting in $0.002ms^{-1}$. In the simulations we thus used 10 values from $0.0002ms^{-1}$ to $0.002ms^{-1}$.

The results are presented in Figure 4. We can clearly see that our model exhibits different behavior than expected. Frequency of rotation is higher. We performed several tests to see whether

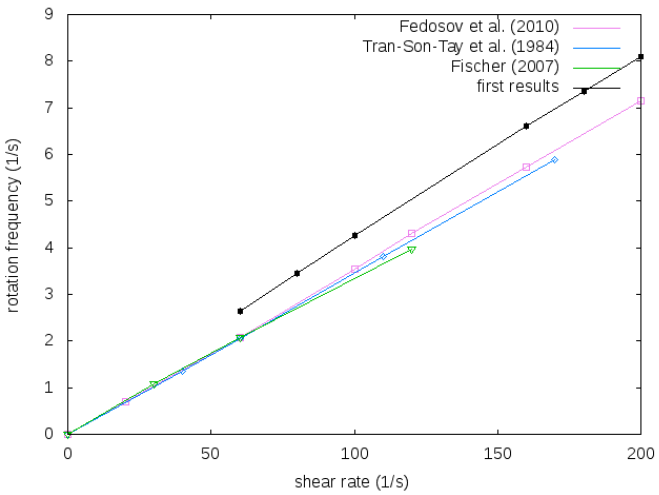


Figure 4. Rotation frequency for the original model (black line) and experimental data (other colors) taken from [6,7,14]. Results from the original model are approximately 20% higher than experimental data.

the model can be adapted to fit the experimental data. We tried to slightly change the elastic coefficients, this however did not have any effect on the rotating frequency.

V. Model Revisited

The biological membrane of a RBC exhibits visco-elastic properties [16]. Our model does not include such properties. In general, visco-elasticity is a property of the material that penalizes fast changes of the shape. Since the tank-treading motion of the cell introduces permanent changing of the membrane's shape, there is a hope that including visco-elasticity could have a effect on rotational frequency of the cell. In [17], an attempt has been made to include visco-elastic module into the model with promising results. In the following, we will test this implementation. The basic principle is to include visco-elastic contribution F_{vis} to the elastic forces. This contribution takes into account how fast the length of an edge of the mesh changes. The force acts against these changes according to the following expression

$$F_{vis} = k_{vis} \frac{dL}{dt} \tag{6}$$

We have performed identical experiments with this extended model. In (6) however, the proper value of k_{visc} needs to be chosen. We performed numerous simulations with different values of k_{visc} ranging from 0 to 2.5. We identified that $k_{visc} = 1.5$ is the optimal value. Using this value, we obtained results presented in Figure 5. We can clearly see that the rotational movement of the cell slowed down and the frequency perfectly fit the experimental data.

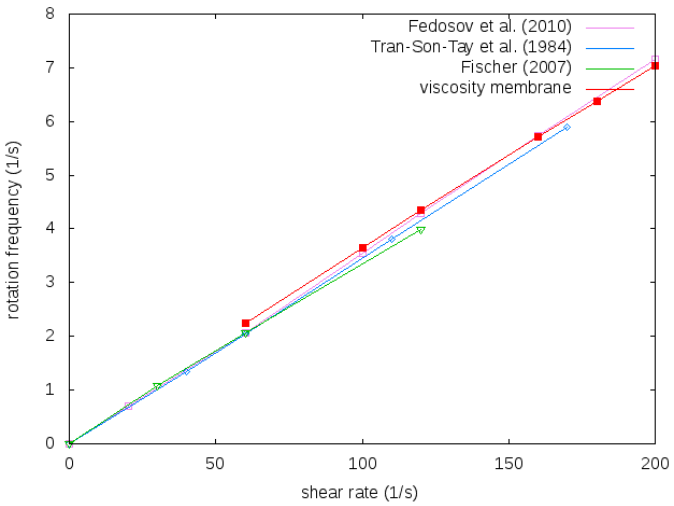


Figure 5. Rotation frequency for adjusted model including visco-elastic modulus (red line) and experimental data (other colors) taken from [6,7,14]. The results from the original model coincide with the experimental data.

VI. Conclusions

In this paper, we tested the model of RBC that has been previously calibrated on static stretching experiments. Our aim was to verify, whether the dynamical properties of the model correspond to the biological behavior of the cell. We considered the rotation frequency of a cell that is immersed in a shear flow. The computational results were compared to the available biological data.

This work presents how the inclusion of the membrane viscosity in the model influences results. We showed, that without including the viscosity, the membrane resistance against the shear flow is too low and thus the frequency of the rotation is higher than in biological experiments. This is shown in Figure 4, where the frequency dependence on the applied shear stress is depicted. The black line stands for the model without viscosity and other lines represent the biological data from references [6,7,14]. We see that the frequency is higher and the elastic properties of red cell membrane must be adapted. The frequency of rotating cell in simulations was roughly 20% higher.

This discrepancy was assigned to visco-elastic properties of biological membranes. The original model did not account for viscous forces. We tested a previously introduced model extension. Including the viscosity, we show that the black line coincides with the biological data, see Figure 5. Therefore, we claim that the inclusion of viscosity improves the model fit to the biological data. Eventually, the viscous modulus added to the model caused that the rotation has been slowed down to fit the data.

The additional term from (6) however requires to set proper coefficient k_{visc}. At the current stage of knowledge, we do not know the connection between the biological values of visco-elasticity of the membrane reported in [16] and k_{visc}. The thorough investigation will be necessary in this matter. Here, we only demonstrate that such model extension does positively influence the dynamical behavior of cell in a shear flow.

Acknowledgment

We would like to thank prof. M.Klimo that he made it possible to use computational resources from his project ITMS 26210120021.

References

[1] M. Abkarian, M. Faivre, and A. Viallat „Swinging of red blood cells under shear flow", Phys. Rev. Lett. 98:188302., 2007

[2] J. P. Mills, L. Qie, M. Dao, C. T. Lim, and S. Suresh, "Nonlinear elastic and viscoelastic deformation of the human red blood cell with optical tweezers," *Molecular & Cellular Biomechanics*, vol. 1, no. 3, pp. 169–180, 2004.

[3] T. Krüger, M. Gross, D. Raabe, and F. Varnik, "Crossover from tumbling to tank-treading-like motion in dense simulated suspensions of red blood cells", Soft Matter 9, 9008, 2013.

[4] Gleghorn et al., "Capture of circulating tumor cells from whole blood of prostate cancer patients using geometrically enhanced differential immunocapture (GEDI) and a prostate-specific antibody. Lab Chip. 10:27–29, 2010.

[5] M. Nakamura, S. Bessho and S. Wada, “Spring-network-based model of a red blood cell for simulating mesoscopic blood flow”, Int. J. Numer. Meth. Biomed. Engng., 29: 114–128, 2013

[6] R. Tran-Son-Tay, S. P. Sutera & P.R Rao, “Determination of RBC membrane viscosity from rheoscopic observations of tank-treading motion”, Biophys. J. 46, 65–72, 1984

[7] T. M. Fischer “Tank-tread frequency of the red cell membrane: Dependence on the viscosity of the suspending medium”, Biophys. J. 93, 2553–2561, 2007

[8] I. Cimrák, M. Gusenbauer, T. Schrefl, “Modelling and simulation of processes in microfluidic devices for biomedical applications”, Computers and Mathematics with Applications, Vol 64(3), pp. 278-288, 2012.

[9] K. Bachratá, H. Bachratý, “On modeling blood flow in microfluidic devices”, ELEKTRO 2014: 10th International Conference, IEEE, Slovakia, ISBN 978-4799-3720-2, pp. 518-521,2014.

[10] R. Tóthová, I. Jančigová, M. Bušík, “Calibration of elastic coefficients for spring-network model of red blood cell”, Information and Digial Technologies (IDT) 2015, International Conference, IEEE, Slovakia, ISBN 978-1-4673-7185-8, pp. 376-380, 2015.

[11] I. Cimrák, M. Gusenbauer, I. Jančigová, “An ESPResSo implementation of elastic objects immersed in a fluid”, Computer Physics Communications, Volume 185, Issue 3, Pages 900-907, 2014.

[12] B. Dunweg, A. J. C. Ladd, "Lattice-Boltzmann simulations of soft matter systems", Advances in Polymer Science 221, 89–166, 2009.

[13] I. Jančigová, R. Tóthová, “Scalability of forces in mesh-based models of elastic objects”, ELEKTRO 2014: 10th International Conference, IEEE, Slovakia, ISBN 978-4799-3720-2, pp. 562-566, 2014.

[14] D. A. Fedosov, B. Caswell, G. E. Karniadakis, “Cell deformation in shear flow” Computational Hydrodynamics of Capsules and biological Cells, CHAPMAN & HALL/CRC Mathematical and Computational Biology Series, pp. 204-209 ISBN 78-1-4398-2006-3, 2010.

[15] M. Bušík, I. Cimrák, “The calibration of fluid-object interaction in immersed boundary method”, Experimental fluid mechanics 2016, 15.-18.11.2016, Mariánské Lázně, Czech Republic, preprint available at cell-in-fluid.fri.uniza.sk/en/content/publications.

[16] R.M. Hochmuth, P.R. Worthy, E.A. Evans, "Red cell extensional recovery and the determination of membrane viscosity". Research paper, Biophys. J. 26 (1), pp. 101–114, 1979.

[17] M. Bušík, "Viscoelasticity in spring network models for proper dynamics of cell membrane", Journal of Information, Control and Management Systems, [S.I.], v. 13, n. 1, apr. 2015.

Statistical Hypotheses Testing for Random and Pseudorandom Generators Based on Statistical Estimators of Entropy

Uladzimir Palukha[1], Yuriy Kharin[2]
1) Faculty of Applied Mathematics and Informatics; 2) Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
Minsk, Belarus
palukha@bsu.by, kharin@bsu.by

***Abstract*—The topical information security problem of development of statistical tests for hypotheses on the discrete uniform distribution ("pure randomness") of output sequences produced by random or pseudorandom generators is considered. For Shannon, Rényi and Tsallis entropy functionals, the point statistical estimators based on the plug-in principle are constructed. The asymptotic probability distributions of constructed point estimators under the "pure randomness" hypothesis are found in the asymptotics, when the number of observed data is comparable with the number of parameters. Interval statistical estimators of considered information entropy functionals are constructed. On the base of interval estimators the decision rules for statistical hypotheses testing on "pure randomness" of observed discrete sequence are developed. The results of computer experiments are given.**

Keywords—Shannon, Rényi and Tsallis entropy, asymptotically normal probability distribution, statistical estimators, hypotheses testing, generators of random and pseudorandom sequences.

I. INTRODUCTION

Generators of random and pseudorandom sequences are the main structural elements of information security systems and play a central role in the construction of encryption schemes [1]. For example, generators are using to initialize parameters of digital signature systems and keys in encryption systems. The reliability of such systems depends on how the properties of generated sequence are close to the properties of "pure random" sequence [2]. To check the quality of generators in the sense of matching of probabilistic properties of their output sequences to probabilistic properties of "pure random" sequence statistical tests are used. The essence of the tests is as follows. The output sequence of the generator is observed and a null hypothesis H_* that the sequence is "pure random" is introduced. Some statistic, the probability distribution of which under the true null hypothesis H_* is known, is calculated. Based on the value of statistic, the hypothesis is either accepted or rejected. In this paper we propose to use estimators of the information entropy functionals as test statistics. The most common entropy functionals are the Shannon, Rényi and Tsallis entropy functionals, for point statistical estimators of which the asymptotic probability distributions under the true hypothesis H_* are found in this paper. The found probability distributions allow us to construct statistical tests and to apply them to analyze the output sequences of random and pseudorandom number generators.

II. MATHEMATICAL MODEL

Let a random variable $x = x(\omega) = \omega$ with the set of states $\Omega = \{\omega_1, \ldots, \omega_N\}$ and with the discrete probability distribution $p_k = P\{x = \omega_k\}$, $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1$, $k = 1, \ldots, N$, be defined on a probability space (Ω, F, P). We define the functional of generalized entropy according to [3]:

$$H_{h,w}^{\varphi_1,\varphi_2}(P) = h\left(\frac{\sum_{k=1}^{N} w_k \varphi_1(p_k)}{\sum_{k=1}^{N} w_k \varphi_2(p_k)}\right), \qquad (1)$$

where $w_k > 0$, $k = 1, \ldots, N$, is the weight of ω_k, φ_1: $[0, 1) \to \mathbb{R}$, φ_2: $[0, 1) \to \mathbb{R}$, h: $\mathbb{R} \to \mathbb{R}$, are given functions. There are different entropy functionals, for example, formulas of 23 functionals are given in [3]. Table 1 shows the most frequently used [3] particular cases of the generalized entropy functional (1), defined by the specification of the functions $h(\cdot)$, $\varphi_1(\cdot)$, $\varphi_2(\cdot)$, $\{w_k\}$, plugging into (1). It is worth to note that the Shannon entropy functional is the limiting case of the Rényi and Tsallis functionals for $r \to 1$ [4] and differs from them by the presence of some additional properties (for example, additivity [2]). Under the true hypothesis H_*, all three functionals reach their maximum value.

TABLE I. BASIC ENTROPY FUNCTIONALS

Type	Formula	$h(x)$	$\varphi_1(x)$	$\varphi_2(x)$	w_k
Shannon entropy	$H(P) = -\sum_{k=1}^{N} p_k \ln p_k$	x	$-x\ln x$	x	$w \equiv 1$
Rényi entropy	$H_r(P) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N} p_k^r\right)$	$\frac{\ln x}{1-r}$	x^r	x	$w \equiv 1$
Tsallis entropy	$S_r(P) = \frac{1}{r-1}\left(1-\sum_{k=1}^{N} p_k^r\right)$	$\frac{x-1}{1-r}$	x^r	x	$w \equiv 1$

A common approach to statistical estimation of entropy is the construction of frequency estimators of the states probabilities and the substitution of the obtained estimates in the entropy functional instead of the true values of probabilities. In this paper, we propose a method for constructing statistical estimators of Shannon, Rényi and Tsallis entropy. Also we give the probabilistic properties of the obtained estimators in the asymptotics, which is more often encountered in practice, and means that the number of observables is comparable with the number of parameters being evaluated. Using the point estimators, interval statistical estimators of entropy are constructed. Based on the interval statistical estimators the decision rules for statistical testing of hypothesis about the closeness of the observed output sequence to the "pure random" sequence are developed.

III. Construction of Statistical Estimators for Entropy

A. Frequency Estimators of Probabilities

Let there be a random sequence of length n, from the probability distribution $\{p_k\}$. We construct the frequency estimators of the probability distribution $\{p_k: k = 1, \ldots, N\}$:

$$\hat{p}_k = \frac{\nu_k}{n}, \quad \nu_k = \sum_{t=1}^{n} I\{x_t = \omega_k\} \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}, \qquad I\{x_t = \omega_k\} = \begin{cases} 1, & x_t = \omega_k; \\ 0, & x_t \neq \omega_k. \end{cases} \tag{2}$$

As it had been already mentioned in the introduction, we introduce the hypothesis H_* = {{x_t} is "pure random" sequence} = {{x_t} are independent identically distributed random variables, $p_k = 1 / N$, $k = 1, \ldots, N$}, and the general alternative $\overline{H_*}$.

Following [5], we will assume that the series scheme holds. In this case, the vector $(\nu_1, \ldots, \nu_N)^T$, composed of the frequencies ν_k from (2), has the multinomial probability distribution Mul(n, N, p_1, …, p_N), and each of the components has the binomial probability distribution $Bi(n, p_k)$. Consider the asymptotics:

$$n, N \to \infty, \; n / N \to \lambda, \; 0 < \lambda < \infty, \tag{3}$$

which differs from the classical one ($n \to \infty$, $N < \infty$) in that the duration of observation n and the number of values N grow synchronously. In the asymptotics (3), the probability distribution of the statistics $\{\nu_k\}$ is approximated by the Poisson distribution $\Pi(\lambda_k)$ with the parameter $\lambda_k = np_k$. Under the true hypothesis H_*, all elementary probabilities are equal: $p_k = 1 / N$, $k = 1, \ldots, N$, therefore, all frequencies $\{\nu_k\}$ have the same Poisson distribution parameter $\lambda = n / N$.

The theorem on the asymptotically normal distribution of statistics that are functions of the frequencies ν_k is proved in [5]. It can be briefly reformulated as follows. Let $f(\cdot)$: $\mathbb{N}_0 \to \mathbb{R}$ be a function; $Z = \sum_{k=1}^{N} f(\nu_k)$, where ν_k, $k = 1, \ldots, N$ are frequencies with the joint multinomial distribution, approximated by the Poisson distribution in the asymptotics (3). Then, under certain regularity conditions, the statistic Z has the asymptotically normal distribution $\mathcal{L}\left\{\frac{Z-\mu}{\sigma}\right\} \to N_1(0,1)$:

$$\mu = \sum_{k=1}^{N} E\{f(\nu_k)\}, \tag{4}$$

$$\sigma^2 = \sum_{k=1}^{N} \mathrm{var}\{f(\nu_k)\} - \left(\sum_{k=1}^{N} \mathrm{cov}\{\nu_k, f(\nu_k)\}\right)^2 \Big/ n, \tag{5}$$

where $N_1(0, 1)$ is the standard one-dimensional normal probability distribution with zero mathematical expectation and variance that equals one, $E\{\xi\}$ and $\mathrm{var}\{\xi\}$, respectively, the mathematical expectation and the variance of the random variable ξ, $\mathrm{cov}\{\xi, \eta\}$ is the covariance of the random variables ξ and η. Under the true hypothesis H_*, the relations (4) and (5) are transformed respectively:

$$\mu = \sum_{k=1}^{N} E\{f(\nu_k)\} = NE\{f(\nu)\},$$

$$\sigma^2 = N\,\mathrm{var}\{f(\nu)\} - N^2\,\mathrm{cov}^2\{\nu, f(\nu)\}/n = N\left(\mathrm{var}\{f(\nu)\} - \mathrm{cov}^2\{\nu, f(\nu)\}/\lambda\right).$$

To apply the results from [5] to the proof of the probabilistic properties of statistical estimators of entropy, it is necessary to express the estimators of the entropy functionals in the terms of frequencies.

B. Statistical Estimation of Shannon Entropy

We take $f(\nu) = \nu \ln \nu$ as a function f. The statistical estimator of Shannon entropy $\hat{H}(n, N)$ is linearly expressed in terms of $Z = \sum_{k=1}^{N} \nu_k \ln \nu_k$ [6]:

$$\hat{H} = \hat{H}(n, N) = -\sum_{k=1}^{N} \hat{p}_k \ln \hat{p}_k = -\sum_{k=1}^{N} \frac{\nu_k}{n} \ln \frac{\nu_k}{n} = \ln n - \frac{1}{n} Z. \tag{6}$$

The theorem on the asymptotic probability distribution of the statistic (6) is proved by the first author in [7].

Theorem 1. *In the asymptotics* (3), *the statistic* (6) *with the true hypothesis* H_* *has the asymptotically normal distribution* $\mathcal{L}\left\{\frac{\hat{H}-\mu_H}{\sigma_H}\right\} \to N_1(0,1)$:

$$\mu_H = \ln n - e^{-\lambda}\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}, \quad (7)$$

$$\sigma_H^2 = \frac{e^{-\lambda}}{n}\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1) - \frac{e^{-2\lambda}}{N}\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2 - \frac{e^{-2\lambda}}{n}\left(\sum_{k=1}^{+\infty}\ln(k+1)\frac{\lambda^k}{k!}(k+1-\lambda)\right)^2. \quad (8)$$

The behavior of the mathematical expectation of the Shannon entropy estimator of a binary sequence, which is divided into fragments of length s (they are called s-grams; in this case $N = 2^s$), is considered in [8].

Knowing the asymptotic probability distribution of the point estimator (6) allows us to construct the interval estimator for the Shannon entropy:

with probability that equals 1 – ε, the entropy estimate

$$\hat{H}(P) \in (H_-, H_+), \quad H_\pm = \mu_H \pm \sigma_H \Phi^{-1}\left(1-\frac{\varepsilon}{2}\right), \quad (9)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The disadvantage of the constructed point estimator is the presence of a bias, as it's demonstrated in [8]. Therefore we consider the construction of the statistical estimators for the Rényi and Tsallis entropy functionals.

C. Statistical Estimation of Rényi and Tsallis Entropy

We consider the Rényi and Tsallis entropy functionals with a parameter $r \in \{2, 3, \ldots\}$. As it can be seen from the Table 1, functionals have a common function $\varphi_1(x) = x^r$. The argument of the function is the probability p_k. It's also seen that the Rényi and Tsallis entropies are functions of the quantity

$$P_r(P) = \sum_{k=1}^{N} p_k^r \,. \quad (10)$$

Consequently, the problem of the statistical estimation of the quantity $P_r(P)$ arises.

It's known [9] that the statistical estimator for (10) $\hat{P}_r(P) = \sum_{k=1}^{N}\hat{p}_k^r = \sum_{k=1}^{N}\left(\frac{\nu_k}{n}\right)^r$, constructed by the plug-in principle, is biased. To construct an asymptotically unbiased estimator, we define the r^{th} descending factorial of x:

$$x^{\underline{r}} = x(x-1)\ldots(x-r+1) = \frac{x!}{(x-r)!} = \sum_{i=0}^{r} s(r,i)x^i, \quad (11)$$

where $s(r, i)$ is the Stirling number of the first kind [10]; by definition, for $x < r$ it is assumed $x^{\underline{r}} ::= 0$. The asymptotically unbiased and consistent statistical estimator for the quantity (10), which is based on (11), is proposed in [9]:

$$\tilde{P}_r(P) = \sum_{k=1}^{N}\frac{\nu_k^{\underline{r}}}{n^r}. \quad (12)$$

Assuming that $f_r(v) = v^{\underline{r}}$,

$$Z_r = \sum_{k=1}^{N} f_r(\nu_k) = \sum_{k=1}^{N}\nu_k^{\underline{r}} = n^r\tilde{P}_r(P). \quad (13)$$

We have the following lemma [11] on the probability distribution of statistic (13).

Lemma. *In the asymptotics* (3), *the statistic* (13) *with the true hypothesis* H_* *has the asymptotically normal distribution* $\mathcal{L}\left\{\frac{Z_r-\mu_r}{\sigma_r}\right\} \to N_1(0,1)$:

$$\mu_r = N\lambda^r = n\lambda^{r-1},$$

$$\sigma_r^2 = N\lambda^r\left(\sum_{i=1}^{r}s(r,i)\sum_{j=0}^{i-1}C_i^j r^{i-j}\sum_{k=1}^{j}S(j,k)\lambda^k - r^2\lambda^{r-1} + r!\right) = n\lambda^{r-1}\left(\sum_{i=1}^{r}s(r,i)\sum_{j=0}^{i-1}C_i^j r^{i-j}\sum_{k=1}^{j}S(j,k)\lambda^k - r^2\lambda^{r-1} + r!\right),$$

where $S(j, k)$ is the Stirling number of the second kind [10].

Corollary 1. *For* $r = 2$, *the following expressions for the parameters of the asymptotically normal probability distribution of the random variable* Z_2 *take place:*

$$\mu_2 = n\lambda, \ \sigma_2^2 = 2n\lambda.$$

The statistical estimators of Rényi and Tsallis entropy are expressed in terms of Z_r [11]:

$$\hat{H}_r(n, N) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{\nu_k^{\underline{r}}}{n^r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln Z_{n,r}\right), \quad (14)$$

$$\hat{S}_r(n, N) = \frac{1}{r-1}\left(1-\sum_{k=1}^{N}\frac{\nu_k^{\underline{r}}}{n^r}\right) = \frac{1}{r-1}\left(1-\frac{Z_{n,r}}{n^r}\right). \quad (15)$$

The theorems on the asymptotic probability distribution of statistical estimators of Rényi and Tsallis entropy, which are based on [5], are proved by the authors [11] and allow us to

construct the interval estimators. We also give corollaries of the theorems for the most commonly used particular case $r = 2$.

Theorem 2. *In the asymptotics* (3), *the statistic* (14) *is a consistent estimator of Rényi entropy and, under the true hypothesis* H_*, *has the asymptotically normal distribution*: $\mathcal{L}\left\{\frac{H_r - \mu_{H,r}}{\sigma_{H,r}}\right\} \to N_1(0,1)$,

$$\mu_{H,r} = \ln N, \tag{16}$$

$$\sigma^2_{H,r} = \frac{\sum_{i=1}^{r} s(r,i) \sum_{j=0}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k) \lambda^k - r^2 \lambda^{r-1} + r!}{(r-1)^2 n \lambda^{r-1}}. \tag{17}$$

Corollary 2. *For* $r = 2$, *the variance of the estimator* (14) *is*

$$\sigma^2_{H,2} = \frac{2}{n\lambda}.$$

Note that $p_k = 1 / N$, $k = 1, \ldots, N$ with the true hypothesis H_*, therefore the value of the Rényi entropy equals $H_r(P) = \frac{1}{1-r} \ln\left(\sum_{k=1}^{N} p_k^r\right) = \frac{1}{1-r} \ln\left(\sum_{k=1}^{N} \frac{1}{N^r}\right) = \ln N$, that is equal to (16).

Knowing the asymptotic distribution of the consistent point estimator (14) allows us to construct the interval estimator for Rényi entropy:

with probability that equals $1 - \varepsilon$, the entropy

$$H_r(P) \in (H_-, H_+),\ H_\pm = \mu_{H,r} \pm \sigma_{H,r} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \tag{18}$$

Theorem 3. *In the asymptotics* (3), *the statistic* (15) *is a consistent asymptotically unbiased estimator of Tsallis entropy and, under the true hypothesis* H_*, *has the asymptotically normal distribution* $\mathcal{L}\left\{\frac{S_r - \mu_{S,r}}{\sigma_{S,r}}\right\} \to N_1(0,1)$:

$$\mu_{S,r} = \frac{1}{r-1}\left(1 - \frac{1}{N^{r-1}}\right), \tag{19}$$

$$\sigma^2_{S,r} = \frac{\lambda^{r-1}}{(r-1)^2 n^{2r-1}}\left(\sum_{i=1}^{r} s(r,i) \sum_{j=1}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k) \lambda^k - r^2 \lambda^{r-1} + r!\right). \tag{20}$$

Corollary 3. *For* $r = 2$, *the mathematical expectation and the variance of statistic* (15) *are respectively:*

$$\mu_{S,2} = 1 - \frac{1}{N},\ \sigma^2_{S,2} = \frac{2}{Nn^2}.$$

Knowing the asymptotic distribution of the consistent point estimator (15) allows us to construct the interval estimator for Tsallis entropy:

with probability that equals $1 - \varepsilon$, the entropy

$$S_r(P) \in (S_-, S_+),\ S_\pm = \mu_{S,r} \pm \sigma_{S,r} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \tag{21}$$

IV. Testing of "Pure Randomness" by Entropy Estimators

The obtained interval estimators (9), (18) and (21) allow us to construct the decision rule for testing hypotheses about whether the observed generator's output sequence is "pure random" sequence: H_* and $\overline{H_*}$. Let $\varepsilon \in (0, 1)$ be a given significance level. We introduce the notation: $\hat{h}$ is the statistical estimate of the Shannon (6), Rényi (14) or Tsallis (15) entropy, μ_h is the asymptotic expectation of the statistical estimator of Shannon (7), Rényi (16) or Tsallis (19) entropy, σ_h^2 is the asymptotic variance of the statistical estimator of Shannon (8), Rényi (17) or Tsallis (20) entropy under the true hypothesis H_*. We compute the statistic $\hat{h}$ for the observed sequence. The decision rule based on statistic $\hat{h}$ has the form:

$$\begin{cases} H_*, & \text{if } t_- < \hat{h} < t_+; \\ \overline{H_*}, & \text{else;} \end{cases} \quad t_\pm = \mu_h \pm \sigma_h \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \tag{22}$$

If we accept the hypothesis H_*, it can be concluded that, at the significance level ε, the analyzed process is indistinguishable from the "pure random" sequence with respect to its entropic properties on the base of the observed output sequence with the length no longer than n.

The advantage of the developed approach in comparison with other statistical tests is that we can specify different dimensions of the alphabet. For example, if we deal with a binary sequence, we can form s-grams from each s neighboring bits, thereby obtaining an alphabet of dimension 2^s. Applying the decision rule for different values of s, we can improve the decision rule: we assume that the sequence is "pure random" if the proportion of rejections of the hypothesis H_* does not exceed a given significance level. In addition, the dependence of the entropy estimate on the length of the fragment s can also be used to analyze the presence of dependencies in the observed sequence.

V. The Results of Computer Experiments

The developed decision rule (22) with the significance level $\varepsilon = 0.05$ was used to analyze the output binary sequence of a real physical binary random sequence generator [12] $\{y_\tau\}$, $\tau = 1, \ldots, T$, with a length of $T = 125 \cdot 2^{25}$ bits. The output sequence was "cut" into non-overlapping consecutive fragments of length s (s-grams): $X^{(t)} = (X_j^{(t)}) = (y_{(t-1)s+1}, \ldots, y_{ts}) \in \{0,1\}^s$, $t = 1, \ldots, n = [T / s]$. From the resulting s-grams, a new sequence $\{x_t\}$ with the dimension of alphabet equals $N = 2^s$ was formed by the rule $x_t = \sum_{j=1}^{s} 2^{j-1} X_j^{(t)} + 1$. The entropy estimates were calculated by the algorithms from [13].

Fig. 1 shows the values of the deviation of Shannon entropy estimate (6) from the mathematical expectation (7) divided by the confidence interval boundaries: $\dfrac{H - \mu_H}{\sigma_H \Phi^{-1}(1-\varepsilon/2)}$, depending on $s \in \{5, \ldots, 24\}$. Fig. 2 shows the values of the deviation of Rényi entropy estimate (14) at $r = 2$ from the mathematical expectation (16) divided by the confidence interval boundaries: $\dfrac{H_r - \mu_{H,r}}{\sigma_{H,r} \Phi^{-1}(1-\varepsilon/2)}$, depending on $s \in \{2, \ldots, 30\}$. If the value does not fall into the confidence bandwith $(-1; 1)$, it means that the statistic does not match the confidence interval and the hypothesis H_* is rejected. As can be seen from Fig. 1, at many values of the order s the hypothesis H_* is rejected, which indicates that the output sequence of the generator is differ from the "pure random" sequence. Therefore, as can be seen from Fig. 2, at values $s \le 25$ the output sequence of the generator is consistent with the "pure random" sequence by the value of Rényi entropy estimate.

The decision rules based on the estimators of Shannon and Rényi entropy have given different results. This means that these tests should be applied in a complex: one test can reveal deviations from the "pure random" sequence that another test has not revealed.

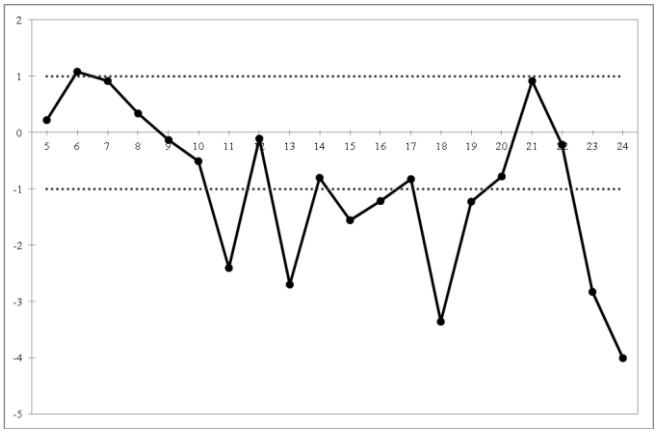


Figure 1. Deviation of Shannon entropy estimate from its expectation for $s \in \{5, \ldots, 24\}$.

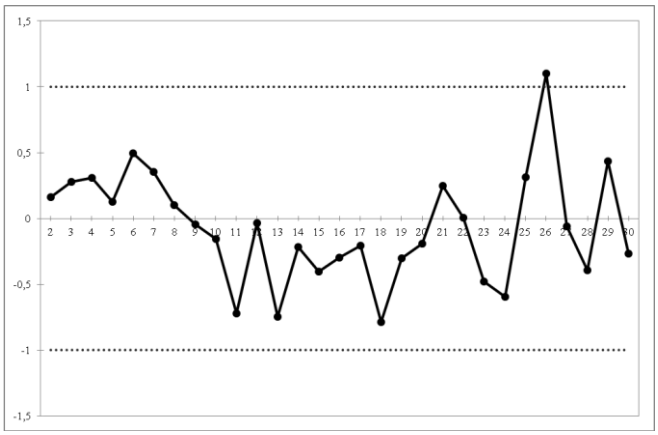


Figure 2. Deviation of Rényi entropy estimate from its expectation for $s \in \{2, \ldots, 30\}$

References

[1] O. Goldreich, "Foundations of Cryptography: Basic Tools". Cambridge: University Press, 2004.

[2] Yu. S. Kharin, S. V. Agievich, D. V. Vasilyev, and G. V. Matveev, "Cryptology" [in Russian]. Minsk: BSU, 2013.

[3] M. D. Esteban and D. Morales, "A summary on entropy statistics", in Kybernetika, 1995, vol. 31, no. 4, pp. 337–346.

[4] P. A. Bromiley, N. A. Thacker, and E. Bouhova-Thacker, "Shannon Entropy, Renyi Entropy, and Information". Available at http://www.tina-vision.net/docs/memos/2004-004.pdf.

[5] L. Holst, "Asymptotic normality and efficiency for certain goodness-of-fit tests", in Biometrika, 1972, vol. 59, no. 1, pp. 137–145.

[6] U. Yu. Palukha, "The probability properties of the multivariate entropy estimator in information security tasks" [in Russian], in Proceedings of the XVII Young Scientists Republican Scientific and Practical Conference, Brest, 2015, Part 1. Brest: BrSU, 2015, pp. 57–59.

[7] U. Yu. Palukha, "Statistical tests based on entropy estimates for checking the hypotheses of the uniform distribution of a random sequence" [in Russian], in Proceedings of the National Academy of Sciences of Belarus: Physics and Mathematics Series, 2017, no. 1, pp. 79–88.

[8] U. Yu. Palukha and Yu. S. Kharin, "Entropy characteristics of binary sequences in cryptography" [in Russian], in Proceedings of the XX Scientific and Practical Conference "Complex Information Protection", Minsk, 2015. Minsk: RIHE, 2015, pp. 99–102.

[9] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "Estimating Renyi Entropy of Discrete Distributions". Available at http://arxiv.org/pdf/1408.1000v3.pdf.

[10] A. Yu Envin, "Discrete mathematics: lecture notes" [in Russian]. Cheliabinsk: YuUrGU, 1998.

[11] Yu. S. Kharin and U. Yu. Palukha, "Statistical estimates of Rényi and Tsallis entropy and their use for testing the hypotheses of "pure randomness"" [in Russian], in Proceedings of the National Academy of Sciences of Belarus: Physics and Mathematics Series, 2016, no. 2, pp 37–47.

[12] speedtest-500MB.bin. Available at http://qrng.physik.hu-berlin.de/files/speedtest-500MB.bin.

[13] U. Yu. Palukha and Yu. S. Kharin, "Calculating of the entropy functionals statistical estimates of binary sequences" [in Russian], in Proceedings of the CSIST'2016. Minsk: BSU, 2016, pp. 472–476.

From Unstructured Data Included in Real-Estate Listings to Information Systems over Ontological Graphs

Krzysztof Pancerz and Piotr Grochowalski
Department of Computer Science, Faculty of Mathematics and Natural Sciences
University of Rzeszów, Poland
{kpancerz,piotrg}@ur.edu.pl

***Abstract*—In the paper, we consider the problem of automatic transformation of unstructured data included in real-estate listings into data arranged in a tabular form, as the so-called information tables. Transformation is an important preprocessing stage enabling us to obtain data in a form accepted by many data mining and machine learning tools. In the presented approach, information tables represent information systems over ontological graphs. In such systems, the domain knowledge covering some data semantics is directly incorporated. In the paper, we describe important aspects of a specialized computer tool created by us to automatize the transformation process.**

***Index Terms*—data preprocessing; data semantics; ontology; real-estate listings**

I. Introduction

On the one hand, in many data mining and machine learning algorithms, input data are given in a tabular form, called information tables. An information table includes vectors of values of attributes describing objects in a given universe of discourse. On the other hand, a lot of data, with which a human being has to deal every day, has an unstructured form, for example textual. Therefore, we need to apply some transformation procedures enabling us to obtain data in a form allowing their further analysis. In recent years, the problem of dealing with unstructured data has become increasingly important in the context of processing and analysis of very large data sets, characterized as *big data*.

In our research, we deal with data coming from real-estate listings. Such data presented in real estate websites are, in many cases, unstructured. Each advertisement consists of loosely coupled words (terms, concepts) rather than full, grammatically correct sentences. The problem of mining data coming from real-estate listings was considered by us in our earlier papers (see [1], [2]). In this paper, we describe important aspects of a specialized computer tool created by us to automatize the transformation process of unstructured data, included in real-estate listings, into data arranged in a tabular form, as the information tables. In the presented approach, information tables represent information systems over ontological graphs defined in [3]. In this case, the domain knowledge in some form of ontologies is directly incorporated into information systems (understood in the Pawlak's sense, i.e., as knowledge representation systems [4]). In the classic Pawlak's information systems, there is a lack of information about data semantics, i.e., with each attribute describing objects of interest, only a set of its values is associated. In order to cover the meaning of data, the so-called simple information systems over ontological graphs and complex information systems over ontological graphs were formally defined in [3]. These definitions are recalled in Section II. In case of simple information systems over ontological graphs, values of a given attribute are concepts from the domain described by this attribute. The domain is modelled using an ontology. The ontology is presented, in a simplified way, by means of a graph structure, called an ontological graph. In such a graph, each node represents one concept from the ontology, whereas each edge represents a relation between two concepts. It is assumed that the ontological graph represents the whole domain of a given attribute, i.e., only concepts present in the ontological graph can become attribute values. In case of complex information systems, values of a given attribute are subgraphs (called local ontological subgraphs) of the ontological graph associated with the attribute. The domain knowledge covering some data semantics (e.g. semantic relations between attribute values) is very important for further analysis of data (cf., [5], [6]).

An important part of our tool is the ontology of places in Poland. A part of this ontology was described in [7]. We are developing the ontology of places in Poland covering variety of aspects of places, mainly administrative and socio-economic. The ontology is implemented using the OWL 2 Web Ontology Language [8]. The created ontology is the knowledge base providing some information about places where real-estates are located. A short description of this ontology is given in Section III.

II. Basic Definitions

As it was mentioned in Section I, we use information systems understood in the Pawlak's sense, i.e., knowledge representation systems [4]. A formal definition is as follows. An information system IS is a quadruple $IS = (U, A, \{V_a\}_{a \in A}, f_{inf})$, where:

- U is the nonempty, finite set of objects,
- A is the nonempty, finite set of attributes,
- $\{V_a\}_{a \in A}$ is the family of sets of attribute values,

- $f_{inf} : A \times U \to \bigcup_{a \in A} V_a$ is the information function such that $f_{inf}(a, u) \in V_a$ for each $a \in A$ and $u \in U$.

In [3], we defined simple information systems over ontological graphs and complex information systems over ontological graphs to directly include the knowledge about some semantics of attribute values describing objects of interest. This semantics is expressed in ontologies associated with attributes. In formal definitions, we use the notion of an ontological graph instead of the notion of ontology. This is justified because the notion of ontology is a much wider notion and a way to present ontology is not restricted to a graph structure describing concepts and relations between them. Formally, ontologies are defined on the basis of description logics, for example, in a form of axioms separated into three groups: terminological (*TBox*), assertional (*ABox*), and relational (*RBox*) (see e.g. [9]). Therefore, ontological graphs are a simplified way to represent a part of the knowledge included in ontologies. Let $\mathcal{O}$ be a given ontology. Formally, an ontological graph is a quadruple

$$OG = (\mathcal{C}, E, \mathcal{R}, \rho),$$

where:

- $\mathcal{C}$ is the nonempty, finite set of nodes representing concepts in the ontology $\mathcal{O}$,
- $E \subseteq \mathcal{C} \times \mathcal{C}$ is the finite set of edges representing semantic relations between concepts from $\mathcal{C}$,
- $\mathcal{R}$ is the family of semantic descriptions (in a natural language) of types of relations (represented by edges) between concepts,
- $\rho : E \to \mathcal{R}$ is the function assigning a semantic description of the relation to each edge.

Let $OG = (\mathcal{C}, E, \mathcal{R}, \rho)$ be an ontological graph. A local ontological subgraph of OG is a graph in a form $LOG = (\mathcal{C}_L, E_L, \mathcal{R}_L, \rho_L)$, where $\mathcal{C}_L \subseteq \mathcal{C}$, $E_L \subseteq E$, $\mathcal{R}_L \subseteq \mathcal{R}$, and ρ_L is a function ρ restricted to E_L.

Formally, a simple information system SIS^{OG} over ontological graphs is a tuple

$$SIS^{OG} = (U, A, \{OG_a\}_{a \in A}, f_{inf}),$$

where:

- U is the nonempty, finite set of objects,
- A is the nonempty, finite set of attributes,
- $\{OG_a\}_{a \in A}$ is the family of ontological graphs associated with attributes,
- $f_{inf} : A \times U \to \bigcup_{a \in A} \mathcal{C}_a$, is the information function such that $f_{inf}(a, u) \in \mathcal{C}_a$ for each $a \in A$ and $u \in U$, $\mathcal{C}_a$ is the set of concepts from the graph OG_a.

It is not necessary for an information function to be a total function. It means that only some of the concepts defined in the ontological graph associated with a given attribute appear as values of this attribute. However, the knowledge about semantic relations between these concepts and other concepts present in the domain described by this attribute is very important from the point of view of the processes of knowledge discovery from data.

Formally, a complex information system CIS^{OG} over ontological graphs is a tuple in the form

$$CIS^{OG} = (U, A, \{OG_a\}_{a \in A}, f_{inf}),$$

where:

- U is the nonempty, finite set of objects,
- A is the nonempty, finite set of attributes
- $\{OG_a\}_{a \in A}$ is the family of ontological graphs associated with attributes,
- $f_{inf} : A \times U \to \bigcup_{a \in A} \mathbb{LOG}_a$ is the information function such that $f(a, u) \in \mathbb{LOG}_a$ for each $a \in A$ and $u \in U$, $\mathbb{LOG}_a$ is the family of all local ontological subgraphs of the graph OG_a.

It is worth noting that we can consider also hybrid information systems, in which, ontological graphs are associated only with selected attributes.

In general, ontologies model varied semantic relations between concepts (cf. [10]). At the beginning, we focus special attention on those fundamental relations considered in linguistics which are called paradigmatic semantic relations (or paradigmatic relations shortly), i.e., synonymy, antonymy, hyponymy, hyperonymy, meronymy, and holonymy (cf. [11]). Such relations are, among others, distinguished in WordNet [12] - a large lexical database of English, as well as in the project called Wikisaurus [13] aiming at creating a thesaurus of semantically related terms. However, research in the areas as knowledge engineering, linguistics, logic, cognitive psychology, has recognized a variety of taxonomies of different types of semantic relations. A comprehensive review of the literature concerning semantic relations is given in [14].

For simplicity, instead of semantic descriptions (in a natural language) of types of paradigmatic relations, we use the following labels: $R_{\sim}$ - synonymy, $R_{\leftrightarrow}$ - antonymy, $R_{\lhd}$ - hyponymy, $R_{\rhd}$ - hyperonymy, $R_{\subset}$ - meronymy, $R_{\supset}$ - holonymy.

III. Ontology of Places

One of the main goals of our research is to build the ontology of places in Poland (see [7]). The created ontology is used in our tool described in Section IV. It is the knowledge base providing some information covering variety of aspects, mainly administrative and socio-economic, of places where real-estates are located. The ontology is created using Protégé [15] - a free, open source, platform-independent environment for creating and editing ontologies and knowledge bases implemented using the OWL 2 Web Ontology Language [8]. We assume that, as a target, the ontology will represent rich information about different aspects of places in Poland. Some examples of classes are collected in Tables I, II, and III. As the ontology is created in Polish, in tables, we have given the English terms corresponding to the Polish terms originally implemented in the ontology.

It is worth noting that, in the ontology of places, we have also distinguished functional types of communes (as classes):

- urban communes,
- urbanized communes,

TABLE I
THE OWL ONTOLOGY OF PLACES IN POLAND: CLASSES REPRESENTING CATEGORIES OF ADMINISTRATIVE DISTRICTS

English Term	Polish Term
Country	Kraj
Voivodship	Województwo
County	Powiat
City with County Status	Miasto na prawach powiatu/Powiat grodzki
Commune	Gmina
Urban Commune	Gmina miejska
Urban-Rural Commune	Gmina miejsko-wiejska
Rural Commune	Gmina wiejska

TABLE II
THE OWL ONTOLOGY OF PLACES IN POLAND: CLASSES REPRESENTING CATEGORIES OF PLACES

English Term	Polish Term
Place	Miejscowość
City/Town	Miasto
Village	Wieś

TABLE III
THE OWL ONTOLOGY OF PLACES IN POLAND: CLASSES REPRESENTING CATEGORIES OF WAYS

English Term	Polish Term
Way	Droga
Road/Route	Droga (publiczna)
Motorway/Freeway	Autostrada
Expressway	Droga ekspresowa
National Road	Droga krajowa
Voivodship Road	Droga wojewódzka
County Road	Droga powiatowa
Commune Road	Droga gminna

- multifunctional transitional communes,
- overwhelmingly agricultural communes,
- prevalently agricultural communes,
- tourism and recreational function communes,
- forestry function communes,
- mixed function communes.

This information is significant for the real-estate market. Other classes concern, among others, public institutions, rail and air transport, rivers, and physical regions of Poland.

In case of the ontology of places, we take into consideration three basic semantic relations between concepts:

- SUBCLASS-OF (hyponymy), also known as IS-A. If c SUBCLASS-OF c, it means that c is a kind of c (c is a more specialized concept than c), for example, *city* is a kind of *place*.
- PART-OF (meronymy). If c PART-OF c, it means that c is a part of c, for example, *commune* is a part of *county*.
- INSTANCE-OF. If i INSTANCE-OF c, it means that i is an instance (example) of c, for example, *Rzeszów* is an instance of *city*.

Basic semantic relations are used in our ontology of places to describe relationships covering the administrative aspects, for example, categories of administrative districts, categories of places, categories of roads, etc. Moreover, we can distinguish many specific semantic relations describing relationships covering economic and social aspects of places. Some important relations from the economic and social points of view are, for example, those representing access to roads as well as possession of airports, railway stations, schools, universities, courts, post offices, shopping centers, cinemas, theatres, concert halls, churches, monuments, etc.

IV. COMPUTER TOOL

A general scheme of the preprocessing procedure, implemented in our computer tool, that transforms unstructured data included in real-estate listings (on the website) into information tables representing simple or complex information systems over ontological graphs is shown in Figure 1 (cf. [1]). In the transformation procedure, we can distinguish several data preprocessing steps:

- *Acquisition and morphological operations* - acquisition of data from web sites, extraction of information concerning advertisements, and defining basic grammatical forms (roots) for particular words existing in advertisements. The words that occur in different grammatical forms are replaced with their invariable parts (roots).
- *Attributation* - assigning concepts (built from words determined on the basis of roots) to proper attributes as their values. A finding of proper attributes is made according to defined ontological graphs associated with attributes.
- *Deinstantiation* - replacing instances existing in advertisements with the most specific concepts (with respect to the hyponymy / hyperonymy relation) whose instances they are.

In case of advertisements, deinstantiation may primarily concern names of places. In our tool, deinstantiation is made on the basis of the knowledge included in the ontology of places. It enables us to replace a name of a given place with a category of an administrative district or a category of a place whose instance is represented by the name. It is worth noting that deinstantiation is an important step if we are interested in a more general knowledge derived from real-estate listings, for example, some client is interested only in houses in a village (not in a particular one). After deinstantiation, we obtain a simple or complex information system over ontological graphs. The proposed idea of data preprocessing steps, allowing transformation of an advertisement in the text format into an object in a simple information system over ontological graphs, is shown schematically in Figure 2.

The created tool is designed for the Java platform. In the first step (*Acquisition and morphological operations*), we have used two available libraries:

- *jsoup* - an open source library working with real-world HTML documents. It provides a convenient API for extracting and manipulating HTML data (see [16]).
- *Morfologik-stemming* - an open source library (see [17]) in which the stemming algorithm, based on the Porter algorithm [18], for Polish language is implemented. The

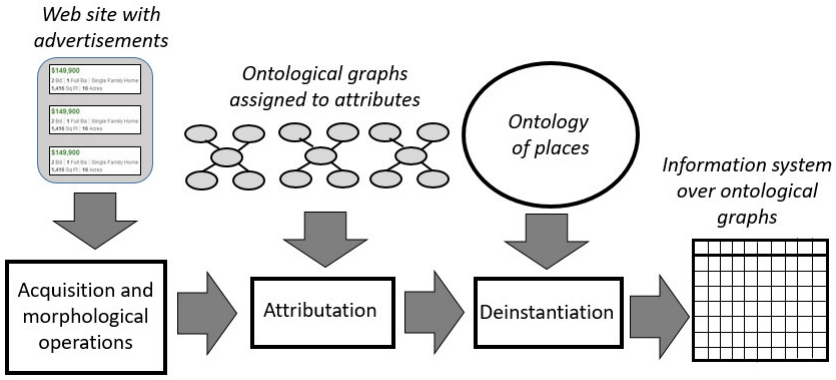


Fig. 1. A procedure for transformation of real-estate listings into simple/complex information systems over ontological graphs

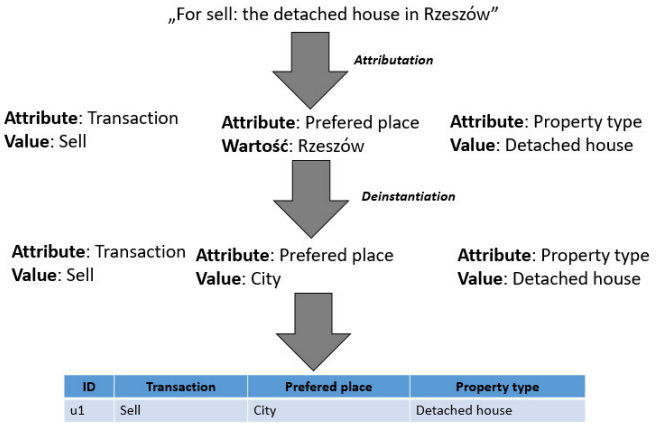


Fig. 2. An idea of data preprocessing steps allowing transformation of an advertisement in the text format into an object in a simple information system over ontological graphs

stemming algorithm enables us to extract basic grammatical forms (roots) for particular words existing in advertisements.

An important issue for the transformation process is to define a set of attributes which will be included in an information system over ontological graphs and to which data from real-estate listings will be assigned. For example, in case of advertisements concerning flats (apartments), we have distinguished the following set of attributes:

- Place (for instance: *Rzeszów*).
- Size (for instance: $75m^2$).
- Material of building (for instance: *brick*).
- Heating type (for instance: *gas heating*).
- Form of ownership (for instance: *property right*).
- The amount of rent (for instance: *350 PLN*).
- Market type (for instance: *secondary market*).
- Level (for instance: *ground floor*).
- State (for instance: *for renovation*).
- Number of bedrooms (for instance: *3*).
- Year of construction (for instance: *1960*).
- Type of building (for instance: *tenement*).
- Type of windows (for instance: *wooden windows*).
- Access date (for instance: *from August 2017*).
- Price (for instance: *180000 PLN*).
- Additional information (for instance *balcony* and *garage*).

As one can see, values of attributes are of different types. In case of selected attributes, we assign ontological graphs to them. Information about domains of attributes is collected in Table IV. One can see where an ontological graph is assigned. A part of the ontological graph describing types of heating is

TABLE IV
ATTRIBUTE DOMAINS IN CASE OF ADVERTISEMENTS CONCERNING FLATS (APARTMENTS)

Attribute name	Attribute domain
Place	Ontological graph
Size	Number
Material	Ontological graph
Heating type	Ontological graph
Form of ownership	Ontological graph
The amount of rent	Number
Market type	Ontological graph
Level	Number
State	Ontological graph
Number of bedrooms	Number
Year of construction	Number
Type of building	Ontological graph
Type of windows	Ontological graph
Access date	Date
Price	Number
Additional information	Text

shown in Figure 3. A part of the ontological graph describing types of windows is shown in Figure 4. It is worth noting that real-life ontological graphs are much more complex. Both ontological graphs are depicted as a hierarchy of classes in Protégé. The hierarchy of concepts presented in ontological graphs is important in case of comparison of attribute values. Information included in real-estate advertisements is given at different levels of abstraction, for example one can find *plastic windows* or *PVC windows*. In this case, the ontological graph provides information that *plastic windows* is a direct hyperonym of *PVC windows*.

In case of complex information systems over ontological graphs, we can omit the deinstantiation step. Instead of it, for an attribute describing a place where a given real-estate is

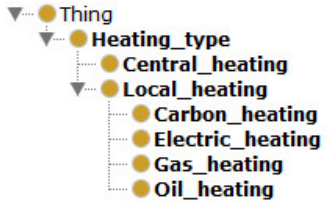


Fig. 3. Hierarchies of classes in Protégé for a part of the ontological graph describing types of heating

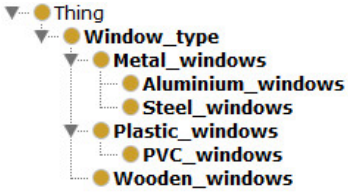


Fig. 4. Hierarchies of classes in Protégé for a part of the ontological graph describing types of windows

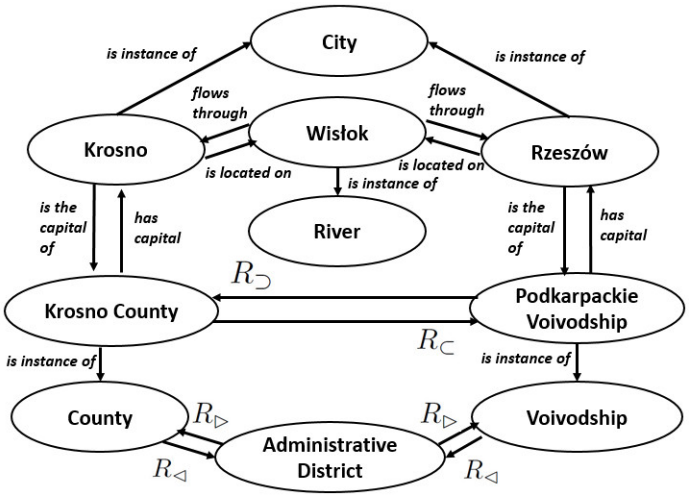


Fig. 5. A fragment of the ontological graph representing the ontology of places

located, we put values which are local ontological subgraphs of the ontological graph representing the ontology of places. Let us consider a fragment of such an ontological graph shown in Figure 5. Then, the values of the attribute can be local ontological subgraphs as shown in Figure 6. One can see, that, in this case, attribute values include more information about places, derived from the ontology of places.

Two examples of the approaches enabling us to compare values of attributes that are local ontological subgraphs were mentioned in [19]. The first approach can be based on different morphisms of local ontological subgraphs including fuzzy morphisms (see [20]). In the second approach, it is possible to find correspondences between local ontological subgraphs using ontology matching algorithms (see [21]).

V. Conclusions and Further Work

In the paper, we have described selected aspects of a computer tool for transformation of unstructured data included in real-estate listings into data arranged in a tabular form, as information tables. The transformation process enables us to prepare data for further analysis. The next step in our research is to design an engine for intelligent searching for advertisements. An important issue for future work is to implement more general mechanisms in the created computer tool enabling us to transform data coming from other short forms of texts, for example, other kinds of advertisements, tweets, microblogs, etc. In each case, a significant task is to define ontological graphs representing domains described by attributes.

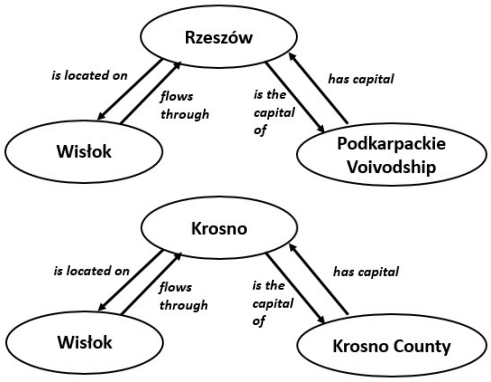


Fig. 6. Examples of local ontological subgraphs which are values of the attribute with which the ontological graph representing the ontology of places is associated

References

[1] K. Pancerz and O. Mich, "Mining real-estate listings based on decision systems over ontological graphs: Extended abstract," in *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2014)*, L. Popova-Zeugmann, Ed., Chemnitz, Germany, 2014, pp. 176–179.

[2] ——, "Numerical data clustering algorithms in mining real estate listings," *Barometr Regionalny. Analizy i prognozy*, vol. 12, pp. 43–50, 2014.

[3] K. Pancerz, "Toward information systems over ontological graphs," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Artificial Intelligence, J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra, and L. Polkowski, Eds. Berlin Heidelberg: Springer-Verlag, 2012, vol. 7413, pp. 243–248.

[4] Z. Pawlak, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991.

[5] S. Bloehdorn and A. Hotho, "Ontologies for machine learning," in *Handbook on Ontologies*, S. Staab and R. Studer, Eds. Berlin, Heidelberg: Springer, 2009, pp. 637–661.

[6] P. Ristoski and H. Paulheim, "Semantic web in data mining and knowledge discovery: A comprehensive survey," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 36, pp. 1–22, 2016.

[7] K. Pancerz, P. Grochowalski, and A. Derkacz, "Towards the ontology of places in Poland: an example of the Mazowieckie Voivodship," *Barometr Regionalny. Analizy i prognozy*, vol. 14, pp. 127–134, 2016.

[8] "OWL 2 Web Ontology Language: Structural specification and functional-style syntax (second edition)," W3C Recommendation, Tech. Rep., 2012.

[9] M. Krötzsch, F. Simancik, and I. Horrocks, "Description logics," *IEEE Intelligent Systems*, vol. 29, no. 1, pp. 12–19, 2014.

[10] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. Swartout, "Enabling technology for knowledge sharing," *AI Magazine*, vol. 12, no. 3, pp. 36–56, 1991.

[11] M. L. Murphy, Ed., *Semantic relations and the lexicon: antonymy, synonymy, and other paradigms*. Cambridge, UK: Cambridge University Press, 2003.
[12] C. Fellbaum, Ed., *WordNet - An Electronic Lexical Database*. MIT Press, 1998.
[13] The Wikisaurus Homepage: http://en.wiktionary.org/wiki/Wiktionary:Wikisaurus.
[14] V. Nastase, P. Nakov, D. O. Séaghdha, and S. Szpakowicz, *Semantic Relations Between Nominals*. Morgan & Claypool Publishers, 2013.
[15] M. Musen *et al.*, "The Protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015.
[16] The jsoup Homepage: https://jsoup.org.
[17] The Morfologik Homepage: https://github.com/morfologik.
[18] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130–137, 1980.
[19] K. Pancerz, "Some remarks on complex information systems over ontological graphs," in *Man-Machine Interactions 3*, ser. Advances in Intelligent Systems and Computing, A. Gruca, T. Czachórski, and S. Kozielski, Eds. Springer International Publishing, 2014, vol. 242, pp. 55–62.
[20] A. Perchant and I. Bloch, "Fuzzy morphisms between graphs," *Fuzzy Sets and Systems*, vol. 128, no. 2, pp. 149–168, 2002.
[21] J. Euzenat and P. Shvaiko, *Ontology Matching*. Berlin Heidelberg: Springer-Verlag, 2007.
[22] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, ser. Intelligent Systems Reference Library. Switzerland: Springer International Publishing, 2015, vol. 72.

Heuristic approach to online purchase prediction based on internet store visitors classification using data mining methods

Uladzimir Parkhimenka
Economics Department
BSUIR
Minsk, Belarus
parkhimenko@bsuir.by

Mikhail Tatur
Intellectual Processors Ltd.
Minsk, Belarus
tatur@i-proc.com

Anna Zhvakina
Electronic Computing Machines Department
BSUIR
Minsk, Belarus
antim07@mail.ru

***Abstract*—Last years research gave some preliminary results in approaches to customer online purchase prediction. However, it still remains unclear what exact set of features of data instances should be incorporated in a model and is enough for prediction, what is the best data mining method (algorithm) to use, how stable over time could be such a model, whether a model is transferable from one online store to another. This study is focused on a heuristic approach to dealing with the problem under conditions of such theoretical and methodological diversity in order to find a quick and inexpensive first approximation to the solution or at least to find useful patterns and facts in the data.**

Keywords—online purchase prediction; statistical classification; feature selection and design; automatic marketing decision-making; data mining & knowledge discovery

I. INTRODUCTION

Online stores provide tremendous amount of data to owners and managers concerning online visitors, their characteristics, clickstream behavior and financial outcomes of their actions. It is obvious to use these data for optimizing of marketing strategy in order to maximize sales and/or profit, but that is still rarely the case in real life especially for small non-global companies. One reason of that is lack of technical, human, time and financial resources to launch a real-time-marketing system or even to implement just a few data mining algorithms. Another reason is lack of certainty in real economic advantages of using data mining methodology over traditional approaches. The third reason is uncertainty with possibility of building robust models in this sphere in principle.

Nevertheless, in the past years much research has focused on the problem of online customer behavior prediction as a first step to the broader task of marketing strategy optimization using clickstream and other data. Factors influencing online purchase have been under intensive experimental research in order to construct minimal set of best predictors. Study have been also conducted on comparison of prediction power and accuracy of different data mining methods with building a combination of different methods for better prediction. Not only the mere fact of purchase / no purchase behavior has been studied, also the item and time of the next purchase.

Contrary to many years of research and some preliminary results it still remains unclear (i) what exact set of features of data instances should be incorporated in a model and is enough for prediction, (ii) what is the best data mining method (algorithm) to use, (iii) how stable over time could be such a model, (iv) whether a model is transferable from one online store to another.

The lack of conventional theory and model means for the companies the necessity of finding their own way to the online customer prediction problem. The purpose of this study was to propose a heuristic approach to dealing with the problem under conditions of such theoretical and methodological diversity in order to find a quick and inexpensive first approximation to the solution or at least to find useful patterns and facts in the data.

II. LITERATURE REVIEW

Zhao and Kulkarni review literature on factors influencing online shopping behavior and propose “a tentative model for consumer online behavior” [1]. In accordance with the model two groups of exogenous factors: customer profile (income, culture, lifestyle, gender, education level, ethnicity etc.) and online shop profile (the brand name of the shop, marketing communications stimuli, firm capability, product characteristics, and virtual store features, which include convenience, simplicity, information, accessibility) are

hypothesized to influence customer online purchasing behavior. Another factor (endogenous) in the model – trust. Finally, clickstream data are emphasized to be the most important and direct endogenous predictor of online purchase decision.

Lim, Byun, and Kim test such (traditional to the literature) predictors of online purchase as total visit time, the average time per page viewed, total number of clicks, total number of product-related pages viewed, as well as a newly constructed predictor "density of a session based on graph analysis" [2]. They found that all predictors except total visit time are useful and statistically significant to differentiate between "purchase" and "not purchase" group of visitors, but density of a session based on graph analysis predicts purchase behavior better compared to other predictors.

Vieira uses for purchase prediction parameters constructed from clickstream data, such as session duration before purchase, click to buy ratio, median number of sessions before buy, number of clicks in a session, number of page views in the last week, etc. and finds improvements of using deep neural networks for purchase prediction over existing algorithms [3].

E. Kim, W. Kim, and Lee propose to predict customer's purchase behavior by combining multiple classifiers based on genetic algorithm. Their model incorporates 10 demographic features of a customer (age, gender, education, occupation, etc.) and 5 transactional features (frequency of purchasing target product, items purchased frequently, items viewed frequently, etc.) [4].

Zhang, Pang, Shi, and Wang focus on prediction of future purchases given the historical data of a user [5]. In their research, they design five groups of features: counting features, ratio features, flag features, and global features. They use regression (Gradient Boosting Regression Tree) and classification (Logistic Regression and Random Forest) methods as single models, which are then aggregated in a two-stage approach, using linear regression for blending, and finally employ a linear ensemble of blended models. The approach is rather accurate in the first 3 days, but predicting difficulty significantly increases after that [5].

Kooti, Lerman, Aiello, Grbovic, Djuric, and Radosavljevic examine demographic, temporal, and social factors affecting online purchases and conclude that "click and browsing features represent only a weak proxy of user's purchase intent" [6]. They also build the model (using 55 features of a data instance) to predict the time of the next purchase, and the amount that will be spent on that purchase. However, they give up predicting the exact time, price of a purchase (due to the task difficulty), and concentrate on the "simpler classification task of predicting the class of the purchase among a finite number of predefined price or time intervals" [6].

Lee, Ha, Han, Rha, and Kwon show that consumers' behavior features (e.g., item viewcount or cart usage) are more important predictors of their purchases than item features (e.g., price or popularity). Their proposed model for predicting the actual purchase of consumers achieved over 80% accuracy across the four shopping sites [7].

Bedi, Kaur, Lal examine the influence of such (non-behavioral and non-demographic) variables as visual design, website interactivity, website privacy and security and ease of use on customer online purchase intention. They find that such influence is mediated by perceived web enjoyment and attitude and is statistically significant with overall model's $R^2 = 0.67$ [8].

Kim, Im, Han criticize clickstream-based prediction models for several limitations. In their study, they propose a new method for purchase prediction that combines information theory with machine learning techniques [9].

III. HEURISTIC APPROACH TO ONLINE PURCHASE PREDICTION

The proposed heuristic approach could be summarized by seven principles:

(1) The purpose of online purchase prediction is not the prediction itself, but marketing strategy optimization based on behavioral data. Therefore, the main performance criteria is not just predictive accuracy, but real possibility to positively influence the revenue stream of an online store with reasonable cost of implementation.

(2) The solution of the prediction problem should be seen as an iteration process, preferably starting from simple tasks, small datasets, few features, simple algorithms and ending with more complicated model and data if needed.

(3) There is a hierarchy of factors influencing online purchase decision. "Behavioral" factors (predictors) are preferable to demographic, psychological, social, temporal and other "descriptive" factors. The latter have indirect influence and are difficult to track for each customer, the former have direct influence, though are not good to track long-term behavior.

(4) Not a "customer", but a "visit" should be taken as a reasonable unit of analysis. This approach lack of opportunity to employ historical data on a customer, but is much simpler and avoid the problem of "identification" of multiple users in data from just one user logged on different devices, cleared web cookies or changed the name.

(5) Quantitative features are preferable to qualitative ones. Qualitative features should be avoided as much as possible due to algorithmic and computational difficulty.

(6) Because of small fraction of positive outcomes ("purchase") in comparison to negative ones ("not purchase") the procedure of oversampling for correct work of data mining classification methods seems to be crucial here [10].

(7) There is a need of taking into consideration computational capability available.

IV. Experimental results

A. Background

In order to verify applicability and usefulness of the proposed heuristic approach it has been tested on real business data which had been kindly provided by an online store from CIS region on the condition of anonymity.

For data processing R language and RStudio have been used.

B. Data

Raw data are presented in several .json files. The data are full, but with very short coverage of online transactions in time (Table I).

TABLE I. Raw data characteristics

File name	File content	Number of instances	Time span
XXX_categories	Product categories	77	–
XXX_offers	Products description	2585	–
XXX_view_page	Data on all pages viewed by a visitor	1877	From 24-08-2016 till 25-08-2016
XXX_view_product	Data on product pages viewed by a visitor	15481	From 31-07-2016 till 26-08-2016

In accordance to the 4th principle of our heuristic approach a "visit" has been taken as a unit of analysis. A "visit" means here a sequence of activities made by a user including product pages view, other (non-product) pages view and target action ("purchase" or "not purchase").

Analysis shows that there are only 428 unique visits in period from 24-08-2016 till 25-08-2016. Among them only 6 visits ended with a purchase (1,4%).

In accordance with 3rd and 5th principles of the heuristic approach 9 features, predominantly behavioral and quantitative, have been selected for the analysis (Table II).

TABLE II. Selected features for the analysis

Feature name	Feature type	Value range	Commentary
Visit duration in seconds	Numeric	$(0, +\infty)$	Calculated as a sum of duration for all page views in the current visit
Hour of visit begin	Numeric	[0, 23]	Taken from raw data
Previous visits count	Numeric	$[0, +\infty)$	Taken from raw data
Previous purchases volume in rubles	Numeric	$[0, +\infty)$	Taken from raw data
Previous purchases count	Numeric	$[0, +\infty)$	Taken from raw data
Time lapsed from the previous visit	Categorical	{no visits previously, less than 1 day ago, from 1 to 2 days ago, from 2 to 3 days ago, from 3 to 5 days ago, from 5 to 9 days ago, from 10 to 30 days ago, more than 30 days ago}	Taken from raw data
Number of pages viewed during the current visit	Numeric	$(0, +\infty)$	Calculated from raw data
Number of products viewed during the current visit	Numeric	$(0, +\infty)$	Calculated from raw data
Current visit outcome	Logical	1 – purchase done 0 – purchase not done	Taken from raw data

C. Preprocessing

Data preprocessing has been done in four steps:

(1) A new dataset of 428 instances (only unique visits in period from 24-08-2016 till 25-08-2016 with only 9 target features) has been constructed from the data.

(2) One instance which had two product page views and only one page view has been deleted from the dataset. (This discrepancy comes from discrepancy in time period covered in different files with data, see Table I).

(3) A feature "Current visit outcome" has been transformed to a class label (purchase / not purchase).

(4) In accordance to the 6th principle of the heuristic approach the procedure of oversampling has been implemented using R function *ovun.sample()* from the package *ROSE* and resulting in a new dataset with 854 instances balanced to equal shares of both classes (purchase / not purchase).

D. Methods and algorithms

The new dataset has been randomly shuffled and divided in two groups: 85% – a training set, 15% – a test set. Then it has been processed using different methods of data mining:

- Rpart (Recursive Partitioning and Regression Trees) from the package *rpart.*
- Ctree (Conditional Inference Trees) from the package *party*.
- randomForest (Random Forest) from the package *randomForest*.
- glm (Generalized Linear Models / Logistic Regression) from the package *stats.*
- kknn (K-neigbours) from the package *kknn*.

Five cycles of cross-validation have been conducted for each method.

E. Results

All five implemented methods have successfully converged with extremely high accuracy of prediction (Table III).

TABLE III. ACCURACY OF METHODS EMPLOYED

Method name in R	**Average accuracy within 5-fold cross-validation**
Rpart	0,9813
Ctree	0,9797
Random Forest	0,9953
Logistic Regression	0,8703
K-neigbours	0,9786

Such a high level of accuracy does not reflect neither real high predictability of online behavior, nor superficial power of used methods, but is rooted as we think in small size of dataset and small time period covered. Nevertheless all methods show some important facts which should be taken into account at least as a first approximation to the final solution:

- There is enough only three features to make rather accurate prediction of an outcome of an e-store visit: Number of pages viewed during the current visit, Number of products viewed during the current visit, Previous visits count.
- There is a high probability of an online purchase (0.97) if at the same time Number of pages viewed during the current visit > 8 and Number of products viewed during the current visit <6 and Previous visits count <2.
- Otherwise a visitor doesn't make a purchase.
- In some cases of the cross-validation procedure, some methods introduced an additional fourth feature to the prediction model – Visit duration in seconds – without any serious subsequences to the results.
- All mentioned above makes possible to talk about implementation of real-time marketing system at the given online store. It seems like such a system would be rather simple because only 3 or 4 features of a visitor are included in the model. The question of what kind of a marketing strategy (decision) should be offered by a system based on the prediction is open and stay out of the scope of this article.

V. CONCLUSIONS AND FUTURE WORK

The main result of this study is applicability of the proposed heuristic approach to find a quick and inexpensive first approximation to the prediction of online customers purchase behavior.

Limitations of the study are obvious: small dataset, low fraction of positive outcomes (just 6 instances), narrow time period (just two days), only one internet store data.

In the future, there should be approbation of the heuristic approach on the real-world data (i) with bigger time coverage, (ii) taken from different online stores. In addition, there is a need of future research of principle stability of online purchase behavior of internet customers.

VI. ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for their constructive comments, which helped us to plan our future research.

REFERENCES

[1] Fan Zhao, and Sagar S. Kulkarni, "Predicting Online Customer Shopping Behavior", Emerging Trends and Challenges in Information Technology Management, Volume 1, 2006.

[2] Miseon Lim, Hyunsoo Byun, and Jinhwa Kim, "A Web Usage Mining for Modeling Buying Behavior at a Web Store using Network Analysis", Indian Journal of Science and Technology; Volume 8, Issue 25, October 2015.

[3] Armando Vieira, "Predicting online user behaviour using deep learning algorithms", arXiv:1511.06247v3 [cs.LG] 26 May 2016.

[4] Eunju Kim, Wooju Kim, Yillbyung Lee, "Combination of multiple classifiers for the customer's purchase behavior prediction", Decision Support Systems, 34 (2002), pp. 167–175.

[5] Yuyu Zhang, Liang Pang, Lei Shi, and Bin Wang, "Large Scale Purchase Prediction with Historical User Actions on B2C Online Retail Platform", arXiv:1408.6515v3 [cs.LG] 4 Mar 2015.

[6] Farshad Kooti, Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, "Portrait of an Online Shopper: Understanding and Predicting Consumer Behavior", WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

[7] Munyoung Lee, Taehoon Ha, Jinyoung Han, Jong-Youn Rha, and Ted "Taekyoung" Kwon, "Online Footsteps to Purchase: Exploring Consumer Behaviors on Online Shopping Sites", WebSci'15, June 28–July 01, 2015, Oxford, United Kingdom.

[8] Sarbjit Singh Bedi, Sukhwinder Kaur, Amit Kumar Lal, "Understanding Web Experience and Perceived Web Enjoyment as Antecedents of Online Purchase Intention", Global Business Review, 2017, Vol 18, Issue 2, pp. 465 – 477.

[9] Minsung Kim, Il Im, Sangman Han, "Purchase Prediction by Analyzing Users' Online Behaviors Using Machine Learning and Information Theory Approaches", Asia Pacific Journal of Information Systems, 2016, Vol. 26, No. 1, pp. 66-79.

[10] Chawla, Nitesh V. (2010) Data Mining for Imbalanced Datasets: An Overview, In: Maimon, Oded; Rokach, Lior (Eds) Data Mining and Knowledge Discovery Handbook, Springer, pp. 875-886.

Statistical Analysis of Utilization of Landsat Data in Observation of Small Inland Water Bodies

Miroslav Pasler, Jitka Komarkova, Ivana Cermakova
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
{miroslav.pasler, jitka.komarkova, ivana.cermakova}@upce.cz

***Abstract*—Spatial data are very important for assuring quality of decision-making. Satellite images are used in many applications as a source of spatial data. Precise knowledge of number of usable (suitable) satellite images suitable for analyses is important. Statistics can be used to predict number of suitable images for a particular time period. The determination which images are suitable enough for analysis is the first step. The next step is to analyze the results of the determination of image suitability, especially with respect to distribution of images in time. This is the main topic of this paper. Statistical analyses methods are used. The main aim is to predict how many suitable Landsat images can be expected in the case of small inland water bodies observation. A part of Pardubice region in the Czech Republic is used within the case study to demonstrate the proposed procedure. The proposed procedure can be applied to other areas and regions as well.**

***Keywords*—Landsat; imagery; small water body; remote sensing; statistics**

I. INTRODUCTION

Water quality monitoring and measuring is historically very significant branch in the field of satellite based remote sensing. From the very beginning of satellite imaging, it is widely used in scientific research in water quality observation. In fact, it is one of the most used method of data collection. Works of Bukata, Harris and Bruton [2, 3] are the very early works in the field dealing still with Landsat1 (ERTS-1) data. There were many of works during 70's and 80's as stated by Middleton and Marcell [4]. According to analysis of publications, the main increase of studies came along with Landsat 5 and later Landsat 7 and 8 launches, e.g. for land use changes detection [8], ice dynamics [9], surface water proportion in inland river basins [10], identification of impervious features [11], vegetative cover loss [12], etc. While new methods including UAV and Lidar are approaching, the satellite based remote sensing of small water bodies still plays significant role in nowadays research. Its role is even more important considering the works dealing with Sentinel satellites [1], [7]. However, according to analysis of keywords, the most popular and the most used system in this branch is still Landsat system with Landsat 8 in the leading position of the current research.

Previous works of authors [5, 6] pointed out that there is a lack of scientific researches focusing on evaluation of suitability of remotely sensed images for small water bodies observation. There is a theoretical research available, which deals mostly with design of models for determination of chosen parameters of water quality. There are also practical applications and case studies available in this field. Today, there is a high need for periodic observation of the areas with small inland water bodies. It means that area of water bodies cannot be covered by clouds or image errors. Clouds and image errors are the most important sources of gaps in time series of satellite imagery. Gaps in the time series can influence observation of an area of interest in time. So, there is a need to calculate in advance how long could be the gap be or how many images will be with the highest probability usable in a specific time period.

The previous works of authors [5, 6] are mostly focused on methodology of an evaluation of suitability of Landsat images, on the determinants of the suitability of images and their influencers. Cloud cover and the black image gaps are identified as the main influencers on the suitability of images in the case of Landsat 7 images. Based on the results and on the nature of observation of small inland water bodies, there is also a difference between influence of these factors to larger water bodies observation and influence of these factors in cases of observation of areas with smaller water bodies, which are differently distributed in an area of interest. This is caused by a possible specific distribution of clouds and the black gaps in the image with respect to distribution of the water bodies themselves. It leads to the fact that the percentage of cloud cover in a whole image is not sufficient enough to determine the suitability of images.

The main aim of the paper is to propose a suitable procedure of calculation of a number of usable satellite images (i.e. with visible small inland water bodies) in advance with taking into account cloud cover and errors in sensing. There is no serious research correlating the cloud cover over an image and usability of the image for small inland water bodies observation. There is no statistical analysis available that allows to determine the time gap between images, probability of having a usable image within the given time period or the most probable number of usable images within a specific time period as well. The proposed procedure will be demonstrated on a case study.

Structure of the paper is as follows: the second chapter describes used data and methods. Next, hypotheses and

research questions are stated. Next chapter describes the proposed procedure. The following chapter provides answers to the research questions and hypotheses. Conclusion follows.

II. Data and Methods

A. Area of Interest and Data

An area of interest is located to the north of the city of Pardubice and it contains several small inland water bodies – ponds and small lakes created by mining of sand. The water bodies are predominantly ponds designed for fish breeding and for outdoor swimming. Total water area is approximately 0.1 km^2 and it is spread over a region with an area of 150 km^2 [5, 6]. The water bodies are shown in Landsat 8 image and in map (see Fig. 1).

Figure 1 Landsat 8 image of the chosen water bodies

Data for analyses presented in this paper consist of Landsat 7 and 8 satellite images of the area of interest. Landsat satellites provide moderate-resolution imagery of the Earth's land surface in several various spectrums.

In the case of the study described by this paper, analyzed data set consists of 215 Landsat 7 and Landsat 8 images recorded from March 29th, 2013 to November 15th, 2015. The time gap between particular images is not constant due to used path/row coordinates, used satellites and their flight offset. The time spacing is neither random. It is periodical set of days between the images lasting from 1 to 8 days (16 days in the case of a missing image in rare cases).

B. Methods

For every image, there is calculated percentage of clear water surface as it was mentioned in the previous work [6]. The percentage is an input for determination if the image is usable or not for small water bodies observation. The minimum percentage of the clear water surface can differ with respect to planned analyses, observed parameters and other factors. In this paper, suitability of images is evaluated from the point of different values of the minimum percentage of the clear water surface. The parameter is called γ further in this paper and its values are as follows:

$$\gamma_a \in 0;1;\ a \in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 75, 80, 85, 90, 95\} \quad (1)$$

The a index represents the minimum required percentage of clear water surface to consider the image as usable. The usability itself (γ_a) is considered as a dichotomic variable where 1 represents usable image and 0 unusable image. It implies that the higher value of α parameter the lower number of usable images.

There are calculated numbers of days between following usable images for different values of index a of parameter γ. The result represents time gap between the usable images and it is called δ_a further in this paper, where index a has the same meaning as in case of γ_a parameter. Parameter δ_a is calculated for every image as well as parameter γ_a so the minimum value is 1 and hypothetic maximum depends on the number of days between the first and the last image in data series.

Except the calculated area of free water surface, γ_a and δ_a parameters, there are used other parameters, which represent metadata for the image. These variables are part of the data set due to their potential correlation to outputs and to each other. The variables, their possible values, their type and their brief description are described by the Table 1.

Common tools of statistical induction are used for the data analysis. The significance value of all tests is set to $\alpha = 0.05$.

Let's assume that the acceptability γ_a is a random variable coming from a set with Bernoulli distribution, where the success of random event is presence of a usable image $\gamma_a = 1$ with probability π. So the point estimation is calculated by (2) and interval estimation is calculated by (3) and (4):

$$\pi = \frac{x}{n} \quad (2)$$

where x is number of usable images and n is number of all images in the set.

TABLE I. Variables of Used Data Set

Variable name	Type	Values	Description
Percentage	numeric, continuous	<0 - 100> [%]	Represents percentage of water surface in the image which is not devaluated by presence of clouds, gaps or other influences
γ_a	dichotomic	{0, 1}	Described above
δ_a	numeric, discrete	<1 – 920> [days]	Described above
Date	date	-	Day, month and year, when the image was taken

Satellite	nominal	{L7, L8}	The satellite taking the image
Path	nominal	{191, 192}	Represents the path of the satellite taking the image
Azimuth	numeric, continuous	<0 - 360> degrees	The angle of sun when the image was taken
Elevation	numeric, continuous	<0 - 90> degrees	The elevation of sun when the image was taken
ACCA	numeric, continuous	<0 – 100> [%]	Percentage of clouds over whole image calculated by Automated Cloud Cover Assessment

$$p - \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < \pi < p + \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad (3)$$

$$p - \frac{1}{2n} - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}} < \pi \quad (4)$$

Equations (3) and (4) represent 95 % estimations of parameter π.

Let's assume that parameter n_k is an average number of images taken in a period indexed with k (1 – month, 2 – quarter, 3 – year) so γ_a is a random variable with binomial distribution B(n,π). So, the estimation of this distribution is calculated using the following equations:

$$E(X) = Np \quad (5)$$

for mean, and:

$$D(X) = Np(1-p) \quad (6)$$

for variance.

Equation (7) represents probability that just r images will be usable:

$$P(X = r) = \binom{N}{r} p^r (1-p)^{N-r} \quad (7)$$

So the values of distribution function are estimated as follows:

$$P(X \leq r) = \sum_{k=0}^{r} \binom{N}{k} p^k (1-p)^{N-k} \quad (8)$$

with an assumption of the equation (9).

$$P(X = s) = 1 - P(X \leq r); s = r + 1 \quad (9)$$

III. Hypotheses and Research Questions

The used methods of statistical induction should give an answer to the question described in the introduction and give quantification of the number of usable images (with visible surfaces of small water bodies). The hypotheses are formulated in the following way according to the problem and the described dataset:

1. There is a correlation between values of variables Satellite and Path to variable Percentage
2. There is a correlation between variables ACCA and Percentage
3. There is a significant trend in the variable Percentage considered as time series.

Except for the hypotheses, there are formulated the following research questions to be quantified using the described methods:

1. How long can possibly be the time period without any usable image?
2. What is the probability that in specific period (month, quarter, year) there will be stated number of suitable (usable) images with respect to parameter a?
3. How many images will be suitable (usable) in specific time period with the highest probability?
4. What is the probability of predetermined minimum number of suitable (usable) images in the chosen time period?

IV. Case Study and Data Processing

The area of interest and used data are described in the previous chapter along with the used methods.

The basic descriptive statistics is calculated for the data set. The results for chosen variables are summarized in the Table 2 where summarized number of γ_a, minimum, maximum and average values of δ_a are shown.

The decreasing linear trend of number of usable images according to the parameter a is shown in the Fig. 2 (the left part). There is also shown dependency between a and average time gap between suitable (usable) images (the right part), which is increasing exponentially.

TABLE II. BASIC DESCRIPTIVE STATISTICS (MINIMUM, MAXIMUM AND AVERAGE NUMBER OF DAYS WITHOUT USABLE DATA IS SHOWN)

a		5	10	15	20	25	30	40	50	60	70	75	80	85	90	95
$\gamma_{a\,(sum)}$		85	80	77	71	65	62	60	45	35	29	25	20	20	16	8
δ_a	**Min**	1	1	1	1	1	1	1	1	1	1	1	16	16	16	16
	Max	39	39	39	55	55	55	56	57	105	113	113	113	113	154	273
	Ave	10.8	11.5	11.9	13	14.2	14.8	15.3	20.4	26.3	30.7	35.4	43.8	43.8	54.1	102

	Average	**Minimum**	**Maximum**	**σ**	**N**
Percentage	21.631	**0**	**100**	32.358	**215**
ACCA	58.869	0	99	29.967	**215**

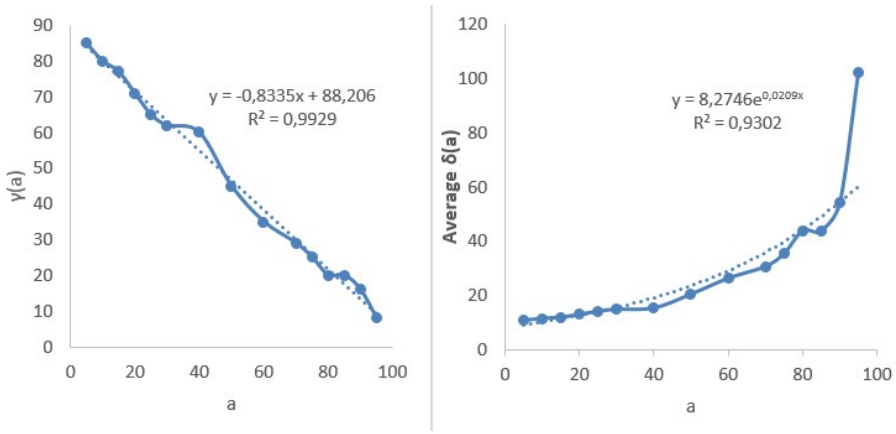


Figure 2 Number of acceptable images and time gap between them dependent on *a* parameter.

Table 3. Parameters of Binomial Distribution of Variable γ_a

a	5	25	50	75	95
			n=7		
E(X)	2.765	2.114	1.463	0.812	0.259
D(X)	1.673	1.475	1.157	0.718	0.249
r	3	2	1	0	0
P(X=r)	0.2889	0.317	0.358	0.422	0.768
s	1	0	0	0	0
P(s=1)	0.971	0.92	0.807	0.578	0.232
			n=21		
E(X)	8.295	6.342	4.389	2.436	0.777
D(X)	5.018	4.427	3.472	2.153	0.748
r	8	6	4	2	0
P(X=r)	0.175	0.187	0.212	0.271	0.453
s	5	3	2	0	0
P(s=1)	0.999	0.999	0.993	0.925	0.547
			n=83		
E(X)	32.785	25.066	17.347	9.628	3.071
D(X)	19.834	17.496	13.721	8.511	2.957
r	33	25	17	9	3
P(X=r)	0.089	0.095	0.107	0.136	0.228
s	26	18	11	5	1
P(s=1)	0.999	0.999	0.999	0.999	0.956

The correlation between variables, namely between ACCA and Percentage, is calculated and the influence of variables Path and Satellite is tested by Kruskal-Wallis test. The correlation coefficient for relation between ACCA and Percentage is -0.769.

There is tested a presence of trend, cyclic and seasonal component in variable Percentage taken as time series. The autocorrelation between values of Percentage is also tested. There is found no autocorrelation as well as no significant component of decomposition of the time series.

According to equations (5) to (9), there are estimated parameters of variable γ_a as a variable with binomial distribution B(n,π). They are calculated for different values of the parameter *a* and for different values of *n*. The parameter *n*

represents an average number of images taken in the time period (7 – month, 21 – quarter, 83 – year). The parameter π is taken from estimation calculated by means of the equation (2). Table 3 shows the results for chosen values of a. There is calculated expected value, variance, the most probable number of usable images (r), probability of occurrence of r (P(X=r)), number of usable images occurred with the probability higher than 0.95 (s) and probability of occurrence of at least one usable image (P(s=1)).

V. Results

Data set analyses provides the following results:

1. Neither variable Satellite, nor variable Path have any significant influence to percentage of clear water surface
2. There is negative relation between the variables. The correlation coefficient is -0.769
3. There is no significant trend in variable Percentage

According to the processing of data, the answers for the questions are following:

1. Minimum, maximum and average length of the time period without usable data is shown in the Table 2 with variable δ_a. The values are in days
2. The probability is shown in the Table 3. The probability can be calculated using parameters of binomial distribution shown in the Table 3
3. Values of r in the Table 3 describe the resulting answer
4. Table 3 describes the probability for at least one usable image in the period (P(s=1))

VI. Conclusion

Problem of a real number of usable satellite images, i.e. without cloud cover, is a significant influencer of inland water bodies observation based on remotely sensed data.

The paper proposes a method for evaluation of suitability (usability) of the time series of remotely sensed images for small water bodies observation. It is based on statistical induction. The statistical methods are demonstrated on the case study (observation of small ponds near Pardubice, the Czech Republic) for calculations over the dataset consisting of images coming from Landsat 7 and Landsat 8.

The calculations also show that the real number of usable images is relatively low and it depends on the minimum percentage of clear water surface necessary to next analyses.

The proposed procedure is generalizable. It can be used for further research based on different data sets, i.e. for other satellite data sets with comparable time resolution, for longer time period or different areas of interest (wit focus on small inland water bodies). The calculated values of γ_a show it is a random variable with binomial distribution. It provides possibility to calculate the probabilities and to give quantitative answers for the given questions.

Acknowledgment

This research is supported by University of Pardubice, SGS_2017_17 project.

References

[1] P.L. Brezonik et al, “Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters,” Remote Sensing of Environment, vol. 157, pp. 199-215, February 01, 2015.

[2] R. P. Bukata, G. P. Harris, and J. E. Bruton, “The detection of suspended solids and chlorophyll-a utilizing digital multispectral ERTS-1 data,” 2nd Can. Symp. Remote Sensing, 1974, pp. 552–564.

[3] R. P. Bukata, G. P. Harris, and J. E. Bruton, “Satellite-observations of water-quality,” Transportation engineering journal of asce, pp. 537-554, 1976.

[4] E. M. Middleton, and R. F. Marcell, [Online], “Literature relevant to remote sensing of water quality,” 1983, Available at: https://ntrs.nasa.gov/search.jsp?R=19830026142 [cited 2017-02-05].

[5] M. Pásler, J. Komárková, P. Sedlák, “Comparison of possibilities of UAV and Landsat in observation of small inland water bodies,” International Conference on Information Society, i-Society 2015, pp. 45-49.

[6] M. Pásler, and J. Komárková, “Utilization of Landsat data for water quality observation in small inland water bodies,” International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, vol. 41, pp. 373-377, 2016.

[7] M. Salama, M. Radwan, and R. van der Velde, “A hydro-optical model for deriving water quality variables from satellite images (HydroSat): A case study of the Nile River demonstrating the future Sentinel-2 capabilities,” Phys. Chem. Earth, Parts, pp. 224-232, 2012.

[8] P. Sedlák, Z. Szczyrba, E. Kudrnovský, “Spatial temporal changes of land use in postcommunist towns with remote sensing data - case study of Olomouc city”, Symposium Remote Sensing of Urban Area. Regensburg, Germany, pp. 176 – 178, 2003.

[9] W. Van Wychen et al., “Variability in ice motion and dynamic discharge from Devon Ice Cap, Nunavut, Canada”, Journal of Glaciology, vol. 63, pp. 436-449, June 2017.

[10] S. D. Wang et al, “A Simple Enhanced Water Index (EWI) for Percent Surface Water Estimation Using Landsat Data”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, pp. 90-97, January 2015.

[11] Z. Q. Wang, C.C. Gang, X.L. Li, Y.Z. Chen, and J. L. Li, “Application of a normalized difference impervious index (NDII) to extract urban impervious surface features based on Landsat TM images”, International Journal of Remote Sensing, vol. 36, pp. 1055-1069, 2015.

[12] Q. Ying et al., “Global bare ground gain from 2000 to 2012 using Landsat imagery”, Remote Sensing of Environment, vol. 194, pp.161-176, June 01, 2017.

Modelling of pedestrian queuing behaviour independent of movement model utilising BDI reasoning in ABAsim architecture

Marek Pecho[1], Norbert Adamko[1] and Michal Varga[2]
[1]Department of Mathematical Methods and Operation Research
[2]Department of Informatics
Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovak Republic
Pecho.Marek@gmail.com
{Norbert.Adamko, Michal.Varga}@fri.uniza.sk

***Abstract*—In a pedestrian simulation models, the waiting for service represents special behaviour that should be carefully modelled. This paper presents an approach to the modelling of queuing behaviour of pedestrians with no modifications to the movement model. The queuing behaviour of a pedestrian is defined on the tactical layer of pedestrian's reasoning, leaving the operational movement model intact. This approach can then be easily adapted to microscopic, mesoscopic as well as hybrid movement models. Two-layer agent based architecture ABAsim with the support of BDI reasoning model of agents is utilized. Plans on the tactical layer of reasoning that define the queuing behaviour of agent will be discussed. Different types of queue creation and its automatic organisation independent of the movement model will be also presented.**

Keywords—pedestrian simulation, queuing model, microscopic movement model, mesoscopic movement model

I. INTRODUCTION

Many systems with high density of pedestrians (such as airports, railway stations or stadiums) contain service places, operations of which is vital to the performance of the whole system. Thus, modelling such systems also requires proper modelling of the behaviour of pedestrians at these service places. Due to usually limited amount of resources that are utilised to serve customers at service places, queues of customers are being formed. Queueing behaviour of people depends on many factors, including local customs, organisation by authorities (e.g. queues at airport security checks), free space available at the location or privacy issues, which all need to be taken into account to guarantee a valid model.

In this paper, we present an approach to modelling of queueing behaviour implemented in our pedestrians simulation model PedSim. This model is based on our ABAsim [1] architecture of simulation models and utilises BDI paradigm to model agents' reasoning. Our approach defines the queuing behaviour on tactical level of agent's reasoning and is independent of the movement model. This supports the ability of PedSim to create hybrid models with various granularity of movement modelling (e.g. combination of microscopic and mesoscopic modelling of movement). In order to explain our approach, we first present basic principles of ABAsim architecture, our model of agent's reasoning that is based on BDI paradigm and also briefly describe the pedestrian model employed.

II. BDI PARADIGM IN ABASIM ARCHITECTURE

The agent based ABAsim architecture supports creation of models of service systems and utilises two distinct types of agents:

1. *Managing agent* is a component responsible for specific part of modelled system; it manages subordinated agents, entities and events of its subsystem. Managing agents are organised in a static hierarchical structure and cooperate to fulfil given system goal. Topology of agent hierarchy usually reflects the structure of modelled system; this is especially useful for modelling of service systems. Managing agent does not need to have a physical representation in modelled system, e.g. it can represent a governing body such as office. Every managing agent can control a group of dynamic agents.
2. *Dynamic agent* represents an autonomous intelligent entity of the model capable to fully control its own behaviour and reach given goals. Dynamic agent is at any time subordinated to exactly one managing agent. During the execution of the model, the dynamic agent can be transferred to other managing agent (e.g. when pedestrian is entering different infrastructure area, he might be transferred to a managing agent that is responsible for this area of the model). The superior managing agent assigns mandatory goals to the dynamic agent. Dynamic agent can also generate its own goals, but these must be in no conflict with the goals given by its managing agent.

Figure 1 depicts the hierarchy of agents used in presented model. For the sake of brevity, submodel of *Infrastructure* agent is simplified (more details can be found in [2]). The crucial submodel is the one of agent *Movement*, which will be described in more detail in chapter III.

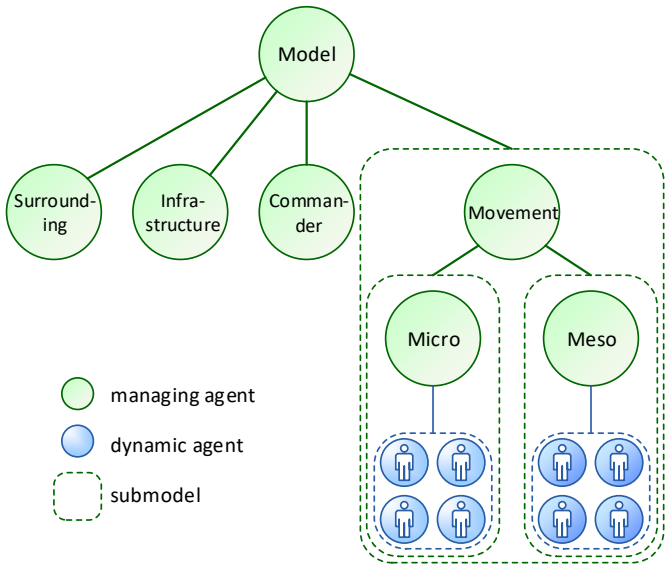


Figure 1. Simplified agent hierarchy of pedestrian simulation model

A. BDI reasoning

In order to support modelling of intelligent entities, the ABAsim architecture has been augmented by BDI (belief – desire – intention) reasoning model [3] [4] that is utilised in dynamic agents (these are used to model individual pedestrians in presented model). The essentials of the BDI paradigm can be summed as follows:

1. *Beliefs* represent agent's knowledge of the world. Beliefs do not have to be necessarily true.
2. *Desires* represent goals, which agent wishes to fulfil. These can be in conflict with each other. Every goal assigned to a dynamic agent by its managing agent generates a *desire* to fulfil it.
3. *Intentions* represent the connection of a *desire* that dynamic agent choses to fulfil with a *plan*. In contrast to *desires*, *intentions* cannot be in contradiction to each other.

B. Plans of BDI agent

Plans are typically stored in a structure called *plan library*. A plan is always bound to a single desire (plan is used to fulfil given desire); however more plans can be assigned to fulfil one desire. Agent then autonomously decides which plan to use to fulfil given desire; some strategies are discussed in [5].

Each plan defines three conditions:

1. *Trigger* – determines if the plan can be used.
2. *Finish* – if satisfied, the plan and respective desire are considered fulfilled.
3. *Abort* – if satisfied, the plan and respective desire are considered failed.

To define the plan, we utilise a special graph inspired by the activity diagram defined in UML 2.0. Each plan contains single *root* node, *finish* node (if this node is reached, the plan is considered fulfilled) and an *abort* node (if this node is reached, the plan is considered failed). Furthermore, a plan can contain other nodes representing basic actions that an agent can perform:

1. *Perform an activity*. After defined time elapses, agent changes its state according to the activity executed. This is the only node with defined duration of execution.
2. *Communicate* with other agents. Agents in the ABAsim architecture communicate via messages. When dynamic agent reaches this node, it will formulate and send a specific message that will be delivered by special delivery mechanism [6].
3. *Create own goal*. Dynamic agent is allowed to create private goals, as long as these are not conflicting with goals given by managing agents. Typically, the goals are formulated as desires, e.g. "*Fulfil my need*". Generating new goal from a plan may cause embedding of new plan into current plan (this mechanism is similar to calling a method from other method in a computer program). Detailed description of generating new goals and embedding of plans in ABAsim can be found in [2].
4. *Decide*. This node enables agent to split the logical flow of the plan into several branches. Decision node is not restricted by only two branches (it is similar to "switch" command).

Specific plan is constructed by first creating nodes of respective types (technically as descendants of abstract class ancestors) and connecting them into a graph. When agent interprets the plan, each node will execute a specific activity. For better understanding of presented graphs, we provide a summarizing table of plan nodes with their purpose and graphical representation.

TABLE I. SUMMARY OF PLAN NODES

Node	Purpose	Graphical representation
Root	Starting node of the plan.	
Finish	If reached, plan is considered done.	
Abort	If reached, plan is considered failed.	
Action	After specified time passes, dynamic agent will change its state.	A
Message	Formulate and send a message.	
Goal	Create own goal (desire).	G
Decision	Split the logical flow of a plan.	D

C. Three layer decision making model

The reasoning of an agent can be separated into three distinct layers (our approach was inspired by [7]). This enables us to separate goal definition (what model designer wants the agent to do), reasoning of agents (which actions the agent has to take to fulfil the goal) and simple operations of the agent (e.g. take a step).

The topmost layer is *strategy* layer. Model designer defines the strategy of an agent by means of a graph called *mission*. Individual vertexes of *mission* represent either

- a place to *spawn* the agent (pedestrian); there is a restriction of one spawn vertex per mission,
- a *goal* for the agent or
- agent's final destination, so called *sink*.

Edges of the graph define sequential order of goals (e.g. pedestrians' path) and can define conditions that must be fulfilled in order to pass the edge. Defined mission is interpreted by specialised managing agent (in our example, the *Commander* agent); dynamic agent just follows orders (goals) given to it by managing agent according to the interpreted mission. After dynamic agent finishes given order, interpreting managing agent decides, which strategy node of the mission should be executed next. This is repeated until the end of the mission (the sink vertex) is reached.

Middle layer is a *tactical* layer. Tactics for dynamic agents (pedestrians) are defined in terms of BDI plans, which are stored in a plan library. Pedestrian can choose a plan that will be used to fulfil the goal given from strategic layer by its managing agent.

The lowest layer of the model is an *operational* layer. Operative actions (e.g. walking) are defined in nodes of BDI plans (see table I). Note that plan defines only the sequence of nodes, not the action itself.

In order to explain our queueing behaviour model (defined at tactical layer) let us first briefly present all three layers of our pedestrian model using very simple example.

III. PEDESTRIAN MODEL

Consider simple open space infrastructure as depicted in figure 2. Two places, called *areas*, are the only entities present. Entity is a common term for some element in the infrastructure (service point, elevator, area, wall, etc.). Area can be understood as a virtual (without physical representation in the real world) place, mainly used as a spawn (pedestrian can appear in the area at the beginning of its lifecycle), sink (pedestrian will be removed in the area) or waypoint (pedestrian can be navigated through some area).

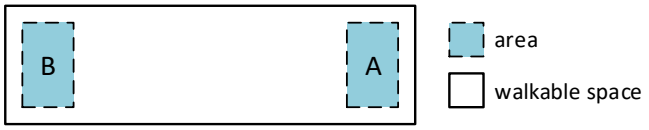


Figure 2. Simple open space infrastructure

A. Mission definition (strategic layer)

Imagine, we want the pedestrian to be spawn in *area A*, next let him walk to the *area B* and then remove him from the model. The respective mission of a pedestrian can be constructed as depicted in figure 3.

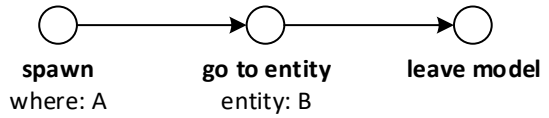


Figure 3. Mission to move pedestrian from area *A* to area *B*.

Recall that a mission is interpreted by managing agents. First, they will cooperate to spawn a pedestrian. Then agent *Commander* starts the mission interpretation, so the pedestrian receives the goal "*Go to entity B*"; this in turn generates a desire to fulfil it. After pedestrian (dynamic agent) finishes his goal, managing agents will remove him from the simulation.

B. Plans definition (tactical layer)

To fulfil a desire, pedestrian must generate an intention – pick a correct plan (from pedestrian's plan library) that will be executed. In our example mission, just one goal is defined, namely "*Go to entity*". The easiest way to fulfil it is by taking steps towards the entity until it is reached. Therefore, a very simple plan "*Go directly to entity*" that guides the pedestrian to walk towards an entity in an infrastructure will be chosen (figure 4a). Consider the action "*Take a step towards entity*" in this plan – it picks the point in the geometric shape of the target. Pedestrian may use different strategies to do so, for example:

- Pick closest point in each step (which is the most precise approach, but has the highest performance impact).
- Pick closest point once and walk towards it.
- Pick random point once and walk towards it.

If the pedestrian will not pick the point in each step (last two options), we may create a new plan "*Go to point on entity*" (figure 4b). In this plan, the pedestrian first executes the action "*Find a point in entity*" and then creates new own goal (with respective desire) to go to this point. This is completely acceptable, since the desire "*Go to point*" in no way collides with the original goal assigned by agent *Commander*. This desire will be fulfilled by means of "*Go directly to point*" plan (figure 4c). The execution of this plan will cause the suspension of the desire "*Go to entity*" until the desire "*Go to point*", generated from the plan "*Go to point on entity*", is fulfilled.

Utilizing the plan "*Go to point in entity*", the performance load is lowered, since the point is computed only once. Notice that the point selection algorithm is encapsulated into single action and can be easily replaced if necessary.

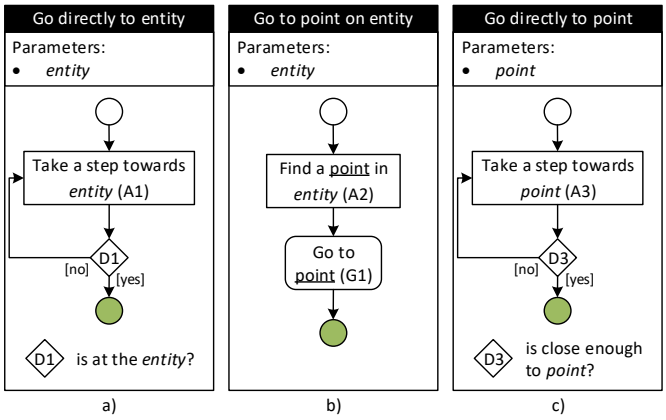


Figure 4. Plans for simple moving. a) "Go directly to entity", b) "Go directly to point on entity", c) "Go directly to point"

C. Movement modelling (operational layer)

The movement algorithm is hidden in the activities "*Take a step towards entity*" or "*Take a step towards point*". This makes the strategic and tactical layer of reasoning completely independent on operative actions. Furthermore, presented approach enables us to combine different movement models on microscopic [8] and mesoscopic [9] levels and even to create models with hybrid granularity of movement [10]. In such models, different movement models might be applied to distinct model areas (e.g., more detailed, but generally slower microscopic model can be used only in certain important areas of the model, and the rest might be modelled utilising faster mesoscopic movement model). Pedestrians are free to move between these areas and their movement model is automatically adjusted, based on their location.

Let us briefly explain basic properties of two movement models we are using in our simulator.

1) Microscopic movement model

At the microscopic level of granularity, each pedestrian is considered autonomous and is completely responsible for his actions. Microscopic movement is based on the precise position within the space. There are plenty of movement models available; most of them are based on social forces model [11] or magnetic force model [12]. In our simulator, we mostly use social forces model, though other models can be utilized as well [13]. Recall that pedestrian is modelled as a dynamic agent of the ABAsim architecture and every dynamic agent is subordinated to one managing agent. Pedestrians located in areas of the model that are modelled at microscopic granularity level belong to the submodel of *Micro* agent (see figure 1). *Micro* agent alters the movement behaviour of pedestrian in no way, it gives the pedestrians a free hand during walking.

2) Mesoscopic movement model

At mesoscopic granularity level, we still see pedestrian as an individual capable of complete control of his own reasoning. Similar to microscopic movement, when pedestrian enters the mesoscopic movement zone [10], he will be transferred under the control of agent *Meso*. In contrast to microscopic level of granularity, mesoscopic level neglects precise position of pedestrian in the model space. The model space is divided into cells and the pedestrian is present in a cell; however, the precise position in a cell is unknown. Therefore, agent *Meso* needs to alter the movement behaviour of pedestrians. Pedestrian's BDI interpreter still works (the agent is able to reason), however when pedestrian reaches "take a step" action in his current plan, he is not allowed to execute this action on his own. The movement of all mesoscopically-modelled pedestrians is executed by *Meso* agent. *Meso* agent moves pedestrians from cell to cell following transfer restrictions between them (agent must meet cell as well as flow capacity restrictions). Deep coverage of this topic can be found in [9].

Our proposed queuing model is independent on the model used for movement modelling; it can be used with microscopic and mesoscopic models without significant modifications. In following chapter, the queuing model will be presented on the example of microscopic movement model. Later we will summarize minor modifications required to use proposed queuing model with mesoscopic movement model.

IV. QUEUING BEHAVIOUR MODEL

Queues are typically formed in front of *service places* (cashiers, ATMs, etc.). Every service place contains a number of *resources* that can simultaneously serve customers. Each time a service is provided, a resource is taken; after the service ends, resource is freed and is ready for another service. If customers come to a service area with all resources taken, they have to wait in a queue. Shape of the queue can be predetermined by defined *queuing line* (used typically in supermarkets, at the airports, etc.); or the customer simply freely stands behind the last customer, so the queuing line is formed dynamically. It is common that service places define a *privacy zone*, which determines minimal distance between the area providing service and the first waiting customer. In real systems, we can observe that customers forming the queuing line are not static, creating so some kind of an obstacle. Usually there is still a chance for other walking pedestrians to pass through the formed queue. The waiting customers take a step aside forming a gap in a queue and then take a step back to their former waiting position.

In order to mimic queues in real service systems, the model of queue forming should respect the shape and the size of the queuing line (if defined), the size of the privacy zone (if defined) as well as possible obstacles located close to the service place or the queue. Presented model tries to reflect all mentioned properties and behaviour of real systems. Furthermore, since our approach requires no modification of the movement model, interactions with obstacles and other pedestrians are properly modelled even for pedestrians waiting in a queue.

We will describe our queuing behaviour model on very simple example – consider (1) an infrastructure containing a service area and (2) a simple mission that contains goal *"Take a service in the service area"*. After a pedestrian receives (from its managing agent) this new goal, he creates respective desire and picks a plan to fulfil it (see chapter III.A). To follow this specific desire, the pedestrian will use plan *"Take a service"* that is depicted in figure 5.

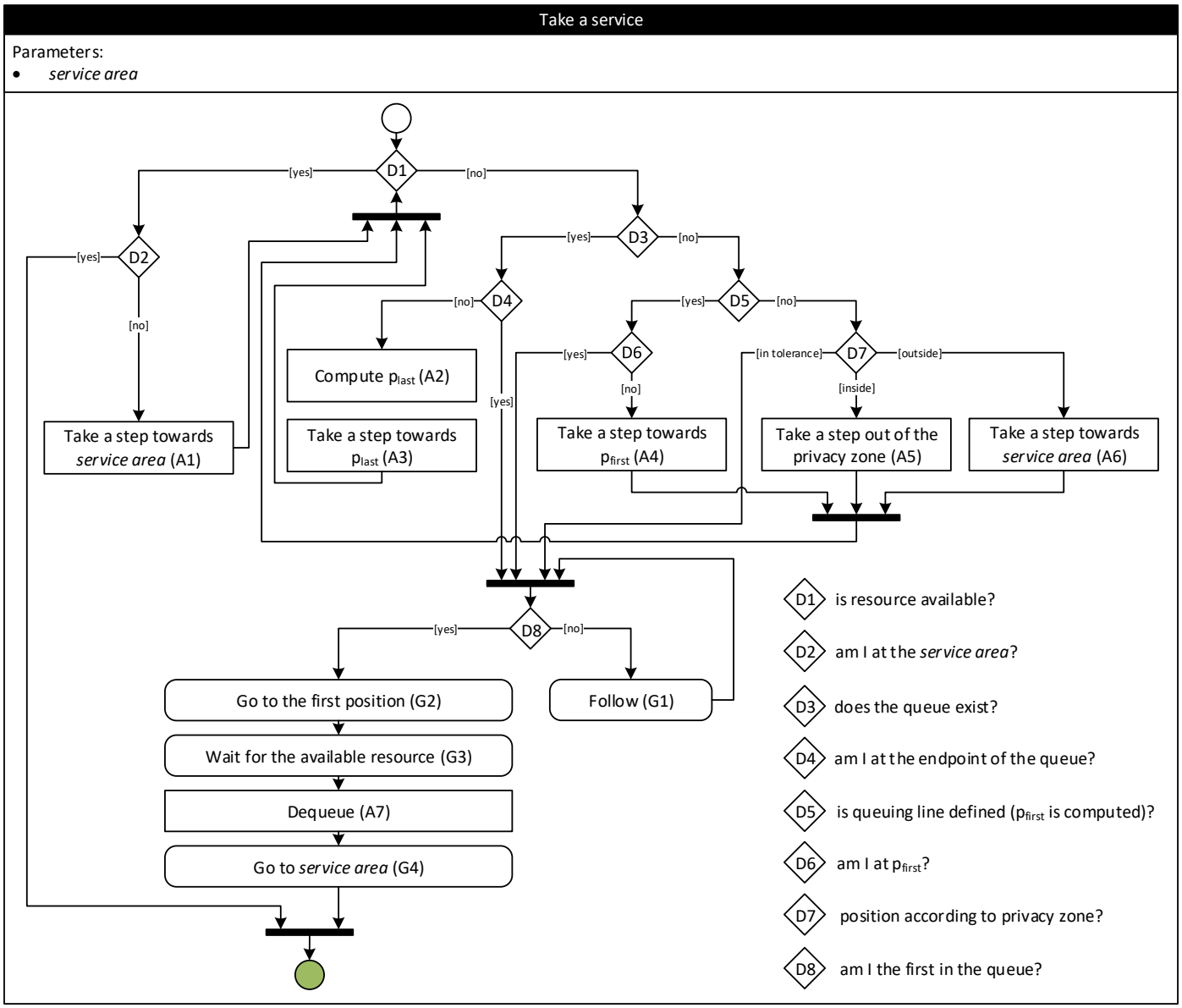


Figure 5. Plan for desire "*Take a service*"

First decision (D1) that pedestrian makes determines whether there is any resource available. If the pedestrian does not possess the information about resources, he beliefs that there is a resource available. This is conform to the BDI paradigm – beliefs can be false, pedestrian (dynamic agent) will alter them when the information will be available for him (e.g. he sees the resource). In addition, since more pedestrians can walk towards the same service area, they all consider it available until one of them reaches it and seizes the resource. Other pedestrians will subsequently alter their beliefs and start forming a queue.

If pedestrian should walk towards a service area, it seems reasonable to utilise the goal "*Go to entity*", plan of which has been presented in figure 4a. However, this would cause the pedestrian to stop evaluating the condition of resource availability, which is necessary in case when more pedestrians walk towards the idle service area. Therefore, we have to modify the plan; we will use the same arrangement of plan nodes (D2, A1 on figure 5) as in original "*Go to entity*" plan, but the execution loop will return to D1 so the pedestrian can adapt to potential changes in resources availability.

In case, when resources of the service place are all occupied (decision D1 in the plan), a queue should start to form. We distinguish two possible situations (D3): (1) no queue is present, i.e. the pedestrian will be the first one in the queue and (2) the queue already exists so the pedestrian should queue behind last person in the queue.

A. *Queuing as first customer (no queue is present)*

First position in the queue depends on the size of the privacy zone and shape of queuing line (if defined). If the queuing line is defined (D5), the point p_{first} can be computed in advance and lies in the distance $d_{privacy}$ from the beginning of the queuing line (figure 6a). The pedestrian that is going to first place in the queue then simply heads towards this point (action A4 in the plan).

If the privacy zone is larger than the length of a queuing line (or the queuing line is missing), we consider the privacy zone to be a circle with centre in the centroid of the service area and radius $d_{privacy}$ (see figure 6b).

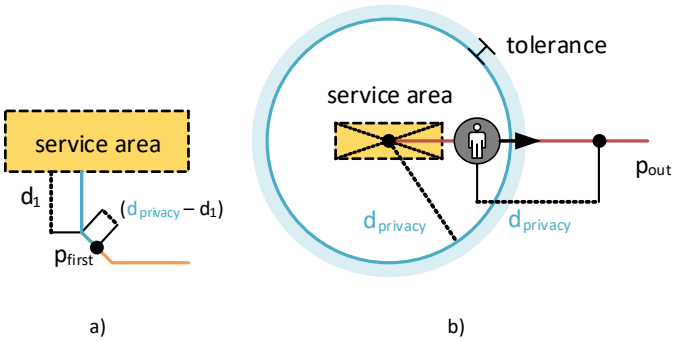


Figure 6. Identification of first place in the queue: (a) queuing line is present; (b) no queuing line is present.

When first pedestrian approaches service area and the first point p_{first} is not known, pedestrian must keep controlling his position according to the privacy zone. This may lead to three outcomes (D7):

- *Pedestrian is inside the privacy zone* – he must leave it and enter the queue behind its borders (action A5 in the plan). To navigate the pedestrian outside of the zone, a point p_{out} is computed. This point lies on the line connecting the centroid of service area and centre of the pedestrian, at the distance $d_{privacy}$ from the current position of a pedestrian in the direction heading outwards the service area (see figure 6b).
- *Pedestrian is outside the privacy zone* – he may continue to approach service area (action A6 in the plan).
- *Pedestrian reached the borders of the privacy zone* with given tolerance – he may enter the queue. In our models, this tolerance is set to 0.1m (equal to the tolerance used for reaching a point).

Notice the loop in the plan – after the pedestrian takes a step, the plan continues to the very first decision D1 of the plan. This enables the pedestrian to detect and adapt to changes, which may have occurred in the model.

B. *Queuing as next customer (last place in queue)*

The part of plan responsible for queuing by the last place in the queue is rather straightforward. The key part is to compute the last point of the queue p_{last}; this is contained in the action "*Compute last point*" (A2). The position of next/last customer in the queue cannot be easily precomputed (even if the queuing line is present). Several factors and conditions influence the computation of desired position of last pedestrian in the queue. Let us evaluate four possible situations:

1) Queue is empty and queuing line is defined

This is the simplest scenario. Position of last (in this case also the first) pedestrian is defined by pre-computed point p_{first}.

2) Queue is not empty and queuing line has enough space

In this case, pedestrian's position p_{last} will be determined similarly to the computation of p_{first}, simply by following the queue line from the position of last waiting pedestrian to the distance d_{queue} (see figure 7). If two pedestrians are approaching the working service area, they both might try to reach the same spot at the end of the queue. In such situations, the pedestrian, which reaches this point first will be enqueued first, the other pedestrian will have to go to the newly computed endpoint of the queue.

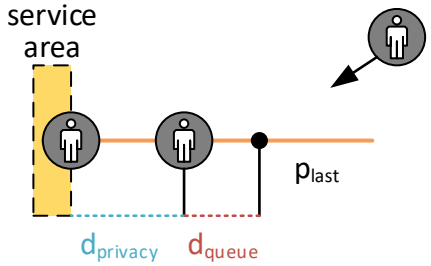


Figure 7. Endpoint detection on the queuing line

3) Queue is not empty and queuing line does not have enough space; or queuing line is not defined

Point p_{last} is determined by the positions of last two pedestrians waiting in the queue. We create a line connecting the centres of last two waiting pedestrians. This line is used as a virtual queuing line and the point p_{last} is then computed as previously described (figure 8).

If no pedestrian is currently waiting and no queuing line is defined, the points defining the virtual queuing line are the centroid of the service area and the centre of the pedestrian to be served (i.e. the point p_{last} is the closest point to the pedestrian that lies outside the privacy zone).

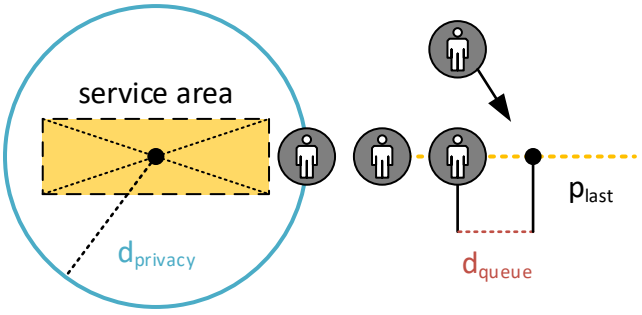


Figure 8. Endpoint detection out of the queuing line

4) Computed p_{last} is not reachable

In certain situations, the computed point p_{last} (target point for new pedestrian approaching the queue) might lie in an obstacle or outside the walkable area. If this situation occurs, pedestrian will simply head towards the last waiting pedestrian in the queue. When reaching the distance d_{queue} from the position of last waiting pedestrian, he will enter the queue (see figure 9). To achieve this behaviour, the goal "*Go to point*" with its respective plan can be used.

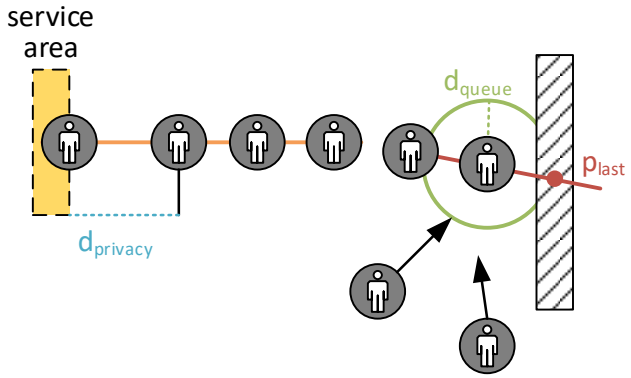


Figure 9. Endpoint detection close to an obstacle.

After identification of last point the pedestrian takes a step towards it and similarly to previous case, the execution flow returns to the D1, so the pedestrian may adapt to potential changes.

C. Waiting in queue

Sequence of waiting actions follows after the pedestrian reached the service area (branch "*in tolerance*" of D7); or the first (branch "*yes*" of D6) or last point in the waiting queue (branch "*yes*" of D4). The behaviour of a pedestrian in the queue depends on his position in the queue:

- At the first position in the queue, pedestrian must actively observe the service area, whether it has resources capable to serve him (D8). If so, pedestrian leaves the queue and walks towards the service area where he will be served.
- Any other position leaves pedestrian actively standing. He follows the pedestrian in front of him and takes care to follow the queue line (if present). Even in queue, the pedestrian still actively reacts to pedestrians near its position (e.g. pedestrians crossing the waiting queue).

The activities of this part of the plan are quite complex, so it is beneficial to define them by creating new goals (G1-4). This approach allows us to customize pedestrian's queuing behaviour simply by changing the plans agent uses to reach those goals. Let us now focus on the plans for these new goals.

1) Plan for goal "Follow"

Following pedestrians in the queue (or we can say standing behind someone's back) is vital part of dynamic queue formations. Every pedestrian except the first one (see D8 in plan "*Take a service*") follows this goal. The plan for this goal is depicted on figure 10.

First, proper position of the pedestrian in the queue must be defined. The pedestrian must follow the queuing line if present; and align according to last two agents in the queue. However, a queue of pedestrians, which would stand precisely on the predefined line with predefined gaps between them, would look very unnatural. Therefore, we permit pedestrians a tolerance 0.5 m from both sides of a queuing line (D2). If pedestrian exceeds this tolerance, he will head back (A2) towards the point computed using algorithms presented in IV.B. The gaps between pedestrians are not constant either; pedestrian randomly (we use continuous uniform distribution with parameters UNI(1, d_{queue}) meters) decides how far behind the back of the pedestrian in front of him he would like to stand (A1).

When proper position is achieved, (branch "yes" of D2) the standing action A3 is commenced. We want to keep the benefits of the movement model such as reactions to obstacles or other pedestrians; therefore, we employ so-called "active standing". This means, we do not stop pedestrian from moving, neither have we altered the movement model. Instead, the "*Stand*" action is leading the pedestrian towards the point he is already standing at. Let us discuss the consequences of such approach on the example of social force movement model [11]. Authors split the walking behaviour into three components (vectors) responsible for (1) taking steps towards target, (2) avoiding pedestrians and (3) avoiding obstacles. If we let the pedestrian walk towards his current position, first vector will have zero length. This eliminates its influence, but the pedestrian and obstacle avoiding part is still present. This allows us to model such phenomena like letting walking pedestrians pass through the queue without additional effort.

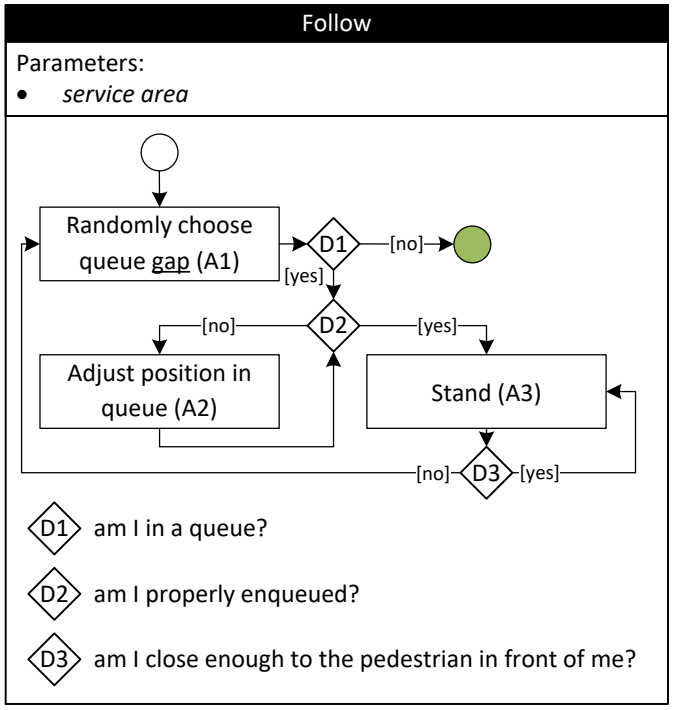


Figure 10. Plan for goal "*Follow*"

2) Plan for goal "Go to the first position"

This plan (depicted on figure 11) is used when the pedestrian that is waiting in queue has become the first one (i.e. the one in front of him just entered the service). The plan guides the pedestrian to the first position in the queue. This position depends on the existence and properties of queuing line and privacy zone. After reaching the desired position, this plan ends.

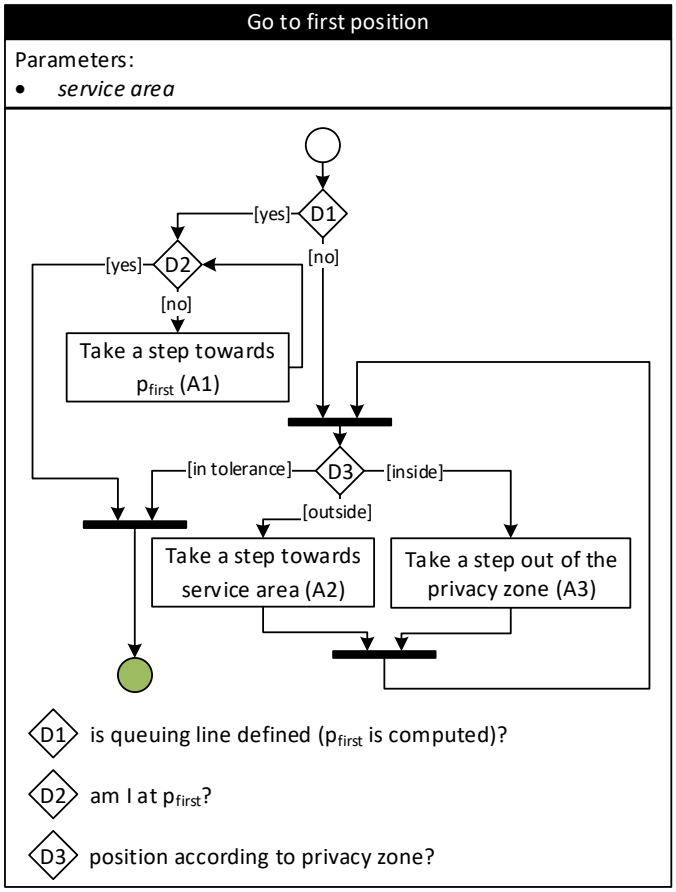


Figure 11. Plan for goal "Go to first position"

3) Plan for goal "Wait for available resource"

The plan to fulfil this goal is quite simple (figure 12), the pedestrian stands and checks whether a resource is available; we can say that the pedestrian is actively standing.

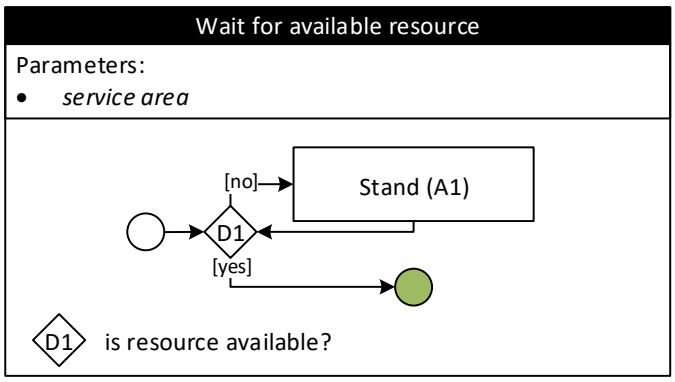


Figure 12. Plan for goal "Wait for available resource"

4) Plan for goal "Go to service area"

The goal "*Go to service area*" is identical to the goal "*Go to entity*". In this case, the entity parameter is set to the service area. Plan to fulfil this goal was already presented in chapter III.

V. QUEUING BEHAVIOUR MODEL WITH MESOSCOPIC MOVEMENT MODEL

Recall, that the movement itself has no influence on the reasoning and queue forming behaviour of pedestrians. However, since if the model uses mesoscopic movement modelling, all movements are performed by managing agent *Meso* (moving pedestrians from cell to cell following model restrictions), slight implementation adjustments are necessary.

If the granularity of simulation model is mesoscopic (the infrastructure model consists of cells), the queuing line has to be mapped to underlying cells; these are then called *queuing cells*. The capacity of these cells is reduced to the half of their original capacity. This is required in order to model sparsely standing pedestrians in a queue as well as to guarantee that transferring pedestrians do not block the full capacity of the cells beneath the queue line, which in turn may cause splitting of the queue (as illustrated in figure 13).

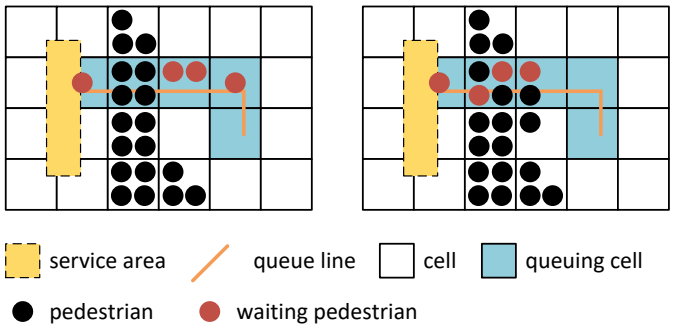


Figure 13. Queuing mesoscopic cells

Furthermore, we need to slightly modify the meaning of some activities used in employed plans. Where navigation towards a point is used, it has to be replaced with navigation towards a cell containing respective point. Other activities remain intact and require no change.

VI. VALIDATION OF THE MODEL

Model has been validated on the example of simple service infrastructure. Since we did not make any modifications to existing valid movement models, there was no need to focus on the movement model validation. Instead, we focused on the formation of waiting queue and behaviour of standing pedestrians. The results of our model were compared to commercial simulation tools (e.g. Viswalk, SimWalk, AnyLogic) and judged by experts with satisfying outcomes. This leads us to the belief that our queuing model can be considered as functional and sufficiently valid. Example of queue forming in our simulation model for both microscopic and mesoscopic granularity level is presented in following pictures.

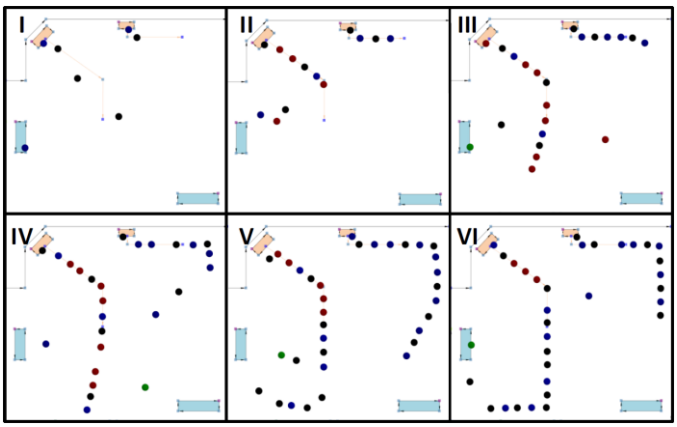


Figure 14. Process of queue forming on microscopic level of movement granularity

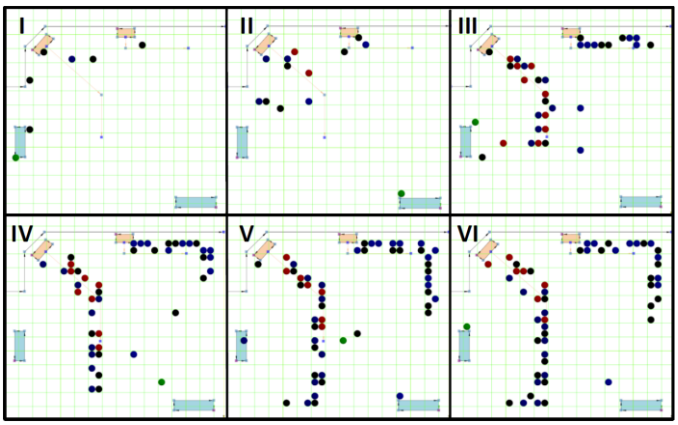


Figure 15. Process of queue forming on mesoscopic level of movement granularity

VII. Conclusion

Proper modelling of queuing behaviour can have significant impact on validity of pedestrian models especially models of service (aptly also known as queuing) systems. The queuing behaviour model presented in this paper reflects the important properties of pedestrians' queuing (such as dynamic queue formation, queue recovery after crossing pedestrians, uneven gaps, etc.), while at the same time it is independent of the movement model employed in the simulation model. This is achieved by utilising BDI paradigm support integrated in our ABAsim architecture, i.e. modelling the queuing behaviour by means of BDI desires and plans. Our approach leads to flexible simulation models and allows for changes in granularity level of the movement model (microscopic vs. mesoscopic using very different methods to model movement) without modification to the model of queuing behaviour.

This paper is the result of the implementation of project "Centre of excellence for systems and services of intelligent transport II.", ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF. "Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

References

[1] N. Adamko and V. Klima, „Agent based simulation architecture augmented by actors," In: *ESM´2006 - The 2006 European Simulation and Modelling Conference.* Toulouse: EUROSIS-ETI. ISBN 90-77381-30-9. pp 305–309.

[2] M. Varga, „*Modeling of information gathering and its utilization in intelligent decision making process in agent oriented simulation models*," PhD. dissertation, Department of Mathematical Methods and Operation Research, Faculty of Management Science and Informatics, University of Žilina, 2014, pp. 136.

[3] M. Varga and N. Adamko, „Integration of BDI paradigm into ABAsim architecture," In. *Journal of Information, Control and Management Systems*, Vol. 12, No. 1, 2014.

[4] A. S. Rao and M. P. Georgeff, „BDI Agents: From Theory to Practice," In. *Proceedings of the First International Conference on Multiagent Systems*, 1995.

[5] A. Casali et al, „A graded BDI agent model to represent and reason about preferences," *Artifical intelligence*, 2011, 175.7, pp. 1486-1478.

[6] A. Kavička et al, „Simulations of transportation logistic systems utilising agent-based architecture," In: *International Journal of Simulation Modelling*. 2007, Vol. 6, No. 1, ISSN 1726-4529, pp. 13 – 24.

[7] S. P. Hoogendoorn and P. H. L. Bovy, „Pedestrian route-choice and activity scheduling theory and models," In. *Transportation research, Part B: Methodological*, 2004, 38.2, pp. 169-190.

[8] A. Kormanová, „*Hybrid simulation models of pedestrian movement and behavior with non-homogenous granularity*," PhD. dissertation, Department of Mathematical Methods and Operation Research, Faculty of Management Science and Informatics, University of Žilina, 2014, pp. 100.

[9] M. Čadecký et al, „Mesoscopic movement model of deliberative pedestrian agents" In. *Information and digital technologies 2015 : the international conference : 7-9 July 2015*, Žilina, Slovakia, 2015.

[10] A. Kormanová et al, „Hybrid model for pedestrian movement simulation," In. *Digital technologies : the 10th international conference : 9-11 July 2014*, Žilina, Slovakia, 2014.

[11] D. Helbing et al, „Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations and Design Solutions," In. *Transportation Science*, 2005, 39.1.

[12] S. Okazaki and S Matsushita, „A study of simulation model for pedestrian movement with evacuation and queuing," In. International Conference on Engineering for Crowd Safety, 1993, pp. 271-280.

[13] M. Varga and M. Mintál, „Microscopic pedestrian movement model utilizing parallel computations," In. *SAMI 2014: IEEE 12th international symposium on Applied machine intelligence and informatics: January 23-25, 2014,* Herľany, Slovakia, IEEE, 2014. ISBN 978-1-4799-3441-6. pp. 221-226.

Multispectral imagery super-resolution with logical reallocation of spectra

Iryna O. Piestova, Sergey A. Stankevich
Scientific Centre for Aerospace Research of the Earth
National Academy of Sciences of Ukraine
Kiev, Ukraine
pestovai@ukr.net

Jozef Kostolny
Department of Informatics
University of Zilina
Zilina, Slovakia
kozef.kostolny@fri.uniza.sk

Abstract — **The multispectral imagery spatial resolution enhancement with logical reallocation of spectra is presented. This method includes preprocessing, subpixel resampling, subpixel neighborhood analysis and subpixel values reallocation using similar spectra spatial cross-coupling. This method intended primarily for European Sentinel-2 multispectral satellite system, but it can be adapted to other multispectral systems with non-uniform spatial resolution too.**

Keywords — multispectral imagery, super-resolution, subpixels cross-coupling, logical reallocation

I. INTRODUCTION

The novel ESA Sentinel-2 multispectral satellite sensor (MSS) equipped with separate 13 spectral bands of 10, 20, and 60 m ground sampling distance (GSD). It is useful for calculating various vegetation and other objects features, but the interpolation for GSD equalizing must be carried out before analysis.

At moment the nearest neighbor dumb interpolation is used during Sentinel-2 data processing. This type of interpolation does not affect the physical spatial resolution of multispectral imagery (MSI); nevertheless, the spatial resolution enhancement can be achieved eventually. Therefore, the physical resolution enhancement of Sentinel-2 MSI is an urgent task now.

II. SUPER-RESOLUTION METHODS OVERVIEW

A method for super-resolution of a set of four subpixel-shifted images is known. This one allows an up-twice resolution enhancement [1]. The most common and powerful tool for image super-resolution is the of least squares method [2]. This one provides high accuracy and can be applied to multispectral imagery. A drawback of this method is the need to solve a high-dimensional system of non-linear equations. Also the iterative back-projection method is known, which is adapted for over-resolution of different-resolution imagery [3]. A number of methods for image super-resolution have been developed on the basis of analysis of local textures parameters [4]. A special method for digital imagery estimating and super-resolution based on analysis of high spatial frequencies spectrum is described in [5]. Parametric model for the continuous version formation of processed image is developed. The model parameters are determined by maximum likelihood estimator. Several effective algorithms for image super-resolution are demonstrated using discrete Fourier transform and non-linear optimization [6]. In [7] the method is proposed for image super-resolution based on the detection of linear edges of segments. Another method for image super-resolution is implemented by means of multi-directional wavelet transform [8].

For hyperspectral satellite imagery, a super-resolution technology without additional images has been pursued [9]. The algorithm is exploited the spatial mismatch between individual band images registered by hyperspectral imager. Alternative method for hyperspectral imagery super-resolution [10] is based on subpixel redistribution of spectral endmember fractions obtained by any possible unmixing technique [11] and does not need additional images. The redistribution of endmember fractions within pixel fulfils the principle of compactness, which states that subpixel fractions of the same material in adjacent pixels are located in adjacent subpixels too. This principle is carried out by a vector field that affects over the pixel fraction of the endmember via one's connections to the pixel fractions of the same endmember in other pixels of image.

The identification of spatial dependencies between the land cover fractions that are interpreted inside image also make possible other method for a super-resolution [12]. The spatial allocation of fractions within several subpixel-shifted low-resolution images are combined into single image with super-resolution

III. CONCEPT AND METHODS

In this paper, the physical spatial resolution enhancement is based on the reference spectra method [13]. This method includes preprocessing, subpixel resampling, subpixel neighborhood analysis and subpixel values reallocation using similar spectra spatial cross-coupling.

A. Pre-processing

Preliminary processing of source multispectral satellite imagery includes georeferencing, geometric, radiometric and atmospheric corrections, as well as noise filtering if need. The next step is 20 m spectral bands resampling up to 10 m GSD.

Reference spectra are selected from the spectral library according to the region of the study or other conditions [14].

The sensor-like spectral signatures of the reference continuous spectra are calculated by convoluting with the target sensor spectral band responses, which is Sentinel-2 in our case.

B. Scanning pattern

Dumb resampling quadruples the pixel by the nearest-neighbor rule. Such procedure generates 4 identical subpixels in case of Sentinel-2. It is necessary to reallocate signals in these subpixels properly to enhance the overall image physical resolution. This must be done taking into account the surrounding area.

The profile of the scanning window from the nearest eight subpixels (are numbered as 1 .. 4 and 6 .. 9) around current one (is numbered as 5) is shown in Fig.1.

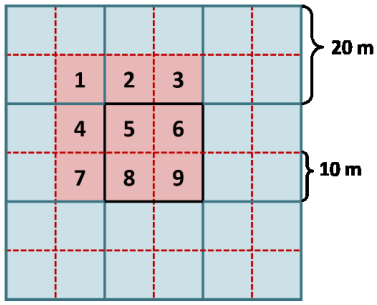


Figure 1. Neighborhood of processed subpixel (5) within scanning window (1 .. 4 and 6 .. 9) including 4 subpixels (5, 6, 8, 9) of one low resolution pixel

For correct spatial reallocation of subpixels during the scanning process the topological descriptions of spectral signatures must be analyzed.

C. Topological descriptions

Because all sensor-registered signals are natively statistical, then spectral signatures analyses will be affected by noise. To exclude the noise influence, the inherent topological properties of signatures should be considered. It is possible using the logical operations. To do this, we engage the statistically adopted special sign() function, taking into account statistical deviations.

Each signature can be described by the X vector of subpixel reflectance values x and its statistical deviation σ – for the current signature and respectively y, ε – for reference signature Y: $X=\{x_1,\sigma_1,x_2,\sigma_2\ldots x_m,\sigma_m\}$, $Y=\{y_1,\varepsilon_1,y_2,\varepsilon_2\ldots y_m,\varepsilon_m\}$, where m is a number of spectral bands.

For each subpixel, similar to the reference spectral signatures, a topological description of x values in m bands is calculated [15] as

$$C(X)=\{\text{sign}(x_1-x_2,\varepsilon_1+\varepsilon_2),\ldots\text{sign}(x_{m-1}-x_m,\varepsilon_{m-1}+\varepsilon_m)\},\quad(1)$$

where

$$\text{sign}(x,\varepsilon)=\begin{cases}1 \text{ if } x>0 \text{ and } x>\varepsilon,\\ 0 \text{ if } |x|\le\varepsilon,\\ -1 \text{ if } x<0 \text{ and } |x|>\varepsilon.\end{cases}\quad(2)$$

The number of items in the topological description for each signature depends on the number of bands and is equal to $C_m^2=m(m-1)/2$.

For each reference spectral signature, an analogous topological description is also defined

$$C(Y)=\{\text{sign}(y_1-y_2,\sigma_1+\sigma_2),\ldots\text{sign}(y_{m-1}-y_m,\sigma_{m-1}+\sigma_m)\}.\quad(3)$$

After determining topological descriptions, it is possible to perform topological classification of subpixels.

D. Topological classification

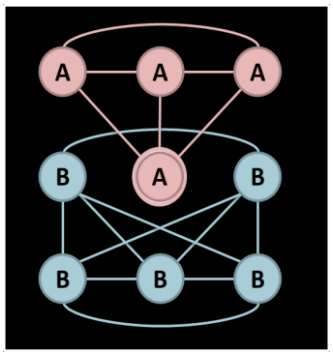
The first stage of the classification is carried out by comparing the topological descriptions in each subpixel X with the topological description of the reference spectra Y as the topological difference $D(X, Y)$:

$$D(X,Y)=|C(X)-C(Y)|,\quad(4)$$

where D value is in [0 .. m (m–1)] interval. Preference is given to the class of spectra with the smallest value of D.

The second stage consists in classes reallocating within the scanning window taking into account neighboring subpixels (Fig. 2). Here m_1 .. m_5 are weights of cross-coupling between these subpixels. Weight values depend on their relative allocation i.e. the spatial distance from each other and the classes' composition within the scanning window.

a



b

Figure 2. Subpixels cross-coupling: *a* – general pattern, *b* – case of two classes A (red) and B (blue)

In the center of the Fig. 2 pattern, the 5 subpixel is processed by its neighborhood analysis to make a decision about current class change or preserve. For any neighboring subpixel numbering order, the weights of their cross-coupling with each other and with the central subpixel will not change, therefore the rotation-invariance is provided.

The common reallocating matrix between nine subpixels within scanning window is presented in Table I.

TABLE I. MATRIX OF SUBPIXELS REALLOCATING

	1	2	3	4	5	6	7	8	9
1	0	m_3	m_2	m_3	m_4	m_1	m_2	m_1	m_1
2	m_3	0	m_3	m_2	m_4	m_2	m_1	m_1	m_1
3	m_2	m_3	0	m_1	m_4	m_3	m_1	m_1	m_2
4	m_3	m_2	m_1	0	m_4	m_1	m_3	m_2	m_1
5	m_4	m_4	m_4	m_4	m_5	m_4	m_4	m_4	m_4
6	m_1	m_2	m_3	m_1	m_4	0	m_1	m_2	m_3
7	m_2	m_1	m_1	m_3	m_4	m_1	0	m_3	m_2
8	m_1	m_1	m_1	m_2	m_4	m_2	m_3	0	m_3
9	m_1	m_1	m_2	m_1	m_4	m_3	m_2	m_3	0

In Table II is shown as the weight for each class is formed from the subpixel reallocating matrix. The cross-coupling between subpixels from different classes does not take into account.

TABLE II. SUBPIXEL CROSS-COUPLING MATRIX BY CLASSES A (RED) AND B (BLUE)

	1	2	3	4	5	6	7	8	9
1	0	m_3	m_2	0	m_4	0	0	0	0
2	m_3	0	m_3	0	m_4	0	0	0	0
3	m_2	m_3	0	0	m_4	0	0	0	0
4	0	0	0	0	0	m_1	m_3	m_2	m_1
5	m_4	m_4	m_4	0	m_5	0	0	0	0
6	0	0	0	m_1	0	0	m_1	m_2	m_3
7	0	0	0	m_3	0	m_1	0	m_3	m_2
8	0	0	0	m_2	0	m_2	m_3	0	m_3
9	0	0	0	m_1	0	m_3	m_2	m_3	0

The next stage will be the selection of a reallocating matrix for a particular subpixel considering its surroundings.

E. Selection of subpixel reallocating matrix

Flowchart of subpixel reallocating matrix selection is shown in Fig. 3. All possible classes K that correspond to the reference spectral signatures (marked with circles) forms C_K^9 class pattern within the 3×3 scanning window (grid). For the reduction of high dimensionality of these patterns, we need to engage special logical operator (rectangle) for optimal selection of subpixel reallocating matrix (squares). In particular, it is convenient and expedient to realize this operator using the logical differential calculus [16, 17, 18]. Then, the final number of selected matrix will be h.

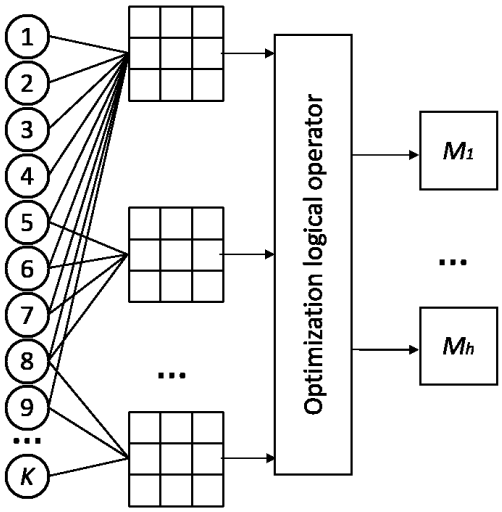


Figure 3. Selection of subpixel reallocating matrix

Selected reallocation matrix is a tool for reclassification of current subpixel under analysis. Its class is considered jointly with near neighborhood, and resulting decision is made to save this one or change it to another class from a set of surrounding subpixels.

F. Weighting operation

Restriction should be placed on subpixels of enhanced resolution output images for preserving the undistorted averaged radiometric value L_0 of source pixel [19]:

$$L_0 \equiv \frac{1}{s}\sum_{q=1}^{s} L_q \,, \tag{5}$$

where L_q is the radiometric value of the subpixel q, s is the number of output image subpixels within the source pixel of the original low-resolution image. In the case under consideration the $s = 4$ (Fig.4).

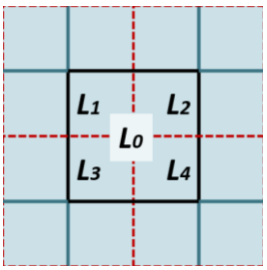


Figure 4. Fore subpixels (L_1 .. L_4) inside one low resolution pixel (L_0)

The restriction (5) must be satisfied for all spectral bands of the original image, for which the super-resolution is performed.

IV. ALGORITHM

In accordance with the methodology described in section 2, the processing of multispectral imagery for spatial resolution enhancement with logical reallocation of spectra is explained by dataflow diagram Fig. 5.

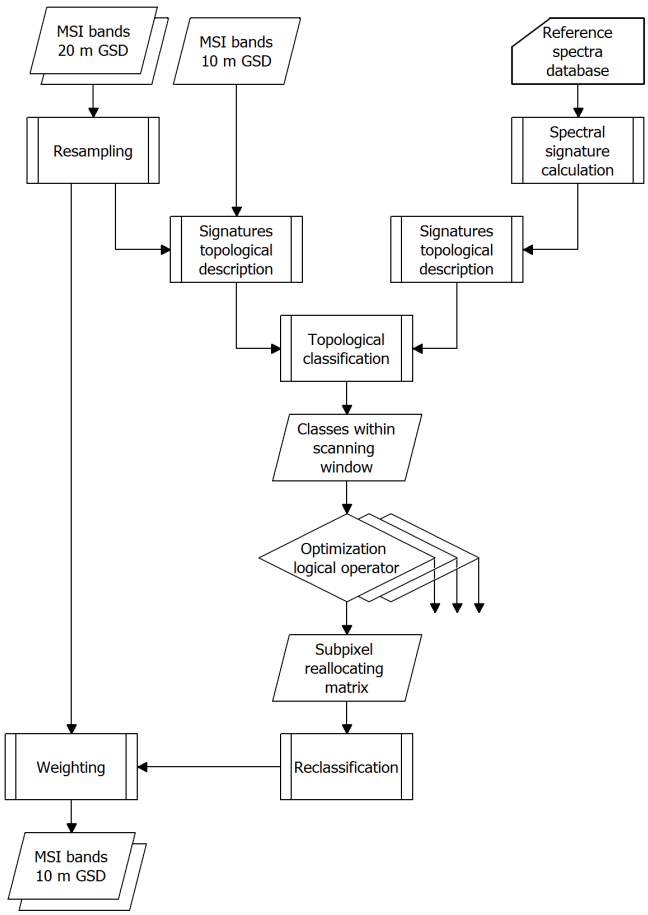


Figure 5. Multispectral imagery super-resolution dataflow diagram

The logical operator's optimization is made by reference to the classes composition and allocation topology within the scanning window current position. The 10 m super-resolution multispectral image bands are the overall processing output.

V. DEMO

The above-described algorithm was implemented as a software module in SciLab numeric computation environment. It was applied to fragment of Sentinel 2A 10 bands multispectral image (Fig.6,7) over the territory near Kyiv acquired at August 27, 2016.

Figure 6. Sentinel-2 MSI 10 m resolution image fragment

Figure 7. Sentinel-2 MSI 20 m resolution image fragment

All 10 spectral bands were used for processing: four 10 m resolution bands as reference and six 20 m resolution bands for super-resolution. A classification based on spectral library spectra of test area (Fig.8) was performed (Fig.9) and spectral signatures of all bands were calculated.

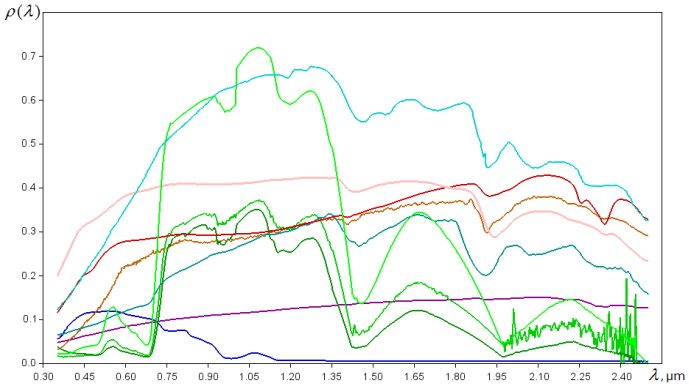
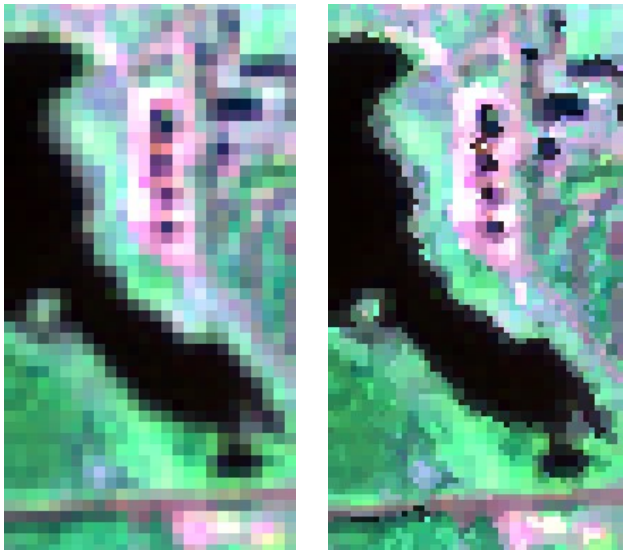


Figure 8. Spectral library spectra of test area

Figure 9. Sentinel-2 MSI test area land cover classification

After super-resolution procedure running the enhanced resolution bands was obtained. Fig. 9 illustrates the super-resolution output's performance.



a *b*

Figure 10. Sentinel 2 MSI fragment with bands combination 705 nm + 740 nm + 1610 nm: *a* – 20 m resolution input, *b* – after super-resolution applaying

According to Fig. 9 analysis, it follows, that the contrast segments borders become more exact. This will be very useful for remote sensing applications.

VI. Conclusion

Thus, a novel processing method and algorithm for multispectral imagery superresolution are developed and presented in this paper. They ones are adapted to Sentinel-2 MSS and will be useful in further analysis and study of various objects and phenomena on the land surface. The preliminary results of developed algorithm testing over actual Sentinel-2 multispectral imagery confirm the efficiency and usefulness of proposed approach.

References

[1] P. Vandewalle, S. Süsstrunk, M. Vetterli, "Superresolution images reconstructed from aliased images", Proceedings of SPIE, vol. 5150, pp. 1398–1405, June 2003.

[2] A.S. Vasileisky, J.-L.R. Casanova, K.R. Al-Rawi, "Automated sub-pixel co-registration of the same sensor images by the least square technology", Proceedings of the 21th EARSel Symposium "Observing our environment from space: new solutions for a new millennium", Paris: Balkema, pp. 221–226, May 2002.

[3] Y. Lu, M. Inamura, "Spatial resolution improvement of remote sensing images by fusion of subpixel-shifted multi-observation images", International Journal of Remote Sensing, vol. 24, no. 23, pp. 4647–4660, June 2003.

[4] S. Gao, X.J. Zhang, W.D. Sun, "Lossless inter-array predictive coding for subpixel-shifted satellite images based on texture analysis", Proceedings of 12th International Conference on Geoinformatics − Geospatial Information Research: Bridging the Pacific and Atlantic (Geoinformatics 2004), Gävle: University of Gävle, pp. 275–282, June 2004.

[5] C.A. Glasbey, G.W.A.M. van der Heijden, "Alignment and sub-pixel interpolation of images using Fourier methods", Journal of Applied Statistics, vol. 34, no. 2, pp. 217–230, July 2007.

[6] M. Guizar-Sicairos, S.T. Thurman, J.R. Fienup, "Efficient subpixel image registration algorithms", Optics Letters, vol. 33, no. 2, pp. 156–158, January 2008.

[7] M.P. Cipolletti, C.A. Delrieux, G.M.E. Perillo, C.M. Piccolo, "Superresolution border segmentation and measurement in remote sensing images", Computers and Geosciences, vol. 40, pp. 87–96, March 2012.

[8] A.P. Reji, T. Tessamma, "Single frame image super resolution using learned directionlets", International Journal of Artificial Intelligence and Applications, vol. 1, no. 4, pp. 29–42, October 2010.

[9] S.-E. Qian, G. Chen, "Enhancing spatial resolution of hyperspectral imagery using sensor's intrinsic keystone distortion", IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 12, pp. 5033–5048, December 2012.

[10] M.A. Popov, S.A. Stankevich, S.V. Shklyar, "Method for spatial resolution enhancement of hyperspectral aerospace imagery using subpixel reallocation of spectral endmember fractions", Ukrainian patent 92541, November 2010.

[11] S.A. Stankevich, S.V. Shklyar, "Land-cover classification on hyperspectral aerospace images by spectral endmembers unmixing", Journal of Automation and Information Sciences, vol. 38, no. 12, pp. 31–41, December 2006.

[12] F. Ling, Y. Du, F. Xiao, H.P. Xue, S.J. Wu, "Super-resolution land-cover mapping using multiple sub-pixel shifted remotely sensed images", International Journal of Remote Sensing, vol. 31, no. 19, pp. 5023–5040.– October 2010.

[13] B.S. Zhukov, M.A. Popov, S.A. Stankevich, "Multispectral multiresolution image synthesis using library object spectra for regularization (in Russian)," Current problems in remote sensing of the Earth from space, vol. 11(2), pp.50-67, November 2014.

[14] S. Stankevych, O. Tytarenko, A. Kozlova, I. Piestova, I. Yushchenko, "Integration of spectral library with reflectance ofl natural and man-made objects into geoinformation systems," Journal of Geodesy and Cartography, vol. 4(97), pp.27-30, October 2015.

[15] S.A. Stankevich, I.A. Piestova, V.N. Podorvan, "Deep learning concept for hyperspectral imagery classification," Central European Researchers Journal, vol. 2(1), pp.30-36, April 2016.

[16] M. Kvassay, E. Zaitseva, V. Levashenko, "Importance analysis of multi-state systems based on tools of logical differential calculus", Reliability Engineering and System Safety, vol. 165, pp.302-316, March 2017.

[17] E. Zaitseva, J. Kostolny, V. Levashenko, "Multi-state system importance analysis based on direct partial logic derivative", in: Proc. of 2012 IEEE International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2012), 2012, pp.1514-1519.

[18] M. Kvassay, E. Zaitseva, V. Levashenko and J. Kostolny, "Minimal cut vectors and logical differential calculus," in Proc. IEEE 44th International Symposium on Multiple-Valued Logic (ISMVL) 2014, pp. 167–172, http://dx.doi.org/10.1109/ISMVL.2014.37

[19] M.A. Popov, S.A. Stankevich, "Multispectral imagery resolution enhancement in environmental remote sensing," Thesis of the International Advanced Research Workshop on Fuzziness and Uncertainty in GIS for Environmental Security and Protection, Kiev: NAUU, pp.9-10, July 2006.

Right-click Authenticate Adoption: *The Impact of Authenticating Social Media Postings on Information Quality*

Pardis Pourghomi, Ahmed Abu Halimeh, Fadi Safieddine, Wassim Masri
College of Engineering and Technology
American University of the Middle East (AUM)
Egaila, Kuwait
{Pardis.Pourghomi; Ahmed.Abu Halimeh; Fadi.Safieddine; Wassim.Masri}@aum.edu.kw

***Abstract*—Getting the daily news from social media has nowadays become a common practice among people. Unreliable sources of information expose people to a dose of hoaxes, rumours, conspiracy theories and misleading news. Mixing both reliable and unreliable information on social media has made the truth to be hardly determined. Academic research indicates an increasing reliance of online users on social media as a main source of news. Researchers found that young users, in particular, are to believe what they read on social media without adequate verification. In previous work, we proposed the concept of 'Right-click Authenticate' where we suggested designing an accessible tool to authenticate and verify information online before sharing it. In this paper, we present a review of the problem of sharing misinformation online and extend our work by analysing how 'Right-click Authenticate' reduces the challenges of while improving key metrics within the Information Quality fields.**

Keywords—social media, misinformation, fake news, information quality.

I. INTRODUCTION

Over the past several years, social media sites, such as Facebook, Twitter, Instagram, and Youtube have drastically changed the social interaction landscape by creating new platforms for communication and information exchange. Organisations and individuals are striving to integrate information from various social media into their daily business practices in recruiting, sales and marketing [1]. Yet, if individuals are to rely on data collected through social media sites, they need to understand the quality of information emanating from these sites. Although there are concerns with the quality of this information, understanding relevant quality attributes and effective means of their assessment is limited. This has raised, for many researchers, the question of the quality of user-generated content in social media [2].

Given the distinctive characteristics of social media such as wide accessibility, permanence, global audience, recentness and ease of use, Information Quality (IQ) in this context is quite unique [3]. Social media has extended knowledge creation boarders across organisational boundaries, therefore unlike traditional information systems, users have no control to influence the quality of the information obtained [4]. According to [5, 6], IQ has been defined from the user's point of view that is the extent to which the information fits for the intended use of the consumer. With the lack of means to verify information, social media has been accused of becoming a hotbed for sharing of misinformation. Facebook as one of the largest social networking services has been facing widespread criticism on how its newsfeed algorithm is designed thus amplifying dissemination of misinformation [7].

II. BACKGROUND

This section describes the problem of misinformation propagation in social media and focuses on how important this problem is. It then provides a summary of IQ dimensions and metrics in addition to discussing their role of IQ on social media. Our previously proposed approach, namely, Right-click Authenticate (RCA) is then briefly explained to ground the further analysis in presented in the next sections.

A. Misinformation spread in online social networks

Several social media outlets such as Facebook, Youtube, and Twitter are becoming the main source of news and information for online users. There is, however, some validated concerns in using these models since there are little accountability and source validation resulting in the spread of misinformation propagation. Social media networks are being blamed for not doing enough to combat the spread of misinformation and allowing the spread of fake news [7]. Limiting the spread of misinformation on the web has so far proved very difficult if not impossible especially when considering the variety of online social media websites [8].

Some researchers have attributed the problem of the spread of misinformation online to the algorithms employed by some social media [9]. Facebook algorithms, for example, do not distinguish what users' feeds get or attempt to validate news posts that are being shared. Hence fake news posts can spread in a very similar way as genuine news. An Ipsos survey [10] of Facebook users demonstrates that individuals using Facebook as their main source of news are more likely to categorise misleading news as correct [10].

A survey conducted by Pew research [11] showed that a 62 percent of U.S. adults depend on the news they see on social

media, and this may include misinformation and fake news. What is more concerning is that this survey shows that even though many social users recognise the presence of fake news on social media, these users still consider social media a main source of reliable news.

B. Information quality

As defined in [12], the meaning of information quality lies in how the information is perceived and used by its consumer. Though absolute attributes are important, it is how those attributes are perceived, now and in the future, that defines information quality. Two stages are involved in recognition of quality information: (1) highlighting which attributes are important and (2) determining how these attributes affect the customers in question. Accuracy can be seen as just one element of IQ but, depending upon how it is defined, it can also be seen as encompassing many other dimensions of quality. In other scenarios, it is often observed that there is a trade-off between accuracy and other dimensions, aspects or elements of the information determining its suitability for a given task. Table I [12] provides a summary of dimensions used in assessing IQ.

TABLE I. Information Quality Categories and Dimensions

Information Quality	Intrinsic IQ	Accuracy
		Believability
		Objectivity
		Reputation
	Contextual IQ	Value added
		Relevancy
		Timeliness
		Completeness
		Amount of data
	Representational IQ	Interpretability
		Ease of understanding
		Representation & consistency
		Conciseness of representation
		Manipulability
	Accessibility IQ	Access
		Security

In [3] authors study the unique characteristics of social media and address how existing methods fall short in mitigating the IQ issues it faces. Despite being extensively studied, IQ theories have yet to be embraced in tackling IQ challenges on social media. They redefined social media challenges as IQ challenges, and they proposed an IQ and Total Data Quality Management (TDQM) approach to the Social media challenges. They mapped the IQ dimensions, social media categories, social media challenges (Table II and III), and IQ tools in order to bridge the gap between the IQ framework and its application in addressing IQ challenges on social media.

The IQ dimensions discussed above can be used to assess IQ of social media. In this paper, these dimensions will be studied since they have been used widely in IQ research and they are the most cited dimensions in IQ literature [3], [6], [13], [14], [15]. They mapped the IQ dimensions to different social media categories as presented in Table II and III. This shows the significance of individual dimension for different types of social media on a scale of high, medium, and low denoted by H, M, and L respectively. Not applicable (NA) is used in the case that the IQ dimension does not apply to a specific social media context.

According to their mapping, Social News is considered to add little value to the needs of its consumers, and not everyone thinks that its content is accurate; for the scales of the significance of Social News, they assigned Low, Medium and High respectively to the IQ dimensions of value-added, accuracy and timeliness. The IQ dimensions discussed above can be used to assess IQ of social media. However, Agarwal and Yiliyasiv [3] found in their study that social media IQ problems do not map to any accessibility and security measures, and they found accuracy is not applicable for media in media sharing and social friendship networks as well. However, this remains a subjective evaluation since social media content can be subjectively and objectively measured using IQ metrics and tools, and very few researchers and social media operators have focused on or utilised IQ frameworks to address the challenges in the social media [3]. Thus, our 'Right-click Authenticate' approach has the potential of playing a significant role in creating and maintaining high-quality social media that deliver high-quality information.

III. RIGHT-CLICK AUTHENTICATE (RCA)

The 'Right-click Authenticate' approach - previously proposed in [16] reviews, ranks, and identifies misinformation. We identified three categories of authentication: textual, imagery, and video misinformation while we concentrated on the first two: textual and imagery authentication. Using this approach, users who want to check the validity of the news could right-click and select authenticate as conceptualised in figure 1 and figure 2. 'Right-click Authenticate' does not prohibit sharing of misinformation; it provides a demonstration of facts in the same layout as Wikipedia [16].

Furthermore, the authors showed that the tools and procedures that might be used to authenticate text and images are available online but may need support among different organisations.

Fig.1. Conceptualising ‘Right-click Authenticate’ option

TABLE II. Mapping Information Quality *Challenges* and Social Media [3]

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Completeness	Amount of Data	Interpretability	Ease of Understanding	Ease of Understanding Consistency	Manipulability	Conciseness	Accessibility	Security
Spam	X	X		X	X	X										
Contextual Relevance						X										
Colloquial Usage and Intentional Misspelling	X				X											
Information Overload									X		X		X	X		
Freshness of Information	X	X		X			X									

TABLE III. Mapping Information Quality *Dimensions* and Social Media [3]

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Completeness	Amount of Data	Interpretability	Ease of Understanding	Ease of Understanding Consistency	Manipulability	Conciseness	Accessibility	Security
Blogs	L	M	L	L	L	M	M	L	M	M	H	L	M	L	L	M
Media Sharing	NA	M	L	L	L	M	M	L	M	L	M	L	L	L	M	M
Micro Blogging	L	L	L	L	L	L	H	L	L	L	M	L	L	H	L	L
Social Bookmarking	M	M	L	L	L	M	M	L	L	L	L	L	L	M	L	M
Social Friendship Network	NA	M	L	M	M	L	M	L	L	L	M	M	L	H	H	L
Social News	M	M	M	M	L	M	H	M	L	L	M	M	L	L	H	H
Wikis	M	M	M	M	L	M	L	M	H	M	H	H	H	M	H	M

'Right-click Authenticate' exemplifies an important step to examine and forecast the dynamic tendency of misinformation dissemination. In [17, 18], we proposed a scheme for combating misinformation online through identifying and demonstrating key variables and factors. The proof-of-concept has been constrained by conventions recognised on remarks of 2D and 3D computer simulation as well as reflective analysis subjective to individual practices of the research group. These simulations provided a way to exam the effectiveness of a control strategy before the actual employment of the control strategy.

IV. Research Objectives

The aim of this paper is to evaluate the RCA approach against the four metrics of IQ and thus to investigate the impact this approach can have on the 'Accuracy', 'Authority/verifiability', 'Validity', and 'Believability' dimensions of IQ. The purpose of this paper is therefore to provide an analysis describing the ways in which RCA impacts the metrics mentioned above. In so doing, we focus on the following research question:

How 'Right-click Authenticate' approach would affect Information Quality metrics for information propagation on social media? To achieve this, the team set out to do the following:

1) Present the current state of each metrics for IQ on Facebook. Facebook has been selected as the case since it has been considered one of the largest social media networks in the world. Although the assessment of each metric is considered subjective, the team will attempt to back their justification with evidence where applicable.

2) The team will then discuss how introducing RCA to Facebook would affect the IQ metrics. This will result in the re-mapping of the metrics presented in Table II and III.

3) Finally, the team will evaluate the impact of the change and its level of significance (if any).

V. Research Outcome

The research outcome follows the research objectives outlined above.

A. Re-evaluating of the IQ metrics for Facebook

In evaluating the IQ challenges of Facebook, the team identified differences when it comes to using Facebook as a source of news. Table IV shows that challenges associated with using Facebook as a source of news have an impact on three IQ metrics. Contextual Relevance of the information with regards to Accuracy, Believability, Value-Added, and Ease of Understanding Consistency became new IQ challenges.

With regards to Information Overload, news sharing on Facebook would have Interpretability challenges. The team did not believe Manipulability of news as an Information Overload challenge, and thus it was removed. Finally, with regards to Freshness of Information, the team agreed that this reflected live news and added Value-Added, Relevancy, and Manipulability as new challenges.

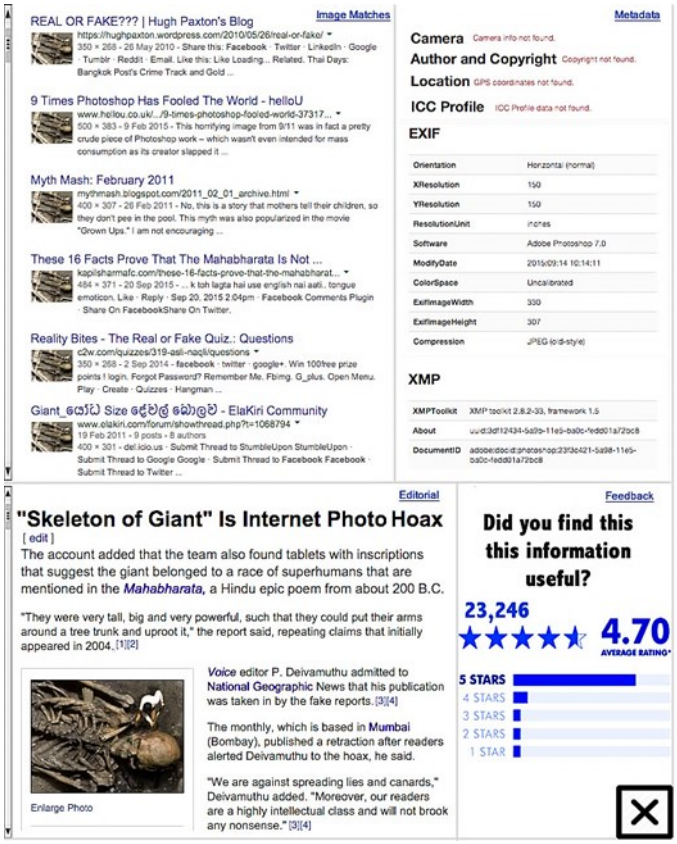


Fig. 2. Conceptualising 'Right-click Authenticate' results

It is evident from this mapping that using Facebook as a source of news introduces new IQ challenges and a sum of 16 challenges.

Table V maps the Information Quality dimensions for Facebook focusing on the one element associated with sharing of fake news and misinformation wherein the original work of [3] it had been identified as 'Social News'.

Our analysis shows that Accuracy should be considered low, Believability should be high, the Amount of Data is medium, Interpretability of information is medium, Ease of Understanding is high, and Conciseness is high. The results represent a fine-tuning that reflects the changes in how Social News is viewed and perceived in recent times. The key problematic areas are the low accuracy, high believability, and low value-added.

B. RCA approach and its impact on IQ challenges

In considering the impact of applying RCA as a tool to help validate news and information online, the team revaluated the IQ challenges. The outcome in Table VI shows a significant reduction in IQ challenges for social media news when a tool for validating misinformation is in use. The team considered that having a tool that would allow means to validate news and information will render concerns about Accuracy, Believability, Objectivity, Reputation, Value-added, Relevancy, and Timeliness in relations to Contextual Relevance and Freshness of Information less challenging. When social media users of Facebook can validate if the news is factual, this will have a cascade effect on all these IQ dimensions. However, RCA does not seem to have much impact on Information Overload as expected earlier in our review since RCA does not stop the spread of information and misinformation. There has been some debate if it could remove the challenge of Amount of Data. Nevertheless, it has been agreed that this would not eliminate this problem.

In evaluating the impact of RCA on Social News dimension, the team found that having means to authenticate news would improve the accuracy of the information provided to a level that is considerably higher than before.

However, this will depend on the number of users who will rely on such approach; as such, the team upgraded accuracy from low to medium.

The team also agreed that reputation would improve as users will trust what they see more and this will upgrade IQ reputation from medium to high. The team also assessed that given how well IQ Accuracy, Believability, and Reputation stand then naturally IQ Value-added should be upgraded to medium.

TABLE IV. Mapping Information Quality challenges and Facebook

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Amount of Data	Interpretability	Ease of Understanding	Ease of Understanding Consistency	Manipulability	Conciseness
Contextual Relevance	X	X			X	X					X		
Information Overload								X	X	X		R	X
Freshness of Information	X	X		X	X	X	X					X	

TABLE V. Mapping Information Quality social news dimension and Facebook

	Accuracy	Believability	Reputation	Value-added	Relevancy	Timeliness	Amount of Data	Interpretability	Ease of Understanding	Consistency	Conciseness
Social News	L	H	M	L	M	H	M	M	H	M	H

TABLE VI. Mapping Information Quality challenges for Facebook using RCA

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Amount of Data	Interpretability	Ease of Understanding	Ease of Understanding Consistency	Manipulability	Conciseness
Contextual Relevance											X		
Information Overload								X	X	X			X
Freshness of Information													

TABLE VII. Mapping Information Quality social news dimension and RCA

	Accuracy	Believability	Reputation	Value-added	Relevancy	Timeliness	Amount of Data	Interpretability	Ease of Understanding	Consistency	Conciseness
Social News	M	H	H	M	M	H	M	H	H	M	H

Finally, Interpretability having means to explain the news and provide its source will most definitely upgrade news dimension to high.

C. *Compare and reflect on the outcomes including further research*

It is evident from comparing Tables IV and VI is that the challenges associated with Accuracy, Believability, Objectivity, Reputation, Value-Added, Relevancy, Timeliness are no longer considered of significant relevance. In fact, the sum of challenges had been reduced from 16 to 5. As for the Social News dimension, the key challenges have been improved including Accuracy, Reputation, Value-added, and Interpretability as shown in Table VII. It is therefore fair to conclude that the introduction of RCA for use by Facebook will improve IQ.

It could also be deduced that improvement to IQ should be expected when applying RCA to other social media networks - although this could be done in future research. What is more, Facebook has introduced its new approach for Fact-Checking approach involving the use of third party checkers [7] with implication on IQ.

VI. Concluding remarks

Managing Information Quality in the age of information and social media has introduced new challenges. Where early research has viewed social media and axillary source of news, evidence from the research shows that this is no longer the case for many online users.

This, in turn, has had an important impact on IQ. This paper has shown that the dimensions of Information Quality can be used to add structure to this inherent complexity. Furthermore, providing means to validate news and information can significantly improve the quality of information users can get.

References

[1] J. Sinclaire and C. Vogus, “Adoption of social networking sites: an exploratory adaptive structuration perspective for global organizations,” Information Technology and Management, vol. 12, 2011, pp. 293-314.
[2] R. Baeza-Yates, “User generated content: how good is it?” in Proceedings of the 3rd ACM workshop on Information credibility on the web, April 2009, pp. 1-2.
[3] A. Nitin and Y. Yiliyasi, "Information quality challenges in social media." International Conference on Information Quality (ICIQ), November 2010.
[4] G. Kane and S. Ransbotham, “Codification and Collaboration: Information Quality in Social Media,” Paper presented at the Thirty Third International Conference on Information Systems, Orlando, 2012.
[5] K. Chai, V. Potdar, and T. Dillon, “Content quality assessment related frameworks for social media,” in Computational Science and Its Applications, ICCSA 2009, pp. 791-805.
[6] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” Communications of the ACM, vol. 40, 1997, pp. 103-110.
[7] P. Pourghomi, F. Safieddine, M. Dordevic, and W. Masri, “How to Stop Spread of Misinformation on Social Media: Facebook Plans vs. Right-click Authenticate Approach,” in IEEE International Engineering and MIS Conference, Monastir, Tunisia, May 2017, in press.
[8] N.P. Nguyen, G. Yan, M.T. Thai, and S. Eidenbenz, “Containment of misinformation spread in online social networks,” in Proceedings of the 4th Annual ACM Web Science Conference, 2012, pp.213-222.
[9] Matt Mahoney “The Limits of Fact-Checking Facebook,” MIT Technology Review, (December 20, 2016).
[10] “Ipsos/BuzzFeed Poll - Fake News,” December 2016, [Online]. Available at: http://ipsos-na.com/news-polls/pressrelease.aspx?id=7497 (Last Accessed on 9 February 2017)
[11] E. Shearer, and J. Gottfried, “News Use Across Social Media Platforms,” Pew Research Center. Washington, USA, 2016.
[12] C. Fisher, E. Lauria, S. C. Smith, and R. Wang, “Introduction to Information Quaity,” in MIT Information Quality Program, 2008.
[13] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,” Information & Management, vol. 40, pp. 133-146, 2002.

[14] S. Besiki, L. Gasser, M. B. Twidale, and L. C. Smith, "A framework for information quality assessment," Journal of the Association for Information Science and Technology, vol. 58, pp. 1720-1733, 2007.
[15] C. C. Chen and T. You-De, "Quality evaluation of product reviews using an information quality framework," Decision Support Systems, vol. 50, 2011, pp. 755-768.
[16] F., Safieddine, W. Masri, and P. Pourghomi, “Corporate responsibility in combating online misinformation,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, 2016, pp.126-132.
[17] M. Dordevic, F. Safieddine, W. Masri, and P. Pourghomi, “Combating Misinformation Online: Identification of Variables and Proof-of-Concept Study,” in Conference on e-Business, e-Services and e-Society, pp. 442-454, Springer International Publishing, September 2016.
[18] F. Safieddine, M. Dordevic, and P. Pourghomi, “Means of Combating the Spread of Misinformation on Social Media: 3D simulation,” In press.

Selection of Appropriate Candidates for a Type Position Using C4.5 Decision tree

Jan Rabcan
Faculty of Management Science and Informatics
University of Zilina
Slovakia, Zilina
jan.rabcan@fri.uniza.sk

Monika Vaclavkova
Faculty of Management Science and Informatics
University of Zilina
Slovakia, Zilina
monika.vaclavkova@fri.uniza.sk

Rudolf Blasko
Faculty of Management Science and Informatics
University of Zilina
Slovakia, Zilina
beerb@frcatel.fri.uniza.sk

Abstract— **Recruitment and selection of new employees rank to the important processes of human potential management and development. Especially the process of employee selection prepares proper conditions for a successful work performance and decides on a future progress-ability of the organizations. In a unique sector of private security, the precise realization of employee selection can solve one of the most frequent problem of the private security organizations: high fluctuation/employee turnover. The paper focuses on an experimental possibility to assign systematically, and with the high measure of exactness, the required competences of candidates to the specification of clients' protected interests or objects. In presented experiment study, built on the basis of benchmarking, the paper presents the concept of well-known decision trees to choose best candidates. Demonstration of proposed system and method is done by public available data, where we achieved 97.27% accuracy of classification.**

Keywords—Decision trees; Data mining; Employee selection; Criteria; Classification;

I. INTRODUCTION

The ability to select, attract and retain the right employees is crucial to the success of any organization [1]. This means the employee selection, i.e. choice of right candidates for a type position, is the fundamental part of many industry areas and represents an important part/subsystem of each sophisticated system of the human resource management. In general, the process of employee selection ranks to the processes with a high organizational and intellectual complexity and needed precision [2-7].

Selecting from the group of candidates involves two main processes: a) listing of suitable candidates; b) decision taking who of the candidates disposes convincingly and truthfully by the presumptions to be the most efficient and successful in a job offered. Decision making on the best candidate (i.e. potential employee or manager) is performed as a sequence of selection procedures (methods) [8], [9] and consists in the comparison of selection criteria [10] with the predictors, i.e. expressed and confirmed abilities, skills, knowledge, competences, and overall creative and capable potential of the candidates. From the viewpoint of responsible selection, "three groups of criteria are important: criteria of company; criteria of section/department; criteria of concrete offered job position" [11]. On the other hand, such a situation can occur that the candidate excellently fulfills a majority of defined criteria compared to predictors but, in some of them, the candidate does not achieve required level. Because of mentioned decisional problem, "it is needed to define basic minimal level in all of the predictors, or, there can be accepted some permissible compensation" [12]. However, the best solution is to apply a multi-criterial decision making. It can be defined as a "collection of methodologies for comparison, ranking and selecting multiple alternatives having multiple attributes" [13]. This is based on the progression of using methods and procedures of multiple conflicting criteria into management processes [14]. When using the multi-criterial approach, such a chance increases, that the company will find, select and employ 'the less as possible improper' candidate.

Several researches have already been carried out in area of supported-by-software choice of suitable candidates to specific job position. In paper [15], the system of HRM which use C4.5 algorithm to detect talent of listed candidates is developed. Also [16] presents the forecast of job candidates' talent by decision trees techniques. In paper [17] several data mining techniques to choose the most talented listed persons are compared. In the concrete, there were used a Multi-Layer Perceptron, Radial Basis Function Network, K-Star, Random forest and C4.5, where C4.5 obtain the best results.

Summing up, the selection of proper applicants, based on data-mining classification, is utilized in many areas, e.g. health care, marketing, finance and many others [17]. Our investigation is focused on choosing the best candidates in a field of private security. This sectors fights with a high employee fluctuation. Mentioned staff turnover is caused by many reasons, e.g. low financial rewards, high responsibility put on the employees in their work performance (when protect the sensitive information or industrial objects, prevent the private or commercial property, guard the clients' health and life, etc.). In addition, the employees are armed – they wear a weapon, are permanently included into various conflict situation, must dispose by a deep expert knowledge (on security and signaling devices, parameters of software used in object protection, etc.). Thereto, it is extraordinarily important to perform correctly the process of new employee selection and judge thoroughly the overall and complicated profile of every applicant's competences, skills, abilities, etc.

Built on mentioned above ideas and presumptions, *the aim of paper* is to search and even confirm the potential possibility to use the proper software methods, tools or techniques in the optimization of employee selection in the private security services. More concretely, the paper deals with the decision-making method for a more improved choice of suitable candidates for a specific employment position based on machine learning methods. A concept of the well-known decision trees to choose best candidates is used when recruiting, selecting and assigning the proper (new) employee to the proper job and thus the 'objects' that have to be protected.

II. Private security in Slovak Republic

Security, i.e. the protection of persons' life, health, and property is an integral part of the security policy of every democratic state. Especially at current, after terroristic attacks in Brussel, London, Manchester, etc., the fight against terrorism and other unlawful conduct play an important role in this respect. The international nature of organized crime requires a common approach from the European Union. This leads to the necessary state that in a combination with the overall society interests, also the protection of private property and information becomes increasingly the front line of national governments.

In December 1997, Slovak Republic adopted Act no. 379/1997 Coll.: The operation of private security services and similar activities [18]. Security services provide privacy protection for individuals and property. The law has created basic legal assumptions and conditions for an acceptable and state-controlled private security service. At present, the security industry has developed dynamically; security service companies employ more than 25,000 service providers. The level of clients' confidence in this service depends on the quality of its service providers. The quality of security services affects the selection of candidates for considered jobs/type positions. An effective system for the selection of private security staff needs improvement. For example, the law has put the performance of a detective service to a level of standard physical protection. However, the requirements for a *detective type position* are much higher (for example, qualification requirements, expertise, and general capability). The described method devalued the system of personnel selection of the detective service providers.

The actual process of employee selection in the private security services has not yet been modified. Efforts to identify core competence requirements for some types of jobs/positions can be assessed positively. The National Occupation System – Employee Card and Integrated System of Types of Positions were created in Slovak republic.

As the text above shows, there is no procedure objectively capable to identify candidates' competences depending on the type of position and the job. This is why the research project has been developed to address these issues. A title of scientific project, those this paper is a result, is *"Optimization of the competences in correlation with the particularity of type positions in security services."* This one is granted by Ministry of Education of Slovak Republic and its purpose consists in helping to solve a relatively unfavorable situation in the private security services. A unique aspect of the project is its multidisciplinarity; the project solvers are considered as the experts in various different areas, e.g. psychology, human potential development, organization behavior, applied informatics, statistics, etc. Mentioned structure of the participants enables the project team to understand the set goals from the different point of views, and in connected effort, work out the new, original known in this area. This means, the part of the project solution is also the design of *a supportive software tool* to eliminate the turnover and quality problems. The role of the instrument will make it easier to the most suitable candidates for the appropriate type position.

In the selection process, employers often look for the most suitable candidate for a given job. The process of selecting candidates for a type of position in security services was analyzed in [19]. It consists of three basic phases:

- Preliminary phase containing the specification and description of the type position, definition of suitability of employees and recruitment.
- Selection phase, which includes the acquisition of backgrounds from potential candidates.
- Evaluation phase in which the process ends with the selection of the most suitable candidates. The evaluation phase of the process has 4 parts that are time-bound [19]:
  - Evaluation of the selection,
  - Selective final interview,
  - Decision to select a particular tenderer,
  - Informing the tenderer of the selection.

The tool we are proposing will be used in the preliminary phase and evaluation phase of the process of finding the most suitable candidates.

Typical positions in security management have predefined certain legal requirements based on the abovementioned laws. It is therefore possible to accumulate them together. With software technologies, the whole process can be greatly simplified in the first phase. However, the selection itself can be carried out by various means. The described evaluation phase also allows solution through machine learning algorithms. In our paper, we propose to use the C4.5 decision tree to select the most proper candidates as possible.

The system for automated resume evaluation (requiting system) in private security in Slovak Republic is developed under the project VEGA 1/0064/15 named as "Optimization of the competencies in correlation with the particularity of the type positions in security services"

The first step in creating the software was to define the *functionality requirements* of the system. The basic requirement was to *create a database of individual types of positions* (jobs) as well as *a database of all attributes and features* of each type of position. This effort was based on:

- The requirements of Act no. 473/2005 Coll. [20], based primarily on integrity, reliability, medical fitness and required professional competence;

- Register of occupations of the National System of Occupations – part Security services.

The second step in project solution consisted in building the group of data and database requirements with the specificities of each object in terms of risk analysis. It means the project team had to work out a specification of external and internal security risks, threat resources in a given object and the overall indoor and outdoor situation specific to the given object in which the type position will be performed. The result of this step will be *Decision Support Making System* (DSMS) to choice a candidate according to indicated requests.

These two steps can be developed at the same time. The background of DSMS and principal tool of this system to choice of a candidate can be proposed and elaborated based on data that are similar to expected data in the first step. Therefore we use typical data to develop and evaluate the planed DSMS.

III. BACKROUND OF PROPOSED METHOD

There are some types of software and information system that can be used in selection of candidates for indicated job position [21, 22, 23, 24] and one of them is decision support system [21]. The prototype of system for choice of candidates for a type position will be developed as decision support making system. The indicated structure of databases (see section III) causes DSMS type that can be classification procedure. The principle of classification is in the learning function that maps (classifies) data records into one of the several predefined classes. The implemented DSMS of candidates' selection for job position can be implemented as a *decision tree* (DT) that is one of the possible approaches for classifying data into known classes. DT are a widespread and useful tool of machine learning designed for classification and prediction. There are several reasons for applying DT technique. The main reason is in its clarity and easy interpretation, which is close to the natural way of human thinking. It allows to users quickly and easily evaluates obtained results, finds key attributes and search for interesting data segments. DT consists of nodes, where the top-level node is referred as a root. Each internal node represents a test on individual input attribute (the root node is also a test) and outcome edges of internal nodes mean a possible test result. External nodes are denoted as a leaf which represents individual classes. Attributes associated with internal nodes can be chosen by numerous criterions. The goal is to identify those attributes that can split the records with a maximum reduction of uncertainty.

There are different types of decision tree and algorithms for their construction as ID3, C4.5, CART, CHAID or fuzzy decision trees based on cumulative mutual information [24]. This idea (proposition) is approbated and evaluated based on typical dataset of ML [25]

One of the first algorithm ID3 for decision tree induction developed by Quinlan in [26] uses information gain which is based on entropy measurement. Shannon entropy of set S of examples is defined as:

$$H(S) = -\sum_{j=1}^{C} \frac{k_j}{|S|} * \log_2 \frac{k_j}{|S|},$$

where $|S|$ is the cardinality measurement of set S and k_j is number of instances belonging to the j-*th* class. This criterion expresses how greatly the uncertainty of the set S is reduced after splitting by attribute A or Information gain $G(S, A)$ is measurement of entropy before splitting and after splitting of set S by attribute A. It is expected reduction of entropy. Information gain $G(S, A)$ is defined as:

$$G(S, A) = H(S) - H(S|A) = H(S) - \sum_{v \in vals(A)} \frac{|S_v|}{|S|} H(S_v),$$

where S_v is the subset of S where samples have value of attribute A equal to v. Information gain tends to prefer attributes with a numerous amount of distinct values. Quinlan improved in [27] this problem in algorithm C4.5 by using gain ratio computed as

$$Gr(S, A) = \frac{G(S, A)}{-\sum_{v \in vals(A)} \frac{|S_v|}{|S|} * \log_2 \frac{|S_v|}{|S|}}$$

To increase the classification performance of decision trees, a technique called tree pruning is used. The idea of this technique is in replacing of sections of the tree by a leaf if one of the pruning criterions is satisfied [28]. Not pruned trees work well in case if trees are used for classification of the training set, but these trees usually contain leaves with an insufficient number of examples given to the size of the dataset. Such decision trees correctly classify training instances, but it is likely that the classification of instances not included in training set will be incorrect. The described algorithm uses three thresholds to check creation of a leaf node.

- *Minimal leaf size.* This criterion indicates the number of instances of a node in its subset S. The tree is inducted in such a way that if subset S of a node contains at least number of instances equal or less to the minimal leaf size, this node is established as a leaf.
- *Minimal gain.* The gain of the node is calculated before dividing the dataset in the node. If this gain is more than the value of this criterion, then the split is executed, otherwise, the node is transformed into a leaf. The increasing value of this criterion will reduce the number of splits. It makes the depth of the tree smaller. If too large value is selected, it will prohibit the execution of any split, and the tree will be created by only one root node.
- *Confidence.* It is statistical pruning technique, which is applied after tree induction (post-pruning). The advantage is that all data from the training set are included in the training process if we do not use some pre-pruning technique. This criterion indicates the confidence level used for the pessimistic error calculation of pruning.

A resulting decision tree is usually obtained by running the induction algorithm several times with different threshold values and analyzing the obtained results [29]. Especially for

predictive methods, threshold values can be obtained in the way shown in the Figure 2.

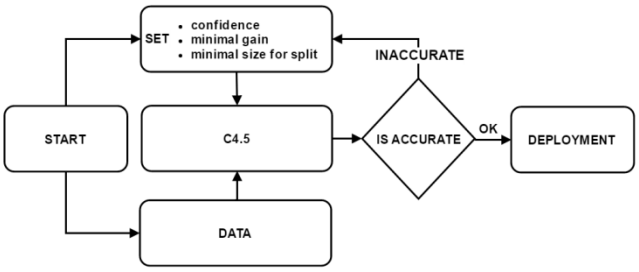


Figure 1. Threshold values for predictive methods (own study)

IV. Experimental Settings

According to requirements of implemented project (VEGA) the information system for selection of candidates for indicated job position is developed. The evaluation of the indicated types of a system based on DT has been realized based on typical data set from ML repositories [30].

We considered algorithm C4.5 for construction of decision trees. For these purposes, the information gain ratio is used. The advantage of decision trees is in their interpretations and the ability to identify the key attributes of the analyzed dataset [30]. The task of the selection of proper candidates to a type position is close related to the human resource (HR) analysis. Therefore, the demonstration of proposed method is done by through public available dataset for HR analysis, which has data of similar structure to data used in the proposed system.

Human resource analysis is free public dataset available at [30]. The dataset contains 15 000 records, where each record corresponds with one employee. It consists of 9 input attributes and one output attribute. In this case, the aim of classification is focused on *identifying key aspects,* why the most valuable employees which are suitable for its positions leave organizations. Detail description of attributes is in Table 1.

Table 1. Table Type Styles

Attribute description		
names	***description***	***Type***
satisfaction_level	Employee satisfaction level	numeric
last_evaluation	The grade the employee got at their last evaluation. Ranges between 0 and 1.	numeric
number_projects	The number of projects the employee is currently working on	numeric
average_monthly_hours	Average monthly hours	numeric
time_spent_company	Time spent at the company	numeric
salary	Level of salary	categorical
sales	Department in which employees work for	categorical
work_accident	Whether they have had a work accident (1 or 0)	categorical
promotion_last_5_years	Whether they have had a promotion in the last 5 years (1 or 0)	categorical
Left	Determine whether the employee left the workplace or not (1 or 0)	categorical (output)

Dataset was randomly divided into training and testing sets in ratio 80:20. Instances from testing sets are used only for training of DT and testing set is used for evaluating of performance of classifier (Figure 3).

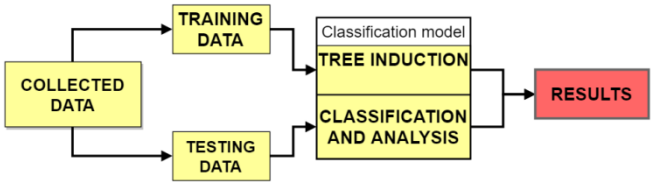


Figure 2. Dataset: training and testing data (own study)

During experiments, we were aimed at determination of the best values of the thresholds. Experiments were performed in the loop, were threshold was changed at begin of each iteration and then was inducted new decision tree. If the model error is high (i.e. its predictions are not correct for the testing data), then the parameters of the algorithm that has been used to create the model are changed and a new model is created. When the loop finished, the best decision tree was chosen. The process of obtaining final model is shown in Figure 4.

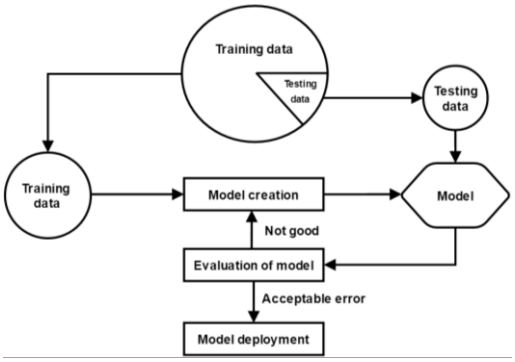


Figure 3. Process of obtaining final model (own study)

V. Evaluation of Proposed Decision Support Making System

The evaluation phase of selecting suitable candidates for a type position is to decide whether the candidate is suitable or not. We use decision tree C4.5 as a support tool in process of candidate selection. The final decision should be always confirmed by a human. This tree has been trained on the public dataset because the structure of used dataset is suitable to validate our method. Detailed description of the dataset is in section IV. The resulting decision tree is shown in figure 5.

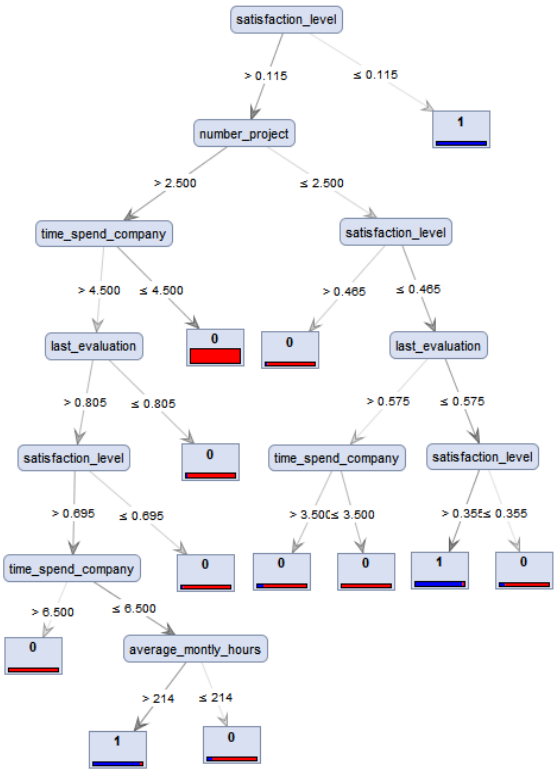


Figure 4. Experimental model (own study)

TABLE 2. TABLE OF CANDIDATES

Candidates example		
names	***Candidate 1***	***Candidate 2***
satisfaction_level	0.3	0.215
last_evaluation	2	1
number_projects	3	1
average_monthly_hours	220	212
time_spent_company	4	8
salary	1600	1900
sales	3	3
work_accident	0	0
promotion_last_5_years	0	0
Left	?	?

Table 2 contains two generated examples of candidates. Using data about candidates contained in Table 2, we can show how classification using decision trees works.

The advantage of decision trees is that they allow an easy-to-understand and simple classification. The new instance passes through decision tree from its root to leaf. For *Candidate 1* from Table 2 is the process of decision taking following. The root of the tree is associated with *satisfaction_level.* The test on this attribute is if instance (*Candidate 1*) has bigger or equals value of this attribute than 0.115. *Candidate 1* has the value of *satisfaction_level* equal to 0.3, therefore we have to continue through the left branch to the node associated with number_project. *Candidate 1* working on 3 projects, so next decision move instance into the node associated with time_spend_comapny. *Candidate 1* spends at company 4 years which is less than 4.5 so instance continues to the leaf, where the proposed decision is 0. It stands for that *Candidate 1* is not leaving the organization according to the classification procedure. The whole movement of instance through decision tree is shown in figure 6.

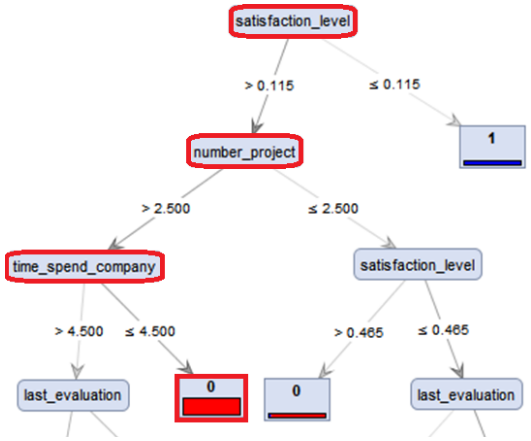


Figure 5. Example of classification (own study)

The same classification steps are applied also for *Candidate 2*. The *satisfaction_level* of *Candidate 2* is equal to 0.215, therefore classified instance (*Candidate 2*) continue to the node associated with *number_project* where we know, that *candidate 2* is currently involved in one project. It causes that instance continues to the right branch. At the end of this branch is node associated with satisfaction level, where the test checks if satisfaction level of an instance is bigger than 0.465. If this level is bigger than this value then classification ends with result that *Candidate 2* is not leaving the organization. The process of classification of *Candidate 2* is shown in figure 7.

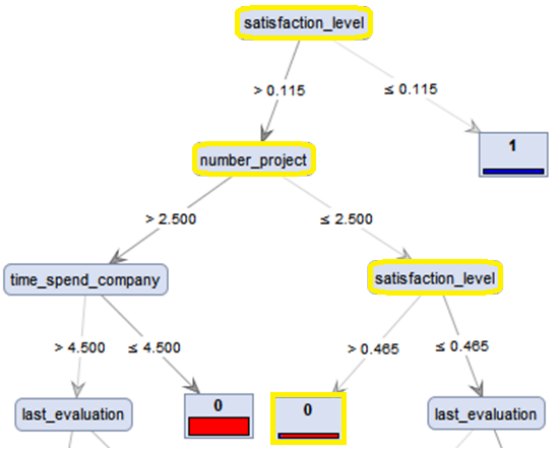


Figure 6. Example of classification (own study)

In this work, we evaluated the quality of the proposed DSMS with the accuracy of the classification. Classification accuracy is the percentage of correctly classified instances. It is the ratio between the count of correctly classified instances and the number of classified instances. We computed the accuracy as:

$$accuracy = \left(\frac{\sum_{i=1}^{k} c(I_i)}{k}\right) * 100,$$

where k is the number of classified instances, I_i is the i-th instance from I, $i = 1, \ldots, k$ and $c(I_i)$ is given by:

$$c(I_i) = \begin{cases} 1, & \text{if } classify(I_i) = \text{class of } x \\ 0, & \text{otherwise} \end{cases},$$

where the function $classify$ returns the resulting class of the classification. The accuracy is also computed for each output class and the result is displayed at confusion matrix also known as error matrix. This matrix is used for evaluating classifiers. The values in the rows indicate the number of instances that have been predicted for each class. The values in the columns represent the actual number of instances belonging to each class. This matrix is in the Table 3. The best achieved threshold values witch the corresponding score of classification accuracy are shown in Table 4.

TABLE 3. CONFUSION MATRIX FOR HR DATASET (OWN STUDY)

Attribute description			
	true 1	***true 0***	***class precision***
pred. 1	3257	96	97.14%
pred. 0	314	11332	97.30%
class recall	91.21%	99.16%	

TABLE 4. TABLE OF RESULTS (OWN STUDY)

Description	***Results***
Total accuracy	97.27%
Confidence	0.17
Minimal leaf size	100
Minimal gain	0.15

VI. CONCLUSIION

The described software tools and algorithms allow user to create the required unified procedure for selecting suitable candidates for security positions.

In this paper decision support making system is developed based on sets of criteria used in security services companies. This criteria set is used for data gathering about individual candidates. The presented software tool is system prototype and has been elaborated based on decision tree. This tree has been inducted according to the algorithm C4.5. The initial data for this tree is the public dataset "Human Resource Analysis" from the repository Kaggle [30]. The application of this dataset allows evaluating of possibility to use the decision support making system for automated resume evaluation tool (requting system). The proposed software tool for the system prototype implements classification with the accuracy that is equal to 97.27%.

The proposed system and method can be used not only in security services. The modification of positional requirements and collection initial data about possible candidates allows us to develop similar system in other areas to select candidates. In feature work, we can use fuzzy decision tree (FDT) instead of C4.5. Usage of fuzzy data can reduce the uncertainty and ambiguity in the original dataset. Also, usage of FDT can increase in better classification accuracy. Fuzzy decision trees are less sensitive to outliers and errors during measurement [31-32].

ACKNOWLEDGEMENT

This work is partly supported by grant VEGA 1/0064/15 named as "Optimization of the competencies in correlation with the particularity of the type positions in security services".

REFERENCES

[1] I. Bakanauskienė, R, Bendaravičienė, and I. Bučinskaitė, "Employer's Attractiveness: Generation Y Employment Expectations in Lithuania," *Human Resources Management and Ergonomics,* vol. 10, no. 1, pp. 6–22, 2016.

[2] W. B. Werther, and K. Davis, *"Lidský faktor a personální management"* (Human Factor and Personnel Management). Praha: Victoria Publishing, 1992.

[3] J. Koubek, *"Řízení lidských zdrojů"* (Human Resource Management), 4th edition. Praha: Management Press, 2007.

[4] M. Armstrong, *"Handbook of Human Resource Practice"*. London: Kogan Page, 2009.

[5] K. Wellington, *"Effective People Management"*. London: Kogan Page, 2011.

[6] S. Borkowski, and J. Rosak-Szyrocka, "Estimation Director's Features Basis on Twelve Golden Principles," S. Borkowski, and R. Stasiak-Betlejewska, (Eds.). *Human Potential Management in a Company. Manager Features.* Zagreb: Damir Jelačić, pp. 93–113, 2011.

[7] S. Lorincová, and G. Giertl, "Proposal for Improvement of Acquisition, Selection and Recruitment of Employees in Small and Medium-Sized Enterprises of Slovak Wood Industry," *Human Potential Development,* Klaipeda 27–28 May, 2015, pp. 127–135.

[8] L. L. Byars, and L. W. Rue, *"Human Resource Management,"* 5th edition. Boston: Irwin/McGraw-Hill, 1997.

[9] M. Blašková, *"Rozvoj ľudského potenciálu. Motivovanie, komunikovanie, harmonizovanie a rozhodovanie"* (Human Potential Development. Motivating, Communicating, Harmonizing and Decision Making). Žilina: EDIS, 2011.

[10] A. Kachaňáková, K. Stachová, and Z. Stacho, "Organizational Culture Affects All Formalised Activities of Human Resources Management," *Human Resources Management and Ergonomics,* vol. 8, no. 2, pp: 61–71, 2014.

[11] C. Lewis, *"Employee Selection."* London: Hutchinson, 1985.

[12] A. Kachaňáková, *"Riadenie ľudských zdrojov"*

(Human Resource Management). Bratislava: Sprint, vfra, 2007.

[13] J. K. Levy, *"Multiple Criteria Decision Making and Decision Support Systems for Flood Risk Management"*. Munchen: Springer-Verlag, 2005.

[14] Umme-e-Habiba, and S. Asghar, "A Survey on Multi-Criteria Decision Making Approaches," *International Conference on Emerging Technologies 2009,* pp. 321–325, 2009.

[15] H. Jantan, A. Razak Hamdan, and Z. Ali Othman, "Human Talent Prediction in HRM using C4.5 Classification Algorithm," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 8, pp. 2526–2534, 2010.

[16] H. Jantan, "Classification for talent management using Decision Tree Induction techniques," *Data Min. Optim. ….*

[17] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Data Mining Classification Techniques for Human Talent Forecasting."

[18] Act No. 379/1997, *On the Operation of Private Security Services and Similar Activities,* National Council of Slovak Republic, 1997.

[19] D. Vidriková and K. Boc, *Personálne aspekty výberu ľudských zdrojov do služieb súkromnej bezpečnosti*. Zilina: Zilinska univerzita, 2014.

[20] Act 473/2005, *On Providing Services in Private Security and Amending and Supplementing Certain Laws (Law on Private Security).* National Council of Slovak Republic, September 23, 2005; January 1, 2013.

[21] N. A. Ruskova, "Decision support system for human resources appraisal and selection", Proceedings First International IEEE Symposium Intelligent Systems, vol. 1, 2002, pp.354 - 357, DOI: 10.1109/IS.2002.1044281

[22] V. Lai, K.J. Shim et al. CareerMapper: An automated resume evaluation tool, 2016 IEEE International Conference on Big Data (Big Data), pp. 4005 - 4007, 2016, DOI: 10.1109/BigData.2016.7841091

[23] J. Malinowski, T. Keim et al., Matching People and Jobs: A Bilateral Recommendation Approach, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), vol 6, 2006, pp. 137c - 137c, DOI: 10.1109/HICSS.2006.266

[24] I. Androulidakis, V. Levashenko, E. Zaitseva, An empirical study on green practices of mobile phone users, Wireless Networks22 (7), 2016, pp. 2203-222 DOI: 10.1007/s11276-015-1097-7

[25] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*

[26] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[27] J. R. Quinlan, "Programs for machine learning. Morgan Kaufmann," p. C4. 5, 1993.

[28] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," *Appl. Rev.*, vol. 35, no. 4, 2005. DOI: 10.1109/TSMCC.2004.843247

[29] J. Rabcan, "Ordered Fuzzy Decision Trees Induction based on Cumulative Information Estimates and Its Application," *ICETA*, p. 6, 2016. DOI: 10.1109/ICETA.2016.7802047

[30] "Human Resource Analysis." [Online]. Available: https://www.kaggle.com/ludobenistant/hr-analytics.

[31] E. Zaitseva, V. Levashenko, Construction of a reliability structure function based on uncertain data, IEEE Transactions on Reliability, 65 (4), 2016, pp. 1710-1723, DOI: 10.1109/TR.2016.2578948

[32] E. Zaitseva, M. Kvassay et al. Introduction to knowledge discovery in medical databases and use of reliability analysis in data mining, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS 20150, 2015, pp. 311-320, DOI: 10.15439/2015F327

FPGA LUTs for a Logic Systems

Tyurin Sergey, Prokhorov Andrey, Vikhorev Ruslan
Faculty of Electrical Engineering; Department of Automation and Telemechanic
Perm National Research Polytechnic University
Perm, Russia
tyurinsergfeo@yandex.ru, npoxop007@yandex.ru, vihrusvla@mail.ru

***Abstract*—FPGA logic based on Look up Tables (LUTs). However, LUT calculates only one logic function in the perfect disjunctive canonical forms (PDCF) for this configuration. The paper proposed the concept of the logic advanced LUTs in three main areas. The first area is Double LUT (DLUT), computes two functions simultaneously with inactive transmission transistors subtree. The second area is DC LUT through inverse this tree to implement the decoder DC, which allows computes whole system of the logic functions. Such technique can significantly reduce hardware expenses for logic systems. The third area is DNF-LUT, which allows the calculations of the system functions in disjunctive normal form (DNF) and else more significantly reduces large number of variables LUTs hardware costs. We explored the models of the proposed concepts in the NI Multisim 10 by National Instruments Electronics Workbench Group. The paper analyzes the assessments of the complexity of the LUT, the conclusions about the effectiveness of the proposed solutions.**

Keywords— look up table; FPGA; perfect disjunctive canonical forms

I. INTRODUCTION

FPGA chips are widely used in computer technology [1-3]. There are quite a large number of energy-saving methods of configuring FPGA [1, 2] for example, such as energy-efficient mapping and clocking, unused blocks power down and others. In Ph.D. Dissertation [4] suggested an FPGA post-fabrication component-specific mapping and an optimized architecture taking into account the characteristics of individual transistors, identified during the operational phase. It uses minimum energy/operation indicator. However, the expansion of the actual FPGA logic capacity for energy efficiency presented in the available sources not full. For this, it is possible to use logic optimization [5-7]. This particularly applies to the implementation of logical systems, an example of which given by CPLD [8].

A goal of the paper is research and development logic elements-LUTs by reducing the complexity of the realization of logic functions of a large number of arguments. It presented LUT PDCF technique – DLUT & DC LUT, computes two and more functions in perfect disjunctive canonical forms (PDCF) simultaneously. Secondly is devoted LUT DNF technique by analogy with Programmable Logic Array (PLA). Described the comparison of the number of transistors to implement system of the logic functions in the known LUT and in all proposed LUT.

II. METHODOLOGY

A. LUT FPGA Calculates Two Functions Simultaneously

Consider the LUT model on the two variables (2-LUT), configured to calculate the exclusive OR function is shown in Fig. 1.

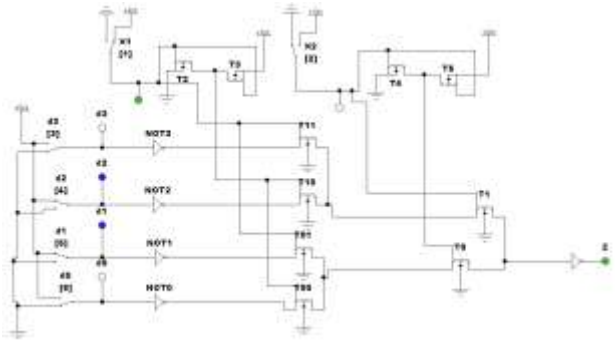


Figure 1. The 2-LUT model configured to calculate the exclusive OR function.

LUT is allocated only one logic function *z*, customizable by the user by downloading the configuration memory SRAM (d0-d3). At the same time during the computation is always activated only half the tree transistors (T0&T00 or T01 when $x_2 = 0$; T1&T10 or T11 when $x_2 = 1$). This creates the conditions for the use of idle half of the transistors with the introduction of another pair of leading variable. However, this requires connectivity configuration memory SRAM, which stores the settings of the second function. The corresponding truth table is stored "backwards" compared to the truth table of the first function, Fig. 2.

The simulation of the Duble2-LUT (D2-LUT) calculates two function $z_1 = x_1 \leftrightarrow x_2$, $z_2 = x_1 \oplus x_2$, $x_1 = x_2 = 0$ in the system NI Multisim 10 is presented in Fig. 3-5. To use the second half of the tree transistors introduced additional transistors leading variable T0.1 and T1.1 and transistors connection settings first function, T00.1, T01.1, T10.1, T11.1, second function T00.2, T01.2, T10.2, T11.2. The setting will simulate constant connection - supply pins V_{cc} and Ground - "zero volts". Simulation confirms the efficiency of the proposed scheme DLUT.

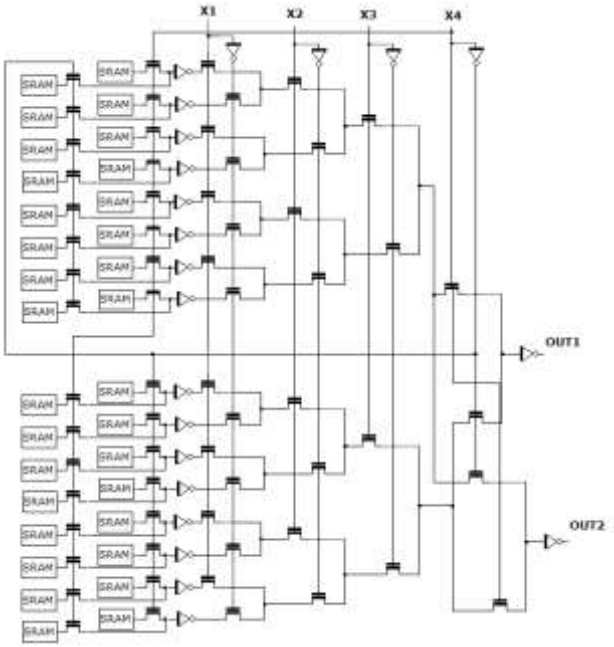


Figure 2. Duble4-LUT (D4-LUT) calculates two functions simultaneously

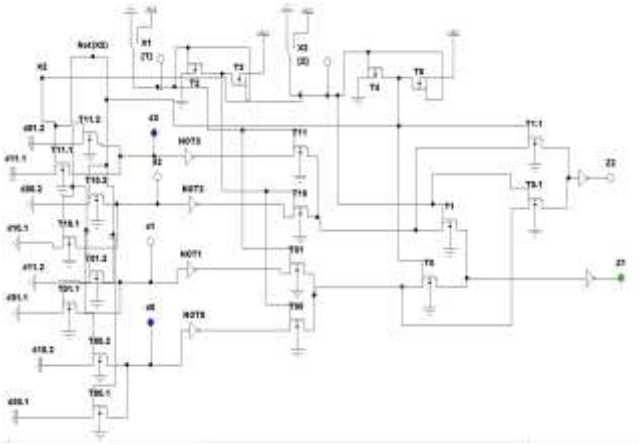


Figure 3. The simulation of the D2-LUT

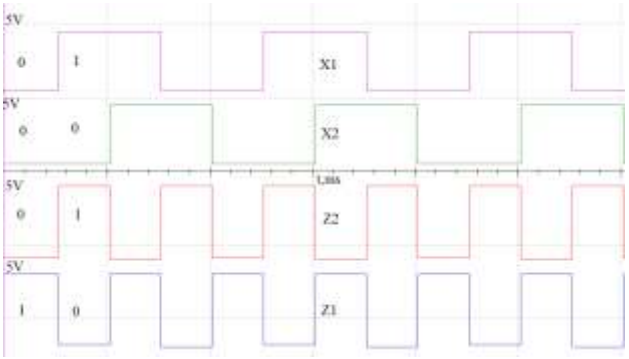


Figure 4. D2-LUT – waveform

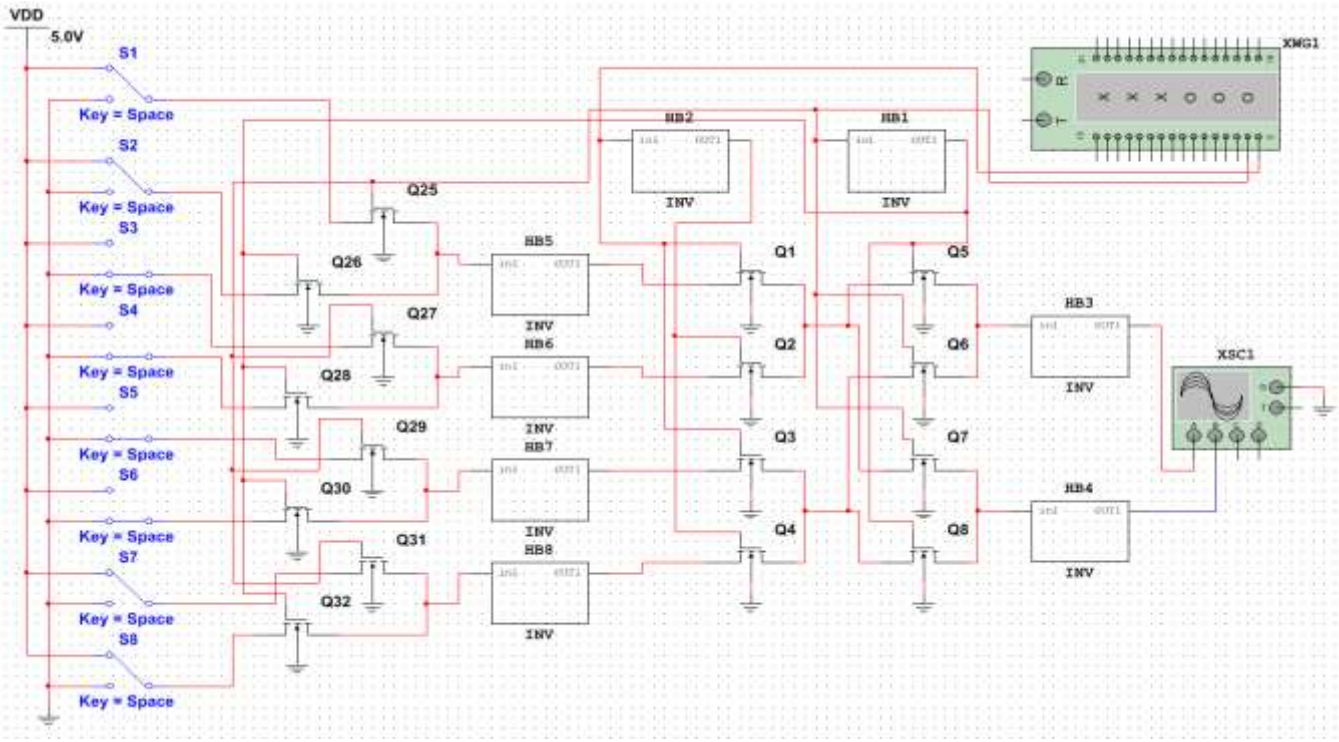


Figure 5. D2-LUT - dynamic model

B. DC LUT FPGA, Calculates System of the Logic Function

The transistor tree "reverse" structure LUT (see. Fig. 1) obtained by “reflection” on the LUT horizontally [9], Fig. 6.

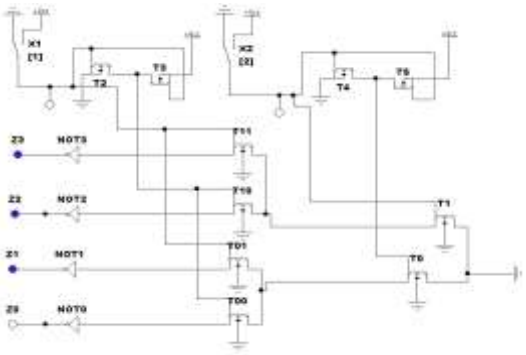


Figure 6. "Reverse" 2-LUT structure

In compliance with the design rules circuits of the transmission transistors required for a drain of each transistor T00, T01, T10, T11 (Fig. 14) to create an alternative chain, transforming its output is guaranteed, for example, in a logic “1”. The best option is to create an alternative transistor for each transmission transistor - as it presented in Fig. 7.

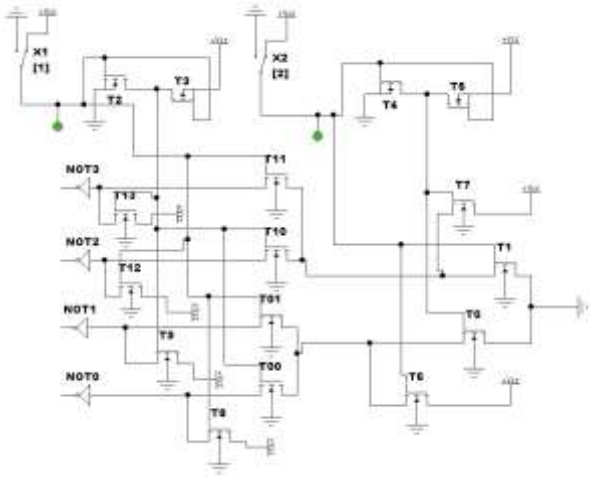


Figure 7. Reverse tree 2-LUT (DC 2-LUT) with alternate transistors T6, 7, 8, 9, 12, 13.

Get the decoder with the output function *z* (without alternative chains). Further, *m* times by combining the OR, the corresponding outputs of the outputs we get the implementation of the system of *m* n-bit logic functions based on perfect disjunctive normal forms (PDNF). DC 2-LUT with one programmable unit disjunctions configured to implement $x_1 \oplus x_2$ show Fig. 8-10.

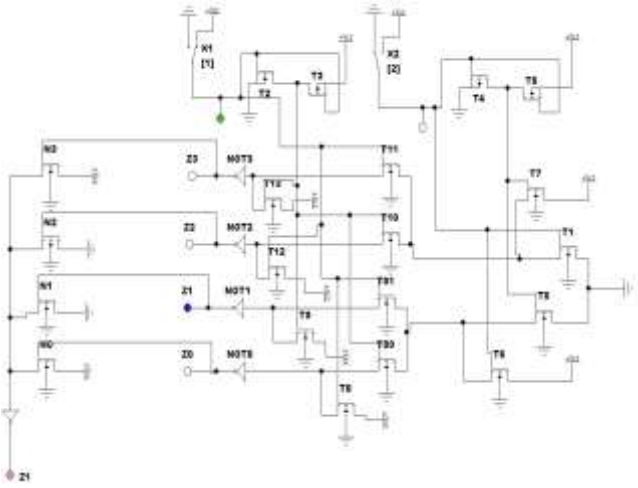


Figure 8. DC 2-LUT - static model

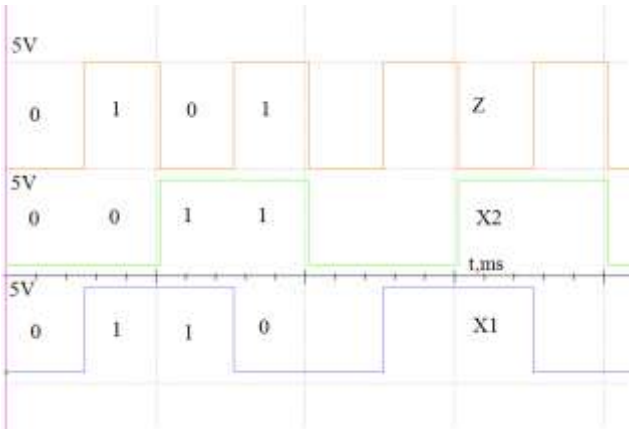


Figure 9. DC 2-LUT - waveform

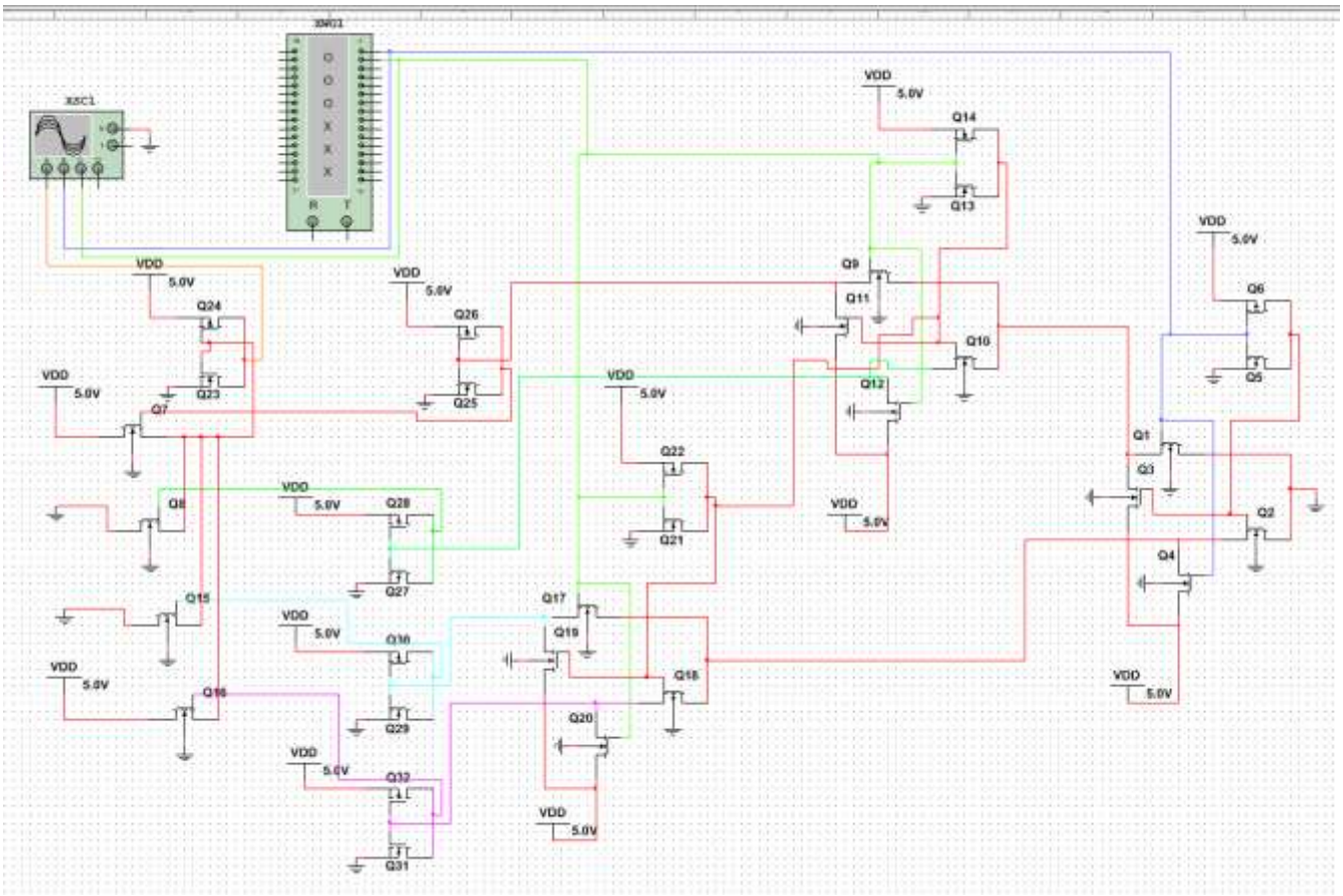


Figure 10. DC 2-LUT - dynamic model

C. Advanced LUT FPGA for Disjunctive Normal Form (DNF) of the Logic Functions Architecture DNF -LUT

The proposed new DNF-LUT [10] is a user-configurable structure similar to a programmable logic array PLA. Architecture DNF-LUT presented on the Fig. 11.

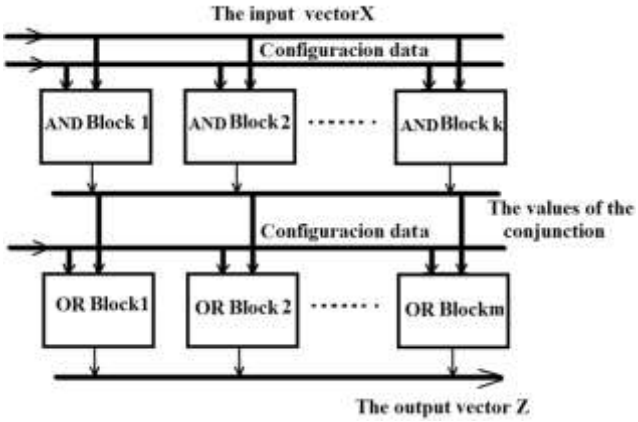


Figure 11. Architecture DNF-LUT

Thus, instead of loading the truth table, it is loading only values programmed conjunctions length *n*, where *n* - the number of variables of *m* logic functions. Occurrences *k* conjunctions in *m* functions are also programmable tuning functions. For a given input set (vector *n* variables *x*) *k* AND blocks calculate value of *k* conjunctions, which then form "an OR" value of *m* logic functions. The proposed structure of the AND DNF-LUT block is shown in Fig. 12.

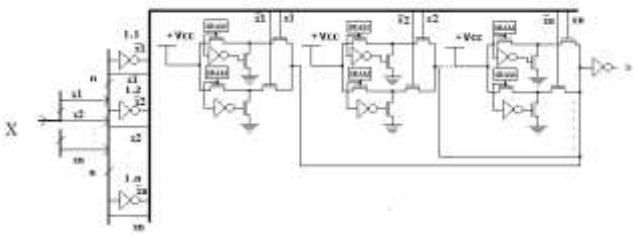


Figure 12. Architecture of the AND DNF-LUT block

One variable SRAM Setup is determined as shown in Fig. 13.

SRAM X	SRAM not X	Output 1	Output 0
1	0	When X	When not X
0	1	When not X	When X
1	1	Anyway	-
0	0	Banned	Banned

Figure 13. One variable AND DNF-LUT SRAM Setup

Thus, if the variable activated, "right", block AND, transmits a logic one signal from the input (left) to the output (right). The same occurs with the immateriality of the variable, i.e., for any value of the variable. If activated "wrong" variable, using inverters and additional transmission transistors supply to the output logical zero. If all variables are the "right" - the

output z_i is a logical zero. The proposed structure of the block OR DNF-LUT show Fig. 14.

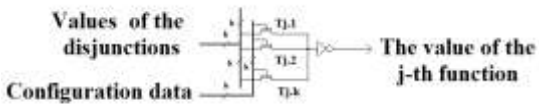


Figure 14. Architecture OR DNF-LUT block

A logical "1" at the output of the corresponding function activated when the inputs given conjunctions are zeros. The simulation of the block AND DNF-LUT executed in the system NI Multisim 10 by National Instruments Electronics Workbench Group show Fig. 15.

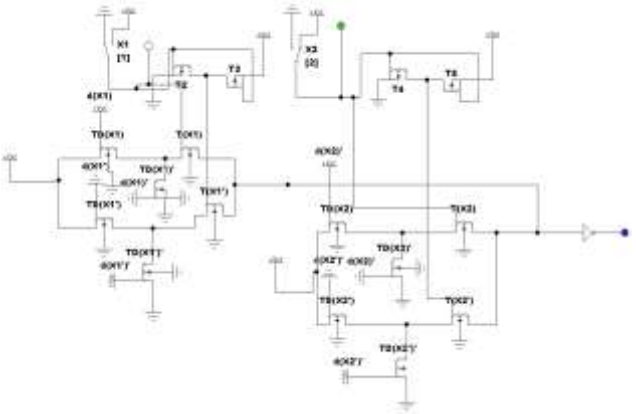


Figure 15. AND DNF-LUT Multisim model: $x_2 x_1 = 0$, since $x_2 = 1, x_1 = 0$

Given a conjunction $x_2 x_1 = 0$, since $x_2 = 1, x_1 = 0$. In case $x_2 = 1,\ x_1 = 1$ a conjunction $x_2 x_1 = 1$ - Fig. 16.

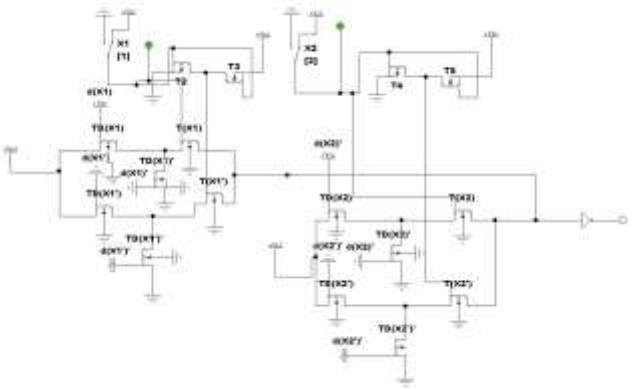


Figure 16. AND DNF-LUT Multisim model: $x_2 x_1 = 1$, since $x_2 = 1, x_1 = 1$

Checking the operation of the rest of the sets also confirms the efficiency of the proposed technical solution. The simulation of the block OR DNF-LUT in the system NI Multisim 10 by National Instruments Electronics Workbench Group presents Fig. 17-19.

Correctly, shaped value function and other combinations of k_1, k_2, k_3, k_4 , modeling confirms the efficiency of the unit disjunctions.

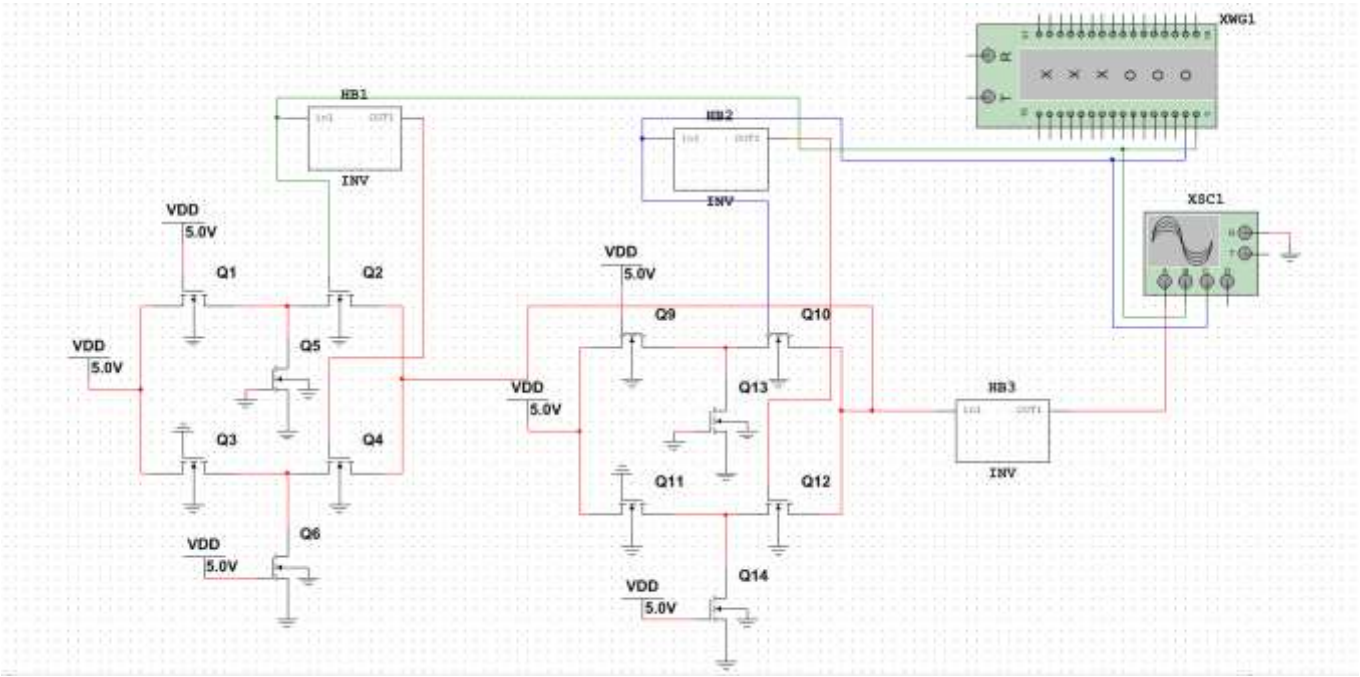


Figure 17. DNF – LUT - dynamic model

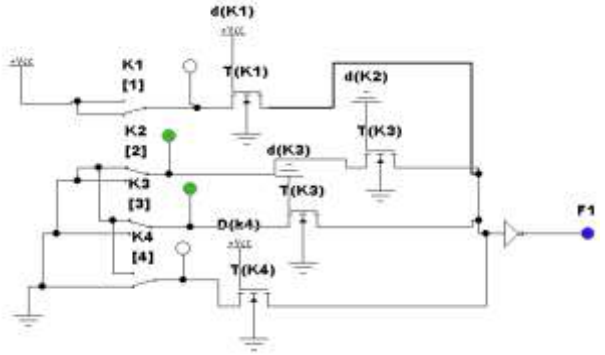


Figure 18. OR DNF-LUT Multisim model: Function $F_1 = 1$ as active k_1 and k_4.

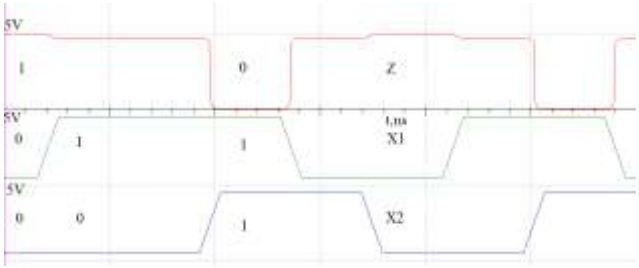


Figure 19. DNF – LUT - waveform

III. Results Of Calculations

Let k - is the dimension of the main (basic) LUT ($k \in \{1, 2, 3, 4\}$). Let us estimate the complexity of the LUT without decomposition. "Ideal" complexity can be only up to $n = 4$, not more than:

$$L_n = 2^n \cdot 8 + 2^{n+1} + 2n \quad (1)$$

With the decomposition n-tree for k LUT, we get "real" complexity:

$$L_{n.k} = 2^n \cdot 8 + (2^{k+1} + 2k) \cdot 2^{n-k} + (2^{2^{n-k}+1} + 2^{n-k+1}) + 2n \quad (2)$$

Then we get the complexity of the two logical functions:

$$2 \cdot L_{n.k} = 2[2^n \cdot 8 + (2^{k+1} + 2k) \cdot 2^{n-k} + (2^{2^{n-k}+1} + 2^{n-k+1}) + 2n] \quad (3)$$

Using Double LUT, we get:

$$L_{d-n.k} = [2^n \cdot 8 + (2^{k+1} + 2k) \cdot 2^{n-k} + \\ + (2^{2^{n-k}+1} + 2^{n-k+1}) + 2n] + 2^{n+3} + 4 \quad (4)$$

In according with the expression (1-5) the complexity of the m known LUT described by:

$$L_{n.k} = m[2^n \cdot 8 + (2^{k+1} + 2k) \cdot 2^{n-k} + (2^{2^{n-k}+1} + 2^{n-k+1}) + 2n] \quad (5)$$

For DC LUT we get:

$$L_{dc-n.k} = [2^n \cdot 8 + 2 \cdot \{(2^{k+1} + 2k) \cdot 2^{n-k} + (2^{2^{n-k}+1} + 2^{n-k+1})\} + 2n] + \\ + \text{ceil}(\frac{n}{4}) \cdot (2k + 4m) \quad (6)$$

Considering expressions (4) and (6), taking into account the discrete parameters we get:

$$L_{dnf} = k(22n + 2) + m(k + 2) + 6m(k + \frac{n}{4} \cdot (2k + 4m)) + 2n \quad (7)$$

Where $k(22n + 2)$ - is a complexity (in the number of the transistors) of the AND blocks; $m(k + 2)$ - is the complexity of the OR blocks; $6mk$ - is the complexity of the OR blocks setting; $2n$ - is the complexity of the inverters for variable inputs; $\frac{n}{4} \cdot (2k + 4m)$ - is the complexity of the blocks in accordance with the restriction [11].

The comparison of the number of transistors to implement system of the logic functions in the known LUT and all proposed LUT at $m = 2$ are presented Fig. 20.

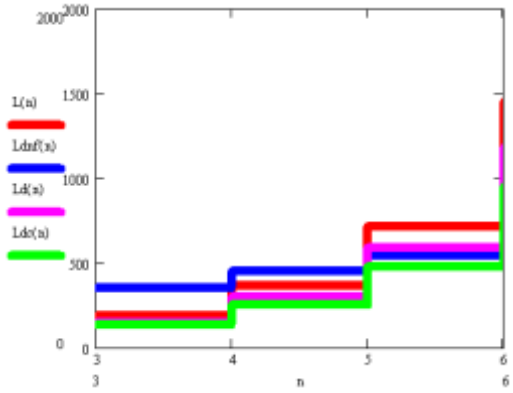


Figure 20. Comparison of the complexity of the m known LUT ($L(n)$) and proposed DLUT ($L_d(n)$), DCLUT ($L_{dc}(n)$), DNF-LUT ($L_{dnf}(n)$), $m = 2$

Thus a small number of functions inefficiently version DNF-LUT ($L_{dnf}(n)$). However, when calculating the 4 functions DNF-LUT ($L_{dnf}(n)$) ahead DCLUT ($L_{dc}(n)$) by 6 variables, Fig. 21.

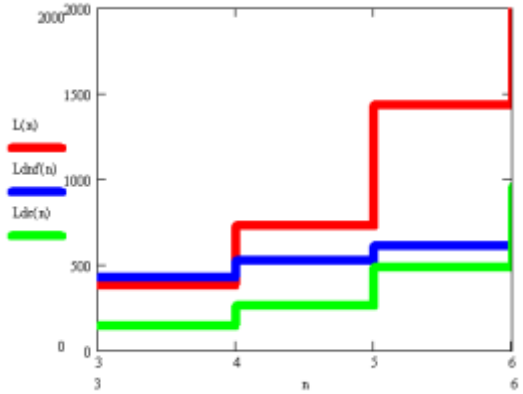


Figure 21. Comparison of the complexity of the m known LUT ($L(n)$) and proposed DLUT ($L_d(n)$), DCLUT ($L_{dc}(n)$), DNF-LUT ($L_{dnf}(n)$), $m = 4$

DNF LUT is a competition about the number of variables is greater than 5, Fig. 22.

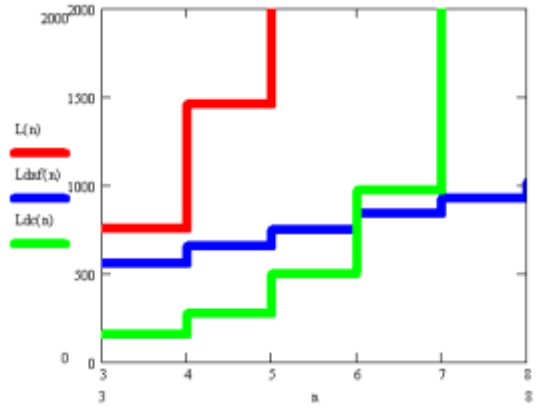


Figure 22. Comparison of the complexity of the *m* known LUT ($L(n)$) and proposed DLUT ($L_d(n)$), DCLUT ($L_{dc}(n)$), DNF-LUT ($L_{dnf}(n)$), $m=4; 3 \le n \le 8$

Comparison of the logical capacity (Number of transistors on a single function) of the *m* known LUT and proposed DCLUT, DNF-LUT presents Fig. 23, 24.

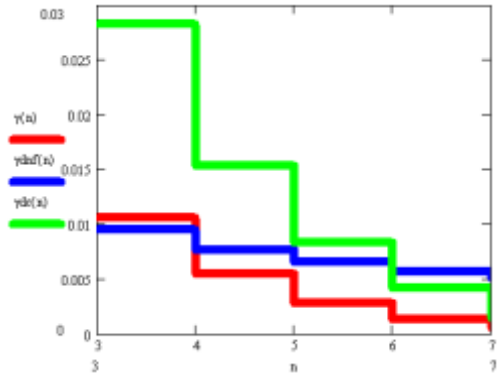


Figure 23. Comparison of the logical capacity of the *m* known LUT and proposed DCLUT, DNF-LUT, $m=4$

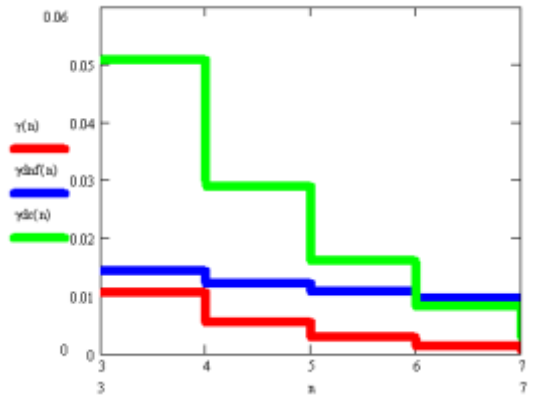


Figure 24. Comparison of the logical capacity of the *m* known LUT and proposed DCLUT, DNF-LUT, $m=8$

IV. Conclusion

We analyze the complexity of the proposed technical solutions and the results of functional simulation. Proposed advanced LUTs for the logic systems significantly reduces hardware costs (from 10% to 60% and more - Fig. 20-23) without essential reducing performance. The simulation of the advanced LUTs executed in the NI Multisim 10 by National Instruments Electronics Workbench Group and confirmed the efficiency of patentable technical solutions. If necessary, a small number of calculating functions appropriate to use DLUT. With an average number of logic functions, it is advisable to use the DCLUT. However, a large number of variables DNF-LUT has the best characteristics as the complexity of the other options is growing exponentially. In the future, it is advisable to consider the integrated use of a variety of solutions in a single FPGA and perform appropriate optimizations. One of the promising areas of research may be the creation of adaptive LUT-based proposed advanced LUT.

References

[1] Tyurin S.F., Green Logic: Models, Methods, Algorithms. Green IT Engineering: Concepts, Models, Complex Systems Architectures,vol. 74, pp.69-86, 2017. DOI: 10.1007/978-3-319-44162-7

[2] Tyurin S.F., Green Logic: Green LUT FPGA Concepts, Models and Evaluations. Green IT Engineering: Components, Networks and Systems Implementation,vol. 105, pp.241-261, 2017. DOI: 10.1007/978-3-319-55595-9

[3] Drozd, A. et al.: The levels of target resources development in computer systems. In: Design & Test Symposium (EWDTS), 2014 East-West, pp. 1-5, IEEE, (2014).

[4] Mehta, Nikil. An ultra-low-energy, variation-tolerant FPGA architecture using component-specific mapping. Dissertation (Ph.D.), California Institute of Technology. (2013) Available at: http://thesis.library.caltech.edu/7226/1/Nikil-Mehta-2013.pdf.

[5] J. Cong, K. Minkovich. Optimality Study of Logic Synthesis for LUT-Based FPGAs. Available at: http://cadlab.cs.ucla.edu/~kirill/tcad06.pdf.

[6] Robert J. Francis A Tutorial on Logic Synthesis for Lookup-Table Based FPGAs. Available at: https://pdfs.semanticscholar.org/e879/aa350c38d381c41eeef0da95b739d2ede6c1.pdf.

[7] S. Brown. FPGA Architectura I Research: A Survey. Available at: http://arantxa.ii.uam.es/~die/%5BLectura%20FPGA%20Architecture%5D%20FPGA%20architectural%20research.%20A%20survey.pdf

[8] S. Brown, J. Rose. Architecture of FPGAs and CPLDs: A Tutorial. Available at: http://malun1.mala.bc.ca:8080/~pwalsh/teaching/355/Toronto_tutorial.pdf

[9] Tyurin S.F., Vikhorev R.V. Programmable logic device: patent RF №2573732; Published 27.01.2016, Bull. №3.

[10] Tyurin S.F. Programmable logic device: patent RF №2544750; Published 20.03.2015, Bull. №8.

[11] Conway, Lynn. Drafts of the Mead-Conway textbook, Introduction to VLSI Systems. University of Michigan. Retrieved 9 June 2015.

Substantiation of the use of viscoelastic material model in numerical analysis the creep of concrete structures

prof. dr hab. ing.Vladimir N. Sidorov
Faculty of Civil Engineering and Architecture
Kielce University of Technology
Kielce, Poland
Department of Architecture and Civil Engineering
RUDN University
Moscow, Russia
Construction Mechanics and Computer Department
Perm National Research Polytechnic University
Perm, Russia
sidorov.vladimir@gmail.com

mr. Katarzyna Nowak
Faculty of Civil Engineering and Architecture
Kielce University of Technology
Kielce, Poland
e-mail: knowak@tu.kielce.pl

Abstract— **The paper presents the results of numerical modeling of concrete rheological manifestations under compression, the properties of which are described by a linear model of a viscoelastic material. The computer model of concrete with viscoelastic properties selected for studies identified by experimental research results of concrete samples. Computer simulation was carried out in a software environment SIMULIA Abaqus. Evaluation of the results of numerical analysis of the concrete elements under the action of constant stress was conducted from their comparison with both the experimental and the normative values of creep strain. The conclusions reached by the results obtained are intended for researchers and designers, performing numerical analysis of concrete and reinforced concrete structures, taking into account creep and relaxation of the material.**

Keywords—viscoelastic model, the Maxwell model, creep strain, concrete sample

I. INTRODUCTION

Concrete is the typical building material whose physical properties change over time. The manifestations of these properties, known as rheological processes cause, including the creep strain or stress relaxation in concrete and reinforced concrete structures. Creep is the progressive increasing in deformation over time in a long period of a construction exploitation, possibly while the load is not changing. Relaxation on the contrary manifest in a change of stress state of the structure, it is possible, when strains remaining unchanged. From the viewpoint of structural mechanics manifestation of the rheological properties of the materials are of particular importance. They can not only lead to a change in the stress-strain state, but also to the fundamental change in the constructive schemes of a structure work, which in turn, may even lead to its destruction. In this connection an important problem in the structures analysis being performed is to determine the regularities of changes in time of both the stress and strain values, and of relations between them. In principle, in numerical studies of structures under conditions of creep and relaxation, the basic rheological properties of the material can be described by a combination of simple mechanical models. In engineering analysis using such kind of models, ones usually come from a simplified representation of the nature of physical phenomena [1-4]. The purpose of this paper is to select, calibrate and evaluate the reliability of the computer rheological model of the material for its reasonable application in the simulation of concrete structures, taking creep into account. In this work, the study of the rheological properties of concrete described by an isotropic linear model of a viscoelastic material is presented. Under conditions of relatively small deformations, the constitutive patterns of behavior of such material, based on the Boltzmann superposition principle, can be represented in the form of an integral equation

$$\sigma(t)=\int_0^t G(t-\tau)\frac{\partial\varepsilon(\tau)}{\partial\tau}d\tau \qquad (1)$$

where G(t) is a time-dependent relaxation function.

The phenomenon of stress relaxation occurring in a viscoelastic material, as a rule, is represented by a mechanical model consisting of two elementary mechanical models: a spring modeled by Hooke's law and a Newton damper. A combination of these elements can create a more complex model, called in this case the Maxwell model (Fig. 1). This model allows to reflect the main features of the rheological manifestation of a viscoelastic material. Its detailed description is given in [1-3, 5-11]. In numerical rheological models, the stress relaxation function of the generalized Maxwell model is represented by exponential functions in the form of the known Prony series [10, 12]

$$G(t)=\frac{\sigma(t)}{\varepsilon_0}=G_\infty+\sum_{i=1}^{n}G_i e^{\left(-\frac{t}{\tau_i}\right)} \qquad (2)$$

where G_∞ is the value of the long-term shear relaxation modulus, determined depending on the instantaneous modulus of elasticity of the material, the initial modulus of shear elasticity, $G_0 = G_\infty + \sum_{i=1}^{n} G_i$, t is time, G_i and τ_i are parameters of the Prony series (i = 1,2,… n) for the material, in this case τ_i is the relaxation time calculated from formula (3), n is the number of elements making up the generalized Maxwell model (Fig. 1).

$$\tau_i = \frac{\eta_i}{G_i} \tag{3}$$

where η_i is the coefficient of viscosity of the material.

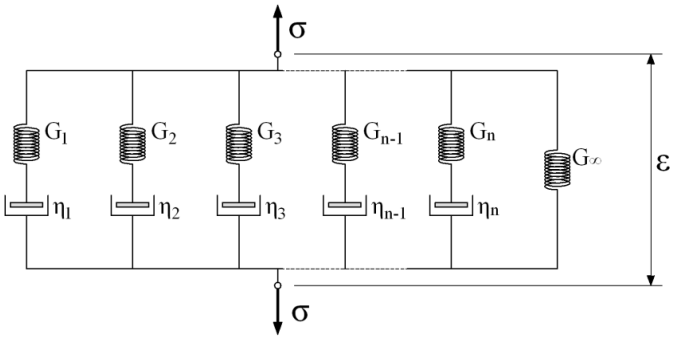


Figure 1. The scheme of the generalized Maxwell model with elements connected in series and parallel: springs with elastic moduli G_i and dampers with viscosity coefficients η_i, and also with an elastic element having a stiffness G_∞.

In this paper, we give examples of numerical simulation of the creep process for viscoelastic concrete samples using the SIMULIA Abaqus program. The parameters of the computer rheological model of the material were set following laboratory tests of concrete samples. To evaluate the operability of the selected model a comparison of the creep strain curves obtained from the results of numerical calculations with the results obtained experimentally [13] is made. The conclusions drawn from the results of these comparisons are recommendations for researchers and designers performing numerical analysis of concrete and reinforced concrete structures taking into account the creep of the material.

II. Numerical Model with Viscoelastic Properties

For performing numerical calculations SIMULIA Abaqus software environment was chosen. To describe the creep of the material a linear model of viscoelastic material consisting of a combination of elastic springs and viscous dampers was adopted. When analyzing a complex stress state, it is possible a separate use of regularities of the change in shape and volume of material. In the SIMULIA Abaqus program, the relaxation function is represented as a dimensionless shear relaxation module [4, 10, 12]

$$gr(t) = \frac{G(t)}{G_0} = 1 - \sum_{i=1}^{n} G_i \left(1 - e^{\left(-\frac{t}{\tau_i}\right)}\right) \tag{4}$$

where G_0 is the instantaneous modulus of shear elasticity, which depends on the instantaneous elasticity modulus E_0 and Poisson's ratio ν_0: $G_0 = E_0/2(1+\nu_0)$.

Similarly, the bulk relaxation function is represented, which can also be present in the form of a dimensionless bulk relaxation module:

$$kr(t) = \frac{K(t)}{K_0} = 1 - \sum_{i=1}^{n} K_i \left(1 - e^{\left(-\frac{t}{\tau_i}\right)}\right) \tag{5}$$

where K_0 is a coefficient of bulk stiffness, depending on the instantaneous modulus of elasticity E_0 and Poisson's ratio ν_0: $K_0 = E_0/3(1-2\nu_0)$, K_i is a parameter of the Prony series, set for the material.

The parameters participated in the expression for the Prony series (4) and (5) needed to describe the dimensionless relaxation function defined in the time domain, can be identified by directly determining the material parameters (Gi, Ki, τi) obtained from the experimental creep or relaxation tests of the material. When using computational computer programs based on the finite element method, it is possible to optimally determine the viscoelastic properties of the material from the results of measurements accumulated during the experimental studies of creep or relaxation of the material.

III. Identification of the Parameters of the Numerical Model of Viscoelastic Material Implemented in the Program SIMULIA Abaqus

The numerical model of concrete with viscoelastic properties chosen for investigation is identified by the results of experimental studies of concrete's samples presented in [13]. The credibility of the results is estimated using the examples of actual measurements of creep strains measured for cylindrical concrete samples having a height of 300 mm and a diameter of 150mm. Their calculated design model, including the conditions for fixing, is shown in Fig. 2. The results of measurements of the active creep strains of concrete's samples are presented in graphs in [13], where together with this the contents of the laboratory test program, their description and implementation were presented in detail.

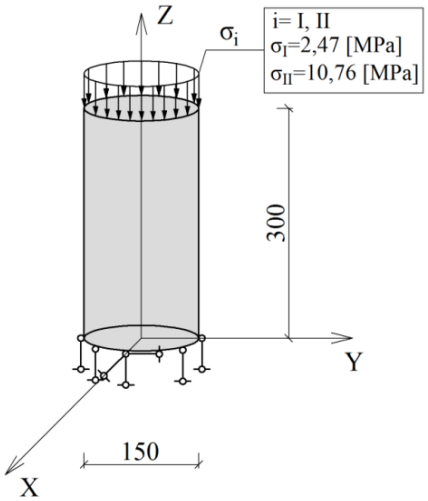


Figure 2. The calculated design model of the concrete sample

The composition of the concrete from which the samples were prepared for the experiment is shown in Table 1. Samples were placed in a hydraulic press when they reached a certain age, and at differentiated values of the external force, they were subjected to loading, uniformly distributed over the upper face of the cylindrical sample. Then creep strains were measured. Also, the results of the experiment are compared with the results obtained from calculations of creep strains made under the Eurocode 2 standard.

TABLE I. COMPOSITION OF THE CONCRETE USED IN THE LABORATORY TESTS

MIX PROPORTION	AMOUNT OF COMPONENT [kg/m^3]
Cement CEM III/A 32,5 N-LH/HSR/NA	350
Water	175
Agreggate 0/2	561
Agreggate 2/8	1309

In the numerical studies described in this paper the results of testing two concrete samples, that are compressed with a force of up to 40 [%] of the value of the destructive load, are used. In the calculations, the results of the tests carried out on the following samples were used:

– I sample: the load 2.47 MPa was applied 1 day after mixing the components, under these conditions the sample was 13 days,

– II sample: the load 10,76 MPa was applied 28 days after mixing the components, under these conditions the sample was 11 days.

In both cases, the samples were in wet conditions of 95% and at a temperature of 20-22° C.

To fully define the parameters of a material model with viscoelastic properties in the ABAQUS Simulia program, the dimensionless relaxation moduli (4) and (5) in the general form of the expression for the Prony series are assigned. Parameters (4), (5) in the said program are specified using the RELAXATION TEST DATA option, and in this case, alternatively, can be interconnected by the COMBINED TEST DATA command. If to assume that the stresses in the concrete remain constant over time (σ (t) = const), it is convenient to use the so-called effective modulus of elasticity [14, 15]:

$$E_{c,eff}=\frac{E_0}{1+\varphi(t,t_0)} \quad (6)$$

where $\varphi(t,t_0)$ is the coefficient of creep, established between the time of application of the load t0 and the age of concrete t being considered and assigned following the rules of Eurokod 2 [16]. Due to the creep coefficient participating in expression (6), the effective relaxation modulus changes its value in time. Thus, it is assumed that the stress relaxation function G(t), determined for the indicated a time period for the calculation of the dimensionless shear relaxation (4), is represented using the effective modulus of elasticity by the formula:

$$gr(t)=\frac{G(t)}{G_0}=\frac{\frac{E_0}{2(1+\nu_0)\left(1+\varphi(t,t_0)\right)}}{\frac{E_0}{2(1+\nu_0)}}=\frac{1}{1+\varphi(t,t_0)} \quad (7)$$

Similarly, a bulk relaxation module can be presented

$$kr(t)=\frac{K(t)}{K_0}=\frac{1}{1+\varphi(t,t_0)} \quad (8)$$

For the correct determination of the value of the creep coefficient $\varphi(t, t_0)$ involved in expressions (7) and (8), the procedure laid down in the rules [16] was used in the work, mainly in its Appendix B. In addition, the age of concrete has been adjusted in calculations, taking into account the type of cement used and the history of temperature changes. The algorithm for determining the coefficient of creep given in the norms was implemented in the Mathcad software environment. When specifying the values of the creep coefficient of concrete, together with the characteristics of the samples described above, the cross-sectional area of the sample Ac = 17671.46 mm^2, the circumference of the sample contacting the external medium u = 471.24 mm was additionally assigned. And also: the average strength of concrete at the age of 28 days fcm = 28,4 MPa, parameter α = -1 for concrete of class C. Tables 2 and 3 show the values of the creep coefficient obtained for the first and second samples.

TABLE II. THE CREEP COEFFICIENT AND THE RELAXATION MODULIES OBTAINED FOR THE I SAMPLE

I SAMPLE		
AGE OF CONCRETE	CREEP COEFFICIENT (corrected in view of respect to temperature and type of cement)	MODULE kr(t) and gr(t)
t [days]	$\varphi(t,t_0)$	$\frac{1}{1+\varphi(t,t_0)}$
1	0	1
2	0,5365	0,650829808
3	0,6143	0,619463545
4	0,6741	0,597335882
5	0,7235	0,580214679
6	0,766	0,566251416
7	0,8036	0,554446662
8	0,8374	0,544247306
9	0,8683	0,535245946
10	0,8968	0,527203712
11	0,9233	0,519939687
12	0,9481	0,513320671
13	0,9714	0,507253728

TABLE III. THE CREEP COEFFICIENT AND THE RELAXATION MODULES OBTAINED FOR THE II SAMPLE

II SAMPLE		
AGE OF CONCRETE	CREEP COEFFICIENT (corrected in view of respect to temperature and type of cement)	MODULE kr(t) and gr(t)
t [days]	$\varphi(t,t_0)$	$\frac{1}{1+\varphi(t,t_0)}$
28	0	1
29	0,3147	0,760629801
30	0,3326	0,750412727
31	0,3486	0,741509714
32	0,3629	0,733729547
33	0,3761	0,726691374
34	0,3882	0,720357297
35	0,3995	0,714540907
36	0,4101	0,709169562
37	0,4201	0,704175762
38	0,4295	0,699545296

From these data, the dimensionless relaxation moduli (7) and (8) were determined for each time interval; these were then set during calculations in the SIMULIA Abaqus software environment. Next, other characteristics of the material were set. Namely, the Poisson's ratio ν = 0.2, and Young's modulus with E = 36200 MPa, obtained from the corresponding experimental tests of concrete at the age of 28 days. Besides, for the I-st sample, the computer simulation was performed on the assumption of a reduced value of the Young's modulus E = 22201.18 MPa, determined for one-day concrete following paragraph 3.1.3. of the Eurocode 2 standard.

IV. THE RESULTS OF THE NUMERICAL ANALYSIS OF CONCRETE SAMPLES CREEP

The section presents the results of numerical modeling of rheological manifestations of deformed concrete, whose properties are described by a linear model of a viscoelastic material. The purpose of the analysis was to justify the acceptability of using such a model by comparing the values of creep strains obtained in the course of computer simulation using the SIMULIA Abaqus program with measurements made during experimental studies of concrete samples [13]. For numerical analysis, the calculated models of two concrete samples (sample I and sample II) were chosen, subjected to the action of the load, different in magnitude and at various times of its application. In work, the parameters of the created model of a material with viscoelastic properties, identified by the results of samples testing were selected. Also, a comparative analysis of the results obtained is supplemented by calculating the normative values of the creep strains $\varepsilon_{cc}(t,t_0)$, determined following the dependence:

$$\varepsilon_{cc}(t,t_0)=\frac{\sigma_c(t_0)}{E_{ci}}\varphi(t,t_0), \tag{9}$$

where: t is the concrete age under consideration, $\varphi(t,t_0)$ is the coefficient of creep, E_{ci} is the modulus of elasticity of concrete at the age of 28 days.

Figure 3 shows the values of creep strains obtained when testing a numerical model of the I-st concrete sample, in comparison with the results of its laboratory tests and with normative values. From a comparison of graphs representing the dependence of the creep strain on time, it can be seen that the creep strain values obtained numerically are much less than those obtained by measurements in experimental studies. In addition, it should be noted that Young's modulus given in computer simulation has a noticeable effect on the numerical results obtained. For the first sample, when the value of the modulus of elasticity E = 22201.18 [MPa], determined in accordance with Eurokod 2, was established, the creep strain values calculated in this way, in accordance with the norms, are less than the corresponding values obtained from the experiments within the limits of 18-49 [%].Figure 3 also shows that the presented numerical simulation results are higher than the calculated normative values. For the same sample, but with a modified Young's modulus revealed in experimental studies [13] for concrete loaded at the age of 28 days, the difference in creep strains measured in the experiment is much larger than the values obtained numerically, namely the difference is within 50-69 [%]. In this case, creep strains obtained in numerical studies are less than normative. The obvious differences in the values of creep strain between the results of laboratory tests and the normative values (Figure 3) follow from the assumptions of the Eurocode 2 standard when predicting the behavior of concrete when exposed to a load at the age of 28 days. Thus, in the design of structures under the above regulatory requirements, taking into account the change in material properties under creep conditions, the parameters of the rheological changes of concrete loaded in the first days of hardening are taken to be much smaller than those obtained from experimental studies [13].

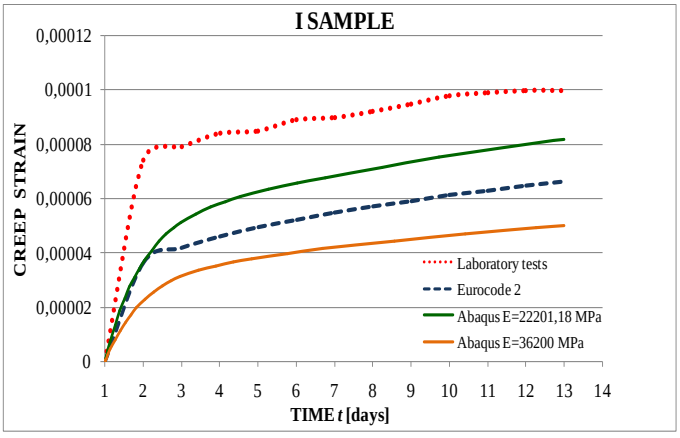


Figure 3. Comparison of the creep strains values obtained for sample I

Figure 4 shows the change in the creep strain in time obtained for the second sample. Comparison of the graphs shows that the results of numerical modeling of creep strains give larger values than the corresponding values obtained as a consequence of the experiment within 1.5-73 [%]. However, the opposite situation occurs in the case of normative strain values - the numerical values of creep strains are less than the normative values by 1-24 [%]. Moreover, based on the graphs shown in Figure 4, it can be concluded that the deformations calculated under regulatory requirements and relating to concrete at the age of 28 days, in comparison with the experimental measurements, correspond to the expected [13].

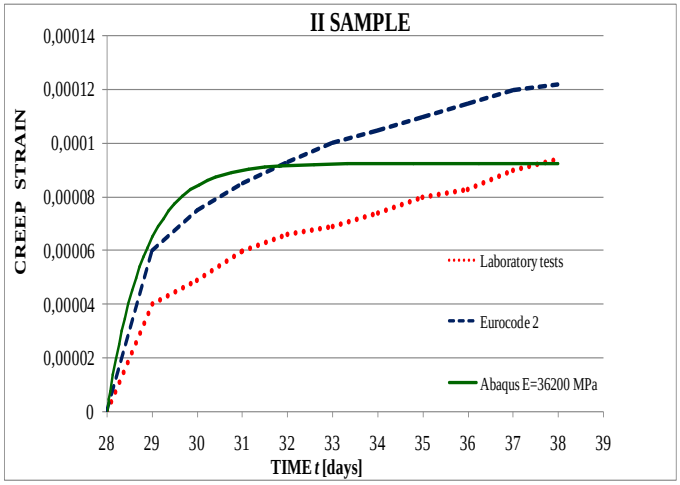


Figure 4. Comparison of the creep strains values obtained for sample II

V. SUMMARY

Thus, the paper presents and analyzes the results of computer simulation of creep strains of concrete samples. In numerical analysis, a linear model of a material with viscoelastic properties was used. Experimental studies determined its parameters. The results of computer simulation were compared with the results of the experiment and with the results obtained under regulatory requirements.

Based on the results of the computer simulation performed for selected concrete samples described by the viscoelastic material model, the following conclusions should be drawn:

- the characteristics of the rheological changes in the properties of concrete depend mainly on the selected initial parameters of the materials, as well as on the magnitude and time of application of the load,

- the key parameter characterizing creep deformation is the creep coefficient, which is established under the regulatory requirements of the Eurokod 2 standard,

- numerical studies of creep of concrete presented in the work performed on the example of two concrete samples, using a linear model of viscoelastic material, showed good agreement concerning concrete loaded at a later age,

- for concrete loaded at the age of 28 days, the values of the numerical analysis of creep strains are in complete agreement with the results of laboratory tests that were presented in [13],

- for concrete loaded in the early stages of hardening (an example is considered in the work when the sample is loaded on the 1st day of hardening), a much lower level of creep strains is obtained from numerical calculations, compared with the deformations measured in the experiment.

Numerical analysis carried out in the program environment of SIMULIA Abaqus on the basis of a linear model of viscoelastic material showed that it can be used to successfully describe the behavior of concrete in time, but taking into account that the coincidence with the results of real tests is achieved for concretes loaded From the 28th day of their hardening. Thus, in the course of computer modeling, it is crucial to carefully analyze the numerical models of the phenomena studied, proposed in modern software, which will make it possible to avoid many errors allowed in the numerical analysis of structures.

REFERENCES

[1] A.Mitzel, “Reologia betonu”, Arkady, Warszawa, 1972.

[2] M.Lachowicz, “Modelowanie ośrodka lepkosprężystego w metodzie elementów czasoprzestrzennych”, Rozprawa doktorska, Uniwersytet Technologiczno – Przyrodniczy w Bydgoszczy, Bydgoszcz, 2015.

[3] A.Bodnar, M.Chrzanowski and P.Latus, “Reologia konstrukcji prętowych”, Politechnika Krakowska, Kraków, 2006.

[4] A.Zbiciak, K.Brzeziński and R.Michalczyk, “Analiza wpływu obciążeń dynamicznych na zachowanie się lepko-sprężystego modelu nawierzchni drogowej”, Logistyka, 3/2014.

[5] Sorvari J., Malinen M., “Numerical interconversion between linear viscoelastic material functions with regularization”, International Journal of Solids and Structures 44, 2007,pp. 1291–1303.

[6] MarquesS.P.C., Guillermo J.C., “Computational Viscoelasticity”, Springer Science & Business Media, 2012.

[7] Lakes R.S,“Viscoelastic Materials”, Cambridge University Press, New York, 2009.

[8] Tscharnuter D., Jerabek M., Major Z., Lang R.W., “On the determination of the relaxation modulus of PP compounds from arbitrary strain histories, Mechanics of Time-Dependent Materials, 15/2011, pp.1-14.

[9] J.Skrzypek, “Plastyczność i pełzanie: teoria, zastosowania, zadania”, PWN, Warszawa, 1986.

[10] R.Michalczyk, “Implementation of generalized viscoelastic material model in Abaqus code”, Logistyka, 6/2011.

[11] H.Schiessel, R.Metzler, A.Blumen and Nonnenmacher T.F., “Generalized viscoelastic models: their fractional equations with solutions”, Journal of Physics A: Mathematical and General United Kingdom, 28/1995, pp.6567-6584.

[12] ABAQUS – Standard, User’s Manual, Theory Manual and Keywords Manual, Hibbitt, Karlsson and Sorensen, Inc., Version 6.12.

[13] J.Błyszko, “Porównawcza analiza pełzania twardniejącego betonu zwykłego oraz modyfikowanego zbrojeniem rozproszonym”, Rozprawa doktorska, Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, Szczecin, 2015.

[14] Sekcja Konstrukcji Betonowych KILiW PAN (2006) Podstawy projektowania konstrukcji żelbetowych i sprężonych według Eurokodu 2. Dolnośląskie Wydawnictwo Edukacyjne, Wrocław.

[15] Ł.Bednarski, R.Sieńko, T.Howiacki, “Oszacowanie wartości i zmienności modułu sprężystości betonu w istniejącej konstrukcji na podstawie ciągłych pomiarów in situ”, Cement Wapno Beton, 19/81, nr 6, 2014, pp.396-404.

[16] Eurokod 2 (2008) Projektowanie konstrukcji z betonu, Część 1: Reguły ogólne i reguły dla budynków, 1992-1-1:2008, PKN, Warszawa.

The Sensitivity of the Statistical Characteristics to the Selected Parameters of the Simulation Model in the Red Blood Cell Flow Simulations

Martin Slavík[1,2,3] Katarína Bachratá[1,2], Hynek Bachratý[1,2], Kristína Kovalčíková,[1,3]
Cell-in-fluid research group, Department of Software Technologies, Faculty of Management Science and Informatics
University of Žilina
Žilina, Slovakia
Email:[Martin.Slavik, Katarina.Bachrata]@fri.uniza.sk

***Abstract*—The main aim of our work is future optimization of microfluidic devices used for capture of rare cells, for instance circulating tumour cells. Since the manufacture of such devices and performing biological experiments are complicated and time consuming, in silico experiments are preferable for this purpose. But first a simulation model has to be properly verified. In this article we propose statistical characteristics for comparison of simulation and biological experiments of red blood cells moving in a fluid. At the same time these characteristics allow better understanding of the course of the simulation and the behaviour of its individual components. In the first part we explore the sensitivity of some of the characteristics formulated on the grounds of the previous research to various topologies of the microfluidic channel with one solid hematocrit.The second part of the article analyzes the entire blood cell group; it focuses on their position (orientation) in the channel. To analyze the orientation, we use the principal component analysis (PCA), by means of which we can determine the statistical distribution of the directions of a blood cell set with a certain channel geometry. This characteristics describes the qualities of blood cells during a simulation, which are interesting when we are analyzing, as a targeted value, the possibility of the largest number possible of blood cells getting to the wall of the channel. We have tested the characteristics for various simulation parameters, and we have obtained interesting results.**

***Keywords*: numerical simulation, red blood cell, statistical characteristics, principal component analysis**

I. Introduction

Recently many research groups have been dedicated to developing numerical models for simulation blood flow and investigation its rheological properties [9], [14], [13]. Every model is based on different method and has its own pros and cons. In this work, we use our previously developed Object-in-fluid framework [8] intended for simulation of flow of elastic objects at the microscopic level. This approach can be used also for a modelling of the blood. The cells are modelled as elastic objects represented by the triangulation of their surface. Plasma or other solution are modelled by the Lattice-Boltzmann method. The immersed boundary method enables the interactions between the two parts. The entire model is described in detail in [6], [7] and it has been implemented in the open source simulation software ESPResSo .

One of the areas to use such simulation model in is designing and optimizing microfluidic devices for the detection, sorting or capturing specific cells, e. g. circulating tumor cells, from blood samples [3], [11]. Nevertheless, producing such devices is very much time and cost demanding [10]. Therefore, to design such devices, it seems appropriate to use computer simulations of the blood flow to be run through the devices. But first of all it is necessary to make sure the computation model is accurate and consistent with reality. The best way to achieve that is to compare the results obtained from the simulation to the results of biological experiments, or to the expected physical behaviour. So far, afew studies have been done on testing and calibration of certain properties and parts of the model, especially red blood cells (RBC). For example, in [12], the elastic behaviour of single RBC has been calibrated and validated. In [4] proper interaction between object and fluid was investigated. In further research, we would like to look at the simulation as awhole, and monitor and describe the behaviour of tens to hundreds of cells contained in the microfluidic device. In a near future, we would like to gain such results from biological experiments and compare them with results from our model. Similar approach was applied in [14]. Unfortunately, due to computational demands we can not yet perform large scale simulation with thousands of RBCs, so we can not repeat this experiment. Moreover, we would like to capture and compare also a dynamic behaviour of RBCs during the experiment. State of the art in this research area is not very advanced, that is why we consider this topic a new one.

From experimental results it is necessary to identify a information on such simulation attributes which are meaningful from the viewpoint of the purpose of the analysis. For instance, the objective function will assign little significance to a blood cell colour, however, it will consider as more significant the

[1] The work of this author was supported by the Slovak Research and Development Agency under the contract No.APVV-15-0751.

[2]The work of this author was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the contract No. VEGA 1/0643/17.

[3]The work of this author was supported by the FVG 2017 grant of Faculty of Management Science and Informatics, University of Žilina.

blood cell velocity or its interacting with other blood cells or the wall of the microfluidic device. The individual characteristics and, consequently, their combinations compounded into an objective function, should serve to compare the simulation experiments to biological ones, as well as for the mutual comparison of the simulation experiments themselves.In [2] we did the first attempt to find such characteristics. On the basis of the data acquired from simulation experiments, we proposed several statistical characteristics describing RBC behaviour in the flow when they are passing through the periodic channel with pillar obstacles. The characteristics have described the behaviour of RBC from the viewpoint of their velocities, their periodic behaviour, rotation and position in the channel, or their skew.

In the first part of this work, we would like to test the velocity characteristics for various obstacle positioning, which causes flow changes inside the channel. Then, in the second part, we want to suggest the characteristics based on more advanced statistical methods. The skew of cells during the simulation can be crucial for the possibility of cancer cells being captured on the pillars in the channel. The skew is measured by three dimensions of the cell. By application of PCA, we can decrease the data dimension and obtain a characteristic which can not only serve for comparison between more simulations or between simulation and biological experiments. At the same time it can also allow better understanding of the course of the simulation and the behaviour of its individual components.

II. Description of the experiments and data

All simulation experiments described in this work were carried out in acuboid-shapedchannel with the length of 100 μm (x-axis), the width of 50 μm (y-axis), and the hight of 30 μm (z-axis). The channel was closed from four sides, i.e. it had its bottom, top and side walls. It was open in the x-direction to let the fluid flow. This was also the direction of the periodic conditions applied which made the experiment less demanding on calculations also for models of larger channels. Cylindrical obstacles (pillars) with the radius of 5 μm were placed variously inside the channel. Again, we used five different topologies, A, B, C, D and E. The dimensions of the topologies A and B were identical (Fig. 1 and Fig. 2) but in the topology B one central obstacle was missing. It was the same for the topologies C and D (Fig. 3 and Fig. 4). There were no obstacles in the topology E (Fig. 5).

With each topology, we carried out the simulation of the flow of 50 RBCs inserted into the moving fluid. At the beginning of each simulation, RBCs had various random initial seedings. Each of the experiments ran for 380,000 simulation steps, which, with atime step of 0.2 μs, corresponds to the real time of the experiment of 0.0762 s. In the time the fastest cells passed the simulated channel section many times, which means they did hundreds of micrometers in total.

A. Experimental setup

All experiments were performed in the simulation software ESPResSo with usage of Object-in-fluid framework for mod-

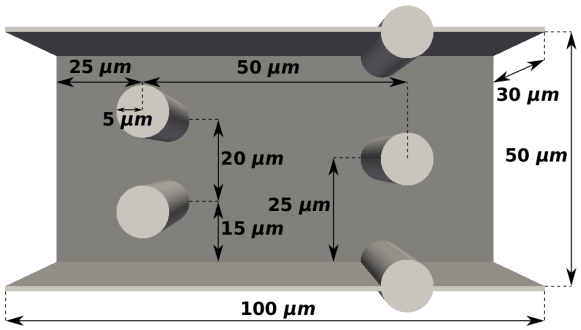


Fig. 1: Empty channel A and its sizes.

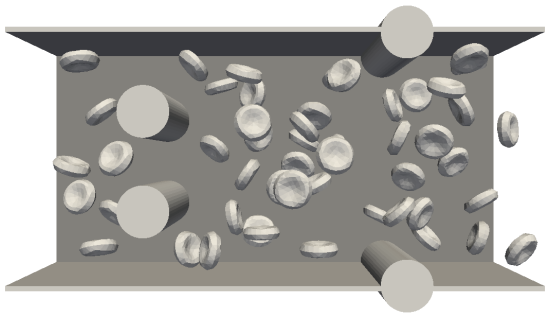


Fig. 2: Channel B with seeded RBCs.

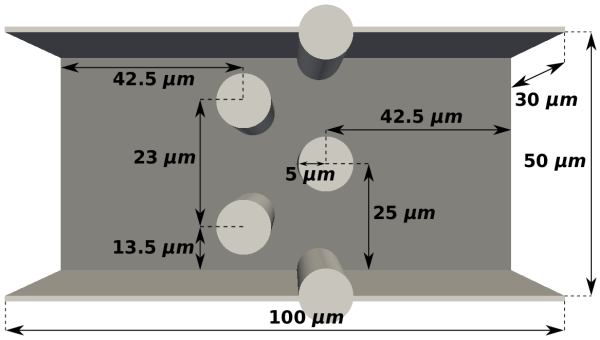


Fig. 3: Empty channel C and its sizes.

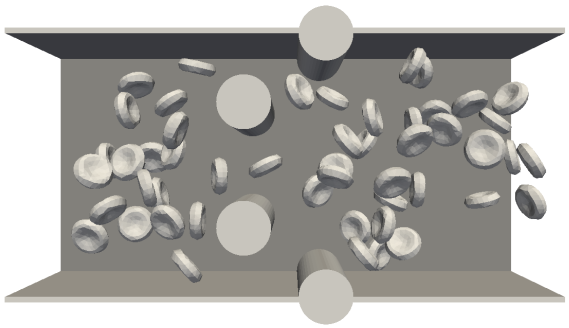


Fig. 4: Channel D with seeded RBCs.

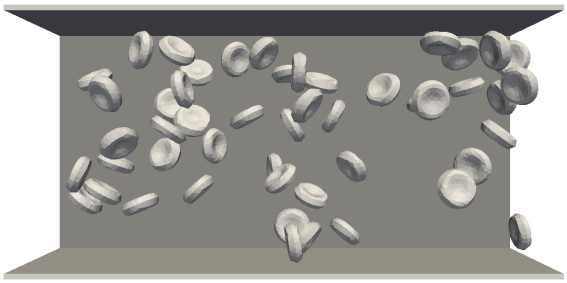


Fig. 5: Channel E with seeded RBCs.

elling cells and LB for modelling fluid. As a model of each of the RBCs we had chosen a surface mesh with 141 nodes and the dimensions of 7.82μm $\times$ 7.82μm $\times$ 2.56μm. The elastic coefficients of the mesh were ks = 0.0044, kb = 0.0715, kal = 0.005, kag = 1, kv =1.25. The simulation mass of every cell was 0.059574468085106386 pg. The interactions among the cells were modelled by membrane collision interaction with the parameters of mc_K = 0.005, mc_n = 2.0, mc_cut = 0.5. The interactions between the cells and surface of the channel were ensured by soft sphere interaction with the parameters of $soft\_K$ = 0.00035, $soft\_n$ = 1.0, $soft\_cut$ = 0.5. The fluid was discretized by a three-dimensional regular lattice with the space step 1 μm. The fluid viscosity was 1.5 mPa.s and the density was 1000 kg/m^3, representing blood plasma at 20 $°C$. The interaction parameter between fluid and object called friction was 0:025865531914893616. Different external volume forces were used to induce the flow in the x-direction.

B. Data description

In this work, we were only interested in two types of data. Firstly, we continuously (every 200 time steps) measured the velocity of the center of each RBC, which is in fact the average velocity of every surface node of the RBC model. Thus, we obtained the vector of absolute velocities for each 50 simulated cells for one simulation experiment. The second type of data were the minimum and maximum values of the x, y, z - coordinates of the cell surface for each RBC. The information enabled us to calculate the dimensions of the cuboid containing the entire RBC and oriented in the direction of the main axes (Fig. 6).

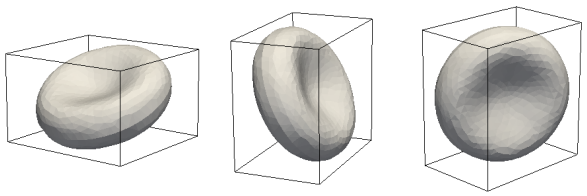


Fig. 6: Cuboid around CTC.

III. Cell velocity analysis

At first, rather simple characteristics of each simulation experiment are determined from the RBCs velocity vector. For this purpose, we took the data from ten simulation experiments, two experiments for every type of topology with the same initial external force applied to the fluid (0,0005). Let us call these experiments (A50a, A50b, B50a, B50b, C50a, C50b, D50a, D50b, E50a and E50b). From each velocity record, we calculated its minimum, maximum, the difference between the minimum and maximum, and the average. Despite the simplification, these characteristics somehow describe the topology of the channel. The minimum value for each RBC captures the deceleration during a collision with an obstacle or when the cell is moving at the widest point of the channel. The maximum captures the high velocity of when the cells pass the narrow parts of the channel. Through the differences of the previous characteristics, the variability in cell movement and the overall ratio of 'slow' and 'fast' parts of its track are assessed. The average can be explained in a similar manner.

For every experiment, each of the calculated characteristic was stored in a separate vector, and its components were sorted by size. The vectors of minimums for all 10 experiments are in Fig 7.

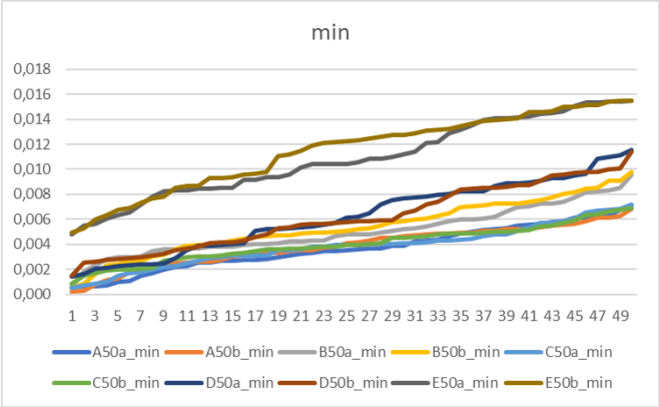


Fig. 7: Plot of sorted minimal velocities of all cells for every simulation.

Even from the plot, we can see some similarities among the minimums from the experiments performed in the channels with the same topologies. And vice-versa, the minimums differ with the experiments with different topologies. To confirm the assumption, we applied the standard Kolmogorov-Smirnov test. The results are shown in Fig 8. In the top right half, "0" means that we did not reject the hypothesis H_0 that the measured data came from the same distribution, using the significance level α = 0.05 against the hypothesis H_1. If we rejected H_0 , we denoted it as "1". In the bottom left half, there are p values.

Kolmogorov-Smirnov tests for other three velocity characteristics resulted similarly. When testing the characteristic for the experiments from the channels with the same topology, H_0 was never rejected. With different topologies, H_0 was rejected

min	A50a	A50b	B50a	B50b	C50a	C50b	D50a	D50b	E50a	E50b
A50a	-	0	1	1	0	0	1	1	1	1
A50b	0.8409	-	1	1	0	0	1	1	1	1
B50a	0.0021	0.0317	-	0	1	0	1	1	1	1
B50b	0.0010	0.0044	0.5077	-	1	1	0	0	1	1
C50a	0.6779	0.3584	0.0317	0.0004	-	0	1	1	1	1
C50b	0.3584	0.9541	0.0951	0.0089	0.5077	-	1	1	1	1
D50a	0.0000	0.0000	0.0089	0.0560	0.0000	0.0000	-	0	1	1
D50b	0.0002	0.0000	0.0317	0.1546	0.0001	0.0000	0.5077	-	1	1
E50a	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-	0
E50b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1546	-

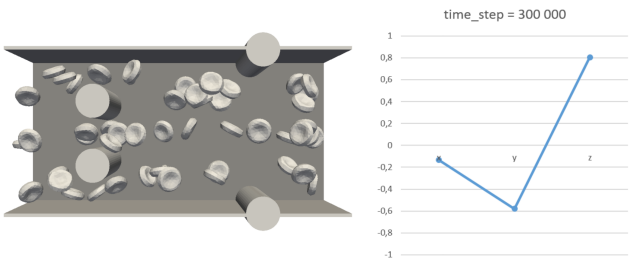
Fig. 8: Table of results of Kolomogorov-Smirnov test for minimal velocities.

in overwhelming majority of cases.

IV. Analysis of the Cell Skew

Naturally, the topologies tested influenced the velocity of the individual cells, which the derived characteristics showed. Velocity is acharacteristics that would be better with flow simulations of spheric objects but with the RBC shape, there is adistortion.

Our previous work [2] dealt with cell rotation, and proposed the characteristics describing such dynamic behaviour. The characteristics, though, would be problematic to obtain from real (biological) experiments which provide only the images in the individual time steps, and processing the video records is technologically demanding. That is why we decided to monitor the data on each cell described in section II-B, representing the cell skew during the simulation. From each simulation experiment carried out, we used aset of data acquired in nine discrete time steps (300000, 310000, 320000, 330000, 340000, 350000, 360000, 370000, 380 000), while each set contained the dimensions of the cuboid (Fig. 6) for each of 50 cells. Then the dimensions of the cuboids were scaled, and PCA was applied to the individual sets. The method is useful, since it preserves alarge amount of information, and, at the same time, it reduces the volume of the original data. In our case we acquired three numbers for each set of the scaled data, representing some kind of the main skew of all cells in agiven time step. The following pictures Fig. 9 shows the graph of the PCA results for the simulation B50a, together with the images from the simulation in the relevant time steps.



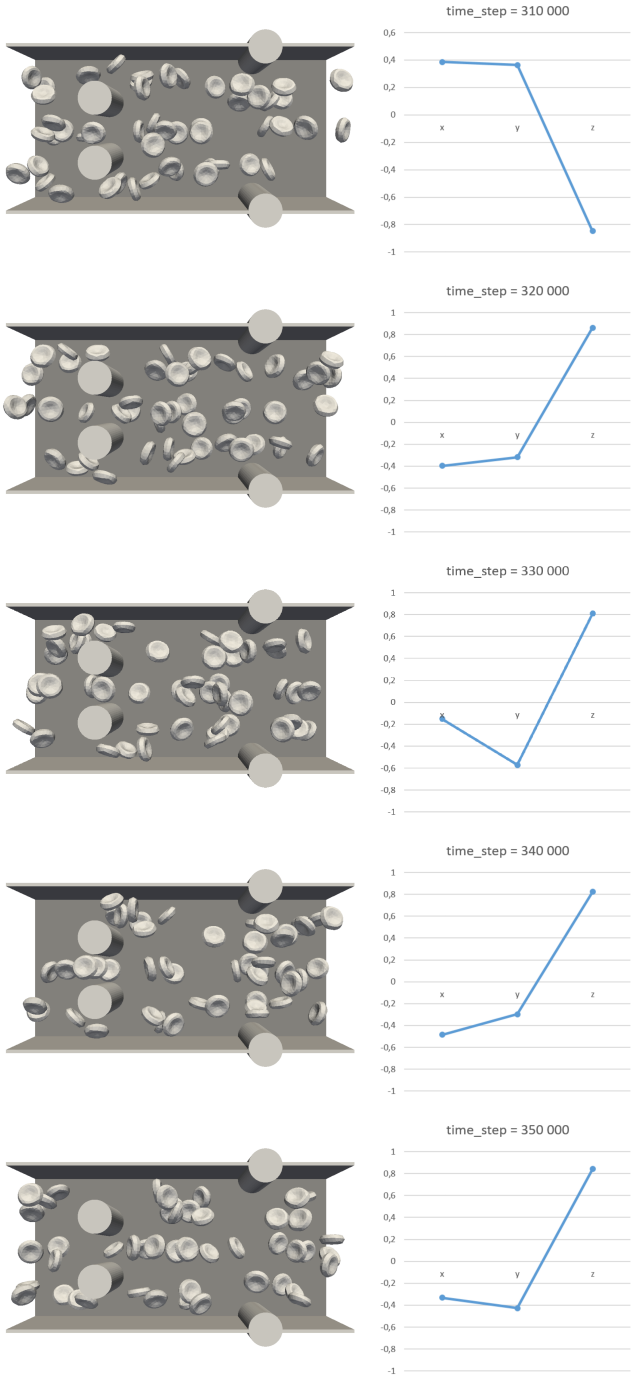


Fig. 9: PCA results compared to the images from the simulation for particular times

The next picture (Fig. 10) compares the plots of the PCA results from all nine time snapshots for the individual simulations (A50a, A50b, B50a, B50b, C50a, C50b, D50a, D50b, E50a and E50b).

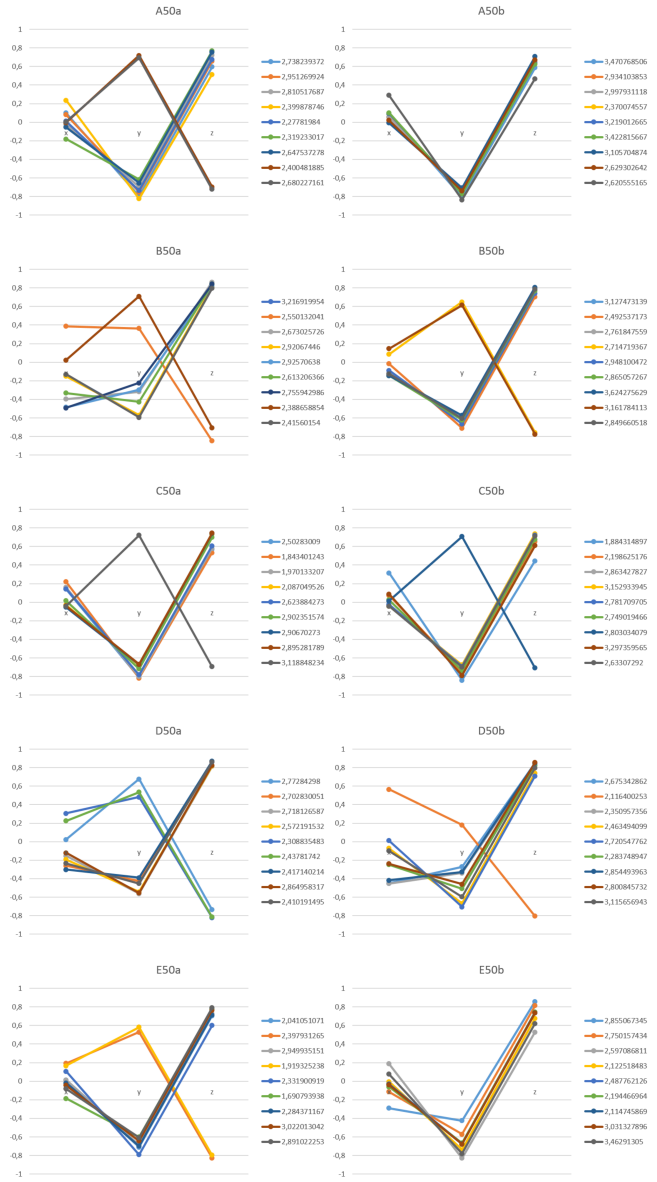


Fig. 10: PCA results from both experiments for each channel for all nine time snapshots. The number for each line represents the importance of the result.

We can see that the characteristics differ for the channels with various types of topology. It may be stated that the characteristics express acertain rate of scattering or various skew of the cells during the simulation. The skew is, of course, influenced by placing the obstacles inside the channel. When the obstacles are close to each other, acell needs to be standing up to get through, i. e. it needs to be parallel to the plane x, z. However, the previous characteristics resulting from the velocity (Section III).) showed that also the fluid velocity, therefore also the cells velocity differed depending on the channel topology.

We decided to use the E-type channel to find out how the fluid velocity itself influenced the cell skew. We ran the simulations with half (0,00025), double (0,001) and quadruple (0,002) external fluid power, compared to the previous simulations. With each, we placed 50 RBCs with different random initial seeding. The PCA analysis results from these three additional simulation experiments, together with the experiment E50a, for data sets and identical time images identical with the previous cases, are shown in the following picture (Fig. 11).

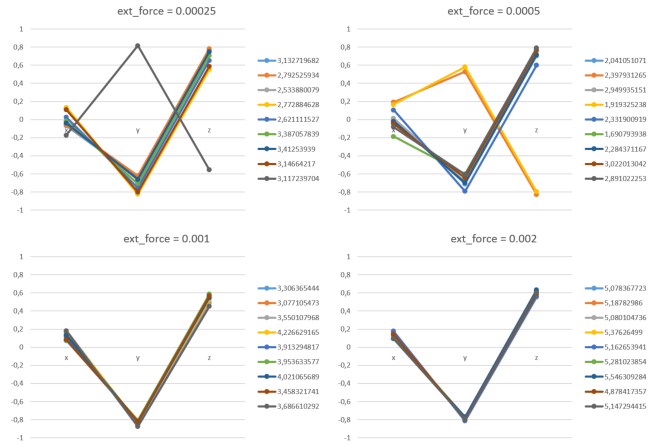


Fig. 11: PCA results for experiments in channel E for different initial fluid external forces.

Obviously, higher fluid velocity results into more tidy simulations, i. e. the cell skew for faster fluid flow differs in the course of the simulation less than with the simulation with slower fluid flow.

Therefore, by analyzing the main components we obtain astatistical characteristics which defines the skew direction of the majority of the cells in the simulation. It appeared that the main directions were different with the different topologies of the channel, and similar with the simulations in the same topologies. Moreover, the main directions get stabilized in the channel without obstacles if we increase the velocity of the fluid in the simulation. Therefore, the characteristics describes well the rotation of the cells in the channel, and distinguishes among various channel topologies.

V. Conclusion

This work is afollow-up to our previous study in which we tried for the first time to propose the statistical characteristics of red blood cells flow simulation in afluid flow. We verified the statistics resulting from RBC velocities during the simulation for various channel topologies. The statistical compliance test carried out confirmed that the characteristics identified well the experiments done in the channels with identical topologies.

In the second part we suggested more advanced characteristics describing cell skew during the simulation. To do so, we used PCA. The results of this analysis also showed similarities among the experiments carried out in the channel with identical topology. We kept testing the characteristics for different fluid velocities in the channel without obstacles. We found out that the fluid velocity influenced significantly the main skew of the cells during the simulation. Not only will such characteristics help us identify the simulation or biological experiments for the purposes of mutual comparisons, it will also help us better understand the cells behaviour during the simulation. Such information might come useful also for our main goal, which is optimizing the microfluidic devices capturing specific blood cells.

REFERENCES

[1] K. Bachratá and H. Bachratý, "On modeling blood flow in microfluidic devices", ELEKTRO 2014: 10th International Conference, IEEE, ISBN 978-4799-3720-2, 2014, pp. 518-521

[2] K. Bachratá, H. Bachratý and M. Slavík, "Statistics for comparison of simulations and experiments of ow of blood cells", International conference Experimental Fluid Mechanics, Mariánské Lázne 2016, efm.kez.tul.cz

[3] M. Bušík, I. Jančigová, R. Tóthová and I. Cimrák, "Simulation study of rare cell trajectories and capture rate in periodic obstacle arrays", Journal of Computational Science, 2016

[4] M. Bušík, I. Cimrák,"The calibration of fluid-object interaction in immersed boundary method", Experimental fluid mechanics 2016, 15.-18.11.2016, Mariánské Lázne 2016, efm.kez.tul.cz

[5] I. Cimrák, M. Gusenbauer, and T. Schrefl, "Modelling and simulation of processes in microfluidic devices for biomedical applications", Computers and Mathematics with Applications, 2012, Vol 64(3), pp.278-288

[6] I. Cimrák, M. Gusenbauer and I. Jančigová, "An ESPResSo implementation of elastic objects immersed in a fluid", Computer Physics Communications, vol. 185, 2014, pp. 900-907

[7] I. Cimrák, I. Jančigová, R. Tóthová, M. Gusenbauer, "Mesh-based modeling of individual cells and their dynamics in biological fluids", In: Applications of Computational Intelligence in Biomedical Technology, Vol. 606 of Studies in Computational Intelligence, Springer International Publishing, 2015, ISBN 978-3-319-19146-1

[8] I. Cimrák, K. Bachratá, H. Bachratý, I. Jančigová, R. Tóthová, M. Bušík, M. Slavík and M. Gusenbauer, "Object-in-fluid framework in modeling of blood flow in microfluidic channels, Communications", Scientific Letters of the University of Zilina, vol. 18/1a, 2016, pp. 13-20

[9] D. A. Fedosov, B. Caswell, G.E. Karniadakis, "A Multiscale Red Blood Cell Model with Accurate Mechanics, Rheology, and Dynamics". Biophysical Journal, 98(10), 2010, 22152225. http://doi.org/10.1016/j.bpj.2010.02.002

[10] M. Figurová, D. Pudiš, P. Gašo and I. Cimrák, "PDMS microfluidic structures for LOC applications", 2016 ELEKTRO, Štrbské Pleso, 2016, pp. 608-611, doi: 10.1109/ELEKTRO.2016.7512150

[11] M. Gusenbauer, G. Mazza, M. Brandl, T. Schrefl, R. Tóthová, I.Jančigová, and I. Cimrák, "Sensing platform for computational and experimental analysis of blood cell mechanical stress and activation in microfluidics", in Procedia Engineering: Proceedings of the 30th anniversary Eurosensors Conference, September 2016, Vol. 168, pp.1390-1393

[12] R. Tóthová, I. Jančigová and M. Bušík, "Calibration of elastic coefficients for spring-network model of red blood cell", International Conference on Information and Digital Technologies (IDT), 2015, pp. 376-380

[13] K. Tsubota, S. Wada, T. Yamaguchi, "Particle method for computer simulation of red blood cell motion in blood flow", Computer Methods and Programs in Biomedicine, Volume 83, Issue 2, 2006, Pages 139-146, ISSN 0169-2607, http://dx.doi.org/10.1016/j.cmpb.2006.06.005.

[14] D. Xu, E. Kaliviotis, A. Munjiza, E. Avital, Ch. Ji, J. Williams, "Large scale simulation of red blood cell aggregation in shear flows", Journal of Biomechanics, Volume 46, Issue 11, 2013, Pages 1810-1817, ISSN 0021-9290, http://dx.doi.org/10.1016/j.jbiomech.2013.05.010.

Control Algorithms for Discrete Delayed Systems with Unknown Inputs and Model Parameters Using Nonparametric Technique

Valery I. Smagin, Gennady M. Koshkin, Konstantin S. Kim
Tomsk State University
Department of Applied Mathematics and Cybernetics
Tomsk, Russia
e-mail: vsm@mail.tsu.ru, kgm@mail.tsu.ru, kks93@rambler.ru

***Abstract*—The paper deals with the control algorithms for discrete delayed systems with unknown inputs (disturbances) and model parameters. The control algorithm is based on the local criterion with using Kalman filters and nonparametric estimators. The example is given to illustrate the proposed approach.**

Keywords—algorithm of control with delay; nonparametric estimator; system with unknown inputs; system with unknown parameters.

I. INTRODUCTION

The locally optimal discrete control systems are a special type of the discrete model predictive control [1, 2] (MPC) with one step forecast. The main advantage of the method of locally optimal control is a significant simplification on the synthesis of the procedure. Last years, the MPC procedures have been applied to technical systems [2], chemical processes [3], inventory control [4], production-inventory system [5–7], and portfolio optimization [8].

In this paper, we study the control-delayed systems with an unknown input and unknown parameters which are interpreted as parameters of the functions of the unknown input. It is important to note that the synthesized control strategy for delayed control systems does not require an extension of the state space that simplifies the synthesis procedure as the matrices used in the algorithm retain their dimensions.

The considered approach generalizes the results of [9] to parametric uncertainty. We propose estimates of the unknown input obtained by making use of the least mean squares (LMS) method [10–13] and nonparametric algorithms [14–18]. The suggested approach allows to improve the estimation accuracy of the state vector and unknown input. There is presented an example illustrating the effectiveness of the proposed control strategies using filtering algorithm with nonparametric estimators in comparison with the known algorithms.

The reported study was funded by RFBR according to the research project № 17-08-00920.

II. MODEL OF DISCRETE SYSTEM WITH UNCERTAINTIES

Consider the model of an object which is described by the discrete equation

$$x(k+1) = (A+\Delta A)x(k) + (B+\Delta B)u(k-h) + Fs(k) + q(k),$$

$$x(0) = x_0,\ u(j) = \psi(j),\ j = -h, -h+1, ..., -1, \qquad (1)$$

where $x(k) \in R^n$ is the state vector, $u(k-h) \in R^m$ is a control vector, h is a time delay, $s(k) \in R^{n_1}$ is a disturbances vector, ΔA and ΔB are matrices of unknown parameters, x_0 and $\psi(j)$ ($j = -h, -h+1, ..., -1,$) are the known vectors. Matrices A, B, and F are the given constant matrices. In (1) $q_x(k)$ are the Gaussian random sequences with

$$\mathrm{E}[q(k)] = 0,\ \ \mathrm{E}[q(k)q^{\mathrm{T}}(j)] = Q\,\delta_{k,j}.$$

Here $\delta_{k,j}$ is the Kronecker symbol, i.e. $\delta_{k,j} = 1$ if $k = j$ and $\delta_{k,j} = 0$ if $k \neq j$, $\mathrm{E}[\cdot]$ denotes the expectation of a random variable, $^{\mathrm{T}}$ denotes the matrix transposition

Introduce the model of observations

$$y(k) = Hx(k) + \eta(k), \qquad (2)$$

where $\eta(k)$ are the Gaussian random sequences with

$$\mathrm{E}[\eta(k)] = 0,\ \mathrm{E}[\eta(k)q^{\mathrm{T}}(j)] = 0,\ \ \mathrm{E}[\eta(k)\eta^{\mathrm{T}}(j)] = V\delta_{k,j}.$$

The model of disturbances contains unknown parameters and is defined by the following difference equation:

$$s(k+1) = (R+\Delta R)s(k) + f + \Delta f + q_s(k),\ s(0) = s_0, \qquad (3)$$

where R is the known matrix, f is the known vector (R and f are nominal parameters of the model of disturbances), ΔR and Δf are an unknown matrix and vector, s_0 is the random vector of initial conditions

$$\mathrm{E}[s_0] = \bar{s}_0,\ \ \mathrm{E}[(s_0 - \bar{s}_0)(s_0 - \bar{s}_0)^{\mathrm{T}}] = P_0,$$

$q_s(k)$ is the Gaussian random sequence with

$$\mathrm{E}[q_s(k)] = 0,\ \mathrm{E}[q_s^{\mathrm{T}}(k)q_s(j)] = Q_s\delta_{k,j}.$$

Matrices ΔA, ΔB, ΔR and vector Δf can be interpreted as parameters of the functions of an unknown input. Then, taking into account these functions, we obtain instead of the models (1) and (3) the following models:

$$x(k+1) = Ax(k) + Bu(k-h) + r_x(k) + Fs(k) + q(k), \quad (4)$$

$$s(k+1) = Rs(k) + f + r(k) + q_s(k), \quad (5)$$

where $r_x(k) = \Delta Ax(k) + \Delta Bu(k-h)$ and $r(k) = \Delta Rs(k) + \Delta f$ are the functions of an unknown input.

Suppose that indirect observations of vector $s(k)$ are described as

$$\omega(k) = \Phi s(k) + \tau(k). \quad (6)$$

Here $\omega(k) \in R^{m_1}$ is a vector of observations; Φ is $(m_1 \times n)$-matrix; $\tau(k)$ are random errors of observations; $\tau(k)$ is the Gaussian random sequence with

$$\mathrm{E}[\tau(k)] = 0,\ \mathrm{E}[\tau(k)\tau^{\mathrm{T}}(j)] = \mathrm{T}\delta_{k,j}.$$

III. Local Criterion and Control

Take the local criterion in the form

$$I(k) = \mathrm{E}[(w(k+1) - z(k))^{\mathrm{T}} C(w(k+1) - z(k)) + u^{\mathrm{T}}(k-h)Du(k-h) / \Omega_0^k, Y_0^k], \quad (7)$$

where $C > 0, D \ge 0$ are weight matrices, $z(k)$ is a tracked vector, $\Omega_0^k = \{\omega(0), \omega(1), ..., \omega(k)\}$, $Y_0^k = \{y(0), y(1), ..., y(k)\}$.

Let us find control by the principle of separation in accordance with [9] using estimates of filtering and forecasting:

$$u(k-h) = -(B^{\mathrm{T}}CB + D)^{-1}B^{\mathrm{T}}C(A^{h+1}\hat{x}(k-h) + \sum_{i=1}^{h} A^i Bu(k-h-i) + A^h(F\hat{s}_f(k-h) + \hat{r}_x(k-h)) + \sum_{i=0}^{h-1} A^i(F\hat{s}_p(k-i) + \hat{r}_{x,p}(k-i)) - z(k)). \quad (8)$$

The estimates in (8) are constructed on the base of the Kalman filter [19] with making use of the LMS method [10–13] and nonparametric algorithms [14–18]. Now obtain the estimate of the state vector (4) by the following algorithm using the vector of estimates of an unknown inputs $\hat{s}_f(\cdot)$ and $\hat{r}_x(\cdot)$:

$$\hat{x}(k) = A\hat{x}(k-1) + Bu(k-h-1) + \hat{s}_f(k-1) + \hat{r}_x(k-1) + K_x(k) \times \{w_x(k) - H[\hat{x}(k-1) + Bu(k-h-1) + \hat{s}_f(k-1) + \hat{r}_x(k-1)]\}, \quad (9)$$

$$K_x(k) = P_x(k/k-1)H^{\mathrm{T}}(HP_x(k/k-1)H^{\mathrm{T}} + V)^{-1},$$

$$P_x(k/k-1) = AP_x(k-1)A^{\mathrm{T}} + Q,$$

$$P_x(k) = (E_n - K_x(k)H)P_x(k/k-1),\ P_x(0) = P_{x,0}. \quad (10)$$

In (10) E_n is the unit $n \times n$ matrix.

The estimate $\hat{r}_x(k)$ is determined by minimum with respect to $r_x(k)$ of the criterion

$$J = \sum_{i=1}^{k} \left\{ \|\chi(i)\|_{W_1}^2 + \|r_x(i-1)\|_{W_2}^2 \right\}, \quad (11)$$

where $\chi(i) = y(i) - H\tilde{x}(i)$, $\tilde{x}(i) = A\hat{x}(i-1) + Bu(i-h-1) + F\hat{s}(i-1)$); $W_1 > 0$, $W_2 \ge 0$ are weight matrices; $\|\chi(i)\|_V^2 = \chi^{\mathrm{T}}(i)W_1\chi(i)$. So, we have the estimate

$$\hat{r}_x(k) = [H^{\mathrm{T}}W_1H + W_2]^{-1}H^{\mathrm{T}}W_1\{y(k) - H[A\hat{x}(k-1) + Bu(k-h-1) + F\hat{s}_f(k-1)]\}. \quad (12)$$

The forecast estimate $\hat{r}_{x,p}(k)$ in the formula (8) is constructed with using methods of time series [20].

Another estimate of $\hat{r}_x(k)$ using nonparametric technique has the form

$$\hat{r}_{x,Np}(k) = [H^{\mathrm{T}}W_1H + W_2]^{-1}H^{\mathrm{T}}W_1\,\tilde{y}(k). \quad (13)$$

In (12) the j-th component of the vector $\tilde{y}$ is defined as the following analog of the known Nadaraya-Watson nonparametric regression estimate [17, 18]:

$$\tilde{y}_j(k) = \frac{\sum_{i=1}^{k} \chi_j(i) K\left(\frac{k-i+1}{h_j}\right)}{\sum_{i=1}^{k} K\left(\frac{k-i+1}{h_j}\right)}. \quad (14)$$

Here $K(\cdot)$ is a kernel function, h_j is a bandwidth parameter. Note, we take

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

because the use of other kernels, as shown by simulation, leads to a deterioration in the quality of estimation.

The estimates $\hat{s}_f(k-h)$ and $\hat{s}_p(k-i)$ are filtering and forecast algorithms, which are based on the optimal Kalman filter [19] using the vector of estimates of an unknown input $\hat{r}(\cdot)$:

$$\hat{s}_f(k-h) = R\hat{s}_f(k-h-1) + \hat{r}(k-h-1) + K_f(k-h)$$

$$\times[\omega(k-h)-\Phi(R\hat{s}_f(k-h-1)+\hat{r}(k-h-1))],$$

$$\hat{s}_f(0)=\bar{s}_0, \quad (15)$$

$$K_f(k-h)=P(k-h/k-h-1)\Phi^{\mathrm{T}}$$

$$\times(\Phi P(k-h/k-h-1)\Phi^{\mathrm{T}}+\mathrm{T})^{-1},$$

$$P(k-h/k-h-1)=RP(k-h-1)R^{\mathrm{T}}+Q_s,$$

$$P(k-h)=(E_{n_1}-K_f(k-h)\Phi)P(k-h/k-h-1),$$

$$P(0)=P_0. \quad (16)$$

To construct the forecast estimates, we have to use the Kalman extrapolator [19], which allows to calculate the estimate of forecasts of disturbances by one step:

$$\hat{s}_p(k-h+1)=R\hat{s}_p(k-h)+f+\hat{r}(k-h)$$

$$+K_p(k-h)(\omega(k-h)-\Phi\hat{s}_p(k-h)),\ \hat{s}_p(0)=\bar{s}_0, \quad (17)$$

$$K_p(k-h)=RP_{pr}(k-h)\Phi^{\mathrm{T}}(\Phi P_{pr}(k-h)\Phi^{\mathrm{T}}+\mathrm{T})^{-1},$$

$$P_{pr}(k-h+1)=(R-K_p(k-h)\Phi)P_{pr}(k-h)$$

$$\times(R-K_p(k-h)\Phi)^{\mathrm{T}}+Q_s+K_p(k-h)TK_p^{\mathrm{T}}(k-h),$$

$$P_{pr}(0)=P_0. \quad (18)$$

The forecasts for the next steps $j=2,...,h-1$ are determined as follows:

$$\hat{s}_p(k-h+j)=R\hat{s}_p(k-h+j-1)$$

$$+f+\hat{r}(k-h+j-1).$$

The estimate of the vector $\hat{r}(k-h)$, obtained by using the LMS method and nonparametric algorithms, is based on the minimization with respect to $r(k-h)$ of the criterion

$$J=\sum_{i=1}^{k-h+1}\left\{\|\bar{\chi}(i)\|^2_{V_1}+\|r(i-1)\|^2_{V_2}\right\}, \quad (19)$$

where

$$\bar{\chi}(i)=\omega(i)-\Phi R\tilde{s}(i-1)\ (\tilde{s}(i-1)=R\hat{s}_f(k-2)+\hat{r}(k-2)),$$

$V_1 > 0$, $V_2 \geq 0$ are weight matrices. So,

$$\hat{r}(k-h)=[\Phi^{\mathrm{T}}V_1\Phi+V_2]^{-1}\Phi^{\mathrm{T}}V_1(\omega(k-h)$$

$$-\Phi(R\hat{s}_f(k-h-1)+\hat{r}(k-h-1))). \quad (20)$$

Another estimate using nonparametric technique has the form

$$\hat{r}_{Np}(k-h)=[\Phi^{\mathrm{T}}V_1\Phi+V_2]^{-1}\Phi^{\mathrm{T}}V_1\Omega(k), \quad (20)$$

where the j-th component of the vector Ω is calculated by the following formula:

$$\Omega_j(p)=\frac{\sum_{i=1}^{p}\bar{\chi}(i)_j(i)K\left(\frac{p-i+1}{h_j}\right)}{\sum_{i=1}^{p}K\left(\frac{p-i+1}{h_j}\right)}. \quad (21)$$

The kernel functions $K(\cdot)$ and bandwidth parameters h_j are defined as in (14).

IV. An Illustrative Example

Calculations are performed for the following data of models:

$$A=\begin{pmatrix}0.75 & 0\\ 0 & 0.79\end{pmatrix},\ R=\begin{pmatrix}0 & 1\\ 0.1 & 0.5\end{pmatrix},$$

$$Q=\mathrm{diag}\{0.05\quad 0.02\},\ Q_s=\mathrm{diag}\{0.07\quad 0.03\},$$

$$W_1=W_2=0.01E_2,\ V=0.05E_2,\ r(k)=\begin{pmatrix}0.8\\ 0.8\end{pmatrix},$$

$$D=0,\ B=H=F=C=P_0=V_1=V_2=E_2.$$

Simulations are carried out with the matrices which are connected with unknown inputs $r_x(k)$ and $r(k)$:

$$\Delta A=\begin{pmatrix}0.05 & 0\\ -0.003 & 0.03\end{pmatrix},\ \Delta B=\begin{pmatrix}0.01 & 0\\ 0 & 0.02\end{pmatrix},$$

$$\Delta R=\begin{pmatrix}0 & 0.2\\ 0.05 & 0.1\end{pmatrix},\ \Delta f(k)=\begin{pmatrix}0.1\sin(0.2k)\\ 0.1\sin(0.2k)\end{pmatrix}.$$

The control and filtering algorithms are compared with the algorithms using the LMS estimates from [10, 11]. These comparisons are given in Figs. 1–7:

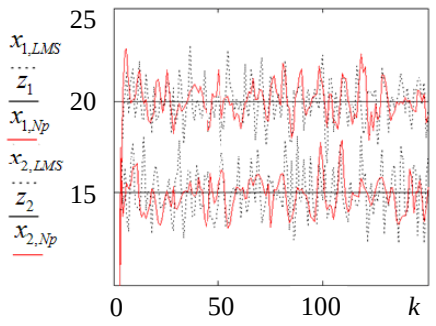


Figure 1. The tracking of components z_1=20 and z_2=15 by x_1 and x_2 with use of the LSM estimates (dotted lines x_{LMS}) and nonparametric estimates (solid line x_{Np})

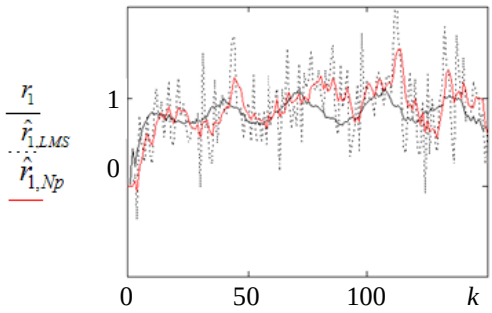


Figure 2. The evaluation of unknown inputs r_1 in model disturbances with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

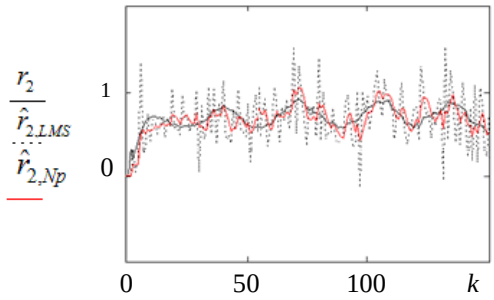


Figure 3. The evaluation of unknown inputs r_2 in model disturbances with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

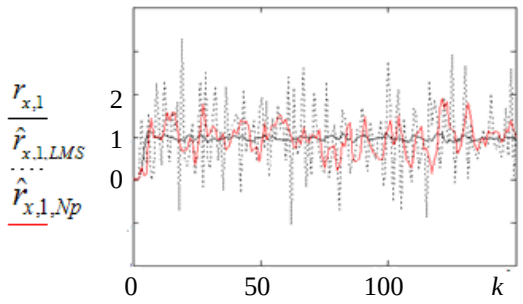


Figure 4. The evaluation of unknown inputs $r_{x,1}$ in model of object with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

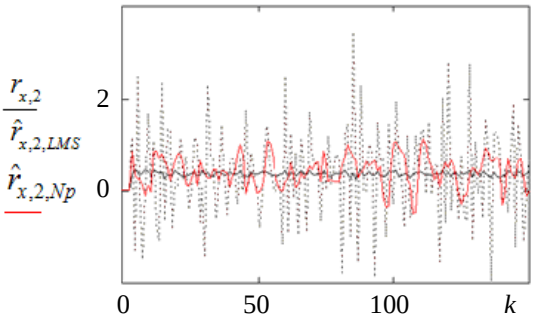


Figure 5. The evaluation of unknown inputs $r_{x,2}$ in model of object with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

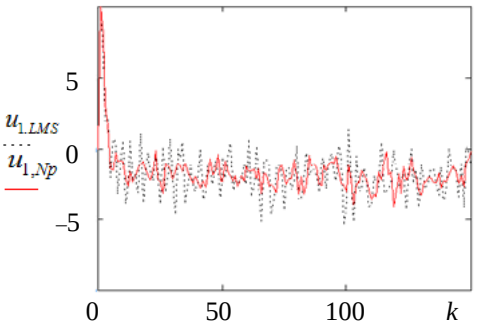


Figure 6. The evaluation of u_1 with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

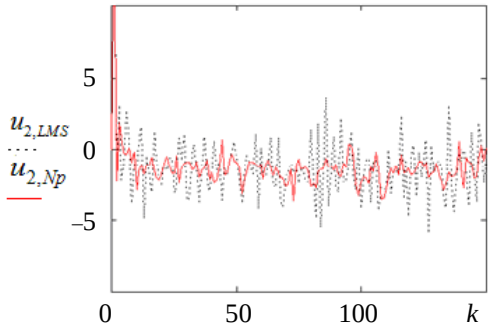


Figure 7. The evaluation of u_2 with use of the LSM-estimates (dotted lines) and nonparametric estimates (solid line)

Below in Tables 1–6, the empirical mean square errors (MSEs) are given for $N = 150$ by averaging 20 realizations:

$$\sigma_{x,i} = \sqrt{\frac{\sum_{k=1}^{N}(x_i(k) - z_i(k))^2}{N-1}},\ \sigma_{\hat{x},i} = \sqrt{\frac{\sum_{k=1}^{N}(x_i(k) - \hat{x}_i(k))^2}{N-1}},$$

$$\sigma_{sf,i} = \sqrt{\frac{\sum_{k=1}^{N}(s_i(k) - \hat{s}_{f,i}(k))^2}{N-1}},\ \sigma_{sp,i} = \sqrt{\frac{\sum_{k=1}^{N}(s_i(k) - \hat{s}_{p,i}(k))^2}{N-1}},$$

$$\sigma_{r_x,i} = \sqrt{\frac{\sum_{k=1}^{N}(r_{x,i}(k) - \hat{r}_{x,i}(k))^2}{N-1}},\ \sigma_{r,i} = \sqrt{\frac{\sum_{k=1}^{N}(r_i(k) - \hat{r}_i(k))^2}{N-1}},$$

$$i = \overline{1,2}\,.$$

TABLE I. EMPIRICAL MSE $\sigma_{x,i}$ OF STATE VECTOR x

Components	LMS	Nonparametric
1	1.36	0.762
2	1.394	0.736

TABLE II. EMPIRICAL MSE $\sigma_{\hat{x},i}$ OF ESTIMATE $\hat{x}$ OF STATE VECTOR

Components	LMS	Nonparametric
1	0.693	0.457
2	0.704	0.474

TABLE III. EMPIRICAL MSE $\sigma_{sf,i}$ OF FILTERING ESTIMATE

Components	LMS	Nonparametric
1	0.219	0.180
2	0.221	0.151

TABLE IV. EMPIRICAL MSE $\sigma_{sp,i}$ OF FORECASTING ESTIMATE

Components	LMS	Nonparametric
1	0.438	0.347
2	0.317	0.241

TABLE V. EMPIRICAL MSE $\sigma_{r_x,i}$ OF UNKNOWN INPUT ESTIMATE

Components	LMS	Nonparametric
1	0.903	0.231
2	0.97	0.225

TABLE VI. EMPIRICAL MSE $\sigma_{r,i}$ OF UNKNOWN INPUT ESTIMATE

Components	LMS	Nonparametric
1	0.387	0.155
2	0.284	0.101

The simulations without using the estimate of unknown input components of the mathematical models of disturbance and object show that there is the breakdown of the tracking vector *z*.

The LMS method for estimating unknown input components of the mathematical models of disturbance and object provides the working capacity of tracking, and application of the nonparametric procedures increase the quality of estimates and the control algorithm.

V. CONCLUSION

In this paper, the algorithms of the Kalman filtering and control for discrete delayed systems with unknown input and unknown parameters are constructed. The proposed method has been verified by simulations. Figures and Tables show that the procedures with nonparametric estimators have the advantages in the accuracy compared to the known algorithms using the LMS estimates. The presented filtering algorithms with nonparametric technique may be used in solving the control problems for the object with time-delay and with uknown parameters for inventory control (see [7]).

ACKNOWLEDGMENT

The authors express their gratitude to the reviewers whose remarks have improved the article.

REFERENCES

[1] J.M. Maciejowski, Predictive control with constraints, Prentice Hall. 2002.

[2] E.F. Camacho and C. Bordons, Model predictive control. London: Springer-Verlag, 2004.

[3] M.M. Arefi and A. Montazeri, “Model-predictive control of chemical processes with a Wiener identification approach”, Industrial Technology, pp. 1735–1740, 2006.

[4] P. Conte and P. Pennesi, “Inventory control by model predictive control methods”, Proc. 16th IFAC World Congress, Czech Republic, Prague, pp. 1–6, 2005.

[5] E. Aggelogiannaki, Ph. Doganis and H. Sarimveis, “An adaptive model predictive control conguration for production-inventory systems”, Int. J. of Production Economics, vol. 114, pp. 165–178, 2008.

[6] J.-C. Henneta, “A globally optimal local inventory control policy for multistage supply chains”, Int. J. of Production Research, vol. 47, issue 2, pp. 435–453, 2009.

[7] V.I. Smagin, G.M. Koshkin and K.S. Kim, “Locally optimal inventory control with time delay in deliveries and incomplete information on demand” Proc. Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management. February 15-18, Beer Sheva, Israel, pp. 570–574, 2016.

[8] V.V. Dombrovskii and T.Yu. Obedko, “Model predictive control for constrained systems with serially correlated stochastic parameters and portfolio optimization”, Automatica, vol. 54. pp. 325-331, 2015.

[9] V.I. Smagin, G.M. Koshkin and K.S. Kim, “Control strategies for discrete delayed systems with unknown input using nonparametric algorithms”, Proc. of The International Conference on Information and Digital Technologies, Rzeszov, Poland, July 5–7, pp. 133-137, 2016.

[10] S. Gillijns and B. Moor, “Unbiased minimum-variance input and state estimation for linear discrete-time systems”, Automatica, vol. 43, pp. 111–116, 2007.

[11] V.I. Smagin, “State estimation for nonstationary discrete systems with unknown input using compensations”, Russian Physics Journal, vol. 58, issue 7, pp. 1010–1017, 2015.

[12] G. Koshkin and V. Smagin, “Kalman filtering and forecasting algorithms with use of nonparametric functional estimators”, Springer Proceeding in Mathematical Statistics. Ricardo Cao et al. Eds. vol. 175, pp. 75–84, 2016.

[13] V.I. Smagin and G.M. Koshkin, “Kalman filtering and control algorithms for systems with unknown disturbances and parameters using nonparametric technique. Proc. 20th Int. Conference on Methods and Models in Automation and Robotics (MMAR), August 2015, Miedzyzdroje, Poland, pp. 247–251, 2015.

[14] A.V. Kitaeva and G.M. Koshkin, “Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression”, Automation and Remote Control, vol. 71, issue 2, pp. 257-274, February 2010.

[15] A. Dobrovidov, G. Koshkin and V. Vasiliev, “Non-parametric state space models”, Heber, UT 84032, USA. Kendrick Press, Inc. 2012.

[16] G. Koshkin and V. Smagin, “Filtering and prediction for discrete systems with unknown input using nonparametric algorithms”, Proc. 10th International Conference on Digital Technologies. Zilina, Slovakia, July 9-11, pp. 120–124, 2014.

[17] E. Nadaraya, “On estimating of regression,” Theory Probab. Appl., vol. 9, pp. 141–142, 1964.

[18] G.S. Watson, “Smooth regression analysis”, Sankhya. Indian J. Statist., vol. A26, pp. 359–372, 1964.

[19] K. Brammer and G. Siffling, “Kalman-Bucy-Filter”, München. R. Oldenbourg Verlag. 1975.

[20] G.E.P. Box and G.M. Jenkins, “Time series analysis: forecasting and control”, San Francisco, CA: Holden-Day, 1976.

Digital Broadband Camera based on a Line Scanning Sensor

Michal Šusta, Pavel Zahradnik, Radek Klof, Petr Záleský and Boris Šimák
Department of Telecommunication Engineering
Czech Technical University in Prague
CZ-16627 Praha, Czech Republic
Email: {sustamic, zahradni, klofr, zaleskp, simak}@fel.cvut.cz

***Abstract*—In this paper we present a concept of a camera based on a line scanning sensor. The core of the camera is a proprietary broadband MEMS-based row imaging sensor and a powerful multiprocessor chip. We present here the architecture of the camera as well as selected subblocks of the system.**

***Index Terms*—digital camera, line sensor, scanning, digital signal processing.**

I. Introduction

One-dimensional (1D) line scanning image sensors are usually less expensive than their two-dimensional (2D) area scanning counterparts, mainly because of a lower number of picture elements and because of a simpler readout electronics resulting in smaller chip area. Frequently, for special purpose cameras, or sensors with special properties, requiring some cutting-edge or even some of not established technologies for their manufacturing, the area sensors are not available. A 1D image sensor completed by a suitable area scanning technology may represent an only viable alternative in constructing a camera. An example of such proprietary line scanning sensor is a 1D broadband MEMS-based microbolometric sensor with structured bridges covered by a carbon nanotube (CNT) absorption layer [1]. We present here a versatile solution of a digital camera which can accept various line scanning image sensors including the proprietary ones. The proposed camera consists of several essential blocks (Fig. 1) which include the optical subsystem, line scanning unit and its controller, line sensor and its controller, analog to digital converter (ADC), central processing unit (CPU), analog and digital video output (AVO, DVO), digital interfaces like GigE, USB etc., flash memory storage system (FMSS), human machine interface (HMI) and wireless data transfer. In particular chapters we describe some of these subblocks.

II. Central Processing Unit

The central processing unit is responsible for the control of all subsystems. Further, it accepts the raw data from the line scanning sensor, processes them digitally and forms a distortion free sequence of images which is output in form of a video stream. The CPU was carefully chosen in order to fulfill all specifications preferably by a programmable technology while avoiding any FPGA platform. The CPU is

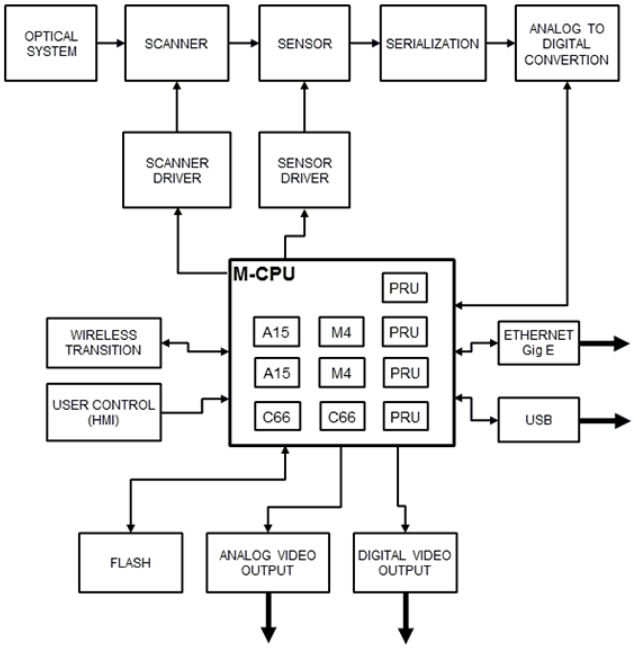


Fig. 1. Block scheme of the camera.

represented by the multiprocessor chip TI Sitara AM5728 [3]. Its block scheme is shown in Fig. 2. From the processing point of view, the core of the AM5728 comprises two ARM Cortex-A15 1.5GHz microprocessors, two ARM Cortex-M4 213MHz co-processors, two C66x floating-point very long instruction word (VLIW) 750MHz digital signal processors (DSP) and two dual-core programmable real-time unit and industrial communication subsystem (PRU-ICSS). They represent a well balanced mixture of computing power both for high-level as well as low-level software tasks. In order to ease the design of the CPU board, we use the BeagleBoard X-15 as a development platform [4]-[5]. The block scheme of the BeagleBoard X-15 platform is shown in Fig. 3.

III. Line Scanning Sensor

The final proprietary line scanning image sensor, which is still under development [1], is based on a row of hundreds of

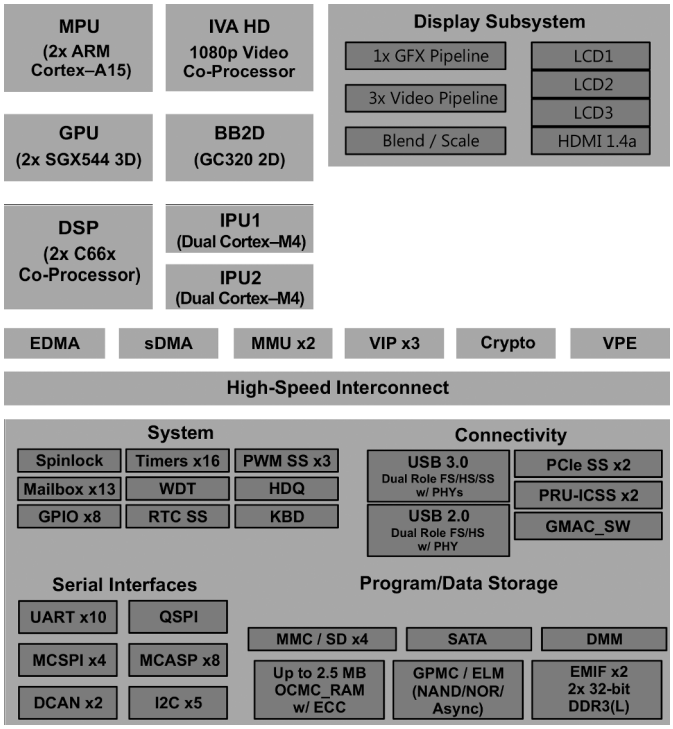


Fig. 2. Block scheme of the AM5728 [3].

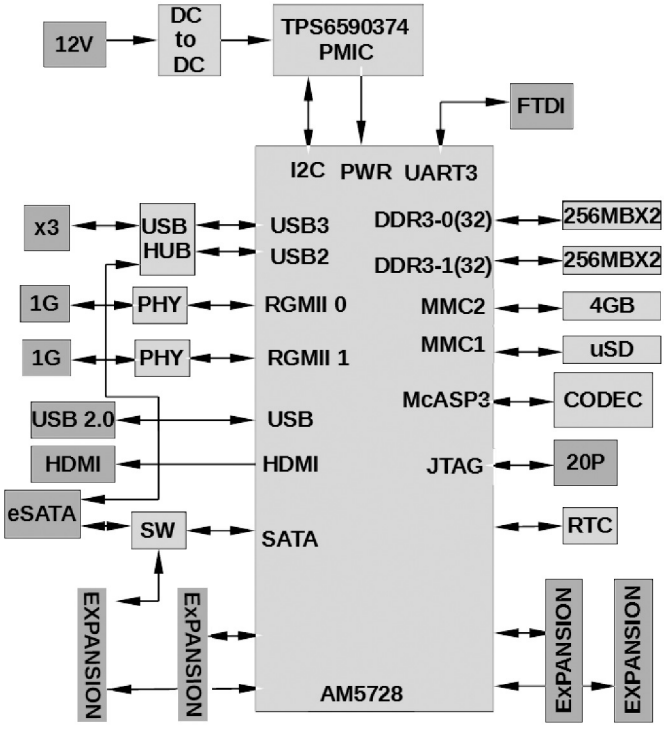


Fig. 3. System block scheme of the BeagleBoard X-15 [5].

ultra broadband CNT MEMS microbolometers. It will cover a broad portion of the electromagnetic spectrum ranging from the deep ultraviolet (DUV) to the long wave infrared (LWIR) region as well beyond it covering a substantial part of the terahertz (THz) region. The camera is purposely designed to be highly flexible in terms of compatibility with line scanning image sensors, even with the proprietary ones. In fact, various image scanning sensors differ by their driving signals as well as by the signals they provide. The flexibility related to various line scanning sensors is obtained due to the programmable PRU-ICSS units which are available inside the multiprocessor AM5728. The PRU-ICSS units are very fast in terms of their responsibility and in terms of pin toggling speed for bit banging purposes. By their proper usage, it is possible to completely avoid the FPGA technology which is frequently used in controlling high speed signals. In the time being, in order to prove the concept of the camera in absence of the anticipated line sensor, we have integrated various sensors, both the CCD as well as the CMOS ones. They include the TSL1402R [10] and the RL0512PAG-021 [11] for the visible region, and the G11135-512DE [12] for the short wave infra red (SWIR) region.

IV. Line Scanners

We have applied two concurrent types of scanners. The first one is based on the actuation of a torsional bridge made of 0.13mm thick FR4 material, see Fig. 5. The torsional oscillation of a triangular form are invoked by a piezoelectric actuator ViVa [6] shown in Fig. 4. The actuator is operated in a closed loop. We applied an optical PSD (Position Sensitive Device) sensor of the type 1L10CP2 [7] in order to get the current angle position of the reflective area. The second scanner is based on a rotational electromagnetic actuator called galvo, namely of the type Compact-506 [8] displayed in Fig. 6. The galvo is excited in order to oscillate the scanning mirror rotationally with a triangular time course of the angle of deflection. The feedback positional signal is obtained by the in-built sensor. Because the reflective area, i.e. the scanning mirror, is required to deflect broadband electromagnetic waves, the temperature of the scanning mirror is of interest. We continuously measure the temperature of the scanning mirror in a contactless manner by a 32×32 array of thermopile cells HTPA32x32d [9]. Fortunately, the emissivity of the mirror depends only slightly of its temperature. For the characterization of scanners, both scanners are evaluated in long-term trials in order to obtain their data in terms of robustness and potential fatigue of materials.

V. Drivers for Line Scanners

The piezo scanner requires a DC voltage of 180V for its full stroke. The piezo scanner represents a capacitive load of approx. 750nF. The corresponding driver (Fig. 7) is in fact a voltage amplifier. The high voltage operational amplifiers,

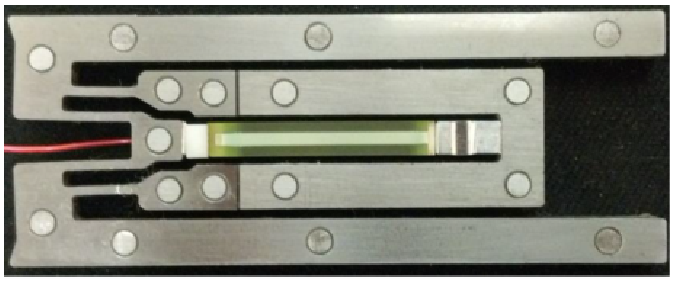


Fig. 4. Piezo actuator [6].

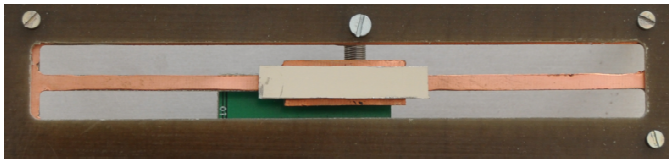


Fig. 5. Piezo based scanner.

typically of Apex, are an excellent option as well. The Compact 506 galvo scanner is driven by the driver displayed in Fig. 8. Due to the nature of the galvo actuator, the driver is designed as a transconductance amplifier resp. like a voltage to current (V-I) converter. The input of the amplifier is in fact the output of the digital to analog converter of the BeagleBoard. Given the optimal configuration of the DAC converters for its voltage range starting at ground potential, the amplifier input is adapted to the input voltage range of from 0 to 5V. The output voltage range of the DAC converter of the BeagleBoard is 0 to 3.3V. Construction of the driver is based on a driven feedback by the output current of the amplifier. The input current into the base of the output transistor is then controlled by an operational amplifier. The transconductance amplifier is controlled by adjusting the trimmers R1 and R2. The output current output of the driver can be designed to an arbitrary

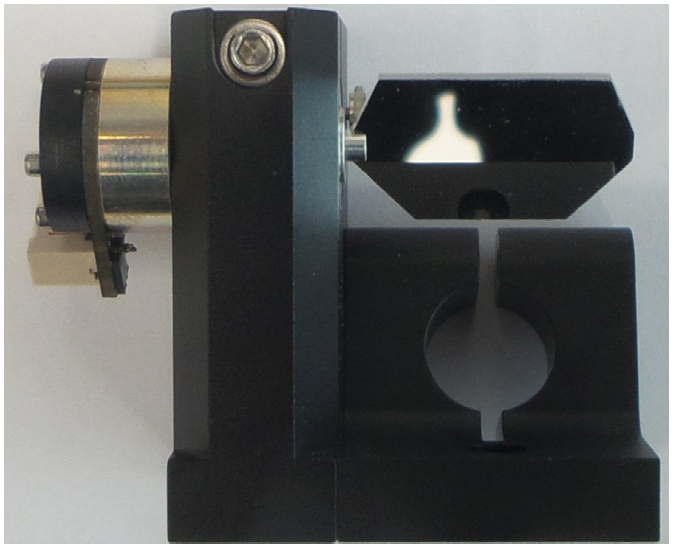


Fig. 6. Galvo scanner [8].

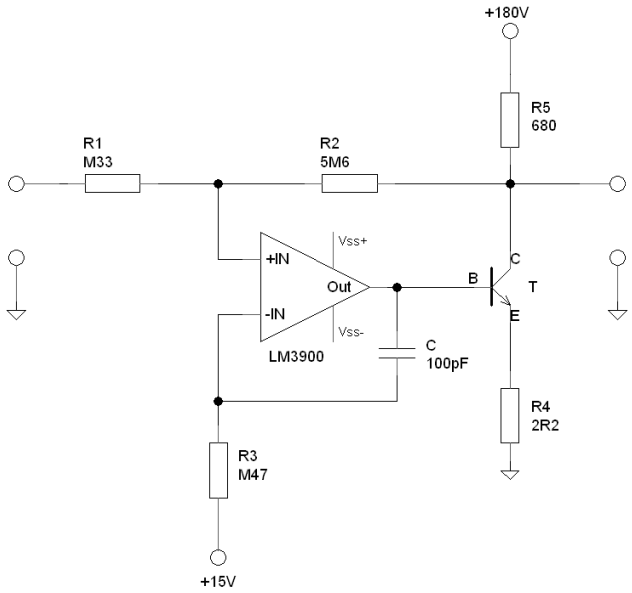


Fig. 7. Driver of the piezo scanner.

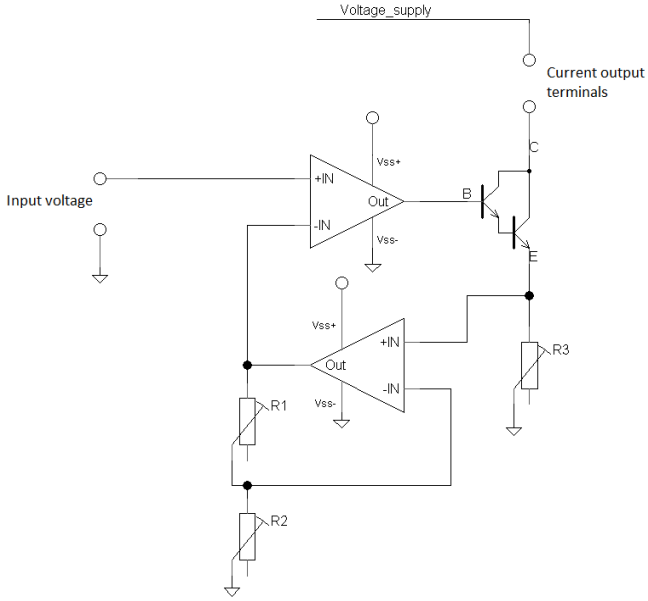


Fig. 8. Driver of the galvo scanner.

current value, which is limited only by the current and power load specification of the output transistor. By an appropriate choice of an operational amplifier, an excellent frequency characteristics of the driver can be achieved. However, due to the inertia of the galvo scanner, the demands on the frequency range of the amplifier are rather limited. The proposed solution provides a highly accurate control of the output current control as well as a high robustness and stability. The main advantage of the driver is its almost total independence in terms of its parameters of the properties of the output transistor. However, the maximum ratings of the output transistor have to be met.

VI. Optical Subsystem

Because the band of sensed wavelengths is unusually wide, the whole band cannot be covered by a single optics, of course. For the DUV, ultraviolet (UV), visible, near infrared (NIR) and mid wave infrared (MWIR) regions the optics consisting of lenses made of suitable materials, e.g. of the BK7, crystal quartz, MgF_2, CaF_2 is adequate. In the LWIR reagion we use ZnSe and Ge optics. In the terahertz region we prefer a mirror based optics sputtered e.g. by gold.

VII. Video Outputs

Essentially, the video outputs are based on the video controller integrated inside the AM5728 chip. In case of the BeagleBoard X-15, which is equipped with the standard HDMI interface, the video output operation is only in terms of customization of software drivers for peripherals of adequate programming graphical interpretation of processed data. We have implemented both the analog VGA as well as the digital HDMI video output.

VIII. Wired Interfaces

The AM5728 chip comprises USB ports, both the host as well as the device, and two Ethernet interfaces. The USB interface allows communication with an external computer system. It is utilized for the user communication and enables the runtime system configuration, e.g. in the testing phase, when it is necessary to flexibly change system parameters and to configure the data acquisition and measurements. At the same time the USB interface provides the ability of a fast data transfer between a host computer system and camera. These data transactions can be conveniently operated via the direct memory access (DMA) controller without adversing the computational capacity of the system. The data can be thus processed in real time, e.g. on a host PC and the system performance can be evaluated effectively. Due to the USB interface the camera can be operated and controlled remotely and share the resources of the host system like storage media, keyboard, touch interface etc. as well. The Ethernet interface has a speed of 1Gbps. This enables to use the data out of the camera at large distances over various networks. An advanced connectivity is good both for the video streaming as well as for the management of the camera, e.g. via a web interface.

IX. Wireless Interface

The camera is equipped with a module ODIN-W262 for wireless connectivity. This module enables wireless transmission in the standard Wi-Fi IEEE 802.11a/b/g/n. The wireless module is connected to the AM5728 via the reduced media-independent interface (RMII). The wireless traffic can be operated in both the 2.4GHz as well as in the 5GHz band. The wireless communication link is used for the data transfer over short and medium distances. The wireless data transfer can occur between the camera and the host PC or another devices with WiFi connectivity. The distance between communication nodes can be increased through additional antennas.

X. Flash Memory Interface

The Beagleboard X-15 module is populated by a slot for uSD card. The flash card can accommodate power-on booting data, configuration data, critical system data as well as various user defined data and image data.

XI. User Interface

The user interface is intended to setup the system parameters as well as for the data output format settings. The system configuration is accessible in several ways, e.g. from the host PC or from another device via USB, Wi-Fi or ethernet connection. Further, the camera signalizes to the user basic status information via LED diodes and via a small in-built display. Finally, a suitable sort of user handheld manual control of the camera can be attached via the USB port.

XII. Conclusion

We have presented a concept of a digital camera based on a line scanning image sensor along with a scanner and with a powerful multiprocessor unit. We have built a prototype of the camera and proved its functionality and usefulness. The camera can be used with various line scanning image sensors. The adjustment to a specific line scanning sensor is obtained by suitable circuitry and mainly by a modification of the software for driving the sensor.

Acknowledgment

This work was supported by the Ministry of the Interior of the Czech Republic under VI20152019043 project.

References

[1] V. Svatoš, I. Gablech, R. B. Ilic, J. Pekárek, P. Neužil, "In-situ observation of a CNT layer growth as IR absorber for bolometer-based sensors on substrates at ambient temperature", Nature Nanotechnology, under review.
[2] http://www.ti.com/product/am5728
[3] http://www.ti.com/product/am5728
[4] http://beagleboard.org/x15
[5] http://elinux.org/Beagleboard:BeagleBoard-X15
[6] http://www.vikingat.com/piezo-products/3mm-piezo-actuator
[7] http://www.sitek.se/pdf/psd/S1-0005-1L10_CP2.pdf
[8] http://www.scannermax.com/products/scanners/compact-506/
[9] http://www.heimannsensor.com/Datasheets/Overview-HTPA32x32d_Rev3.pdf
[10] http://ams.com/eng/Products/Light-Sensors/Linear-Array/TSL1402R
[11] http://www.excelitas.com/Pages/Product/High-Speed-High-Sensitivity-Linescan-Imagers-for-Machine-Vision.aspx
[12] http://www.hamamatsu.com/jp/en/product/alpha/I/4005/4208/4121/G11135-512DE/index.html

Impact of R/X Ratio of Distribution Network on Selection and Control of Energy Storage Units

Agata Szultka, Robert Malkowski, Stanislaw Czapp, Seweryn Szultka
Faculty of Electrical and Control Engineering
Gdansk University of Technology
Gdansk, Poland
agata.szultka@pg.gda.pl, robert.malkowski@pg.gda.pl, stanislaw.czapp@pg.gda.pl, seweryn.szultka@pg.gda.pl

***Abstract*—The interest in energy storage is still increasing. Energy storage units are installed in high voltage networks, medium voltage networks and low voltage distribution networks as well. These units are often used to improve power quality. One of the criteria for improving power quality is reducing voltage deviations. Depending on the type of network and specifying its *R/X* ratio, this criterion can be fulfilled by control of active or reactive power. This paper presents the selection of both location and rated power of energy storage units for two distribution networks with different *R/X* ratios. The selection is performed with the use of genetic algorithm. Effect of this selection on voltage levels in the networks is presented as well.**

Keywords—distribution networks; energy storage; genetic algorithm; voltage control

I. INTRODUCTION

Energy Storage Units (ESU) in the distribution networks are applied for many purposes. Among others, it is possible to use ESU both for the marketing of active power/energy and to improve the quality of delivered power [1]. In the distribution network, one of the important issues is limiting voltage deviations, especially if there are additional sources of renewable energy [2]. As Energy Storage Units, Battery Energy Storage Systems (BESS) are often used. BESS are connected to the power system through an inverter, therefore it is possible to deliver both active and reactive power to the network [3], [4], [5], [6].

The difference in *R/X* ratio of distribution network (network short-circuit resistance to reactance) depends on whether the network is rural (mainly overhead lines) or urban (mainly cable lines), and whether the network is low voltage or medium voltage. The voltage in the nodes is determined both by the network *R/X* ratio and the transferred level of active power *P* and reactive power *Q* (Fig. 1). The relationship between voltages V_j and V_i can be expressed by the following formula [7]:

$$V_j = V_i - \frac{1}{V_j}(R_{i,j} \cdot P_j + X_{i,j} \cdot Q_j) \quad (1)$$

One of the methods of ESU selection, taking into account limitation of voltage deviations, is to create the sensitivity matrix, which presents the derivative of the voltage in every node as a function of the delivered active power $\partial V_i/\partial P_{i,j}$ and reactive power $\partial V_i/\partial Q_{i,j}$.

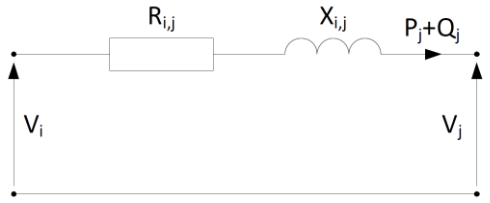


Figure 1. Power flow through the power system components.

It is possible to find the location/placement and rated power of BESS basing on the sensitivity matrix and the daily variability of voltage in every node. Such a problem is considered in the publications [6], [7], [8], [9], [10] – the placement and rated power of BESS is determined by analysis of the sensitivity matrix. The analysis is easy to perform when the analysed power network is not too large and complex.

This paper presents a Genetic Algorithm (GA) for selection of the best location and rated power of BESS in distribution power network, taking into account its *R/X* ratio. In addition, basing on the algorithm, it is possible to indicate the type of control of the BESS (control active or reactive power). The main purpose of the optimal selection is to reduce the voltage deviations by using the minimum power of the BESS. The results obtained using the Genetic Algorithm are compared with sensitivity matrix. The analysis has been performed for two power networks with different *R/X* ratio.

II. METHODOLOGY

A. Genetic Algorithm

The algorithm is prepared for selection of the location and rated power of BESS for two cases. Case 1 represents the situation where the control criterion of BESS is the active power (P_{crit}), Case 2 − control criterion of BESS is the reactive power (Q_{crit}). The flow chart of the GA is presented in Fig. 2.

The first step is to initiate the population, by function *Random*. The individual of the population consists of *n* gens, where *n* represent all nodes in which BESS could be installed. The selection is done from 8 types of BESS (Tab. I). The gens can adopt values (0, 1, 2, ... 8). The value 0 represents no BESS in node.

The population *P* consist of 400 individuals and 200 generations:

$$P = \begin{bmatrix} i_{1,1} & \dots & i_{1,200} \\ \dots & \dots & \dots \\ i_{400,1} & \dots & i_{400,200} \end{bmatrix} \quad (2)$$

The best individuals from each generation are selected by roulette method. Then they are subjected to crossover and mutation processes. A new generation is created in such a way. The coefficient of crossover is equal to 0.5, the coefficient of mutation is equal to 0.05 (the new population is formed mainly by crossing, the mutation only helps in exploring the function peak).

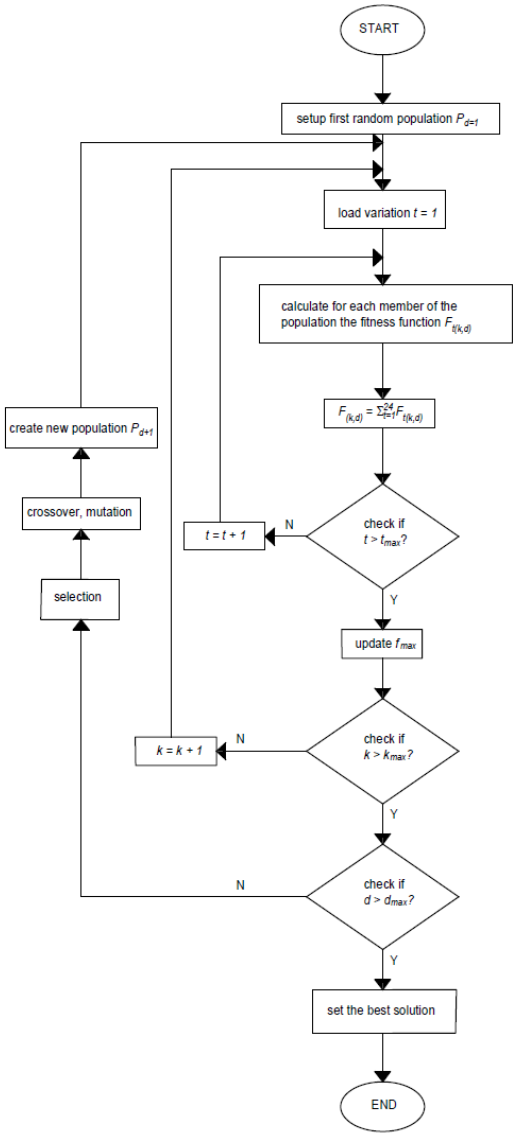


Figure 2. The Genetic Algorithm of finding the best BESS location and type.

TABLE I. BESS LIST ACCEPTED FOR SELECTION BY GENETIC ALGORITHM

No.	Active Power [kW]	Capacity [Ah]	Cost [€]
1	22.08	415	4610
2	46.22	683	7297
3	74.74	1104	10394
4	142.51	2106	19561
5	238.25	3520	32402
6	328.94	4860	41746
7	465.65	6748	60734
8	526.33	7778	68423

B. *Fitness Function*

The values of fitness function indicate the best individuals. They are calculated on the base of the entire period of the analysis. The fitness function consists of two factors. The first is responsible for minimizing voltage deviations (voltage should be in a defined range). The second is responsible for minimizing the overall power of BESS. Both factors are related to costs.

The cost related to voltage deviations is defined according to [11], [12]. The cost depends on the value of voltage. If the voltage is in an acceptable range, the cost is equal to 0. If the voltage is within the range $\pm 10\% V_n$ (V_n − nominal voltage), but does not fulfil a contractual value (V_{min}, V_{max}), the cost is described by the following formula:

$$W_{UT} = \frac{\Delta v}{10\%} A_t \cdot C_t \quad (3)$$

where Δv − is calculated as (V_{max} - V) or (V - V_{min}), A_T is the energy (in MWh) supplied in the analysed period, C_t − the price of energy (€/per unit of energy).

When the voltage in not within the range $\pm 10\% V_n$, the cost is defined by the formula:

$$W_{UT} = A_t \cdot C_t + b_{rT} \cdot t_r \quad (4)$$

where W_{UT} is the cost (in € − the average exchange rate of the euro in the 2016 year was 1 € = 4.4 PLN). In the calculation, an average value of energy price has been adopted, which was C_t = 38.9 €/MWh in 2016 in Poland, b_{rT} − is the bonus for a breach of voltage deviations limits, the value of b_{rT} is 2.27 €/h, t_r − the duration of the deviation in h (hours) [11], [12].

The cost K related to installed power of BESS is adequate to selected type of BESS and their number. The cost of BESS depending on their type is presented in Tab. I.

The fitness function consist of two components: W_{UT} which is related to the cost of over or under voltage problems and K which is related to the cost of installation of energy storage units. The fitness function is described by the formula:

$$F_{k,d} = \sum_{t=1}^{t_{max}} \left(C - \sum_{n=1}^{n_{max}} \left(W_{UT(n,t)} - \frac{K_{(n)}}{D} \right) \right) \quad (5)$$

where C is a constant value, needed for proper GA operation, D is a constant value associated with return investment time of BESS, t_{max} is the last (in the analysed period) load change in

the network, n_{max} is the number of nodes in the power network, d_{max} (see Fig. 1) is the number of the last generation.

C. Selection the Type of Control

The selection of the type of control of BESS by GA is performed two times for each grid (network). Once for the criterion of control BESS with active power − Case 1 (P_{crit}) and another one for the criterion of control BESS with reactive power − Case 2 (Q_{crit}). For the first criterion, it is important to maintain the zero energy balance at the end of the period (energy for charging needs to be equal to the energy of discharging). For the second criterion, the BESS control with maintaining constant value of voltage in the connection node. The reference value of voltage was set to V_{ref} = 0.95. In this case the energy balance needs not to be equal to zero.

III. Description of the Analysed Power Networks

The analysis was performed with regard to two power network models. Both of them are medium voltage power networks but Energy Storage Units are connected at low voltage side of MV/LV transformers. The first is proposed by CIGRE for tests and is mainly composed of cable lines. 14 medium voltage nodes and 14 low voltage nodes are situated in this grid. The second power network is a modified version of the first power network. The *R/X* ratio was changed by the modification of parameters of lines and transformers MV/LV.

The loads have two types of active power patterns (Fig. 3). Reactive power is modelled by assuming constant power factor from the following range: cos(φ) = (0.85÷1).

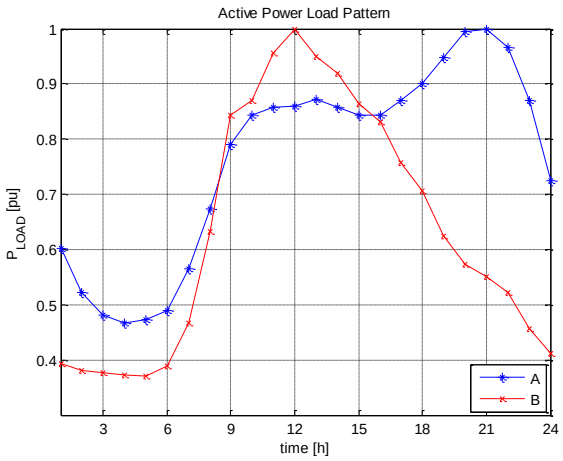


Figure 3. Two types of loads – A and B – considered in the ESU selection process.

IV. Results of Calculations

Properties of the power networks are presented by the sensitivity matrix. This matrix shows the dependence of the voltage change at each node from the injected active or reactive power at subsequent nodes. Dependencies of the sensitivity matrix are described by the formula (6) and (7):

$$\frac{\partial V_{i,j}}{\partial P_j} = \frac{V_{i,j}(\partial P_j) - V_{0\_i}}{V_{0\_i}} \quad (6)$$

$$\frac{\partial V_{i,j}}{\partial Q_j} = \frac{V_{i,j}(\partial Q_j) - V_{0\_i}}{V_{0\_i}} \quad (7)$$

where $V_{0\text{-}i}$ − the voltage in node i when no BESS is in the power network, $V_{i,j}$ − the voltage in node i, when the active or reactive power is injected in node j, ∂P_j − the injected active power (0.1 MW), ∂Q_j − the injected reactive power (0.1 Mvar).

By changing the elements of the network, it is possible to influence directly the values of the sensitivity matrix. When the *R/X* ratio is increased (Grid 1 (*R/X* <= 0.5) vs. Grid 2 (*R/X* >> 0.5)), the effect of reactive power on the voltage in the nodes is considerably lower. Fig. 4 and 5 (Grid 1) indicate that the application of active (Fig. 4) or reactive power (Fig. 5) at a node affects not only the voltage at the node in which the power is applied, but also the voltage at other nodes (many columns representing results are relatively high).

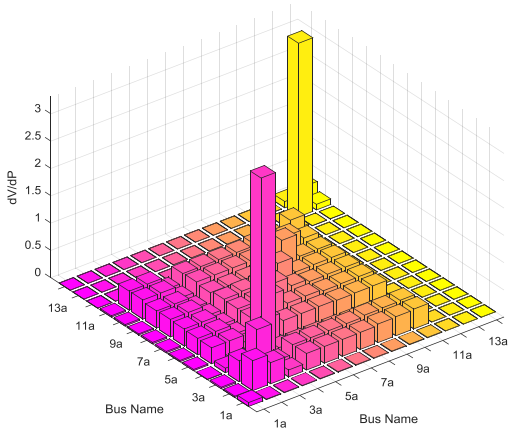


Figure 4. Sensitivity matrix of Grid 1 − $\partial V/\partial P$ in low voltage nodes.

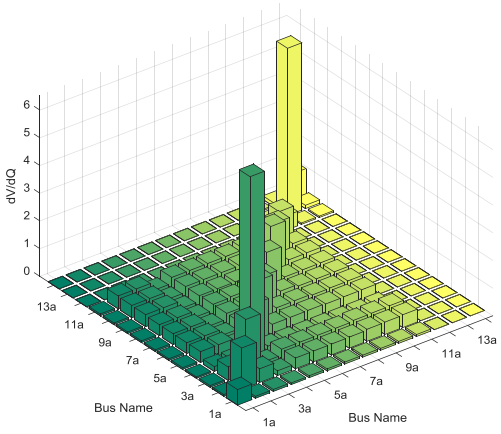


Figure 5. Sensitivity matrix of Grid 1 − $\partial V/\partial Q$ in low voltage nodes.

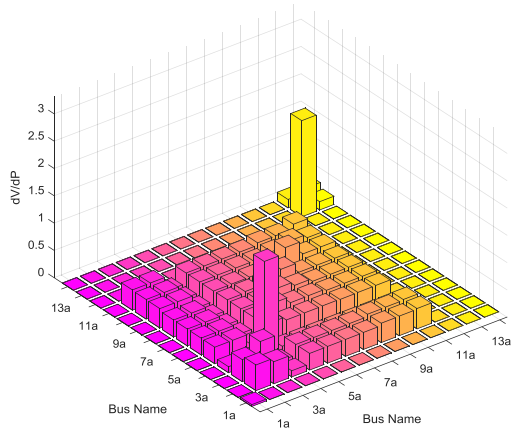


Figure 6. Sensitivity matrix of Grid 2 − $\partial V/\partial P$ in low voltage nodes.

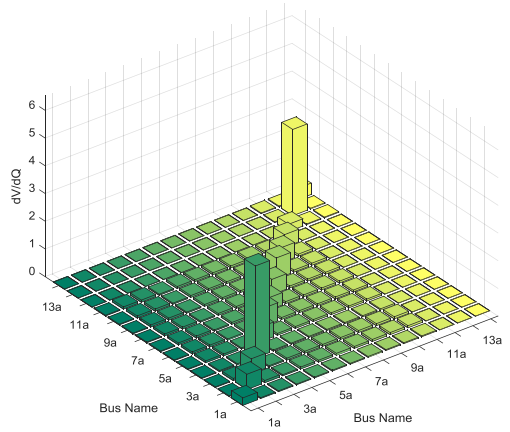


Figure 7. Sensitivity matrix of Grid 2 − $\partial V/\partial Q$ in low voltage nodes.

The graph in Fig. 7 (Grid 2) clearly indicates that application of reactive power at a given node will affect the voltage only at that node (high columns representing results are located only on the diagonal of the matrix), opposite to active power (Fig. 6 – results similar to those presented in Fig. 4 and 5).

It is better to use the criterion 1 (P_{crit}) for improving voltage in Grid 2, and in Grid 1 the results of improved voltage are comparable for the criterion 1 (P_{crit}) and the criterion 2 (Q_{crit}). This attribute was analyzed and verified by the Genetic Algorithm.

The Genetic Algorithm was used for verification of selection of the location and rated power of BESS. Fig. 8 presents the combination of BESS', which represents the best individuals from the whole population for Grid 1 and Grid 2, both for Case 1 (BESS operate with active power P_{crit}) and Case 2 (BESS operate with reactive power Q_{crit}). The sum of installed power of BESS P_{sum} for subsequent variants is equal to:

- P_{sum}*(Case 1, Grid 1)* = 2506.63 kW,
- P_{sum}*(Case 2, Grid 1)* = 2227.07 kW,
- P_{sum}*(Case 1, Grid 2)* = 1495.83 kW,
- P_{sum}*(Case 2, Grid 2)* = 2030.82 kW.

The amount of money which can be saved (because the guaranteed voltage level in the network is easy achieved) in the analysed period is:

- *F (Case 1, Grid 1)* = 3569.73 €,
- *F (Case 2, Grid 1)* = 3586.01 €,
- *F (Case 1, Grid 2)* = 1871.60 €,
- *F (Case 2, Grid 2)* = 1360.42 €.

The best solution supported by the Genetic Algorithms for every variant is presented in Fig. 8. The results show that there is a difference in proportions of delivered active and reactive power. For Grid 1 more benefits are able to get by providing reactive power to improve the voltage level with less active power of BESS (the Fitness Function F has also gained higher value for Case 2 − Q_{crit}). For Grid 2, it turned out to be better to apply $P_{crit.}$

In Fig. 9÷12 the daily variability of voltage in selected nodes is presented. Four cases were selected:

- Fig. 9 − the node where is no BESS in any case,
- Fig. 10 − the node, where for Grid 1: $P_{sum(Pcrit)} > P_{sum(Qcrit)}$, for Grid 2: $P_{sum(Pcrit)} < P_{sum(Qcrit)}$,
- Fig. 11 − the node, where for Grid 1: $P_{sum(Pcrit)} < P_{sum(Qcrit)}$, for Grid 2: $P_{sum(Pcrit)} < P_{sum(Qcrit)}$,
- Fig. 12 − the node, where for Grid 1: $P_{sum(Pcrit)} = P_{sum(Qcrit)}$, for Grid 2: $P_{sum(Pcrit)} < P_{sum(Qcrit)}$.

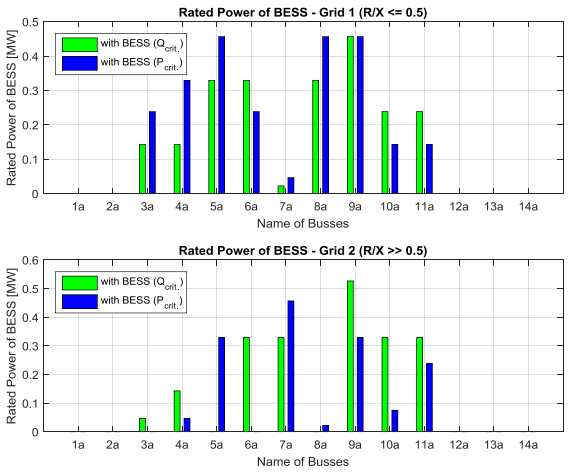


Figure 8. Rated power of BESS for Grid 1 and Grid 2.

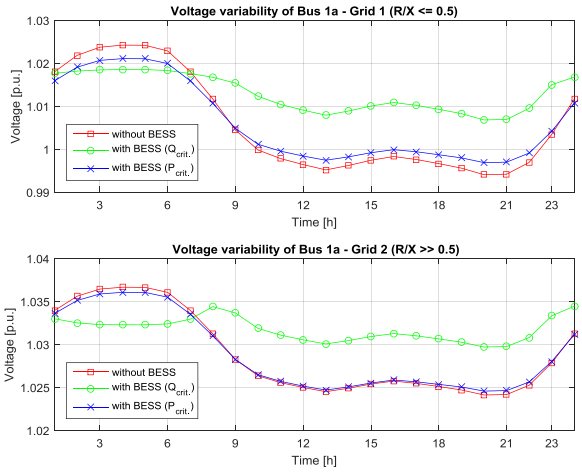


Figure 9. Daily voltage variability in node 1a for Grid 1 and Grid 2.

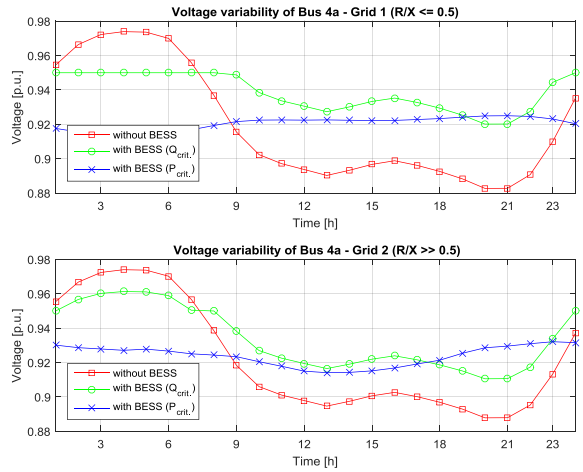


Figure 10. Daily voltage variability in node 4a for Grid 1 and Grid 2.

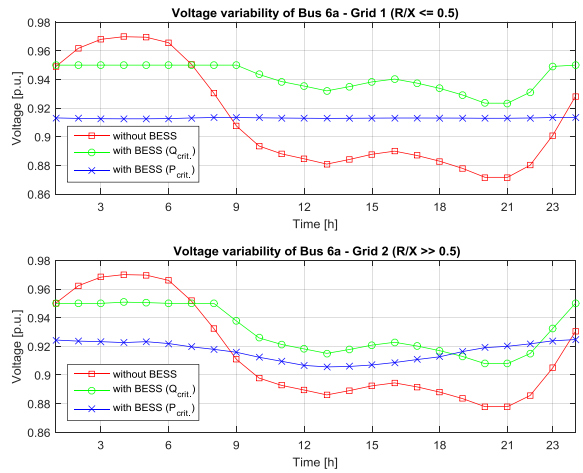


Figure 11. Daily voltage variability in node 6a for Grid 1 and Grid 2.

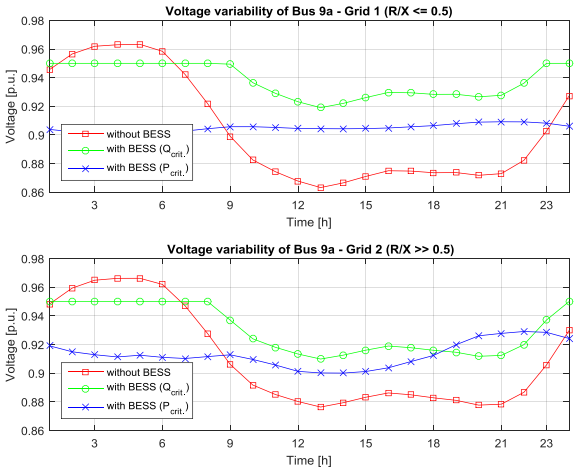


Figure 12. Daily voltage variability in node 9a for Grid 1 and Grid 2.

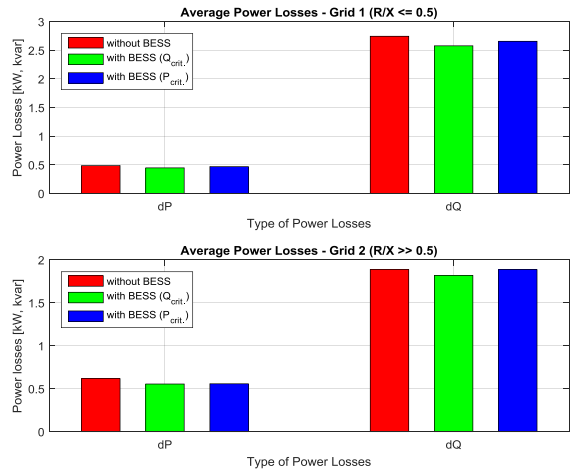


Figure 13. Power losses for Grid 1 and Grid 2 for critera P_{crit} or Q_{crit}.

The voltage lower than 0.92 or 0.9 V_n generates costs described by the formula (3) or (4), respectively. For example, in node 4a (Fig. 10), for Grid 1 $P_{sum(Pcrit)} > P_{sum(Qcrit)}$, the voltage level is better (not lower than $0.92V_n$, which results in no cost generation related with voltage level) for Q_{crit}. For Grid 2 in node 4a (Fig. 10) despite $P_{sum(Pcrit)} < P_{sum(Qcrit)}$, the voltage level got worse for Q_{crit} and improved for P_{crit}. (closer to $0.92V_n$)

The active and reactive power losses in the power network (total losses in lines and in transformers) are presented in Fig. 13. The use of BESS mainly affects the limitation of the reactive power losses. In Grid 1 the lowest active dP and reactive dQ power losses are for Q_{crit}. In Grid 2 the active power losses are almost the same for P_{crit} and Q_{crit}, the reactive power losses are lower for Q_{crit}.

V. CONCLUSIONS

The Genetic Algorithm is used for selection of the location and rated power of BESS in distribution power networks. The heuristic methods are suitable for solving problems with many

variables. They can find a solution much faster than by using traditional methods.

In order to improve the voltage level in distribution networks, it possible to deliver both active and reactive power. The *R/X* ratio of power network affects the influence of the delivered active or reactive power on the voltage level. Basing on the value of the *R/X* ratio or sensitivity matrix ($\partial V/\partial P$ and $\partial V/\partial Q$), it is possible to specify which of the two powers (active vs. reactive) should be delivered for improving the voltage level. This effect can be also achieved using Genetic Algorithm two times for different criteria of BESS (P_{crit} or Q_{crit}). It is important to pay attention to the *R/X* of a power network, because this parameter influences also location, overall power and effects of control of BESS.

Future research should include an extended algorithm for evaluation of the ability to operate BESS with the use of both the criteria P_{crit} and Q_{crit}. The proportion of introduced active and reactive power is another optimization problem, which can be solved using heuristic methods.

REFERENCES

[1] M. Zidar, P. S. Georgilakis, N. D. Hatziargyriou, T. Capuder, and D. Škrlec, "Review of energy storage allocation in power distribution networks: applications, methods and future research," IET Gener. Transm. Distrib., vol. 10, no. 3, pp. 645–652, 2016.

[2] L. Li, L. Wang, C. Sheng, W. Sun, and Y. Li, "Analysis on voltage deviation inactive distribution network and active voltage management," 2014 China International Conference on Electricity Distribution (CICED), 2014, pp. 1610–1614.

[3] M. P. Behera, P. K. Ray, and G. H. Beng, "Single-phase grid-tied photovoltaic inverter to control active and reactive power with battery energy storage device," 2016 IEEE Region 10 Conference (TENCON), 2016, pp. 1900–1904.

[4] X. Xu, M. Bishop, D. G. Oikarinen, and C. Hao, "Application and modeling of battery energy storage in power systems," CSEE J. Power Energy Syst., vol. 2, no. 3, pp. 82–90, 2016.

[5] L. Rouco and L. Sigrist, "Active and reactive power control of battery energy storage systems in weak grids," 2013 IREP Symposium Bulk Power System Dynamics and Control-IX, Optimization, Security and Control of the Emerging Power Grid, 2013, pp. 1–7.

[6] M. N. Kabir, Y. Mishra, G. Ledwich, Z. Y. Dong, and K. P. Wong, "Coordinated control of grid-connected photovoltaic reactive power and battery energy storage systems to improve the voltage profile of a residential distribution feeder," IEEE Trans. on Ind. Inform., vol. 10, no. 2, pp. 967–977, 2014.

[7] L. Chunlai, H. Wei, Z. Lili, and L. Xijun, "Study active distribution network voltage control technology based on the photovoltaic and energy storage," 2016 International Conference on Smart City and Systems Engineering (ICSCSE), 2016, pp. 445–448.

[8] R. Zamora and A. K. Srivastava, "Multi-layer architecture for voltage and frequency control in networked microgrids," IEEE Trans. on Smart Grid, 2017.

[9] S. R. Choudhury, A. Gupta, and S. Anand, "Simulation of low voltage ride through scheme for inverters connected to distribution system with high R/X ratio," 10th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG), 2016, pp. 202–207.

[10] H. Nazaripouya, Y. Wang, P. Chu, H. R. Pota, and R. Gadh, "Optimal sizing and placement of battery energy storage in distribution system based on solar size for voltage regulation," 2015 IEEE Power & Energy Society General Meeting, 2015, pp. 1–5.

[11] "Information of the President of the Energy Regulatory Office No. 45/2016 the average selling price of electricity on the market competitive in the second quarter of 2016 year."

[12] "INDUSTRY BULLETIN URE – Electricity" (in Polish: "BIULETYN BRANŻOWY URE – Energia elektryczna,") no. 187 (2213), 2016.

Implementation of Quality Principles for IT service Requirements Analyse

Stanislava Simonova
Institute of System Engineering and Informatics
University of Pardubice
Pardubice, Czech Republic
Stanislava.Simonova@upce.cz

Nikola Foltanova
Institute of System Engineering and Informatics
University of Pardubice
Pardubice, Czech Republic
Nikola.Foltanova@upce.cz

***Abstract*— An information service (IT service) in the context of the company informatics supports to execute business processes. Business processes need to be supported by targeted and relevant information services to ensure high performance. Company management perceives the information services as a necessary part of the business process and expects their error-free performance. An IT service does not represent a concrete program application but it represents the final value given to the user, i.e. user (worker) within business process. Quality IT service means that the user gets such information service that he/she needs. Generally, business process and IT service create an interconnected unit, and this whole serves customer. IT service development life cycle is supported by the sophisticated and in practice proven methods and tools; procedures of the functional and data modelling are being applied, using a structured approach or an object oriented approach. But "voice of customer" identification exist especially on the level of business process whereas their propagation to IT service definition is not so precisely defined. Article focuses on the application of the business process tools and principles for identifying and modelling of individual IT services requirements with the aim to increase a particular information service quality as a part of quality business performance.**

Keywords—IT service; software requirement; Voice of customer; IT requirement analyzes

I. INTRODUCTION

Organization identifies its business objectives, which proceeds priorities as for the current business processes as for their means of support that includes gaining and processing information (data). Business economics performs significant part during ensuring the comprehensive management of enterprise (including finance, production, logistics, human resource management, etc.) and also during promoting market relations (specific applications, electronic communication, continuous evaluation of situation on the market and so on) [1]. Company management perceives information services as a necessary part of the business process and expects their continuous performance. Therefore, the cost of IT technologies represents a significant item in company budget. Developing of internal information environment by own resources is demanding and it takes sources of the company which could be devoted to its primary activities. One of the solutions for this situation is outsourcing.

Whether a company develops own internal IT services or decides for outsourcing of IT services, it always faces a major problem, such as the initial identification of requirements – functional requirements, data requirement, parameters of IT services etc. The users' requirements should be essential for IT services development. This is an obvious assumption and, of course, there are procedures and recommendations for getting the users' requirements. We get requests from the users, but our crucial questions are - Are the identified requirements really major for the developed IT service?; Did the user identify really the essential requirements necessary for optimal IT service? Didn't the user forget about some important requirements? In other words - can we distinguish the significance of the acquired users' requirements?

II. FORMULATION OF THE PROBLEM

The degree of requirements on IT management in a company is directly proportional to economic and operational needs of the company; at the same time, it is determined by progress and abilities of information and communication technologies. The main goals for IT management are [2]:

- to ensure high functionality of information system; this means not only functions of keeping records and transactions but also analytic, functions for decision support and control functions;
- to achieve high rate of application and technological availability, i.e. security, accessibility, reliability and flexibility;
- to monitor continuously minimization of the cost compared to economic and non-economic effects.

IT management is performed within a particular framework or model which include the ITIL standards (Information Technology Infrastructure Library) [3] or COBIT (Control Objectives for Information and Related Technology) [4]. ITIL is a framework mostly for company IT management which offers guidelines, musters, diagrams and other well-proven methods for management of IT services [5]. COBIT is a framework for IT Governance, it is a strategic framework for management of IT environment with the aim to harmonize company IT management and the goals and company management. Mentioned approaches or frameworks solve

The work reported in this paper was conducted with the kind support of the University Pardubice grant No SG470019.

mostly management of IT processes in relation with business process support to fulfil company goals.

In order to define the relations of the service to business processes and other information elements, it is possible to use the SPSPR model (S, Strategy; P, Business Processes; S, ICT Services; P, ICT Processes; R, ICT Resources), which defines the responsibilities of business and ICT manager for managing the "business - company IT" relationship [6] (see Figure 1).

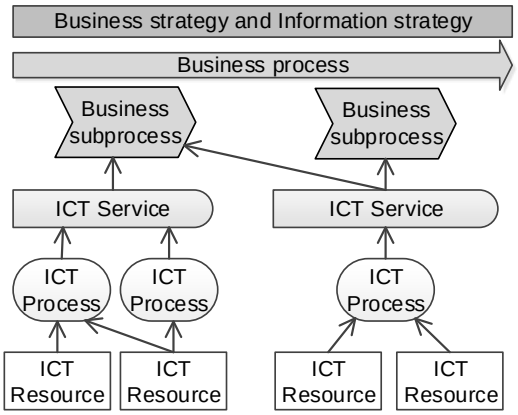


Figure 1. Model SPSPR; source: own, prepared based on [7]

Software development is a technical as well as creating process. The selection of the method of software application development management, i.e. suitable methodology selection, is crucial for project success. Methodologies can be categorized based on different criteria, however, the basic separation is to rigorous and agile methodologies [8] [9]. Rigorous methodologies use the sequence model more frequently and work with formal documentation such as diagrams, lists, tables etc. The main idea is that information system development can be described, planned, managed and measured. Rigorous approaches can be used especially at larger projects where SW processes can be described, where requirements can be defined in advance (and where there is danger of lack of communication with the sponsor/customer) and where it is necessary to plan everything well [10]. The methodologies ensure sufficient contact with the customer/sponsor and react to the need of project planning. Nowadays for example, rigorous approaches are used to support design of eGovernment services and for big data applications [11] [12]. Agile methodologies enable flexible solution adjustment. The main idea is that the only way to prove the correctness of the designed system is to develop it as fast as possible (or develop its parts), present it to the customer and adjust it to customer's feedback. Agile approach is based only on roughly stated requirements without SW process descriptions and is suitable for use in small teams which can regularly communicate with the customer/sponsor. It is suitable to select such processes and tools from the given methodologies that are easily comprehensible for the modeling and understandable for the users. Let us use diagram BPMN (Business Process Model and Notation) as an example for process activities, Activity diagram for expressing work flow, Data Flow diagram for expressing data flow relations, Use Case diagram for capturing required system functionalities, Sequence diagram to visualize data objects communication, Class diagram or Entity Relationship diagram to express the composition of data structures and relationships between them. It is obvious from this list that tools can be combined from different methodologies.

These sophisticated and in practice proven methods and procedures work with significant inputs that are the user requirements. The specification of user requirements to the system, i.e. requested functionalities specification, is one of the first steps within IT service development. The user of the IT service therefore plays an important role throughout the process and has its rights and obligations.

A software customer has the right to [13]:

- Expect analysts to speak your language.
- Expect analysts to learn about your business and your objectives for the system.
- Expect analysts to structure the requirements information you present into a software requirements specification.
- Have developers explain requirements work products.
- Expect developers to treat you with respect and to maintain a collaborative and professional attitude.
- Have analysts present ideas and alternatives both for your requirements and for implementation.
- Describe characteristics that will make the product easy and enjoyable to use.
- Be presented with opportunities to adjust your requirements to permit reuse of existing software components.
- Be given good-faith estimates of the costs, impacts, and trade-offs when you request a requirement change.
- Receive a system that meets your functional and quality needs, to the extent that those needs have been communicated to the developers and agreed upon.

A software customer has the responsibility to [13]:

- Educate analysts about your business and define jargon.
- Spend the time to provide requirements, clarify them, and iteratively flesh them out.
- Be specific and precise about the system’s requirements.
- Make timely decisions about requirements when requested to do so.
- Respect developers’ assessments of cost and feasibility.
- Set priorities for individual requirements, system features, or use cases.
- Review requirements documents and prototypes.
- Promptly communicate changes to the product’s requirements.

- Follow the development organization's defined requirements change process.
- Respect the requirements engineering processes the developers use.

Although it seems that the definition of the software requirement is simply understandable, in practice it is the opposite. The word "demand/request" is a frequent barrier between the customer as an IS user and an IT analyst. If the customer has the task of specifying requirements, he/she usually does not even know what is required. The need to obtain customer requests can be very frustrating for the customer, as he/she subconsciously perceives his role in developing the IS only in two steps. In the first step, the customer communicates his / her requirements, which usually demands vaguely formulated and objectively not measurable. In the second step, the customer expects the final solution. But just this point / process, i.e. getting the requirements, is very important. It is at this stage that the "future" problems and inaccuracies in the context of the development of the entire information system are financially and temporarily least demanding [13], [14].

Customers' requests or "Voice of Customers" and their acquisition and identification, all this is well developed and defined in the area of Business Processes. However, this does not apply to IT. Therefore, it is useful to use the characteristics and principles that work in the business process area.

III. VOICE OF THE CUSTOMER

Business process management relies on time-proven methods and norms of quality management such as Six Sigma, Kaizen, Lean [15], [16]. All of these methods focus on Voice of Customers. The "voice of the customer" is a process used to capture the requirements/feedback from the customer (internal or external) to provide the customers with the best in class service/product quality. The "voice of the customer" is the term used to describe the stated and unstated needs or requirements of the customer [17]. The Voice of customers is primarily applied in the business area, specifically in various business processes or products. This concept can be illustrated, for example, by determining the customer's voice requirements for specific products, the quality of services or the individual machines employed by company employees. The voice of the customer has a great deal of importance here, because it is based on individual innovations and the development of new products and services. And the same view is also offered in the area of the IT service, as the user who serves it is also the customer of the product, i.e. the IT service [18].

The Voice of the customer can be captured in a variety of ways: direct discussion or interviews, surveys, focus groups, customer specifications, observation, warranty data, field reports, complaint logs, etc. [17]. All business process methods recommend the use of support tools and techniques for identifying and characterizing the voice of the customer. These include, for example - Ishikawa fishbone diagram (A cause and effect diagram), Kano model, 5xWhy, CTQ (Critical to Quality), Flowchart, Mental map, etc. Critical to Quality (CTQ), or key quality criteria, are the measurable characteristics of a process or service from the customer's point of view. When using the CTQ method, it actually translates the voice of the customer into these critical characteristics. These characteristics are already applicable since they mostly define clearly the boundaries and specific limits of the information system [19], [20].

A simple example could be the following statement: The customer requests a faster information system (Voice of customers) request. This requirement is not complete; it is difficult to measure it and can not be derived from any specific information [21]. Gradual inquiry and decomposition of this requirement will then result in measurable requirements. In this case, it could be: "the system must provide feedback within 24 hours", "the processing of the loan application will take up to a maximum of 10 minutes". Also brainstorming is useful for using this tool, where basic requirements are found [20].

In practice, it is recommended to use a tree diagram to apply the CTQ method. CTQ tree is a graphical tool for transforming customer's voice to critical values where customer needs are gradually decomposed into individual measurable parameters (see Figure 2).

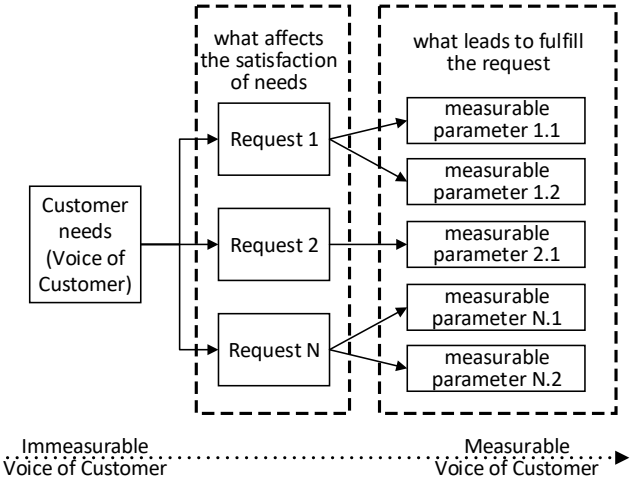


Figure 2. CTQ tree; source: own

Causes and indicator Critical to Quality are analyzed by help of the Cause and Effect diagram, definition and effects has to capture the root cause (Figure 3).

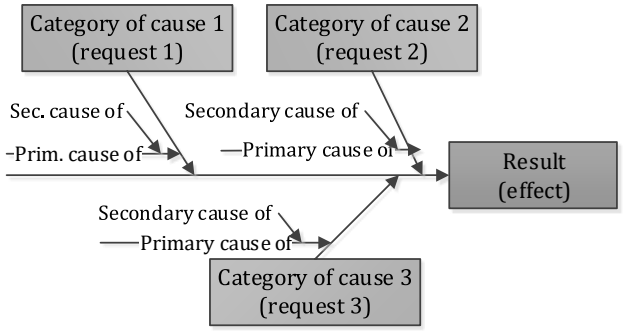


Figure 3. A Cause and Effect diagram; source: own

The Kano model provides a useful method of answering the question of whether measures for increasing product quality have an effect on customer satisfaction [22]. Of course, the product can also be an IT service. Traditional ideas about quality have often assumed that customer satisfaction was simply proportional to how functional the product or service was. This would mean that an IT service with more fulfilled functionalities is more satisfying for users, and an IT service with fewer implemented functionalities less satisfies users. But this is not true, because it depends on which functionalities are implemented and also which functionalities in the IT service are missing. The Kano Model is an insightful way of understanding and categorizing three (or five) types of Customer Requirements (or potential features) for new products and services [23]. (see Figure 4).

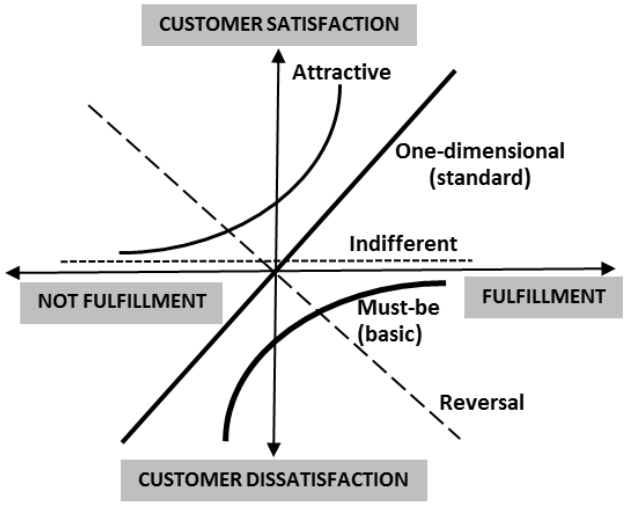


Figure 4. Categorization of requirements byKano model; source: own, prepared based on [22]

Based on the Kano model, requirements are distinguished:

- Must-be (basic) requirements: are usually not explicitly demanded by the customer, the customer considers them as a matter of course. Requirements are basic criteria of a product and fulfilling the must-be requirements will only lead to a state of "not dissatisfied".
- One-dimensional (standard) requirements: are usually explicitly demanded by the customer. The customer satisfaction is proportional to the level of fulfilment.
- Attractive requirements: are neither explicitly expressed nor expected by the customer. If they are not met, there is no feeling of dissatisfaction.
- Indifferent requirements: the customer is neither satisfied nor dissatisfied whether they are met, or not met.
- Reversal requirements: this requirement is not only not wanted by the customer but they even expect the reverse.

Such division of requirements leads to different considerations. For example, on the one hand, it is clear that the most important is meeting the Must-be requirements, but the customer/user often does not explicitly address these requirements. On the one hand, it is not useful to invest in improving Must-be requirements which are already at satisfactory level, because the overall satisfaction of users will not increase. However, one thing is necessary for such considerations - it is necessary to distinguish between the identified requirements, which are significant (must-be, one dimensional) and which are less important (attractive) and which are insignificant (indifferent) and which are even bad (reversal).

The obtained user requirements are divided into requirements types using a special pair of Kano questions. Each requirement is tested by two questions:

- Positively formulated question: How do you feel if that feature is present in the IT service?
- Negatively formulated question: How do you feel if that feature is not present in the IT service?

The customer can answer in one of five different ways: Very Satisfied (++), Satisfied (+), Neutral (0), DisSatisfied (-), or Very DisSatisfied (- -). Based on a pair of responses, the request type is then identified for each request (see Table I). The letters M-O-A-I-R indicate the type of request, such as the letter M means Must-be requirements, O means One-dimensional, A means Attractive, I means Indiffierent, R means Reversal requirements. The letter Q means Questionable result. In this case, there is a contradiction in the customers answers to the questions, the question was phased incorrectly, or the person interviewed misunderstood the question or crossed out a wrong answer by mistake.

TABLE I. CLASSIFICATION OF THE REQUIREMENTS BY THE KANO MODEL

REQUEST OF THE CUSTOMER/USER		***NEGATIVELY formulated question:*** *IT service lacks functionality. How do you feel?*				
		Very Satistied	Satiesfied	Neutral	Dis-satisfied	Very Dis-Satistied
POSITIVELY formulated quest.: *ITs functionality is implemen-tented. How do you feel?*	Very Satistied	Q	**A**	**A**	**A**	**O**
	Satiesfied	R	I	I	I	**M**
	Neutral	R	I	I	I	**M**
	Dis-satisfied	R	I	I	I	**M**
	Very Dis-Satistied	R	R	R	R	Q

Only when such a classification of requirements will make it possible to draw conclusions on:

- whether the identified requirements also include significant requirements;
- whether the identified requirements contribute to increasing user satisfaction with the service;
- whether the identified requirements are not only marginal and therefore whether it is necessary to re-identify the requirements.

IV. Implementation of Quality Principles for IT service Requirements Analyse

The tools and techniques used in the business process management for requirements classification were used within the trial IT project.

A. Preparatory phase

A new IT service should be developed in this project. A new IT service has been set up to support the selected part of business process. The process-related agenda had been dealt with outside IT by then, so it was not possible to build on any previous experience with the IT service. Twelve staff members were assigned to the project team. They were process managers, staff involved in process activities and, of course, the representative of the future IT service users.

The determination of the requirements was commenced by identifying the main interest areas. Four questions were asked by help of an opening questionnaire:

- What do you think about using this IT service?
- What problems, mistakes or shortcomings are you posing in connection with a process that will be supported by the IT service?
- What criteria do you have for a given IT service?
- What would you change to a process that will be supported by the IT service?

The identified topics were discussed within the brainstorming, and were specified and graphically presented using CTQ tree. The requests specified within the CTQ tree then were developed and detailed using the Ishikawa diagram where the root causes of the requests were searched. Functional requirements have been identified (for example: creating an input form, cost calculation, searching for availability of financial resources, etc.). The result was that twenty five basic requirements were identified - the initial desired functionalities.

B. Classification of the requirements by the Kano model

Let us to recall again the sense of the presented IT project. User requirements for the information system are identified and collected (in this case, there are 25 requests). The task is – how to distinguish the significance of the acquired requirements.

Therefore the next step of the IT project was - to evaluate the acquired requirements and classify them into the corresponding types (requirements Must-be, One-dimensional, Attractive, Indifferent, Reversal). Twenty-five future users of the service were asked to express their views on the individual requirements. There were fifty questions in total because each of the twenty-five requests was tested by a couple of questions - How do you feel if that request "is present" / "is not present" in the IT service? The user should answer in one of five different ways - Very Satisfied, Satisfied, Neutral, Dissatisfied, or Very Dissatisfied. Fifty responses were expected, more precisely twenty-five evaluation pairs of responses from each of twenty users. However, four users did not understand negatively formulated questions, so their responses were discarded from further evaluation. Therefore, responses from twenty-one users were used for the final classification of requirements. The type of request was determined on the basis of each pair of responses in accordance with the Kano model (illustration – see Table II).

TABLE II. Sample of Classification by the Kano model

REQUEST OF THE CUSTOMER/USER		***NEGATIVELY formulated question:*** *Information system lacks functionality. How do you feel?*				
		Very Satistied	Satiesfied	**NEUTRAL**	Dis-satisfied	Very Dis-Satistied
POSITIVELY formulated quest.: *IS functionality is implemen-tented. How do you feel?*	Very Satistied	Q	A	A	A	O
	SATIESFIED	R	I	**I**	I	M
	Neutral	R	I	I	I	M
	Dis-satisfied	R	I	I	I	M
	Very Dis-Satistied	R	R	R	R	Q

The evaluation of all 25 requests (P1 – P25) was arranged in the resulting table (see Table III).

Coefficients were also calculated for each tested request. Customer satisfaction coefficient (CS coefficient) state whether satisfaction can be increased by meeting a product requirement, or whether fulfilling this product requirement merely prevents the customer from being dissatisfied. The CS coefficient indicates how strongly the product can affect customer satisfaction or, in case of it "non-fulfillment", customer dissatisfaction (CN means negative coefficient of dissatisfaction) [24].

Calculation of the satisfaction coefficient (CS coefficient):

$$CS = (A + O) / (A + O + M + I) \quad (1)$$

Calculation of the negative satisfaction coefficient (CN coefficient):

$$CN = (O + M) / (A + O + M + I) * (-1) \quad (2)$$

Classification of requirements based on responses was thus determined in the first proposal (see Table III).

Explaining the symbols and values in the Table III:

- P1 - P25: are twenty five acquired users requests to the information system.
- M-O-A-I-R: indicates the type of request, such as the letter M means Must-be requirements, O means One-dimensional, A means Attractive, I means Indifferent, R means Reversal request.
- Q: means Questionable result.
- Sum 21: means that twenty-one participants answered a pair of questions for each requirement (more precisely, twenty-five users were asked to express their opinion, but four users did not understand the

negatively formulated questions, so their responses were discarded from further evaluation).

- Type: is the assignment of the resulting classification. For example, 'P2' request: 2 users evaluated the request as a 'Must-be' type, 3 users evaluated the request as an 'One-dimensional' type, 10 users evaluated the request as an 'Attractive' type and 6 users evaluated the request as an 'Indifferent' type. In this case, the prevailing evaluation is Type A-Attractive.
- CS: means satisfaction coefficient, calculation – see Eq. (1). The positive CS coefficient ranges from 0 to 1, the closer the value to 1, the higher the influence on customer satisfaction. A positive CS coefficient which approaches 0 signifies that there is very little influence.
- CN: means negative satisfaction coefficient, calculation – see Eq. (2). The negative CN-coefficient ranges from 0 to -1. If the CN coefficient approaches minus 1, the influence on customer satisfaction is especially strong if the analyzed unit feature is not fulfilled. A value of 0 signifies that this feature does not cause dissatisfaction if it is not met.

TABLE III. THE RESULTING TABLE OF REQUESTS CLASSIFICATION

Request	M	O	A	I	R	Q	Sum	Type	CS	CN	CS	CN
P1	9	7	3		2		21	M	0,5	-0,8		
P2	2	3	10	6			21	A	0,6	-0,2		
P3			6	8	7		21	I	0,4	0,0		
P4		4	12	5			21	A	0,8	-0,2		
P5		8	9	4			21	A	0,8	-0,4		
P6	3		8	10			21	I	0,4	-0,1		
P7		9	10	2			21	A	0,9	-0,4		
P8	2	9	7	3			21	O	0,8	-0,5		
P9		8	4	9			21	I	0,6	-0,4		
P10		1	9	11			21	I	0,5	0,0		
P11	5	4	5	7			21	I	0,4	-0,4		
P12	1	11	9				21	O	1,0	-0,6		
P13	8	10		3			21	O	0,5	-0,9		
P14		8	6	7			21	O	0,7	-0,4		
P15		13	8				21	O	1,0	-0,6		
P16			14	7			21	A	0,7	0,0		
P17		4	10	7			21	A	0,7	-0,2		
P18		5	6	5	5		21	A	0,7	-0,3		
P19	5		12	4			21	A	0,6	-0,2		
P20		2	4	6		9	21	Q	0,5	-0,2		
P21	14	1	2	4			21	M	0,1	-0,7		
P22	6	15					21	O	0,7	-1,0		
P23	4	2	15				21	A	0,8	-0,3		
P24		2	3	6	10		21	R	0,5	-0,2		
P25	4	10	5	2			21	O	0,7	-0,7		
	12%	26%	34%	22%	5%	2%	100%					

Classification of requirements P1 - P25:	Sum	25
Must-be (basic) requir.:	M	2
One-dimensional (standard) requir.:	O	7
Attractive requir.:	A	9
Indifferent requir.:	I	5
Reversal requir.:	R	1
Questionable result:	Q	1

Classification of requirements based on CS and CN calculation was also presented in graphical visualization (see Table III and see Figure 5).

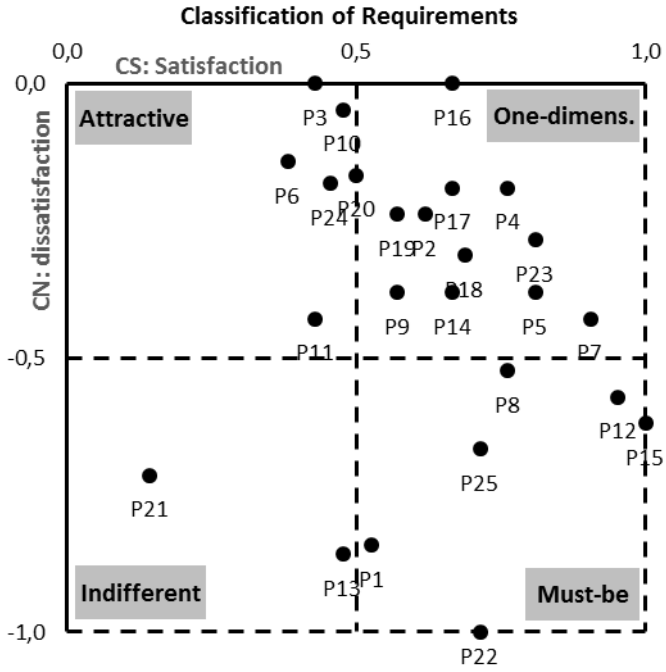


Figure 5. Classification based on CS and CN calculations; source: own

C. Evaluation of outputs

The third phase of the project was the interpretation of the results. The easiest procedure is evaluation and interpretation according to the frequency of answers. Selecting the classification that would have the greatest impact on the product (ordering: M>O>A>I). From this point of view, it is not good in this IT project that only two Must-be requirements have been identified (P1 and P21). The Must-be requirements are basic criteria of a product (IT service), the customer/user regards the Must-be requirements as prerequisites, and therefore does not explicitly demand them. Only two identified Must-be requirements means, that it is necessary to re-identify the requirements. Most of the requirements were classified as Attractive requirements (9 requirements, it is 36%). Discovering and fulfilling attractive requirements creates a wide range of possibilities for user satisfaction, but attractive requirements are not basic and essential, and do not contribute to IT service quality.

It was also clear from the results that some inquiries about the requirements were not comprehensible to the users. The requirement P20 was classifies as Questionable result. Questionable scores signify that the question was phased incorrectly or that the person interviewed misunderstood the question or crossed out a wrong answer by mistake. It is necessary to formulate the question in a different way so as to make it more comprehensible to respondents.

The classification of some requirements has been ambiguous P24 je reversal. For example, the answers to requirements P3 or P18 do not give a convincing

categorization. Both requirements even include the Reversal category, when a priori judgment of functional and dysfunctional was the reverse what the user feels and this feature is not only not wanted by the customer but they even expect the reverse.

The other procedure of evaluation is an interpretation according CS and CN coefficients of satisfaction. The positive CS coefficient ranges from 0 to 1, the closer the value to 1, the higher the influence on customer satisfaction. A positive CS coefficient which approaches 0 signifies that there is very little influence. The negative CN-coefficient ranges from 0 to -1. If the CN coefficient approaches -1, the influence on customer satisfaction is especially strong if the analyzed product feature is not fulfilled. A value of 0 signifies that this feature does not cause dissatisfaction if it is not met. The greatest positive coefficient, i.e. the greatest impact on user satisfaction, have the requirements P12 (CS=1), P7 (CS=0,9), and P4+P5+P8+ P23 (CS=0,8). The greatest negative coefficient, i.e. the greatest impact on user dissatisfaction, have the requirements P22 (CN=-1) and P13 (CS=-0,9). The requirements P11 or P25 have equally large positive and negative coefficients. That is, the user is neither satisfied nor dissatisfied whether the feature is dysfunctional or fully functional within IT service. However, it is important to note that higher user satisfaction does not necessarily imply a higher quality of IT service, as, for example, basic Must-be requirements have no effect on user satisfaction, but they are necessary for IT service quality

Let us remind that the IT project was trial and its aim was to test the accuracy or inaccuracy of the proposed software requirements. However, the conclusion of this project was that the user requirements were set not ideally and that it is therefore necessary again to carry out the whole process of the requirements identification for IT service.

V. CONCLUSION

IT service development is based on sophisticated procedures using different methods. All methods are based on the fact that user software requirements are available. Thus, user requirements are a significant input (starting point), and then the methods are followed by development of analytical and design models, leading to the final implementation. Tools for creating analytical and design models are known and sophisticated (for example use case, class model, ERD, RMD etc.). But the question is – do these tools really model the proper entry requirements for the system / IT service? A simple idea may be that the user will be more satisfied, the more his requirements will be implemented. But this simple idea is not true. Customers of IT services, that is, IT service users, unknowingly separate the content of the IT service and the satisfaction with the IT service. Some requirements, sometimes marginal requirements have a great impact on overall user satisfaction. Conversely, the requirements important for quality content of the IT service are taken by the user for granted and their fulfilment will not increase his satisfaction. But if requirement not fulfilled (or even undetected, not requested, not spoken), the customer will be extremely dissatisfied.

Such customer behavior is well known and described in the business area. So it is useful to use the characteristics and principles that work in the business process area, it means the principles of Voice of customers. The business process methods recommend the use support tools and techniques for identifying and characterizing the Voice of the customer, such as CTQ, Ishikawa cause and effect diagram, Kano model. The purpose of the described IT project was – an implementation of these tools for IT service requirements analyze. Twenty-five requirements were identified for the planned IT service. The goal of the IT project was to analyze the requirements, i.e. whether it is a requirement with an impact on the quality of the content and also whether it is a requirement with an impact on user satisfaction.

Requirements were analyzed and classified, first in terms of the type of request and its relation to the quality of the content of the IT service, then in terms of the impact of the user satisfaction requirement. However, the conclusion of this IT project was that the user requirements were set not ideally and that it is therefore necessary again to carry out the whole process of the requirements identification for IT service. The purpose of a Kano method was - to better understand the characteristics of customer´s requirements. They do not provide exact answer as to which features must be included in the product or which requirements need not be fully satisfied. The responses can be seen as a base for other discussion and requirements analyzes.

ACKNOWLEDGMENT

The work reported in this paper was conducted with the kind support of the University Pardubice grant No SG470019.

REFERENCES

[1] O. Sprync. et al., “Proactive IT / IS Monitoring for Business Continuity Planning”, in E+M Ekonomie a Management, vol. XIV, Issue 3, 2011, p. 57-65.

[2] L. Gála, Z. Šedivá and J. Pour, P. Podniková informatika. Praha: Grada Publishing, 2015.

[3] A. Cartlidge, S. Ashley, C. Rudd, I. Macfarlane, J. Windebank and S. Rance. Úvodní přehled ITIL V3. Praha: Hewlett-Packard, 2007. 56 s. ISBN 0-95551245-8-1.

[4] Isaca. COBIT 5 Framework – A Business Framework for the Governance and Management of Enterprise IT. USA: IT Governance Institute, 2012.

[5] ITIL - IT Infrastructure Library [online]. [cit. 2017-03-03]. Available from WWW < https://www.axelos.com/best-practice-solutions/itil/what-is-itil >.

[6] T .Bruckner, T., J. Voříšek, A. Buchalcevová, I. Stanovská, D. Chlapek and V. Řepa, Tvorba informačních systémů. Praha: Grada, 2012.

[7] S. Simonova, “Identification of IT-Service Metrics for a Business Process when Planning a Transition to Outsourcing”, In International Conference on Information and Digital Technologies (IDT), 2016, pp. 274 - 279, 5-7 July 2016 DOI: 10.1109/DT.2016.7557186.

[8] T. Teorey, S. Lightstone, T. Nadeau, and H. V. Jagadish. Database Modeling and Design. Elsevier, 2011. 304 p. ISBN 978-0-12-382020-4.

[9] S. Simonova. Identification Of Data Content Based On Measurement Of Quality Of Performance. In E+M Ekonomie a Management, 2012, XV, Issue 1, p. 128-138.

[10] S. Simonova and A. Hudec, “Enterprise content management based on identified requirements”, In International Conference on Information and Digital Technologies (IDT), 2015, pp. 324 - 329, 7-9 July 2015 DOI: 10.1109/DT.2015.7222991.

[11] H. Kopackova, “*The issue of measuring e-government success in context of the Initiative 202020”*. In Proceedings of the 21th International

Conference Current Trends in Public Sector Research. Brno: Masarykova univerzita, 2017. pp. 41-49. ISBN 978-80-210-8448-3.

[12] M. Lnenicka and J. Komarkova, "*The Impact of Cloud Computing and Open (Big) Data on the Enterprise Architecture Framework*", In International-Business-Information-Management-Association Conference, 2015, pp. 1679-1683, 11-12 November 2015.

[13] K. E. Wiegers. Požadavky na software. Brno: Computer Press, 2008.

[14] K. Schwalbe. Information Technology Project Management. Course Technology, 2015.

[15] I. Massaki. Gemba Kaizen: A Commonsense Approach to a Continuous Improvement Strategy. McGraw Hill Professional, 2012.

[16] A. Topfer. Six sigma Koncepce a příklady pro řízení bez chyb. Brno: Computer Press, 2008.

[17] iSixSigma: Voice Of the Customer [online]. [cit. 2017-04-04]. Available from WWW <https://www.isixsigma.com/ dictionary/voice-of-the-customer-voc/ >.

[18] A. Svozilová. Zlepšování podnikových procesů. Praha: Grada, 2011.

[19] Critical to Quality (CTQ) Trees. Mind Tools [online]. [cit. 2017-03-22]. Available from WWW:< https://www.mindtools.com/ pages/article/ctq-trees.htm>

[20] S. Simonova. Procesní řízení. Pardubice: Univerzita Pardubice, 2014. ISBN 978-80-7395-766-7.

[21] N. Motycakova. Identifikace a modelování požadavků na informační systém. Pardubice: Univerzita Pardubice, 2015.

[22] G. J. Goddard, G. Raab, R. Ajami and V. B. Gargeya. Customer Relationship Management: A Global Perspective. USA: Gower Pubishing Company, 2008.

[23] What is the Kano Model? [online]. [cit. 2017-03-22]. Available from WWW:< http://www.kanomodel.com/>

[24] E. O. C. Mkpojiogu and N. L. Hashim, Understanding the relationship between Kano model's customer satisfaction scores and self-stated requirements importance, in SpringerPlus 2016 5:197, doi: 10.1186/s40064-016-1860-y.

Mixture Failure Rate: A study based on cross-entropy and MCMC method

Tien Thanh Thach
Department of Applied Mathematics
VSB - Technical University of Ostrava
The Czech Republic

Radim Bris
Department of Applied Mathematics
VSB - Technical University of Ostrava
The Czech Republic

Frank P. A. Coolen
Department of Mathematical Sciences
Durham University
Durham, UK

***Abstract*—In this paper, the parameters and reliability characteristics of the mixture of the failure time distribution are estimated based on a complete sample using both Markov chain Monte Carlo (MCMC) method and maximum likelihood estimation via cross-entropy (CE) algorithm. While maximum likelihood estimation is the most frequently used method for parameter estimation, MCMC has recently emerged as a good alternative. The most popular MCMC method, called the Metropolis-Hastings algorithm, is used to provide a complete analysis of the concerned posterior distribution. A simulation study is provided to compare MCMC with CE, and differences between the estimates obtained by the two approaches are evaluated.**

***Index Terms*—Failure time distribution, failure rate, Markov chain Monte Carlo, Metropolis-Hastings algorithm, maximum likelihood estimation, Cross-Entropy algorithm**

I. Introduction

Engineering systems, while in operation, are always subject to environmental stresses and shocks which may or may not alter the failure rate function of the system. Suppose p is the unknown probability that the system is able to bear these stresses and its failure model remains unaffected, and q is the probability of the complementary event. In such situations, a failure distribution is generally used to describe mathematically the failure rate on the system. To some extent, the solution to the proposed problem is attempted through the mixture of distributions ([1], [2], [3]). However, in this regard, we are faced with two problems. Firstly, there are many physical causes that individually or collectively cause the failure of the system or device.

At present, it is not possible to differentiate between these physical causes and mathematically account for all of them, and, therefore, the choice of a failure distribution becomes difficult. Secondly, even if a goodness of fit technique is applied to actual observations of time to failure, we face a problem arising due to the non-symmetric nature of the life-time distributions whose behaviour is quite different at the tails where actual observations are sparse in view of the limited sample size ([2]). Obviously, the best one can do is to look out for a concept which is useful for differentiating between different life-time distributions. Failure rate is one such concept in the literature on reliability. After analyzing such physical considerations of the system, we can formulate a mixture of failure rate functions which, in turn, provide the failure time distributions. In view of the above, and due to continuous stresses and shocks on the system, let us suppose that the failure rate function of a system remains unaltered with probability p, and it undergoes a change with probability q. Let the failure rate function of the system in these two situations be in either of the following two states ([5]):

1. State 1

Initially it experiences a constant failure rate model and this model may (or may not) change with probability $q(p = 1-q)$.

2. State 2

If the stresses and shocks alter the failure rate model of the system with probability q, then it experiences a wear-out failure model. In comparison with [5], this study brings distinctive generalization of the state by implementation of new parameters, which enables to take into account also more general Weibull model.

In probability theory and statistics, the Weibull distribution is a continuous probability distribution, which is named after the Waloddi Weibull. This is the most commonly used distribution to model times until failure and provides a good description of many types of lifetimes ([1], [12]). The Weibull distribution has two parameters, a shape parameter β and a scale parameter. Only shape parameters $\beta > 1$ correspond to an increasing failure rate, implying that ageing processes can be intensively studied. We will therefore not consider cases with $\beta \leq 1$. Recent studies that adopt a Weibull lifetime distribution include [13], [14], [15].

As a result of flexibility in time-to-failure of a very widespread diversity to versatile mechanisms, the two-parameter Weibull distribution has been recently used quite extensively in reliability and survival analysis particularly when the data are not censored. Much of the attractiveness of the Weibull distribution is due to the wide variety of shapes which can assume by altering its parameters.

Using such a failure rate pattern, the characterization of life-time distribution in the corresponding situation is given. Various inferential properties of this life-time distribution along with the estimation of parameters and reliability characteristics is the subject matter of the present study. Since the estimates based on the operational data can be updated by incorporating past environmental experiences on the random variations in the life-time parameters ([4]), therefore, the Bayesian analysis

of the parameters and other reliability characteristics is also given.

The remainder of this article is organized as follows. Section II introduces the intended mixture of failure rate model including basic characteristics of the corresponding life-time distribution as well. Section III brings maximum likelihood estimators for three unknown parameters of the model using cross-entropy algorithm and Bayes estimation using Markov chain Monte Carlo method. Section IV shows short illustration how estimation procedures work. Section V reports simulation study on special selected situations where both MCMC and CE are confronted.

II. Characteristics of the life-time distribution

Notations:
Let

T:	the random variable denoting life-time of the system.
$h(t)$:	the failure rate function.
$f(t)$:	the probability density function (p.d.f.) of T.
$F(t)$:	the cumulative distribution function of T.
$R(t) = \mathbb{P}(T > t)$:	the reliability/survival function.
$\mathbb{E}(T) = \int_0^\infty R(t)dt$:	mean time to failure (MTTF).

A. The mixture of failure rates model

Let

p:	the probability of the event A, that the system is able to bear the stresses and shocks and its failure pattern remains unaltered.
$q = 1 - p$:	the probability of the complementary event A^c.

Further, let, the mixture failure rate function be

$$h(t) = p\lambda_1 + (1-p)\lambda_2 t^k, \quad \lambda_1, \lambda_2, k, t > 0, 0 < p < 1 \quad (1)$$

for

1) $p = 1$; represents the failure rate of an exponential distribution.
2) $k = 1$ and $p = 0$; represents the failure rate of the Rayleigh distribution or Weibull distribution with shape parameter 2.
   Note: In our context $\beta = k + 1$
3) $k = 1$; represents the linear ageing process.
4) $0 < k < 1$; represents the concave ageing process.
5) $k > 1$; represents the convex ageing process.

In Weibull reliability analysis it is frequently the case that the value of the shape parameter is known ([4]). For example, the Raleigh distribution is obtained when $k = 1$. The earliest references to Bayesian estimation of the unknown scale parameter are in [6]. Since that time this case has been considered by numerous authors, see [5], [7], [8], [9], [10], [11]. This study is free continuation and generalization of the research originally introduced by [5].

B. Characteristics of the life-time distribution

Using the well-known relationship between p.d.f. and failure rate function

$$f(t) = h(t) \exp\left\{-\int_0^t h(x)dx\right\}$$

and in view of (1), the p.d.f. of the life-time T is

$$f(t) = \begin{cases} h(t) \exp\left\{-\left(p\lambda_1 t + \frac{\lambda_2(1-p)}{k+1} t^{k+1}\right)\right\}, & t > 0 \\ 0, \qquad \text{otherwise.} \end{cases} \quad (2)$$

The reliability function is

$$R(t) = \exp\left\{-\left(p\lambda_1 t + \frac{\lambda_2(1-p)}{k+1} t^{k+1}\right)\right\}, \quad t > 0.$$

The MTTF is given by

$$\begin{aligned} MTTF = \mathbb{E}(T) &= \int_0^\infty R(t)dt \\ &= \int_0^\infty \exp\left\{-\left(p\lambda_1 t + \frac{\lambda_2(1-p)}{k+1} t^{k+1}\right)\right\} dt \end{aligned} \quad (3)$$

This integral can be obtained by using some suitable numerical methods.

III. Estimation of parameters and reliability characteristics

Let $D : t_1, ..., t_n$ be the random failure times of n items under test whose failure time distribution is as given in (2) and $\boldsymbol{\theta} = (p, \lambda_1, \lambda_2)$. Then the likelihood function is

$$\begin{aligned} L(D|\boldsymbol{\theta}) = &\left[\prod_{i=1}^n \left(p\lambda_1 + (1-p)\lambda_2 t_i^k\right)\right] \\ &\times \exp\left\{-\sum_{i=1}^n \left(p\lambda_1 t_i + \frac{(1-p)\lambda_2}{k+1} t_i^{k+1}\right)\right\}. \end{aligned} \quad (4)$$

A. The Cross-Entropy Method for Continuous Multi-Extremal Optimization

The main idea for the CE method for optimization can be stated as follows: Suppose we wish to maximize some "performance" function $S(\mathbf{x})$ over all elements/states $\mathbf{x}$ in some set $\mathcal{X} \subset \mathbb{R}^n$. Let us denote the maximum by γ^*, thus

$$S(\mathbf{x}^*) = \gamma^* = \max_{\mathbf{x} \in \mathcal{X}} S(\mathbf{x}) \quad (5)$$

To proceed with CE, we first randomize our deterministic problem by defining a family of pdfs $\{f(\cdot; \mathbf{v}), \mathbf{v} \in \mathcal{V}\}$ on the set $\mathcal{X}$. Next, we associate with (5) the estimation of

$$\ell(\gamma) = \mathbb{P}_u\left(S(\mathbf{X}) \geq \gamma\right) = \mathbb{E}_u I_{\{S(\mathbf{X}) \geq \gamma\}} \quad (6)$$

the so-called associated stochastic problem (ASP). Here, $\mathbf{X}$ is a random vector with pdf $f(\cdot; \mathbf{u})$, for some $\mathbf{u} \in \mathcal{V}$ (for example, $\mathbf{X}$ could be a normal random vector) and γ is a known or unknown parameter. Note that there are in fact two possible estimation problems associated with (6). For a given γ we can estimate ℓ, or alternatively, for a given ℓ we can estimate γ, the root of (6). Let us consider the problem of estimating ℓ for a certain γ close to γ^*. Then, typically

Algorithm Generic CE Algorithm for Optimization

1: Choose some $\hat{\mathbf{v}}_0$. Set $t = 1$.
2: Generate a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the density $f(\cdot; \hat{\mathbf{v}}_{t-1})$ and compute the sample $(1-\varrho)$-quantile $\hat{\gamma}_t$ of the performances according to Eq. (7).
3: Use the same sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and solve the stochastic program (9). Denote the solution by $\tilde{\mathbf{v}}_t$.
4: Apply (10) to smooth out the vector $\tilde{\mathbf{v}}_t$.
5: Repeat steps 2-4 until a pre-specified stopping criterion is met.

$\{S(\mathbf{X}) \geq \gamma\}$ is a rare event, and estimation of ℓ is a non-trivial problem. The CE method solves this efficiently by making adaptive changes to the probability density function according to the Kullback-Leibler CE, thus creating a sequence $f(\cdot; \mathbf{u}), f(\cdot; \mathbf{v}_1), f(\cdot; \mathbf{v}_2), \ldots$ of pdfs that are "steered" in the direction of the theoretically optimal density $f(\cdot; \mathbf{v}^*)$ corresponding to the degenerate density at an optimal point. In fact, the CE method generates a sequence of tuples $\{(\gamma_t, \mathbf{v}_t)\}$, which converges quickly to a small neighborhood of the optimal tuple $(\gamma^*, \mathbf{v}^*)$. More specifically, we initialize by setting $\mathbf{v}_0 = \mathbf{u}$, choosing a not very small quantity ϱ, say $\varrho = 10^{-2}$, and then we proceed as follows:

1) **Adaptive updating of** γ_t. For a fixed $\mathbf{v}_{t-1}$, let γ_t be the $(1-\varrho)$-quantile of $S(\mathbf{X})$ under $\mathbf{v}_{t-1}$. That is, γ_t satisfies

$$\mathbb{P}_{\mathbf{v}_{t-1}}(S(\mathbf{X}) \geq \gamma_t) \geq \varrho,$$
$$\mathbb{P}_{\mathbf{v}_{t-1}}(S(\mathbf{X}) \leq \gamma_t) \geq 1 - \varrho,$$

where $\mathbf{X} \sim f(\cdot; \mathbf{v}_{t-1})$.
A simple estimator of γ_t, denoted $\hat{\gamma}_t$, can be obtained by drawing a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from $f(\cdot; \mathbf{v}_{t-1})$ and evaluating the sample $(1-\varrho)$-quantile of the performances as

$$\hat{\gamma}_t = S_{\lceil (1-\varrho)N \rceil}. \tag{7}$$

2) **Adaptive updating of** $\mathbf{v}_t$. For fixed γ_t and $\mathbf{v}_{t-1}$, derive $\mathbf{v}_t$ from the solution of the program

$$\max_{\mathbf{v}} D(\mathbf{v}) = \max_{\mathbf{v}} \mathbb{E}_{\mathbf{v}_{t-1}} I_{\{S(\mathbf{X}) \geq \gamma_t\}} \ln f(\mathbf{X}; \mathbf{v}). \tag{8}$$

The stochastic counterpart of (8) is as follows: for fixed $\hat{\gamma}_t$ and $\hat{\mathbf{v}}_{t-1}$ (the estimate of $\mathbf{v}_{t-1}$), derive $\hat{\mathbf{v}}_t$ from the following program

$$\max_{\mathbf{v}} \hat{D}(\mathbf{v}) = \max_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \hat{\gamma}_t\}} \ln f(\mathbf{X}_i; \mathbf{v}). \tag{9}$$

Instead of updating the parameter vector $\mathbf{v}$ directly via the solution of (9) we use the following smoothed version

$$\hat{\mathbf{v}}_t = \alpha \tilde{\mathbf{v}}_t + (1-\alpha)\hat{\mathbf{v}}_{t-1}, \quad i = 1, \ldots, n, \tag{10}$$

where $\tilde{\mathbf{v}}_t$ is the parameter vector obtained from the solution of (9), and α is called the smoothing parameter, with $0.7 < \alpha \leq 1$.

Using normal updating, the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are sample from an n-dimensional multivariate normal distribution with independent components, $\mathcal{N}(\hat{\boldsymbol{\mu}}_{t-1}, \hat{\boldsymbol{\sigma}}^2_{t-1})$. While applying CE algorithm, the mean vector $\hat{\boldsymbol{\mu}}_t$ should converge to $\mathbf{x}^*$ and the vector of standard deviations $\hat{\boldsymbol{\sigma}}_t$ to the zero vector. In short, we should obtain a degenerated pdf with all mass concentrated in the vicinity of the point $\mathbf{x}^*$. More detail for such explanation can be found in [18] and [19].

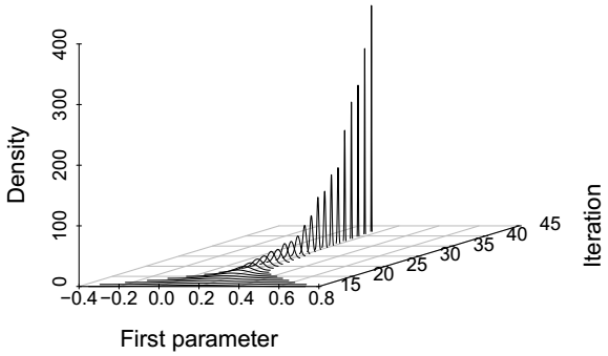


Fig. 1. The evolution of the sampling pdf for the first parameter

B. Maximum likelihood estimation

In our study, we maximize the likelihood function (4) using CE algorithm and obtain the maximizer $\hat{\boldsymbol{\theta}} = (\hat{p}, \hat{\lambda}_1, \hat{\lambda}_2)$. By using the invariance property of MLE's,

1) The MLE for $R(t)$, say $\hat{R}(t)$, will be

$$\hat{R}(t) = \exp\left\{ -\left(\hat{p}\hat{\lambda}_1 t + \frac{(1-\hat{p})\hat{\lambda}_2}{k+1} t^{k+1} \right) \right\}.$$

2) The MLE for $h(t)$, say $\hat{h}(t)$, will be

$$\hat{h}(t) = \left(\hat{p}\hat{\lambda}_1 + (1-\hat{p})\hat{\lambda}_2 t^k \right).$$

3) The MLE for MTTF will be

$$\hat{MTTF} = MTTF(\hat{p}, \hat{\lambda}_1, \hat{\lambda}_2),$$

which can be obtained by installing into formula (3) and integrating.

C. The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is the most popular MCMC method. The basic problem is that it provides a method for sampling from some generic distribution $p(x)$, say target distribution. The idea is that in many cases, we know how to write out the equation for the target distribution $p(x)$, but we don't know how to generate a random number from this target distribution. This is the situation where MCMC is very useful. In fact, for the Metropolis-Hastings algorithm we don't even need to know how to calculate $p(x)$ completely.

The basic idea behind MCMC is to define a Markov chain over possible x values, in such a way that the stationary distribution of Markov chain is in fact $p(x)$. That is, what we are going to do is to use a Markov chain to generate a sequence of x values, denoted $(x_0, x_1, x_2, \ldots)$, in such a way that as $n \to \infty$, we can guarantee that $x_n \sim p(x)$. There are many different ways of setting up a Markov chain that has this property, one of which is the Metropolis-Hastings algorithm.

Here is how Metropolis-Hastings algorithm works. Suppose that the current state of the Markov chain is x_n, and we want to generate x_{n+1}. In the Metropolis-Hastings algorithm, the generation of x_{n+1} is a two-stage process. The first stage is to generate a candidate, which we will denote x^*. The value of x^* is generated from the proposal distribution, denoted $q(x^*|x_n)$, which depends on the current state of the Markov chain, x_n. There is a few minor technical constraints on what we can use as a proposal distribution.

The second stage is the accept-reject step. Firstly, what we need to do is calculate the acceptance probability $A(x_n \rightarrow x^*)$, which is given by:

$$A(x_n \rightarrow x^*) = min\left(1, \frac{p(x^*)}{p(x_n)} \times \frac{q(x_n|x^*)}{q(x^*|x_n)}\right)$$

There are two things to pay attention to here. Firstly, notice that the ratio $\frac{p(x^*)}{p(x_n)}$ does not depend on the normalizing constant for the target distribution $p(x)$. The second thing to pay attention to is the behavior of the other term, $\frac{q(x_n|x^*)}{q(x^*|x_n)}$. What this term does is correct for any biases that the proposal distribution might induce. In this expression, the denominator $q(x^*|x_n)$ describes the probability of generating a x^* as the candidate given that the current state is x_n (i.e., what actually happened), whereas the numerator describes the probability that the "opposite" event would have occurred: that is, if the current state had actually been x^*, what is the probability that you would have generated x_n as the candidate value? If the proposal distribution is symmetric, then these two probabilities will turn out to be equal, $\frac{q(x_n|x^*)}{q(x^*|x_n)} = 1$. This special case of the Metropolis-Hastings algorithm is called the Metropolis algorithm.

Having proposed the candidate x^* and calculated the acceptance probability, $A(x_n \rightarrow x^*)$, we now either decide to "accept" the candidate (in which case we set $x_{n+1} = x^*$) or we decide to "reject" the candidate (in which case we set $x_{n+1} = x_n$). To make this decision, we generate a (uniformly distributed) random number between 0 and 1, denoted u. Then:

$$x_{n+1} = \begin{cases} x^* & \text{if } u \leq A(x_n \rightarrow x^*) \\ x_n & \text{if } u > A(x_n \rightarrow x^*) \end{cases}$$

More detail for such explanation can be found in [17].

D. Bayesian estimation

For our mixture model, the Bayesian model is constructed by specifying a prior distribution for $\boldsymbol{\theta} = (p, \lambda_1, \lambda_2)$, and then multiplying with the likelihood function to obtain the posterior distribution. Given a set of data $D : t_1, ..., t_n$, the likelihood function is

$$L(D|\boldsymbol{\theta}) = \left[\prod_{i=1}^{n} \left(p\lambda_1 + (1-p)\lambda_2 t_i^k\right)\right] \times \exp\left\{-\sum_{i=1}^{n}\left(p\lambda_1 t_i + \frac{(1-p)\lambda_2}{k+1} t_i^{k+1}\right)\right\}.$$

Denote the prior distribution of $\boldsymbol{\theta}$ as $\pi(\boldsymbol{\theta})$, the posterior distribution of $\boldsymbol{\theta}$ given $D : t_1, \ldots, t_n$ is given by

$$\pi(\boldsymbol{\theta}|D) = \frac{L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{11}$$

Because the denominator in (11) is a normalizing constant, Bayes' theorem is often expressed as:

$$\pi(\boldsymbol{\theta}|D) \propto L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

Here the prior distribution is given beforehand, usually based on prior information of the parameters, such as that from historical data, previous experiences, expert suggestions, even wholly subjective suppositions, or simply from the point of mathematical conveniences.

The proposed priors for parameters p, λ_1, and λ_2 may be taken as

$$\pi(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}, \quad a, b > 0.$$

$$\pi(\lambda_1) = \frac{\alpha_1^{\beta_1}}{\Gamma(\beta_1)} \lambda_1^{\beta_1 - 1} e^{-\alpha_1 \lambda_1} \quad \alpha_1 > 0,\ \beta_1 > 0.$$

$$\pi(\lambda_2) = \frac{\alpha_2^{\beta_2}}{\Gamma(\beta_2)} \lambda_2^{\beta_2 - 1} e^{-\alpha_2 \lambda_2} \quad \alpha_2 > 0,\ \beta_2 > 0.$$

For all parameter, we assume independent priors. Then the joint prior distribution for $\boldsymbol{\theta} = (p, \lambda_1, \lambda_2)$ is

$$\pi(\boldsymbol{\theta}) \propto p^{a-1}(1-p)^{b-1} \lambda_1^{\beta_1 - 1} \lambda_2^{\beta_2 - 1} e^{-(\alpha_1 \lambda_1 + \alpha_2 \lambda_2)}.$$

Then, under the square error loss function, the Bayes estimate of p, λ_1, λ_2, failure rate function $h(t)$ and reliability function $R(t)$ are given by

$$p^* = \mathbb{E}(p|D) = \int_\Theta p.\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

$$\lambda_1^* = \mathbb{E}(\lambda_1|D) = \int_\Theta \lambda_1.\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

$$\lambda_2^* = \mathbb{E}(\lambda_2|D) = \int_\Theta \lambda_2.\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

$$h^*(t) = \mathbb{E}(h(t)|D) = \int_\Theta h(t).\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

$$R^*(t) = \mathbb{E}(R(t)|D) = \int_\Theta R(t).\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

In our study, we use adaptive Metropolis-Hastings sampling ([16]) to generate sample $\boldsymbol{\theta}^{(i)} = (p^{(i)}, \lambda_1^{(i)}, \lambda_2^{(i)}), i = 1, \ldots, n$ from the posterior distribution $\pi(\boldsymbol{\theta}|D)$. Then, Monte Carlo integration estimates p^*, λ_1^*, λ_2^*, $h^*(t)$ and $R^*(t)$ by calculating the means:

$$p^* = \mathbb{E}(p|D) \approx \frac{1}{n}\sum_{i=1}^{n} p^{(i)}$$

$$\lambda_1^* = \mathbb{E}(\lambda_1|D) \approx \frac{1}{n}\sum_{i=1}^{n} \lambda_1^{(i)}$$

$$\lambda_2^* = \mathbb{E}(\lambda_2|D) \approx \frac{1}{n}\sum_{i=1}^{n} \lambda_2^{(i)}$$

TABLE I
GENERATED DATA IN CASE $p = 0.3, \lambda_1 = 0.1, \lambda_2 = 0.2, k = 2$ AND $n = 30$

3.697	1.001	4.112	4.322	2.192	0.844
3.314	2.085	0.259	3.746	2.599	2.899
2.353	2.824	2.774	2.361	2.183	2.910
2.382	2.790	2.371	2.239	1.012	1.570
3.795	4.352	2.722	1.625	3.277	1.770

TABLE II
POINT ESTIMATES AND TWO-SIDED 90% BAYES CREDIBLE INTERVAL FOR p, λ_1, λ_2 AND $MTTF$

	True value	MCMC	90% BCI
p	0.3	0.3886	$[0.1229, 0.6693]$
λ_1	0.1	0.0998	$[0.0101, 0.2485]$
λ_2	0.2	0.1921	$[0.1093, 0.3294]$
$MTTF$	2.3789	2.5778	$[2.3064, 2.8717]$

$$h^*(t) = \mathbb{E}(h(t)|D) \approx \frac{1}{n}\sum_{i=1}^{n} h(t;\boldsymbol{\theta}^{(i)})$$

$$R^*(t) = \mathbb{E}(R(t)|D) \approx \frac{1}{n}\sum_{i=1}^{n} R(t;\boldsymbol{\theta}^{(i)})$$

IV. ILLUSTRATIVE EXAMPLE

In this section, we present an example to illustrate the estimation procedures discussed in this paper. We consider data given in Table I, which were generated in case $p = 0.3, \lambda_1 = 0.1$, $\lambda_2 = 0.2$ and $n = 30$ and the shape parameter k is considered to be fixed to two. In this study, we use both CE and MCMC method to estimate the parameters and reliability characteristics. In order to obtain MCMC estimators, prior parameters are arbitrarily taken as $a = b = 2$ and $\alpha_1 = \alpha_2 = 0.1, \beta_1 = \beta_2 = 1$ and we ran the Metropolis-Hastings algorithm to construct Markov chain of length 50000 with burn-in of 1000 and reduced the autocorrelation by retaining only every 5 iterations of the chain and obtain 9801 samples. Table II shows our MCMC point estimates and two-sided 90% of Bayes credible interval for p, λ_1, λ_2 and $MTTF$, and Table III shows our CE point estimates and two-sided 90% bootstrap confident interval BCa (bias corrected and accelerated) for p, λ_1, λ_2 and $MTTF$. Figs. 2-3 shows posterior distributions and traces of each parameter of the Bayesian model obtained by Metropolis-Hastings algorithm, and Figs 4-5 show time courses of reliability function and failure rate function obtained by both CE and MCMC method. On the basis of these results, we see that MCMC and CE gave comparable results relatively small sample size.

V. SIMULATION STUDY

A Monte Carlo simulation study is conducted to compare the performance of CE and MCMC estimators for the parameters of mixture failure rate. For each of the following choice of parameters, we simulate 1000 sets of data with sample size

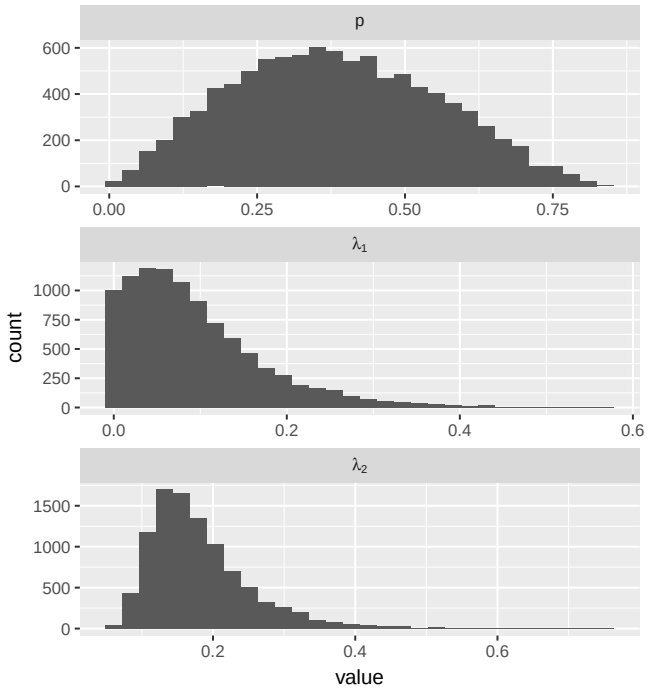


Fig. 2. Histograms of each parameter of the Bayesian model

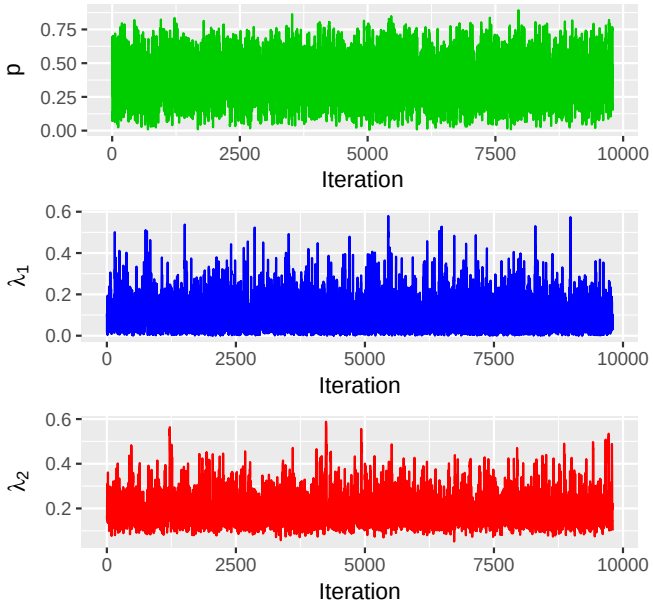


Fig. 3. Traces of each parameter of the Bayesian model

$n = 25$, 50, 100 and 200, respectively, and based on each set of data we computed the CE and MCMC estimator for the model parameters, $\boldsymbol{\theta} = (p, \lambda_1, \lambda_2)$. In order to obtain MCMC estimators, prior parameters are taken as in section IV, and we ran the Metropolis-Hastings algorithm to construct Markov chain of length 20000 with burn-in of 1000 and reduced autocorrelation by retaining only every 5 iterations of the chain and obtain 3801 samples. Note that as discussed earlier,

TABLE III
POINT ESTIMATES AND TWO-SIDED 90% BOOTSTRAP CONFIDENT INTERVAL BCA (BIAS CORRECTED AND ACCELERATED) FOR p, λ_1, λ_2 AND $MTTF$

	True value	CE	90% BCa
p	0.3	0.4267	[0.1036, 0.6631]
λ_1	0.1	0.0966	[0.0000, 0.3637]
λ_2	0.2	0.1938	[0.1138, 0.2804]
$MTTF$	2.3789	2.5187	[2.2340, 2.8050]

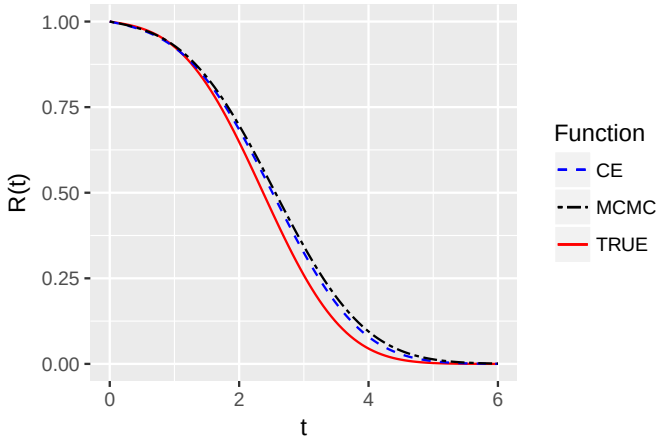


Fig. 4. The time courses of reliability functions

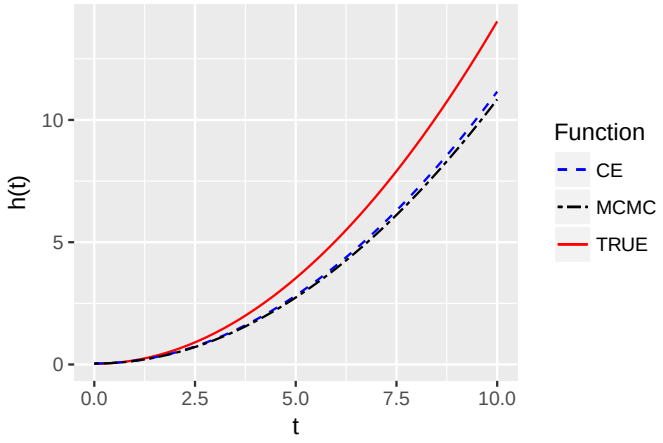


Fig. 5. The time courses of failure rate functions

when $p = 1$, the hazard rate function $h(t)$ represents the failure rate of an exponential distribution, while when $p = 0$, it represents the failure rate of the Rayleigh distribution or Weibull distribution with shape parameter 2.

1) $k = 1$
   - $p = 0.1,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
   - $p = 0.5,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
   - $p = 0.9,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
2) $k = 2$
   - $p = 0.1,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$

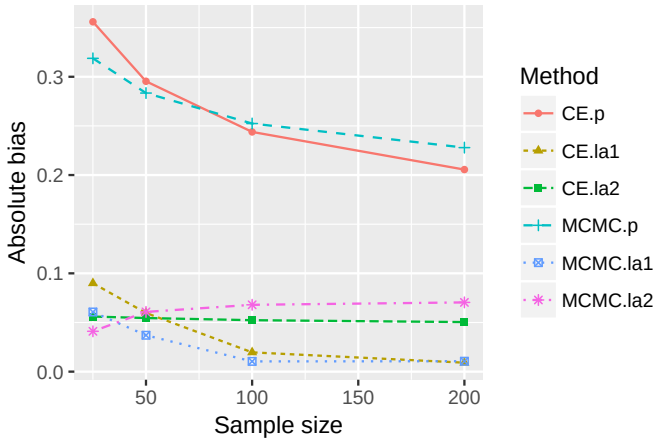


Fig. 6. Comparison of bias of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$, $k = 1$

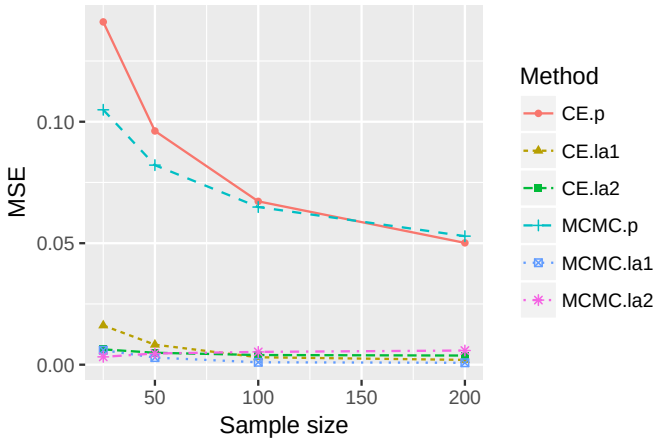


Fig. 7. Comparison of MSE of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$, $k = 1$

   - $p = 0.5,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
   - $p = 0.9,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
3) $k = 3$
   - $p = 0.1,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
   - $p = 0.5,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$
   - $p = 0.9,\ \lambda_1 = 0.1,\ \lambda_2 = 0.2$

The Tables IV-XII list the results of the simulation study. Denote $\hat{\boldsymbol{\theta}}$ as the CE and $\boldsymbol{\theta}^*$ as the MCMC of $\boldsymbol{\theta}$. Bias is calculated as the mean of 1000 estimates minus the true value, and MSE is the mean square error, the mean of the squared differences between 1000 estimators and true value. And Figs. 6-11 show the bias and MSE obtained from Tables IV-XII.

From the comparison of estimates, we observe the following:

- When p is close to 0 and sample size is small, say less than or equal to 50, MCMC performs a bit better than CE with respect to bias and MSE. But when sample is larger than 50, CE performs better than MCMC with respect to bias and MSE. In this case, the bias for p is quite high for both methods.
- When p equals to 0.5, both CE and MCMC method gave small bias and MSE. In this case, MCMC is more stable

TABLE IV
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$ WHEN $k = 1$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.3559	0.1411	0.0900	0.0161	0.0560	0.0063
	MCMC $\boldsymbol{\theta}^*$	0.3187	0.1049	0.0607	0.0060	0.0410	0.0032
50	CE $\hat{\boldsymbol{\theta}}$	0.2954	0.0962	0.0597	0.0082	0.0545	0.0049
	MCMC $\boldsymbol{\theta}^*$	0.2834	0.0821	0.0370	0.0029	0.0608	0.0046
100	CE $\hat{\boldsymbol{\theta}}$	0.2439	0.0672	0.0196	0.0031	0.0523	0.0040
	MCMC $\boldsymbol{\theta}^*$	0.2524	0.0649	0.0104	0.0010	0.0680	0.0053
200	CE $\hat{\boldsymbol{\theta}}$	0.2056	0.0501	−0.0091	0.0020	0.0504	0.0038
	MCMC $\boldsymbol{\theta}^*$	0.2279	0.0529	−0.0106	0.0008	0.0704	0.0058

TABLE V
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.5, 0.1, 0.2)$ WHEN $k = 1$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.0293	0.0105	0.0848	0.0122	−0.0194	0.0026
	MCMC $\boldsymbol{\theta}^*$	−0.0244	0.0039	0.0622	0.0061	−0.0234	0.0016
50	CE $\hat{\boldsymbol{\theta}}$	−0.0360	0.0105	0.0553	0.0065	−0.0244	0.0018
	MCMC $\boldsymbol{\theta}^*$	−0.0523	0.0053	0.0409	0.0035	−0.0110	0.0010
100	CE $\hat{\boldsymbol{\theta}}$	−0.0581	0.0117	0.0316	0.0035	−0.0228	0.0015
	MCMC $\boldsymbol{\theta}^*$	−0.0648	0.0067	0.0279	0.0024	−0.0014	0.0020
200	CE $\hat{\boldsymbol{\theta}}$	−0.0706	0.0177	0.0217	0.0026	−0.0179	0.0019
	MCMC $\boldsymbol{\theta}^*$	−0.0662	0.0069	0.0205	0.0018	0.0109	0.0052

TABLE VI
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.9, 0.1, 0.2)$ WHEN $k = 1$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	−0.2182	0.0587	0.0311	0.0034	−0.1326	0.0182
	MCMC $\boldsymbol{\theta}^*$	−0.2920	0.0895	0.0385	0.0033	−0.0850	0.0129
50	CE $\hat{\boldsymbol{\theta}}$	−0.2243	0.0610	0.0306	0.0026	−0.1317	0.0178
	MCMC $\boldsymbol{\theta}^*$	−0.2688	0.0760	0.0449	0.0037	−0.0621	0.0150
100	CE $\hat{\boldsymbol{\theta}}$	−0.2205	0.0604	0.0295	0.0020	−0.1288	0.0173
	MCMC $\boldsymbol{\theta}^*$	−0.2465	0.0638	0.0471	0.0037	−0.0300	0.0223
200	CE $\hat{\boldsymbol{\theta}}$	−0.2227	0.0842	0.1130	0.4046	−0.0503	0.3029
	MCMC $\boldsymbol{\theta}^*$	−0.2249	0.0539	0.0525	0.0046	0.0032	0.0575

TABLE VII
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$ WHEN $k = 2$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.3307	0.1172	0.0648	0.0095	0.0870	0.0108
	MCMC $\boldsymbol{\theta}^*$	0.2822	0.0811	0.0381	0.0028	0.0636	0.0056
50	CE $\hat{\boldsymbol{\theta}}$	0.2554	0.0727	0.0242	0.0038	0.0673	0.0065
	MCMC $\boldsymbol{\theta}^*$	0.2535	0.0654	0.0143	0.0012	0.0719	0.0061
100	CE $\hat{\boldsymbol{\theta}}$	0.2004	0.0464	−0.0066	0.0019	0.0525	0.0039
	MCMC $\boldsymbol{\theta}^*$	0.2292	0.0534	−0.0080	0.0007	0.0721	0.0058
200	CE $\hat{\boldsymbol{\theta}}$	0.1744	0.0371	−0.0294	0.0020	0.0478	0.0033
	MCMC $\boldsymbol{\theta}^*$	0.2093	0.0445	−0.0249	0.0011	0.0692	0.0052

TABLE VIII
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.5, 0.1, 0.2)$ WHEN $k = 2$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.0130	0.0083	0.0640	0.0091	0.0001	0.0019
	MCMC $\boldsymbol{\theta}^*$	−0.0530	0.0050	0.0458	0.0041	−0.0115	0.0011
50	CE $\hat{\boldsymbol{\theta}}$	−0.0431	0.0094	0.0386	0.0049	−0.0153	0.0016
	MCMC $\boldsymbol{\theta}^*$	−0.0673	0.0066	0.0319	0.0029	−0.0085	0.0008
100	CE $\hat{\boldsymbol{\theta}}$	−0.0686	0.0123	0.0225	0.0030	−0.0197	0.0013
	MCMC $\boldsymbol{\theta}^*$	−0.0748	0.0074	0.0227	0.0020	−0.0054	0.0009
200	CE $\hat{\boldsymbol{\theta}}$	−0.0723	0.0148	0.0204	0.0022	−0.0186	0.0015
	MCMC $\boldsymbol{\theta}^*$	−0.0689	0.0063	0.0230	0.0016	−0.0010	0.0010

TABLE IX
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.9, 0.1, 0.2)$ WHEN $k = 2$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	−0.1841	0.0416	0.0335	0.0048	−0.1221	0.0155
	MCMC $\boldsymbol{\theta}^*$	−0.2806	0.0818	0.0460	0.0049	−0.0774	0.0111
50	CE $\hat{\boldsymbol{\theta}}$	−0.2050	0.0502	0.0274	0.0030	−0.1269	0.0166
	MCMC $\boldsymbol{\theta}^*$	−0.2627	0.0720	0.0457	0.0047	−0.0554	0.0111
100	CE $\hat{\boldsymbol{\theta}}$	−0.1992	0.0483	0.0278	0.0021	−0.1248	0.0162
	MCMC $\boldsymbol{\theta}^*$	−0.2369	0.0588	0.0504	0.0044	−0.0195	0.0174
200	CE $\hat{\boldsymbol{\theta}}$	−0.1944	0.0562	0.0589	0.2819	−0.0920	0.1130
	MCMC $\boldsymbol{\theta}^*$	−0.2130	0.0486	0.0533	0.0049	0.0137	0.0338

TABLE X
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$ WHEN $k = 3$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.3115	0.1058	0.0480	0.0066	0.0928	0.0123
	MCMC $\boldsymbol{\theta}^*$	0.2698	0.0740	0.0288	0.0018	0.0685	0.0064
50	CE $\hat{\boldsymbol{\theta}}$	0.2442	0.0660	0.0089	0.0027	0.0680	0.0065
	MCMC $\boldsymbol{\theta}^*$	0.2442	0.0605	0.0062	0.0008	0.0718	0.0060
100	CE $\hat{\boldsymbol{\theta}}$	0.1875	0.0414	−0.0177	0.0018	0.0505	0.0038
	MCMC $\boldsymbol{\theta}^*$	0.2218	0.0499	−0.0142	0.0007	0.0704	0.0055
200	CE $\hat{\boldsymbol{\theta}}$	0.1655	0.0339	−0.0382	0.0024	0.0464	0.0031
	MCMC $\boldsymbol{\theta}^*$	0.2028	0.0418	−0.0306	0.0013	0.0677	0.0049

TABLE XI
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.5, 0.1, 0.2)$ WHEN $k = 3$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	0.0108	0.0090	0.0538	0.0078	0.0059	0.0023
	MCMC $\boldsymbol{\theta}^*$	−0.0586	0.0054	0.0406	0.0035	−0.0100	0.0011
50	CE $\hat{\boldsymbol{\theta}}$	−0.0523	0.0113	0.0322	0.0046	−0.0155	0.0015
	MCMC $\boldsymbol{\theta}^*$	−0.0717	0.0071	0.0273	0.0025	−0.0086	0.0008
100	CE $\hat{\boldsymbol{\theta}}$	−0.0751	0.0135	0.0234	0.0031	−0.0201	0.0013
	MCMC $\boldsymbol{\theta}^*$	−0.0747	0.0074	0.0234	0.0021	−0.0053	0.0007
200	CE $\hat{\boldsymbol{\theta}}$	−0.0678	0.0128	0.0177	0.0021	−0.0194	0.0014
	MCMC $\boldsymbol{\theta}^*$	−0.0690	0.0062	0.0234	0.0016	−0.0040	0.0010

TABLE XII
COMPARISON OF $\hat{\boldsymbol{\theta}}$ AND $\boldsymbol{\theta}^*$ FOR $\boldsymbol{\theta} = (0.9, 0.1, 0.2)$ WHEN $k = 3$

n	Method	Bias p	MSE p	Bias λ_1	MSE λ_1	Bias λ_2	MSE λ_2
25	CE $\hat{\boldsymbol{\theta}}$	-0.1571	0.0311	0.0297	0.0045	-0.1136	0.0136
	MCMC $\boldsymbol{\theta}^*$	-0.2817	0.0824	0.0475	0.0056	-0.0792	0.0109
50	CE $\hat{\boldsymbol{\theta}}$	-0.1846	0.0418	0.0254	0.0032	-0.1211	0.0152
	MCMC $\boldsymbol{\theta}^*$	-0.2648	0.0729	0.0466	0.0049	-0.0618	0.0103
100	CE $\hat{\boldsymbol{\theta}}$	-0.1851	0.0424	0.0268	0.0022	-0.1212	0.0154
	MCMC $\boldsymbol{\theta}^*$	-0.2392	0.0599	0.0504	0.0046	-0.0317	0.0146
200	CE $\hat{\boldsymbol{\theta}}$	-0.1889	0.0514	0.0411	0.0645	-0.1085	0.0273
	MCMC $\boldsymbol{\theta}^*$	-0.2173	0.0504	0.0537	0.0048	0.0025	0.0367

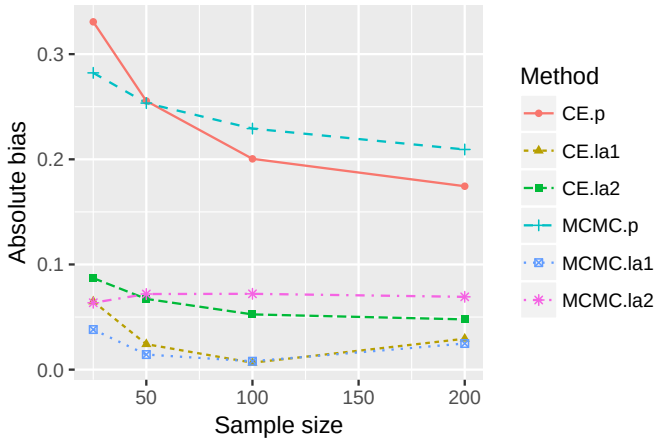


Fig. 8. Comparison of bias of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$, $k = 2$

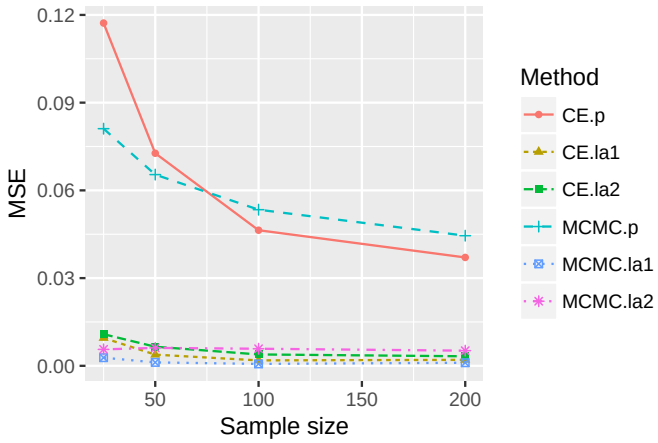


Fig. 9. Comparison of MSE of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$, $k = 2$

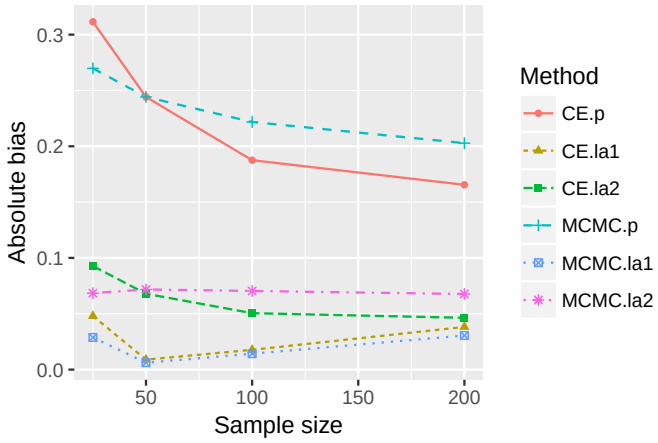


Fig. 10. Comparison of bias of $\hat{\theta}$ and θ^* for $\theta = (0.1, 0.1, 0.2)$, $k = 3$

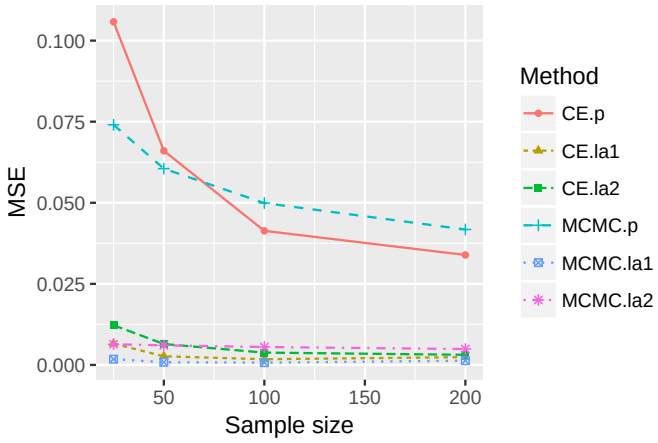


Fig. 11. Comparison of MSE of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta} = (0.1, 0.1, 0.2)$, $k = 3$

than CE.

- When p is close to 1, MCMC is quite stable, but CE's MSE tends to be high when sample size is large. This is an unstable of CE method. In this case, the bias for p is quite high for both methods.
- The bias in p becomes less with increasing k.

VI. CONCLUSION

This study brings two interested approaches to the mixture failure rate, the cross-entropy and the MCMC method. In this study, MCMC is found to be more stable than CE in every aspect of the simulation study. This advantage of MCMC partly dues to the fact that we know the true parameter values.

What would happen if the prior expected values differ quite a bit from those used in the simulation. And as we saw in the simulation study, the biases of both method when p is close to 0 or 1 are still high to ignore. All of these problems are some aspects of interest for further investigation.

ACKNOWLEDGMENT

This article has been elaborated in the framework of the SGS grant from VSB - Technical University of Ostrava (No. SP2017/56).

REFERENCES

[1] J. F. Lawless, *Statistical Models and Methods for Life-time Data*, 2nd ed. Wiley-Interscience, 2002.
[2] N. Mann and E. Schaffer and N. Singpurwalla, *Methods for Statistical Analysis of Reliability and Life Data*, New York: Wiley, 1974.
[3] S. K. Sinha, *Reliability and Life Testing*, USA: Wiley Eastern Ltd, 1986.
[4] H. F. Martz and R. A. Waller, *Bayesian Reliability Analysis*, New York: John Wiley and Sons, 1982.
[5] K. K. Sharma, H. Krishna and B. Singh, *Bayes estimation of the mixture of hazard rate model*, Reliability Engineering and System Safety, vol. 55, pp. 9–13, 1974.
[6] C. M. Harris and N. D. Singpurwalla, *Life Distributions Derived from Stochastic Hazard Functions*, IEEE Trans, Reliab., vol. 17, pp. 70–79, 1968.
[7] G. C. Canavos, *On the Robustness of a Bayes Estimate*, Annual Reliability and Maintainability Symposium, pp.432-435, 1974.
[8] A. H. Moore and J. E. Bilikam, *Bayesian Estimation of Parameters of Life Distributions and Reliability from Type II Censored Samples*, IEEE Trans, Reliab., vol. 27, pp. 64–67, 1978.
[9] V. M. R. Tummala and P. T. Sathe, *Minimum Expected Loss Estimators of Reliability and Parameters of Certain Lifetime Distributions*, IEEE Trans, Reliab., vol. 27, pp. 283–285, 1978.
[10] A. Alexander Aron, H. Guo and A. Mettas and D. Ogden, *Improving the 1-Parameter Weibull: A Bayesian Approach*, IEEE, 2009.
[11] Muhammad Aslam, S. M. A. Kazmi, I. Ahmad and S. H. Shah, *Bayesian Estimation for Parameters of the Weibull Distribution*, Sci.Int.(Lahore), vol. 26(5), pp. 1915–1920, 2014.
[12] H. Rinne, *The Weibull Distribution: A Handbook*,1st ed., Chapmann and Hall/CRC, 2008.
[13] T. Xia, X. Jin, L. Xi and J. Ni, *Production-driven opportunistic maintenance for batch production based on mam-apb scheduling*, European Journal of Operational Research, vol. 240(3), pp. 781–790, 2015.
[14] B. Zhou, J. Yu, J. Shao and D. Trentesaux, *Bottleneck-based opportunistic maintenance model for series production systems*, Journal of Quality in Maintenance Engineering, vol. 21(1), pp. 70–88, 2015.
[15] W. Xu and L. Cao, *Optimal tool replacement with product quality deterioration and random tool failure*, International Journal of Production Research, vol. 53(6), pp. 1736–1745, 2015.
[16] C. Chivers (2012, Feb.) *MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling*, R package version 1.1-8. [Online]. Available: https://CRAN.R-project.org/package=MHadaptive
[17] D. Navarro and A. Perfors, *The Metropolis-Hastings Algorithm*, . [Online]. Available: http://www.compcogscisydney.com/ccs-class.html
[18] D. P. Kroese, S. Porotsky and R. Y. Rubinstein, *The Cross-Entropy Method for Continuous Multi-Extremal Optimization*, Methodol Comput Appl Probab, vol. 8, pp. 383–407, 2006.
[19] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*, Springer Science & Business Media, Inc, 2004.

Researches of transition and quasi-steady state processes in a shunt active power filter

Sergey German – Galkin, Dariusz Tarnapowicz
Maritime University of Szczecin, Poland
Institute of Marine Electrical Engineering and Automation
Szczecin, Poland
s.german-galkin@am.szczecin.pl, d.tarnapowicz@am.szczecin.pl,

***Abstract*— One of the methods to improve the power quality in electric power networks is the use of active power filters.**
In the article, a method of analysis of transition and quasi-steady state processes in a shunt active three-phase power filter (SAPF) was presented. The analysis of electromagnetic processes was carried out on the basis of a mathematical description of the active filter. Structural models for the simulation of dynamic processes in the shunt active three-phase power filter were developed.
The results of studies of dynamic states obtained for the SAPF with independent control and current control were shown. A study of the active filter was performed with the use of Matlab-Simulink program.

Keywords— quasi-steady state processes of active filter; shunt active power filter; transition processes of active filter

I. INTRODUCTION

The theory of instantaneous power (the p-q power theory) [1, 2] allows to determine active and distortion power in passive, asymmetric and non-linear three-phase electrical circuits. On this basis, it is possible to make a mathematical description of the control algorithm for PWM inverter working as an shunt active power filter (SAPF), which compensates all distortion power. In the control algorithm of the active filter used, the power factor (PF) is calculated. For the system of electricity network, this power factor becomes close to unity.

Functional diagram of the three-phase network along with the SAPF is presented in Fig. 1.

Diagram and description of the SAPF is known and presented in the literature [3]. Total power of reactive, nonlinear and asymmetric load consists of:

$\bar{p}$ - mean value of instantaneous real power

$\tilde{p}$ - alternating value of instantaneous real power

$\bar{q}$ - mean value of instantaneous imaginary power

$\tilde{q}$ - alternating value of instantaneous imaginary power

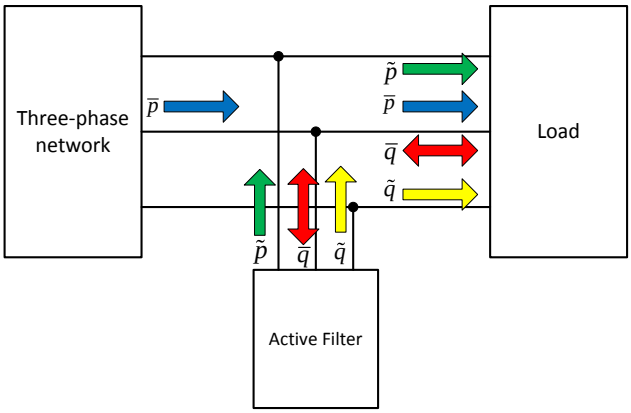


Fig. 1. Diagram of three-phase electricity network along with the SAPF

Active filter makes it possible to generate the load for the entire range of non-active power elements. Therefore, it can reduce (by 20-25 percent) the load of the power line and energy generation systems [4]. The paper presents a SAPF, which is intended to generate passive currents in the symmetrical three-phase resistive and inductive load. A method for the analysis of dynamic states in a SAPF, based on a joint consideration of electromagnetic transition processes and quasi-steady state processes was developed. It should be noted that the subject of dynamic states in active filters in the literature is hardly discussed.

II. ANALYSIS OF ELECTROMAGNETIC PROCESSES IN NETWORKS WITH SHUNT ACTIVE POWER FILTER

Synthesis of the control algorithm, providing the generation in a SAPF distortion power of load currents is solved by the respective conversion of coordinates. The basic coordinate systems and their properties are presented in Table 1 and they are used for a mathematical description of electromagnetic processes in three-phase electrical circuits [5].

Classical transformation (Park transform) of a stationary system of coordinates a, b, c, to the rotating system d, q 0 requires the calculation of the real and instantaneous powers. Furthermore, it requires the separation of continuous and pulsating power component (mean values and alternating value), calculation of currents corresponding to these powers and ensuring the compensation for distortion power of these

currents [6, 7]. All these conversions are carried out for instantaneous values of currents and voltages.

TABLE 1. BASIC COORDINATE SYSTEMS USED IN THE CONTROL OF SAPF

coordinate system	number of coordinate axes	angle between the coordinate axes	the movement of the axes
on the plane			
$a\ b\ c$	3	120^0	stationary
$\alpha\ \beta$	2	90^0	stationary
$d\ q$	2	90^0	rotating
in the space			
$\alpha\ \beta, 0$	3	90^0	stationary
$d\ q, 0$	3	90^0	rotating
$p\ q, r$	3	90^0	rotating

A completely different problem has been solved during the analysis of electromagnetic processes in the network with active power filter. The aim of such an analysis is to determine the relations between real and amplitude values of state variables and calculate (on this basis) the basic properties of transition and quasi-steady state in SAPF. The foundation for the analysis of electromagnetic processes is the method of space vector [8] and method of the first harmonic [9].

The analysis of electromagnetic processes with the use of a space vector was carried out on the basis of a functional diagram shown in Fig. 2.

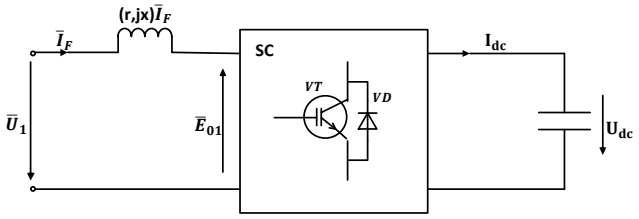


Fig. 2. Functional diagram of SAPF

The presented functional diagram is a special case of an active converter [5]. On one hand, the semiconductor converter (SC) is connected to the three-phase alternating voltage with amplitude $\bar{U}_1$, and on the other hand – to the circuit intermediate with a condenser. Input inductance is a storage for "kinetic" energy, and the condenser collects "potential" energy. The exchange of energy between the inductance and the condenser is carried out via a semiconductor converter of high frequency of valves' switching.

From the side of power network, a mathematical description of SAPF based on the first harmonic method in a rotating coordinate system can be presented in the form of (1):

$$\bar{U}_1(t) = \bar{E}_{01}(t) + L\frac{d\bar{I}_F(t)}{dt} + r\bar{I}_F(t) + jx\bar{I}_F(t) \qquad (1)$$

where:

$\bar{U}_1$ – vector of mains electricity,

$\bar{E}_{01}$ – vector of electromotive force of SAPF

$\bar{I}_F$ – resultant current vector of SAPF

r – active resistance, taking into account the resistance of source, choke and semiconductor elements SAPF

L - inductance on the input SAPF

x - reactance of coil for the network's frequency

Note that the vector of voltage amplitude $\bar{U}_1$, electromotive force $\bar{E}_{01}$ and current $\bar{I}_F$ depends on the time – their record in the function of time.

With regards to the alternating current network, SAPF generates an electromotive force. The vector of this force is equal to [10] :

$$\bar{E}_{01}(t) = \frac{mU_{dc}(t)}{2}\exp(j\varphi_m) \qquad (2)$$

where:

U_{dc} - capacitor voltage (DC bus)

m - modulation index

φ_m – phase of voltage control signal with respect to the mains voltage

III. ANALYSIS OF QUASI-STEADY STATE ELECTROMAGNETIC PROCESSES IN THE NETWORK WITH ACTIVE POWER FILTER

From the side of the network, a mathematical description of SAPF for a quasi- steady state mode in accordance with the equation (1) can be presented:

$$\bar{U}_1 = \bar{E}_{01} + r\bar{I}_F + jx\bar{I}_F \approx \bar{E}_{01} + jx\bar{I}_F \qquad (3)$$

Assuming that x>>r in the further analysis of quasi-stationary processes, the final quotation will take the form (3).

If the network voltage vector $\bar{U}_1$ is directed along d-axis, the vector diagram SAPF on the complex place in accordance with the equitation (3) can be presented as in Fig. 3.

Network current $\bar{I}_1$ includes active $\bar{I}_a$ and passive (inductive) $\bar{I}_r$ component. In order to ensure that the network current is only active $\bar{I}_F$ the active filter should compensate the passive component $\bar{I}_F = -\bar{I}_r$, i.e. current of the filter must be capacitive.

The vector diagram (Figure 3) shows that current $\bar{I}_F$ SAPF is directed along the q-axis $\bar{I}_F = \bar{I}_q$

For the filter's current directed along the q-axis, the voltage drop across the inductance $jx\bar{I}_F$ is opposite to the electromotive force $\bar{E}_{01}$ coinciding with the d-axis, so $\varphi_m = 0$.

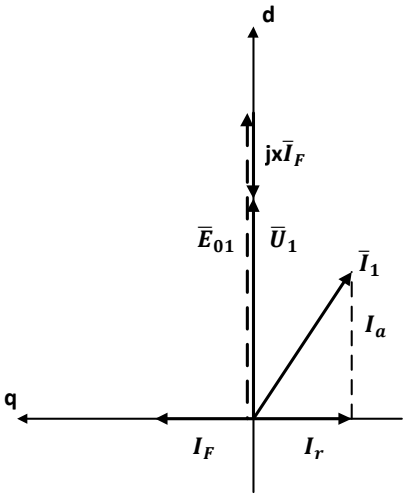


Fig. 3. Vector diagram SAPF

Relations between the mains electricity U_1, electromotive force E_{01} and current I_q of the active filter are determined in accordance with the vector diagram.

$$E_{01} = E_d = U_1 + xI_F = U_1 + xI_q \tag{4}$$

The equation (4) presents a key conclusion that electromotive force should exceed the network voltage in order to obtain the capacitive nature of the current in SAPF. The higher network voltage, the higher inactive component of the load current

IV. Analysis of Electromagnetic Dynamic Processes with Independent Control with the Use of an Shunt Active Power Filter

During the independent operation of SAPF, the control signals are a modulation factor and modulation phase. Simultaneously, in a synchronously rotating coordinate system (d – real axis and q – imaginary axis), in the directing of a voltage vector of the network $\bar{U}_1$ on the real d-axis, modulation phase is zero (φ_m=0), and the equation (1), (2) including the DC circuit, can be written in the following way:

$$\begin{aligned} U_1 &= \frac{mU_{dc}(t)}{2} + L\frac{dI_d(t)}{dt} + rI_d(t) - xI_q(t), \\ 0 &= L\frac{dI_q(t)}{dt} + rI_q(t) + xI_d(t), \\ C\frac{dU_{dc}(t)}{dt} &= I_{dc}(t), \end{aligned} \tag{5}$$

where:

C – condenser capacity SAPF,

$I_{dc}(t)$ - charging current of the condenser.

On the basis of equations (5), it is possible to obtain a differentia equation SAPF in relations to the current $I_q(t)$:

$$\frac{d^2I_q(t)}{dt^2} + \frac{2r}{L}\frac{dI_q(t)}{dt} + \omega_1^2 I_q(t) = \frac{\omega_1}{L}\left(U_1 - \frac{mU_{dc}(t)}{2}\right) \tag{6}$$

SAPF characteristic equation can be written in the following way:

$$s^2 + \frac{2}{\tau}s + \omega_1^2 = 0 \tag{7}$$

Roots of the equation (7):

$$s_{1,2} = -\frac{1}{\tau} \pm \sqrt{\left(\frac{1}{\tau}\right)^2 + \omega_1^2} \approx -\frac{1}{\tau} \pm \omega_1 \tag{8}$$

were:

$\tau = \frac{L}{r}$ - time constant

The equation (7), (8) shows that in the independent control, SAPF occurs an oscillating link connected with the mains frequency, and the attenuation coefficient determines the time constant.

V. Study of the Model of Shunt Active Power Filter with Independent Control

Relation between the value of active filter's current and the condenser's charging current should be taken into account in order to build the structural model of the active filter with independent control on the basis of the system of equations (5).

Charging of the condenser with DC voltage is connected with the active power consumption from the network during the transition process. In the combination of the real d-axis of the rotating coordinate system with the network voltage network, the active power is determined by the current $I_d(t)$, which implies the following equation: $I_{dc}(t) = I_d(t)$. This current charges the condenser of the circuit SAPF.

Taking into account the fact that the structural SAPF model with independent control is built in accordance with equations (5), implementations in the programmable environment MATLAB-SIMULINK (Fig. 4.) were made.

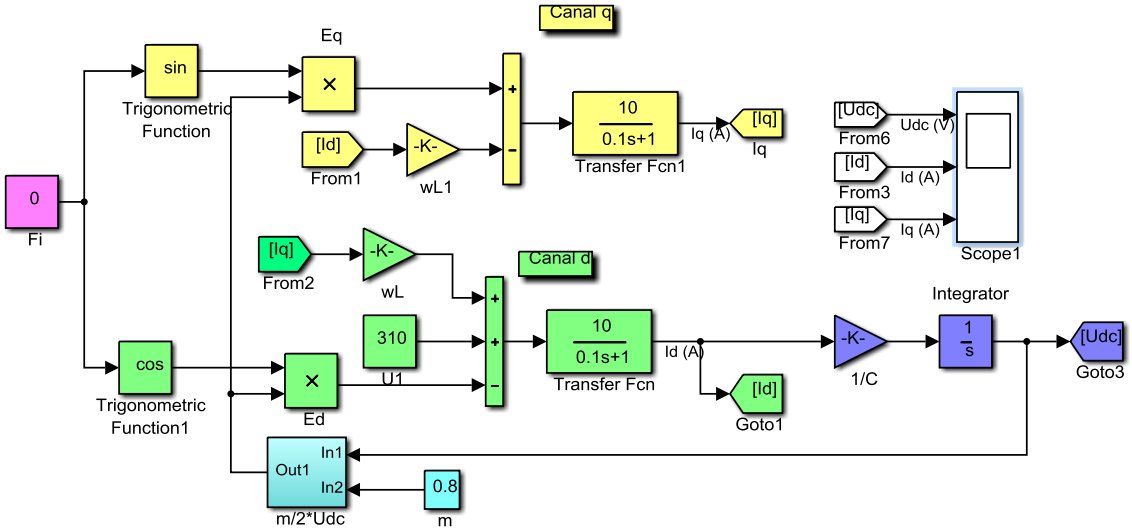


Fig. 4. Structural SAPF model with independent control.

Electromagnetic transition processes in SAPF were analyzed with the following SAPF settings:

- modulation index m=0.8

- output inductance $L = 0.01H (x = 3.14\Omega)$

- active resistance $r = 0.1\Omega$

– capacity of DC link $C = 320\mu F$

Detection of control signals in the system of coordinates d, q implemented in the program environment in a three-phase electrical network with parameters:

$U_1 = 310V$,

$\omega_1 = 314rad / s$,

$I_r = -40A$

The results of the analysis are shown in Fig. 5. These results confirm theoretical considerations obtained in equations (6) and (7), (8).

VI. Analysis of Electromagnetic Transients with Current Control of Shunt Active Power Filter

To compensate the reactive current, a SAPF must be a current source. Current control in a stationary coordinate system is realized with the use of relay superiors [5, 10, 11]. This control in the converter systems is referred to as hysteresis or space control. Linear PID controllers can be used as current regulators in the rotating coordinate system.

During the current control, the inverter is a pulse generator with a high carrier frequency in relation to the set point value. The set point value is marked as I_q^*. During the period $T = \frac{1}{f}$ two commutating intervals are created.

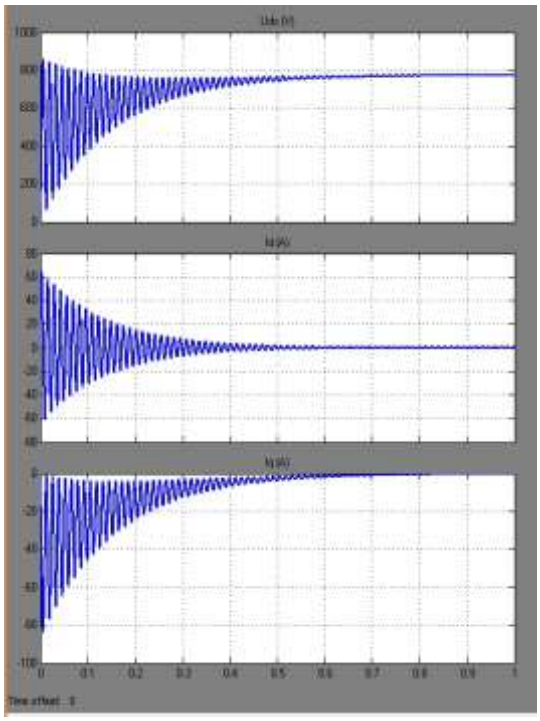


Fig. 5. Electromagnetic transition processes in SAPF with independent control

In the first commutating interval, the condenser (through the respective transistors) is connected to the network through the choke. The duration of this interval is marked by t_1. In this interval, the current flowing through the circuit enables the exchange of passive energy between the choke, condenser and the network. In the second commutating interval, the choke is directly connected to the network and the exchange of energy occurs between the choke and the network. The duration of this interval is marked $t_2 = T - t_1$. The ratio of the maximum

duration of the first commutating interval to the period of carrier frequency is a modulation factor $m = \frac{t_1}{T}$.

In the rotating system of coordinates d, q, the set point value of current is constant. Carrier frequency and duration of switching times are also constant, and the duration of the first commutating interval is equal to the modulation depth.

The equivalent SAPF with the first and second commutating interval was shown in Fig. 6.

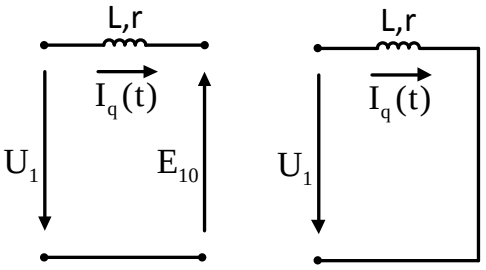


Fig. 6. Equivalent diagram of SAPF – first and second commutating interval.

Let's consider the electromagnetic processes occurring during commutation SAPF. Differential equations of the first and second commutating interval with actual limit values in these intervals can be written in the following manner:

$$\begin{aligned} &L\frac{di_q}{dt} + ri_q = U_1 - E_{10}(t) = -xI_q^*, \\ &i_q(0) = I_q^* + \Delta I_q, \;\; i_q(t_1) = I_q^* - \Delta I_q, \\ &L\frac{di_q}{dt} + ri_q = U_1, \\ &i_q(0) = I_q^* - \Delta I_q, \;\; i_q(t_2) = I_q^* + \Delta I_q, \end{aligned} \tag{9}$$

where:

i_q - instantaneous current during commutation,

ΔI_q - amplitude of SAPF current's pulsation, which is determined by the time constant $\tau = \frac{L}{r}$, width of the hysteresis loop of the current regulator and the set point value of the SAPF current.

Duration of the first commutation interval is determined by the solution of the differential equation (9) in this range:

$$t_1 = mT = \frac{2\Delta I_q}{\omega_1 I_q^*} \tag{10}$$

Duration of the second commutation interval is determined by the solution of the differential equation (9) in this range:

$$t_2 = (1-m)T = \frac{2\Delta I_q}{\omega_1 I_b} \tag{11}$$

where:

$I_b = \frac{U_1}{x}$ - short-circuit current

It should be noted that the current SAPF cannot exceed the value of short-circuit current, usually $I_q^* \le 0.5 - 0.7 I_b$.

The duration of a carrier frequency is determined from equations (10) and (11).

$$T = t_1 + t_2 = \frac{2\Delta I_q}{\omega_1 I_b}\left(1 + \frac{I_b}{I_q^*}\right) \tag{12}$$

Fig. 7. presents, in accordance with the equation (12), dependence of the carried frequency's period on the filter's relative current for three values of the current relative pulsation. Usually, ΔI_q is two times (or times) smaller than the short-circuit current.

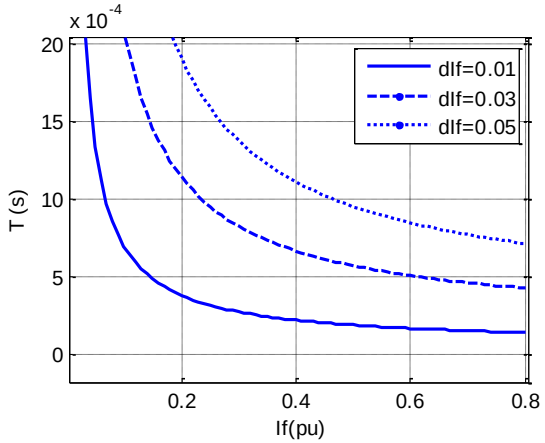


Fig. 7. Dependence of the commutation period on the relative current SAPF

The solution of equations (10), (11) causes that it is possible to find the size of SAPF modulation factor.

VII. Simulation Tests of the Shunt Active Power Filter with Current Control

A full mathematic description of dynamic processes based on the first harmonic method with current control SAPF is carried out on the basis of equations (5) and (12), including equality . In the operational form, equations can be presented as follows:

$$\begin{aligned} &(I_d^* - I_d(t)) \cdot W(s) = \frac{mU_{dc}(t)}{2} + r(s\tau + 1)I_d(t) - xI_q(t), \\ &(I_q^* - I_q(t)) \cdot W(s) = r(s\tau + 1)I_q(t) + xI_d(t), \\ &sCU_{dc}(t) = I_d(t). \end{aligned} \tag{13}$$

where:

$W(s)$ - operational function of PID regulator,

I_d^*=0 and $I_q^* = I_F$ - set currents

Structural model for the analysis of dynamic properties of the shunt active power filter in the rotating system of coordinates d, q constructed on the basis of the system (13), (14) is presented in Fig. 8.

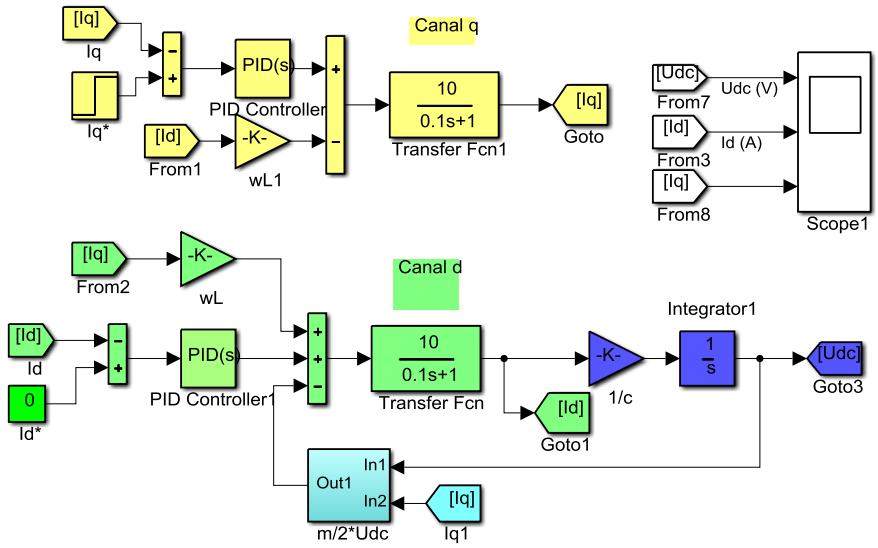


Fig. 8. Structural model of SAPF with current control

The synthesis of PID regulator was carried out on the technical optimum for a dynamic member described by the equation (6).

Transition processes of the voltage and condenser's current, as well as current SAPF are shown in Fig. 9.

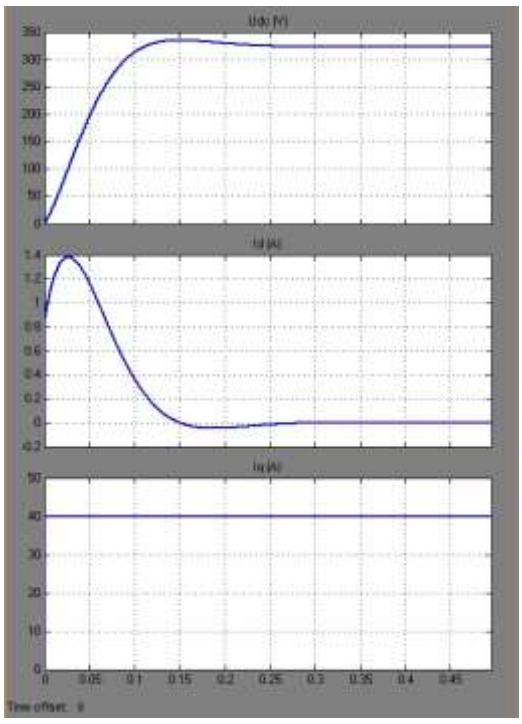


Fig. 9. Transition processes in the active filter with current control

Parameters of SAPF were selected as follows:

$$L = 0.01H\,(x = 3.14\Omega),$$
$$r = 0.1\Omega,\ C = 320\mu F,$$
$$I_q^* = 40A,\ \ I_d^* = 0A.$$

CONCLUSIONS

The main goal presented in this article was to develop a method for the electromagnetic analysis of transient processes in a parallel active filter. This method is based on classic control algorithms: the spatial vector method and the first harmonic analysis.

The method of analysis is based on the joint consideration of electromagnetic transition processes, quasi-determined processes and commutating processes. Electromagnetic and quasi-determined processes describe changes in the amplitude value of variables in time function, and commutating processes demonstrate changes of instantaneous values in time function. Thanks to this approach, it was possible to obtain a mathematic description of the active filter in the current control of SAPF inverter. On the basis of the above-mentioned mathematical description, structural models for the simulation of dynamic processes in the active three-phase filter designed to the compensation of reactive current (at three-phase balanced load) were developed.

The obtained results of our studies confirmed the theoretical assumptions that the active current (causing the charging of the condenser) occurs only in a transition state. In the determined and quasi-determined state, it does not exist.

REFERENCES

[1] H. Akagi, A. Nabae, and Y. Kanazawa, “Generalized theory of instantaneus reactive power in three-phasecircuits”. IPEC Conf. 1983, Tokyo, S. 1375–1386

[2] H. Akagi, “New trends in active filters”. Conf. Proc. EPE'95, Sevilla, 1995, pp.0.017-0.026.

[3] SG. German-Galkin, “Transient processes in active power filter’ School MATLAB. Lesson 19. Power electronics. 2015. № 2, pp 34-39. (in Russian)

[4] Ye.Ye. Chaplygin, and N.G. Kalugin, “Teoriya moshchnosti v silovoy elektronike”. Uchebnoye posobiye dlya studentov, obuchayushchikhsya po spetsial'nosti «Promyshlennaya elektronika». Moskva 2006. (in Russian)

[5] S.G. German-Galkin, Virtual'nyye laboratorii poluprovodnikovykh sistem v srede Matlab-Simulink. Izd-vo «Lan'», Sankt Peterburg, 2013. (in Russian)

[6] K. Kim, F. Blaabjerg, B. Bak Jensen, and J. Choi, “Instantaneous power compensation in three-phase system using p-q-r theory”, IEEE Trans. Pow. Elect., Vol. 17 (2002), pp. 701-710.

[7] R. Strzelecki, H. Supronowicz, „Filtracja harmonicznych w sieciach zasilających prądu przemennego”. Wydawnictwo „Adam Marszałek ”, Toruń 1998.

[8] K.P. Kovach, and I. Rats, “Perekhodnyye protsessy v mashinakh peremennogo toka”/ Per. s nem. M.-L.: Gosenergoizdat, 1963. (in Russian)

[9] A. Bulgakov, “Novaya teoriya upravlyayemykh vypryamiteley”. M. Nauka, 1970 (in Russian).

[10] YU.K. Rozanov, Silovaya elektronika. Moskva. Izdatel'skiy dom MEI, 2007. (in Russian)

[11] G. Zhemerov, V. Kolesnik, SHCH. Il'ina, Sootnosheniya dlya preobrazovaniy koordinat obobshchennykh vektorov napryazheniy i tokov trekhfaznoy sistemy elektrosnabzheniya. NTU. «Khar'kovskiy politekhnicheskiy institut». Khar'kov. 2009. (in Russian)

[12] G.S. Zinov'yev, Osnovy silovoy elektroniki, CH.2, Novosibirsk, 2000. (in Russian)

This research outcome has been achieved under the research project: Nowoczesne technologie w systemach "Shore to Ship" No 2/S/IEiAO/16 financed from a subsidy of the Ministry of Science and Higher Education for statutory activities

COMPARISON OF VARIOUS ERROR-DETECTING AND ERROR-CORRECTING ENCODINGS OF REVERSIBLE AUTOMATA BUILT FROM IRREVERSIBLE STATE TABLES USING EPOE CIRCUITS WITH EXOR LATTICES

Linh Tran, Bruce Yen, Marek Perkowski
Department of Electrical and Computer Engineering,
Portland State University , Portland OR, USA
linht@pdx.edu, bruceyen@paulhastings.com

***Abstract*—This paper presents a new approach to synthesize reversible (quantum permutative) automata from standard state tables. It uses various variants of error–detecting and error-correcting encodings. Next, a combination of recently introduced "EPOE circuits" as well as EXOR Lattices are used for synthesis. The goal is to compare the costs of various encodings and synthesis methods for the same machine, especially strongly unspecified one.**

Keywords— Reversible, EPOE, ESOP, Product of EXOR Sums, Synthesis, Template, Minimization, Quantum Permutative Circuits, Quantum Cost

I. INTRODUCTION

This paper focuses on the realization of reversible (quantum) automata [18,19] with different encodings and realized with different numbers of ancilla bits. For logic synthesis of binary quantum reversible circuit [17] we use the combination of EPOE circuit [15] with a new EXOR Lattice concept introduced here. EPOE circuits are generalizations of circuits from [9] realized in reversible logic and optimized to reduce quantum costs [6]. Because the machine can have state and output don't cares and some encodings also bring don't cares, and finally the Miller method brings don't cares, our logic synthesis method must assume a high percent of don't cares, which distinguishes it from most of the reversible logic synthesis algorithms. Because of these properties the method can be applied not only to synthesis of quantum circuits, but also in Quantum Machine Learning, where a minimal machine/mapping is designed from a set of positive and negative samples – supervised learning in which a classifier is realized as a hardware quantum computer [18]. Paper [16] presents Machine Learning application of ESOP circuits, and the circuits discussed here are a generalization of a reversible version of ESOP (Exclusive-Or-Sum-of-Products) circuits.

Several methods for synthesis of binary reversible/quantum circuits are known [1,2,4,6,7,10,12,13,14,23,24]. For instance, the well-known Miller-Maslov-Dueck (MMD) method minimizes ancilla bit/garbage bit introduction and using gate reduction techniques such as template matching. However, the method suffers in several respects. The MMD method is unable to generate circuits for non-reversible functions. By adding as many ancilla bits as outputs of the initial irreversible specification it is possible to extend the MMD method to non-reversible functions by converting the initial input/output specification to a reversible function specified as a permutation vector. However, as MMD method can only be applied to completely specified functions, the task of optimally specifying the added ancilla bits is non-trivial. Irreversible state machines have don't cares in transitions and outputs, can use codes that introduce additional don't cares, and lead to multi-output excitation/output functions that when using Miller's Conversion Method lead to even more don't cares. Taking it all together, one has to deal with synthesis of a reversible state machine with extremely many don't cares and few ancilla bits. We want to investigate the role of encodings and ancilla bits and the tradeoffs between Quantum Costs introduced in [6] and numbers of ancilla bits and garbage bits (all bits used for encoding).

In the method presented here, garbage bits are considered to be non-restrictive, and are introduced freely to reduce the complexity of synthesis. Future quantum computers (such as those used for integer factorization) will be useful only for a large number of lines (qubits). Some techniques such as mirroring and constant overlap [1] will allow location of reversible permutative oracles inside a larger two-dimensional space similarly as it is done in

contemporary VLSI: the specification is partitioned for logic minimization and the synthesized circuits are placed and routed. This general methodology justifies our more generous use of ancilla bits than in MMD and previous synthesis algorithms.

The paper is organized as follows: Sec. 2 presents the concept of EXOR Lattices. Sec. 3 briefly introduces EPOE circuits and presents EPOEM-1-DC algorithm for synthesis of single output Boolean functions. Sec. 4 introduces EPOEM-MO algorithm for the synthesis of multi-output functions using EXOR Lattices. Sec. 5 describes the entire automata realization system. Sec. 6 discusses experimental results and Sec. 7 concludes the paper.

II. Synthesis based on Exclusive-OR Lattices

The nature of an EXOR logic gate gives the following useful property: given a set of three logic functions A, B, and C, such that A ⊕ B = C, it is obvious that B ⊕ C = A and A ⊕ C = B. Therefore, given any two functions, the third of this closed set of functions can be uniquely determined. The presented below algorithm exploits this property by performing a logical EXOR between all output functions to search for functions that are commonly repeated or easily implemented. Furthermore, the property scales to any number of functions, introducing a new concept of an EXOR lattice, as shown in Fig. 1.

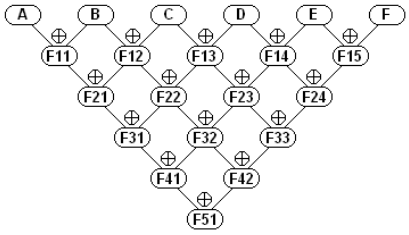


Figure 1: Flattened EXOR Tree (Case of EXOR Lattice)

The top row of nodes in Fig. 1 represents outputs of 6 different functions. Each subsequent row represents the logic functions obtained by logically EXOR'ing the previous row's functions with one another. Therefore, by selecting functions A, F11, and F21, we can implement B directly with a logical EXOR of F11 with A. Performing an EXOR of F11 and F21 we obtain F12, which can be used to obtain output C, as well. In this case, this map is 2-dimensional (planar). However, when creating such a DAG, the first and third node may have some logical EXOR that can be used instead to obtain the third node, rather than having to implement the second to implement the third. This would result in a multi-dimensional DAG (which is a lattice in which the nodes are EXORs of functions being all subsets of the set {A,B,C,D,E,F}).

An example of the implementation is given in Figures 2, 3 and 4. In this example, there is an initial phase to create the setup functions, which will be treated as the functions found in the top row of Figure 1. Figure 2a is the set of Karnaugh maps of desired output functions A, B, and C. The setup functions A', B', and C' need to be created as a result of the method of synthesis via quantum cascades (note that A' does not denote a negation of variable A). These setup functions are created by EXORing the desired output function with the function initially on the input line. For example, because function A will be implemented on the line that has initially value *a*, we perform a logical EXOR of the desired function A, with the initial function A to generate setup function A', as shown in Fig. 2b. Figure 3 shows the logical functions in their EXOR Lattice form. By implementing functions F1 and A', B' can be realized by F1 ⊕ A'. Furthermore, F3 ⊕ F1 gives a function F2, such that F2 ⊕ B' realizes C'. Therefore, by implementing A', F1, and F3, functions A', B' and C' can be realized through a series of EXORs. Notice that F1 and F3 are both rather easily implemented, whereas implementation of B' and C' would have been more costly. Coverings are first found for A', F1, and F3 by using an EXOR synthesis tool such as EXORcism. Such coverings are shown in Fig. 3.

Let us denote by ~x the negation of variable x. The A' is given by a⊕~abc, implementation is straightforward, and given in Fig. 4a. Figs 4(b, c) implement F1 and F3, respectively. These can be trivially cascaded, as shown in Fig. 4(d). To implement the remainder of the lattice, the lowest row is implemented first, as in Fig. 4(e), where F2 is implemented from an EXOR of F1 and F3. As functions A', F1, and F2 now exist, the next row to be implemented is B', and C', which are implemented in Figs. 4f and 4g. Finally, Fig. 4h realizes the final expected functions A, B, and C,

from setup functions A', B', and C'. The complete cascade is shown as Fig. 4i.

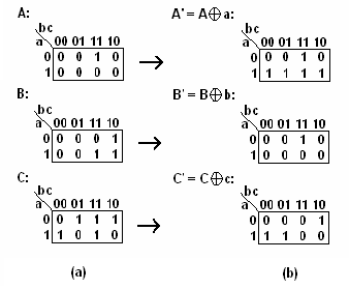


Figure 2: Setup Functions

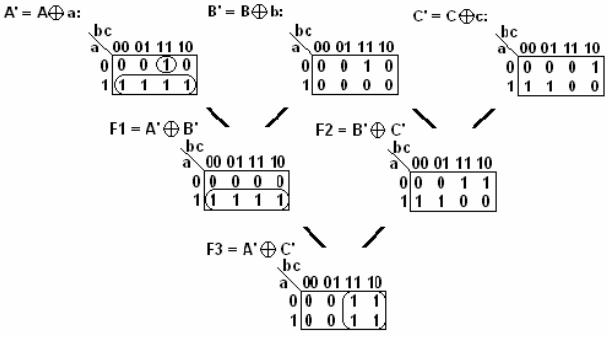


Figure 3: EXOR Lattice

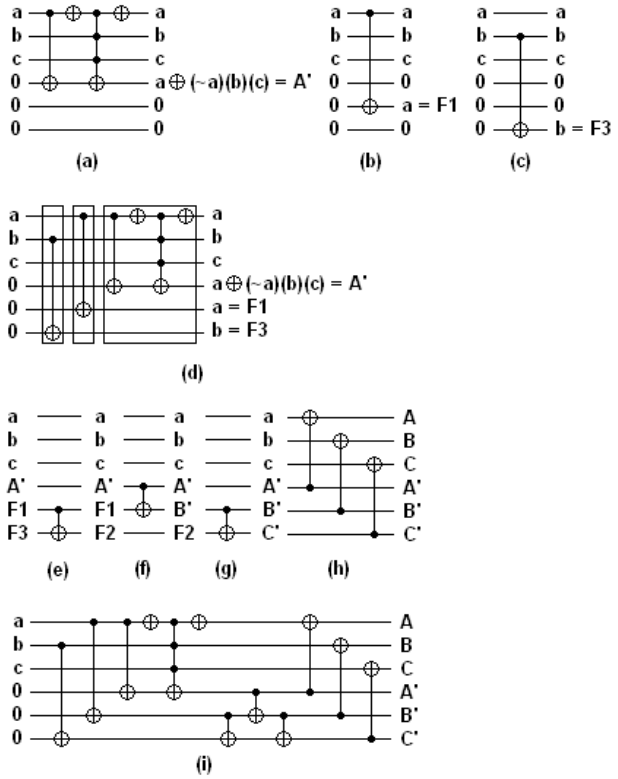


Figure 4. Implementation of EXOR Lattice.

The EXOR lattice can be implemented as several smaller lattices, rather than one large lattice as has been shown in Fig. 1. Note that the maximum necessary garbage inputs correspond directly to the number of the necessarily implemented lattice nodes. Under worst case, this will result in N additional garbage inputs for an N input, N output vector. In Fig. 4a, a lattice is shown where black and grey nodes represent functions that are either easily implementable or are repeated functions. Each shaded node selected as an initialization node (i.e., A', F1, F3 nodes, as from the previous example) either results in a reduction of complexity in implementation of the said node, or even allows ignoring the node altogether, if the node is either easily implemented or repeated, respectively.

In the example of Fig. 5a two smaller trees are selected over the larger tree in Fig. 5b. The solution can be further reduced, as in Fig. 5c by recognizing that overlapping lattices result in no added overhead. This reduces the solution found in Fig. 5b by one node, replacing the node by a shaded node. Choosing the smaller trees give the benefit of less nodes to traverse in order to implement the highest row nodes. The special case of no easily implemented function nodes or no repeated function nodes would result in a special case where the initial nodes may be best selected as the top row. In this case, the setup functions are directly implemented; this case is equivalent to a direct implementation of logic functions onto constant input lines.

Clearly, choosing the best EXOR lattices to implement is a non-trivial task, but can be simplified with good representation that can identify functions that are similar or dissimilar to one another. By grouping similar terms together, good heuristics can be used to relate similar node functions in an EXOR lattice.

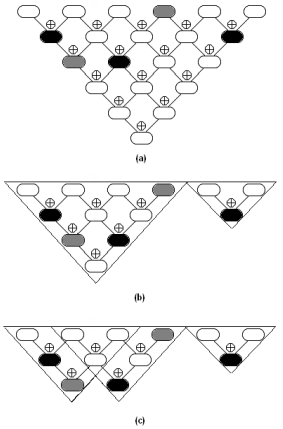


Figure 5: EXOR Lattice Selection

If adapting the MMD algorithm, incompletely specified functions must be first transformed to completely specified due to the necessity of ordering in the output vector. Optimization of the complete specification is a non-trivial and difficult task. However, in the EXOR lattice method, it is possible to trivially synthesize quantum cascades for incompletely specified functions without specifically defining don't-cares. An example is given in Fig. 6. Furthermore, by introducing the N constant-input lines, it is also trivial to implement irreversible functions. For example, Fig. 6, if we remove the variable "c" altogether (thereby removing the don't-care states), output function "A" is irreversible. Adding no additional overhead, the output vector can be implemented in exactly the same fashion. Final quantum array is shown in Fig. 7.

As the method has no dependency on ordering, the output vector needs not be of a fixed order; it can be any permutation of given outputs. This freedom introduces yet another degree of optimization; certain inputs selected for particular output lines may result in further optimization possibilities by changing the setup functions set. This problem, however, is, as of yet, beyond the scope of this work, and should be addressed in the future. It should be noted, in the example of Fig. 6, that the implementation of each setup function is not explicit; as each EXOR minterm of the setup function is realized, it is directly EXOR'ed to the expected output line. This is demonstrated as a different 'flavor' of synthesis, which may be necessary in some cases, when some function needs to be copied several times.

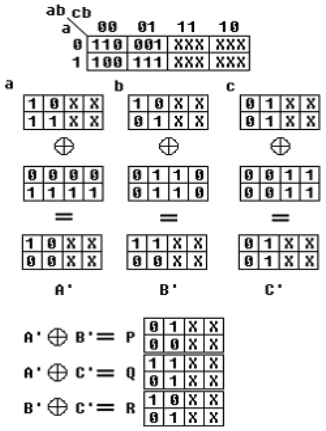


Figure 6: Incompletely Specified Function.

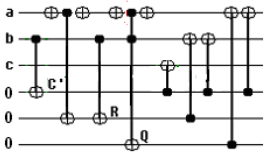


Figure 7. Quantum Circuit for the incompletely specified function from Figure 6

Lastly, it should be noted that, though we previously mentioned that this method assumed N input, M output, $1 \le M \le N$, it is possible to use the ancilla bits to synthesize as many outputs M such that $1 \le M \le 2N$. One special case worth mentioning is when the first N output functions ($N \le M \le 2N$) of M output functions are a reversible vector of outputs, and implemented on the non-constant input lines, and the remaining M-N outputs on the constant input lines. In this special case, because the first N outputs are a reversible set, it will be possible to directly implement the remaining M-N output set on the constant input lines. And, because this method allows for freedom of output permutation, it is a viable method of synthesis to select outputs in such a way.

III. EPOE AND EPOEM SYNTHESIS FOR SINGLE OUTPUT BOOLEAN FUNCTION

EXOR-sum of Products-of-EXOR-sums (EPOE) structures for single-output reversible logic and the synthesis methods for them are introduced in [15]. Although the concept of Product of EXOR-sums has been introduced by Luccio et al [20], the first algorithm for reversible circuit synthesis based on EXORs of such sub-functions comes from [8, 15].

A. EPOE synthesis for EPOEM-1s Algorithm for completely specified Boolean function:

As introduced in [8,15], EPOEM-1s is a synthesis algorithm for EPOE reversible circuits with no ancilla bit but the output line. EPOEM-1s algorithms convert a completely specified Boolean function of *N* input variables *(a,b,c,d, ...)* into EPOE form using a template-matching method. The templates are specific ordered sets of minterms that can be used in EPOE synthesis as in Table 1. Every template represents either an EXOR-sum of literals or a POE. The algorithms use a strategy of searching for a template T_i which intersects the function's ON-set in over *two-thirds* of the template's minterms $M(T_i)$,

more formally $|M(T_i) \cap ON| > \frac{2}{3}|M(T_i)|$ (1). The best matching template is the template that satisfies (1) and EXORed with the function to produce a remainder function with the smallest ON-set. The procedure of selecting the best matching template that satisfies (1) is iteratively applied until the remainder function becomes an empty set of minterms (which means the remainder function becomes 0). The algorithm requires a library of POE templates, calculated in advance and grouped together by the number of product terms employed.

TABLE 1. EXAMPLES OF POES TEMPLATES FOR FUNCTIONS OF THREE VARIABLES

Level	Minterms Covered (Template)	POE Template expressions	
		Single expression template library	Full expressions template library
0	{000,001,010,011,100,101,110,111}	1	1
1	{100,101,110,111}	a	a
1	{010,011,100,101}	a ⊕ b	a ⊕ b
1	{001,010,100,111}	a ⊕ b ⊕ c	a ⊕ b ⊕ c
2	{100,101}	a (a ⊕ b)	a (a ⊕ b) ; a (b ⊕ 1) ; (a ⊕ b) (b ⊕ 1)
2	{100,011}	(a ⊕ c) (a ⊕ b)	(a⊕c)(a⊕b) ; (a⊕c)(b⊕c⊕1) ; (a⊕b)(b⊕c⊕1)

For example, an EPOEM-1s synthesis will be perform on function $F(a,b,c,d) = \sum(0,1,6,10,11,12)$ as shown in Fig. 8.

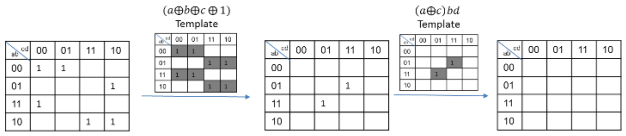


Figure 8. Karnaugh maps that illustrate an EPOEM-1s synthesis for function F

$$F = (a \oplus b \oplus c \oplus 1) \oplus b(a \oplus c)d$$

The corresponding circuit is given in Figure 9.

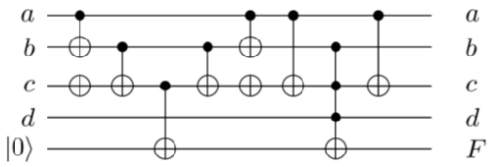


Figure 9. Circuit realization of function F using EPOEM-1s algorithm.

Compared with Exorcism-4 output, $F = b' \oplus c'd' \oplus a'd' \oplus ab'c'd$ and its circuit realization that has quantum cost of 48 (see Figure 10), it can be appreciated that EPOEM-1s result has a cost reduced by 54.1% with respect to the Exorcism-4 result. Exorcism-4 has a higher cost because of using multi-input Toffoli gates.

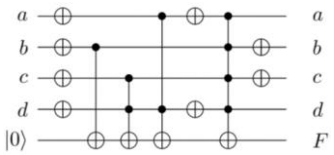


Figure 10. Standard ESOP-like circuit realization of function F(a, b, c, d) = b' ⊕ c'd'⊕ a'd' ⊕ ab'c'd with only Toffoli and NOT gates

B. *EPOE synthesis for EPOEM-1-DC Algorithm for incompletely specified Boolean function:*

The EPOEM-1-DC algorithm is a modification of EPOEM-1s algorithm so it can also applied to an incompletely specified Boolean function. The EPOEM-1-DC converts a Boolean function of *N* input variables *(a,b,c,d, ...)* into EPOE form using a template matching method. It uses a strategy of searching for a POE template (T_i) which satisfies two conditions:

1. The template intersects the function's ON-set and don't care set (DC_set) in over *1/2* of the template's minterms (M(T_i)), more formally: $|M(T_i) \cap (ON \cup DC)| > \frac{1}{2}|M(T_i)|$,
2. The matching template is EXOR-ed with the function's ON-set and DC_set to produce a remainder function with **a smaller ON-set** (the template chosen must have the number of '1' (minterms in the ON_set) covered in the template to be greater than the number of '0' in that template).

The best matching template is one that satisfies both conditions and creates a smallest ON_set. The procedure of selecting the best matching template that satisfies both conditions is iteratively applied until the ON_set becomes an empty set. The algorithm requires a library of POE templates calculated in advance like in EPOEM-1[15]. Below a complete EPOEM-1-DC synthesis will be performed on function $F_1(a,b,c,d) = \sum(0,2,4,5,9,12) + d(3,6,15)$, assuming the 4-variable POE single expression template library calculated in advance.

1.The *ON_set* is initialized to the ON-set = {0000, 0010, 0100, 0101, 1001, 1100}.
2.DC_set = {0011, 0110, 1111}.
3.The Checkset = ON_set EXORed DC_set = {0000, 0010, 0100, 0101, 1001, 1100, 0011, 0110, 1111}

4. At *level* = 0 the *Checkset* has 9 minterms. Even though $9 > \left(\frac{1}{2}\right)2^4$, but the new ON_set becomes {0001, 0111, 1000, 1010, 1101, 1110} which cannot made the ON_set smaller , so function is not negated. Consequently a *"1"* is not appended to the *result*.

5. At *level* = 1 a POE template library search is performed and the templates $(a\oplus b\oplus c\oplus d\oplus 1)$ is found to intersect with 7 minterms in the *Checkset*, and the templates $(a\oplus 1)$ and $(a\oplus b\oplus c)$ are found to intersect with 6 minterms in the *Checkset* which is shown in Figure 11. Because 7 *and* $6 > \left(\frac{1}{2}\right)2^3$ so all templates are satisfied the cover condition.

- With templates $(a\oplus b\oplus c\oplus d\oplus 1)$, $(a\oplus 1)$ and $(a\oplus b\oplus c)$ been selected, *(ON_set – template)* is EXOR-ed with *(template – Checkset)*, the new *ON_set* becomes {0010, 0100, 1010}, {0001, 0111, 1001, 1100} and {0000, 1000, 1100} respectively. Both template $(a\oplus b\oplus c\oplus d\oplus 1)$ and $(a\oplus b\oplus c)$ are produce a new ON_set with three minterms (smaller than the old ON_set with six minterms), so both templates are acceptable. Therefore $(a\oplus b\oplus c)$ is randomly selected and its expression is appended to the *result*.
- The new *Checkset* becomes {0000, 1000, 1100, 0011, 0110, 1111}
- At *level* = 2 a POE template library search is performed and the template $(a\oplus b\oplus 1)(c\oplus d\oplus 1)$ is found to intersect with 4 minterms in the *Checkset*, and the templates $(c\oplus 1)(d\oplus 1)$ and $(a)(c\oplus d\oplus 1)$ are found to intersect with 3 minterms in the *Checkset* which is shown in Figure 12. Because 4 *and* $3 > \left(\frac{1}{2}\right)2^2$ so all templates are satisfied the cover condition.
- With templates $(a\oplus b\oplus 1)(c\oplus d\oplus 1)$, $(c\oplus 1)(d\oplus 1)$and $(a)(c\oplus d\oplus 1)$ been selected, *(ON_set – template)* is EXOR-ed with *(template – Checkset)*, the new *ON_set* becomes {1000}, {0100} and {0000, 1011} respectively. Both template $(a\oplus b\oplus 1)(c\oplus d\oplus 1)$ and $(c\oplus 1)(d\oplus 1)$ are produce a new ON_set with one minterms (smaller than the old ON_set with three minterms), so both templates are acceptable. Therefore $(a\oplus b\oplus 1)(c\oplus d\oplus 1)$ is randomly selected and its expression is appended to the *result*.
- The new *Checkset* becomes {1000, 0111, 1110, 0101}.
- At *level* = 3 a POE template library search is performed and many acceptable templates are found with equal quantum cost with 2 intersecting minterms and create the empty *ON_set*. The template $(a\oplus b)(a\oplus c)(d\oplus 1)$ is randomly selected among them, as shown in Figure 12. Its expression is appended to the *result*.
- The new *ON_set* becomes {} which completes the synthesis. The final value of the *result* is as follows: $F_1(a,b,c,d) = (a\oplus b\oplus c)\oplus (a\oplus b\oplus 1)(c\oplus d\oplus 1)\oplus(a\oplus b)(a\oplus c)(d\oplus 1)$

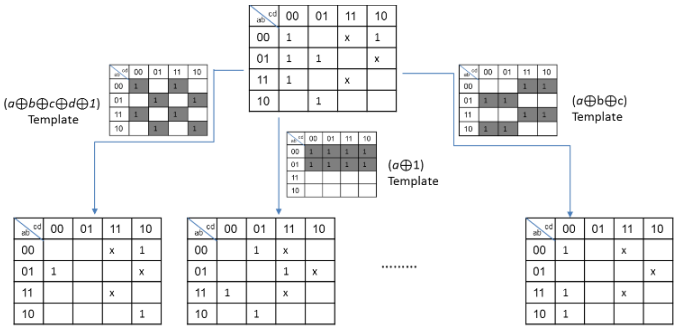


Fig. 11. Karnaugh maps that illustrate the Checkset function comparisons of different templates in EPOEM-1-DC at *level* = 1.

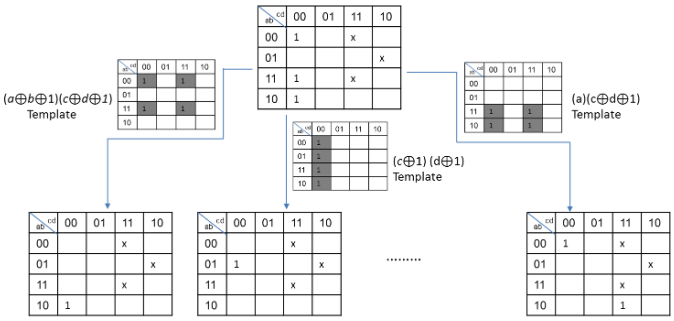


Fig. 12. Karnaugh maps that illustrate the Checkset function comparisons of different templates in EPOEM-1-DC at *level* = 2.

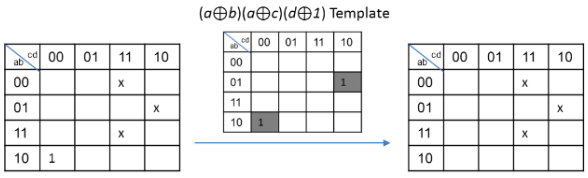


Fig. 13. Karnaugh maps that illustrate the Checkset function evolution in EPOEM-1-DC.

- The corresponding circuit is given in Figure 6

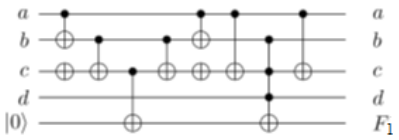


Fig. 14. Circuit realization of function F_1 using EPOEM-1-DC algorithm.

IV. THE EPOEM-MO ALGORITHM

A. *EPOEM-MO-1 Algorithm:*

The EPOEM-MO-1 Algorithm which is a combination of EPOEM-1s (or EPOEM-1-DC) with EXOR-lattice concept for the synthesis of multiple output Boolean functions is shown below. An example of EPOEM-MO-1 synthesis is presented subsequently.

EPOEM-2-MO algorithm for m-input n-output Boolean function:

Stage 1: Create the EXOR lattice for n given output functions $(A, B, C, \dots, A\oplus B, A\oplus C, \dots)$

Stage 2: Apply EPOEM-1s for each function in the EXOR lattice to find the EPOE expression and its quantum cost.

Stage 3: Select n nodes which have lowest cost to realize n given output functions

B. *Example: EPOE synthesis for EPOEM-MO-1 algorithm*

Below a complete EPOEM-MO-1 synthesis will be performed on function as shown in Figure 15.

abcd	ABCD
0000	0100
0001	1111
0010	0110
0011	1101
0100	0000
0101	1110
0110	1010
0111	1001
1000	1100
1001	0101
1010	0001
1011	0111
1100	1000
1101	1011
1110	0010
1111	0011

A

ab \ cd	00	01	11	10
00		1	1	
01		1	1	1
11	1	1		
10	1			

B

ab \ cd	00	01	11	10
00	1	1	1	1
01		1		
11				
10	1	1	1	

C

ab \ cd	00	01	11	10
00		1		1
01		1		1
11		1	1	1
10			1	

D

ab \ cd	00	01	11	10
00		1	1	
01			1	
11		1	1	
10		1	1	1

Fig. 15. Example function that was used for EPOEM-MO-1 synthesis

Stage 1: Create the EXOR lattice for n setup functions $(A' = A\oplus a, B' = B\oplus b, \dots)$ as shown in Figure 16

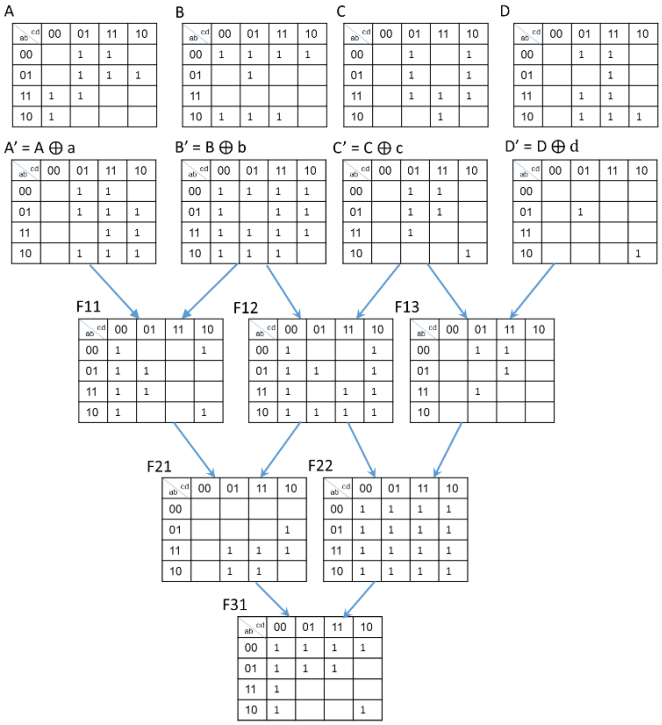


Figure 16. EXOR-lattice of function given in Figure 15

Stage 2: Apply EPOEM-1s for each of the node to find the EPOE expression and the quantum cost of each node.

Node	Cost	EPOE expressions
A'	27	$d\oplus c(d\oplus 1)\oplus(a\oplus c)(b\oplus c)(c\oplus d)$
B'	20	$1\oplus(a\oplus d)(b\oplus c)(c\oplus d)$
C'	24	$(a\oplus 1)d\oplus a(b\oplus c)(c\oplus d)$
D'	19	$(a\oplus d)(b\oplus c)(c\oplus d)$
F11 $(A'\oplus B')$	10	$b(c\oplus d)(d\oplus 1)$
F21 $(A'\oplus C')$	20	$ad\oplus bc(d\oplus 1)$
F12 $(B'\oplus C')$	23	$1\oplus(a\oplus b\oplus 1)d\oplus bcd$
F13 $(C'\oplus D')$	22	$(a\oplus 1)d\oplus b(c\oplus 1)d$
F22 $(B'\oplus D')$	1	1
F31 $(A'\oplus B'\oplus C'\oplus D')$	25	$(a\oplus 1)\oplus a(d\oplus 1)\oplus bc(d\oplus 1)$

Table 2. EPOE expression and quantum cost of each node in EXOR-lattice

Stage 3: Select n nodes which has lowest cost to generate n setup functions if the setup functions do not have the lowest cost.

- From the four single nodes (A', B', C', D'), node D' has a lowest cost so it is selected.
- Base on Table 3, node B' can be generated by using node F22 $(B' \oplus D')$ and node F22 has lower cost than node B', so node F22 is selected.
- Node B' can be generated by using node F22 $(B' \oplus D')$, so node F11 $(A' \oplus B')$ is selected to generate node A' because it has the lower cost than node A'
- Node A' can be generated by using node F11 $(A' \oplus B')$, so node F21 $(A' \oplus C')$ is selected to generate node C' because it has the lower cost than nodes C', F13 $(C' \oplus D')$, F12 $(B' \oplus C')$ and F31 $(A' \oplus B' \oplus C' \oplus D')$
- So, the four nodes which have the lowest cost to generate A', B', C', D' and E' are: D', F_{11}, F_{21} and F_{22}.

Node	Cost
F22 $(B' \oplus D')$	1
F11 $(A' \oplus B')$	10
D'	19
B'	20
F21 $(A' \oplus C')$	20
F13 $(C' \oplus D')$	22
F12 $(B' \oplus C')$	23
C'	24
F31 $(A' \oplus B' \oplus C' \oplus D')$	25
A'	27

Table 3. Sorted nodes list based on its cost.

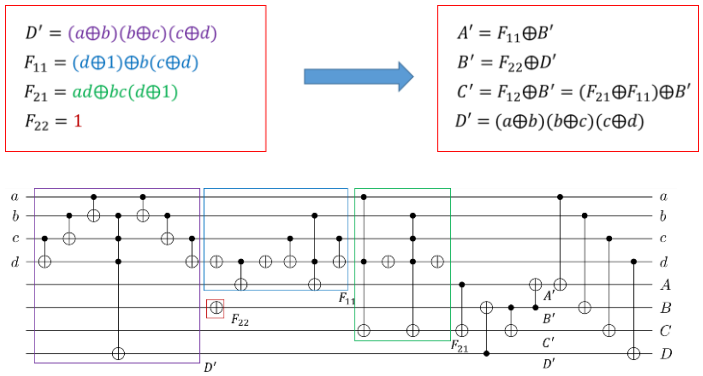


Figure 17. Circuit realization of function given in Figure 15.

The corresponding circuit is given in Figure 17 with quantum cost of 58.

- Given the state encoding table as in Table 5:

Table 5. States encode with 1-out-of-4 code

V. Automata Realization System

Given an example system as shown in Table 4

Present state	Next state				Output
	00	01	10	11	
S0	S0	S1	S2	S3	1
S1	S0	S3	S1	S3	0
S2	S1	S3	S2	S0	1
S3	S1	S0	S0	S3	0

Table 4. Example system for synthesis with EPOEM-MO-1

State	Encode
S0	1000
S1	0100
S2	0010
S3	0001

Table 5. States encode with 1-out-of-4 code

- Replace internal states in Table 4 with codes define in Table 5 which gives the new Table 6

Present state	Next state				Output
	00	01	10	11	
1000	1000	0100	0010	0001	1
0100	1000	0001	0100	0001	0
0010	0100	0001	0010	1000	1
0001	0100	1000	1000	0001	0

Table 6. Example system with new encoding states

- Apply EPOEM-MO-1 algorithm for this function (6-inputs 5-outputs):

In		Out	
Present state (abcd)	Input (ef)	Next State (ABCD)	Output (E)
1000	00	1000	1
1000	01	0100	1
1000	10	0010	1
1000	11	0001	1
0100	00	1000	0
0100	01	0001	0
0100	10	0100	0
0100	11	0001	0
0010	00	0100	1
0010	01	0001	1
0010	10	0010	1
0010	11	1000	1
0001	00	0100	0
0001	01	1000	0
0001	10	1000	0
0001	11	0001	0

In		Out	
Present state (abcd)	Input (ef)	Next State (ABCD)	Output (E)
0000	--	----	-
11--	--	----	-
1-1-	--	----	-
1--1	--	----	-
-11-	--	----	-
-1-1	--	----	-
--11	--	----	-
111-	--	----	-
1-11	--	----	-
11-1	--	----	-
-111	--	----	-
1111	--	----	-

Table 7. Example system for synthesis with new encoding state

Lists of the EPOE expressions and quantum costs of each node after applying the first two stage of EPOEM-MO-1, as in Table 8.

- Select 5 nodes from Table 6 that have lowest cost and can generate $A', B', C', D'$and E'.

Node	Cost	EPOE expressions
A'	48	$1\oplus(b\oplus c\oplus e\oplus f\oplus 1)\oplus(a\oplus b\oplus e\oplus 1)(e\oplus f\oplus 1)\oplus(a\oplus b\oplus c\oplus 1)(e\oplus 1)(f\oplus 1)$
B'	17	$1\oplus(b\oplus 1)\oplus(b\oplus e\oplus 1)(a\oplus f\oplus 1)$
C'	29	$1\oplus(c\oplus 1)\oplus(b\oplus d\oplus 1)(a\oplus c\oplus e\oplus 1)(f\oplus 1)$
D'	46	$1\oplus(d\oplus f\oplus 1)\oplus(c\oplus e\oplus 1)(a\oplus c\oplus d\oplus f\oplus 1)\oplus(a\oplus c\oplus d\oplus 1)(e\oplus 1)(f\oplus 1)$
E'	8	$1\oplus(a\oplus c\oplus e\oplus 1)$
F12 ($A'\oplus B'$)	51	$1\oplus(d\oplus e\oplus 1)\oplus(a\oplus 1)(f\oplus 1)\oplus(a\oplus c\oplus 1)(e\oplus 1)\oplus(a\oplus b\oplus c\oplus 1)(e\oplus 1)(d\oplus f\oplus 1)$
F13 ($A'\oplus C'$)	59	$1\oplus(b\oplus c\oplus e\oplus 1)\oplus(c\oplus 1)(b\oplus d\oplus f\oplus 1)\oplus(c\oplus d\oplus 1)(b\oplus e\oplus 1)\oplus(a\oplus c\oplus d\oplus 1)(b\oplus e\oplus 1)(f\oplus 1)$
F15 ($A'\oplus E'$)	46	$1\oplus(a\oplus b\oplus f\oplus 1)\oplus(a\oplus b\oplus e\oplus 1)(e\oplus f\oplus 1)\oplus(a\oplus b\oplus c\oplus 1)(e\oplus 1)(f\oplus 1)$
F23 ($B'\oplus C'$)	46	$1\oplus(d\oplus f\oplus 1)\oplus(c\oplus 1)(a\oplus d\oplus e\oplus 1)\oplus(a\oplus c\oplus 1)(d\oplus e\oplus 1)(b\oplus d\oplus f\oplus 1)$
F24 ($B'\oplus D'$)	59	$1\oplus(b\oplus c\oplus e\oplus 1)\oplus(c\oplus 1)(b\oplus d\oplus f\oplus 1)\oplus(c\oplus d\oplus 1)(b\oplus e\oplus 1)\oplus(a\oplus c\oplus d\oplus 1)(b\oplus e\oplus 1)(f\oplus 1)$
F34 ($C'\oplus D'$)	51	$1\oplus(d\oplus e\oplus 1)\oplus(a\oplus 1)(f\oplus 1)\oplus(a\oplus c\oplus 1)(e\oplus 1)\oplus(a\oplus b\oplus c\oplus 1)(d\oplus f\oplus 1)(e\oplus 1)$
F35($C'\oplus E'$)	31	$1\oplus(a\oplus e\oplus 1)\oplus(b\oplus d\oplus 1)(a\oplus c\oplus e\oplus 1)(f\oplus 1)$
F45 ($D'\oplus E'$)	52	$1\oplus(a\oplus c\oplus d\oplus e\oplus f\oplus 1)\oplus(c\oplus e\oplus 1)(a\oplus c\oplus d\oplus f\oplus 1)\oplus(a\oplus c\oplus d\oplus 1)(f\oplus 1)(e\oplus 1)$
F1234 ($A'\oplus B'\oplus C'\oplus D'$)	19	$(a\oplus b\oplus c\oplus 1)\oplus(a\oplus b\oplus 1)\oplus(a\oplus b\oplus d\oplus 1)$
F2345 ($B'\oplus C'\oplus D'\oplus E'$)	46	$1\oplus(a\oplus b\oplus f\oplus 1)\oplus(a\oplus b\oplus e\oplus 1)(e\oplus f\oplus 1)\oplus(a\oplus b\oplus c\oplus 1)(e\oplus 1)(f\oplus 1)$

Table 8. List of EPOE expressions and theirs quantum costs of each node

Node	Cost
E'	8
B'	17
F1234 ($A'\oplus B'\oplus C'\oplus D'$)	19
C'	29
F35($C'\oplus E'$)	31
D'	46
F15 ($A'\oplus E'$)	46
F23 ($B'\oplus C'$)	46
F2345 ($B'\oplus C'\oplus D'\oplus E'$)	46
A'	48
F12 ($A'\oplus B'$)	51
F34 ($C'\oplus D'$)	51
F45 ($D'\oplus E'$)	52
F13 ($A'\oplus C'$)	59
F24 ($B'\oplus D'$)	59

Table 9. Sorted nodes list by cost

From Table 7, nodes B' and E' have the lowest cost (in top 5 lowest cost), so nodes B' and E' are selected.

- Nodes B' and E' are selected so node F25($B'\oplus E'$) is not needed and removed because nodes B' and E' are already available. Now node C' have the cost in top 5 lowest cost and it cannot be generated from nodes F1234 and nodes F1345 which have lower cost than it, so node C' is selected.
- Two more nodes have to be selected so it can generate A' and D'.
  - To generate A', node with A' need to be selected, for e.g. F12, F13, F15, F1235 and node A'. Node F15 ($A'\oplus E'$) is selected with the lowest cost of 46. As node A' can generate from node F15, node D' can be generate by selecting node F1234 ($A'\oplus B'\oplus C'\oplus D'$) with the cost of 19
  - To generate D', node with D' need to be selected, for e.g. F24, F34, F45, F2345 and node D'. Node D' is selected with the lowest cost of 46. As node D' is selected, node A' can be generate by selecting node F1234 ($A'\oplus B'\oplus C'\oplus D'$) with the cost of 19.
- The way of selecting D' is chosen because it can save the cost of generating A' from node F15 (the way of selecting F15 to generate A').

So, the five nodes which have the lowest cost to generate $A', B', C', D'$and E' are: $F_{1234}, B', C', D'$and E'.

$$F_{1234} = (a \oplus b \oplus c \oplus 1) \oplus (a \oplus b \oplus 1) \oplus (a \oplus b \oplus d \oplus 1)$$
$$B' = 1 \oplus (b \oplus 1) \oplus (b \oplus e \oplus 1)(a \oplus f \oplus 1)$$
$$C' = 1 \oplus (c \oplus 1) \oplus (b \oplus d \oplus 1)(a \oplus c \oplus e \oplus 1)(f \oplus 1)$$
$$D' = 1 \oplus (d \oplus f \oplus 1) \oplus (c \oplus e \oplus 1)(a \oplus c \oplus d \oplus f \oplus 1) \oplus (a \oplus c \oplus d \oplus 1)(e \oplus 1)(f \oplus 1)$$
$$E' = 1 \oplus (a \oplus c \oplus e \oplus 1)$$

$$A' = F_{1234} \oplus B' \oplus C' \oplus D'$$
$$B' = 1 \oplus (b \oplus 1) \oplus (b \oplus e \oplus 1)(a \oplus f \oplus 1)$$
$$C' = 1 \oplus (c \oplus 1) \oplus (b \oplus d \oplus 1)(a \oplus c \oplus e \oplus 1)(f \oplus 1)$$
$$D' = 1 \oplus (d \oplus f \oplus 1) \oplus (c \oplus e \oplus 1)(a \oplus c \oplus d \oplus f \oplus 1) \oplus (a \oplus c \oplus d \oplus 1)(e \oplus 1)(f \oplus 1)$$
$$E' = 1 \oplus (a \oplus c \oplus e \oplus 1)$$

The corresponding circuit is given in Figure 18 with quantum cost of 127.

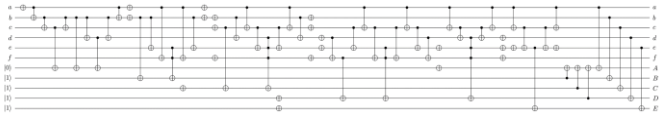


Figure 18. Circuit realization of permutative quantum automaton from Table 4. A feedback loop from outputs A, B, C, D to inputs a, b, c, d, respectively is not shown. This loop includes potentially some flip-flops. Signals e and f are primary inputs and signal E is an output of the machine

VI. Experimental results

EPOEM programs have been implemented in Python and tested extensively on Unix and Windows workstations. The experimental results below were obtained on a 2.9 GHz Intel Core i7 PC under Microsoft Windows 8.1. To verify and compare EPOEM-MO-1 algorithm, multiple-output benchmark functions were taken from Revlib's and Maslov's pages [6,7]. The method is compared with recent works, [20, 21,22,23,24]. Comparisons have been made using 32 benchmarks, out of which EPOEM-MO-1 method provides better result for 28, Table 10. The results show that the proposed technique results in a very significant reduction in quantum cost.

VII. Conclusion and future work

We presented a new approach to synthesize reversible/quantum automata using various error-detecting and error-correcting codes, which lead to irreversible excitation and output functions with many don't cares. These functions are next synthesized to optimized EPOE circuits with EXOR Lattices and compared for their quantum cost. Interfacing 'real world'-interpretable logic with probabilistic spirit of quantum computing requires some form of intermediate description language, for instance multiple-valued logic [3,5]. In this paper, binary logic is used as the most important, developed, and approachable, both by reason of the reduction of complexity in only 2-states, and due to the maturity of binary logic in standard logic synthesis. However, all the presented methods are general and can be relatively easily adapted to ternary or quaternary logic. The reversibility aspect of binary quantum circuits is a necessary characteristic for interfacing between quantum logic and real-world interpretation, as quantum operators (which are what each quantum binary gate is composed of) are intrinsically reversible. A new synthesis method for strongly irreversible incomplete functions to reversible automata is presented that combines binary EPOE circuits and EXOR Lattices. This synthesis method, at the cost of introducing at most N additional garbage bits, can make up for several drawbacks of MMD and similar methods. Non-reversible functions and incompletely specified functions can be realized using the EXOR lattice method of synthesis. Additionally, there are no ordering limitations. The method in essence uses EXOR logic to conglomerate all of the output functions and chooses some optimal subset of functions with which to implement them. The setup functions can be generated randomly among easily realizable high-exor-component functions and the algorithm may be iteratively applied for various setup functions. This is one of topics for future research.

References

[1] V. S. Shivgand, A. Aulakh, and M. Perkowski, "Quantum Circuit Layout," Proc. RM 2005.

[2] W. Hung, X. Song, M. Perkowski, "Reachability Analysis for reversible minimization." Proc. DAC 2004. June 2004.

[3] M.H.A. Khan, M.A. Perkowski, M.R. Khan, and P. Kerntopf, "Ternary GFSOP Minimization using Kronecker Decision Diagrams and Their Synthesis with Quantum Cascades", Accepted to Journal of Multiple-Valued Logic and Soft Computing: Special Issue to Recognize T. Higuchi's Contribution to Multiple-Valued VLSI Computing.

[4] M. Lukac, M. Perkowski, H. Goi, M. Pivtoraiko, C.H. Yu, K. Chung, H. Jee, B-G. Kim, and Y-D. Kim,

"Evolutionary approach to Quantum and Reversible Circuits synthesis", Artificial Intelligence Review, 20, pp 361-417, 2003.

[5] N. Denler, B. Yen, M. Perkowski, "Synthesis of Reversible Circuits from a Subset of Muthukrishnan-Stroud Quantum Multi-Valued Gates", IWLS 2004.

[6] D. Maslov, Improved quantum cost for n-bit Toffoli gates, Electronic Letters, 2003, IET.

[7] W. Hung, X. Song, M. Perkowski, "Reachability Analysis for reversible minimization." Proceedings of DAC 2004. June 2004.

[8] Schaeffer, B., Tran, L., Gronquist, A., Perkowski, M., Kerntopf, P.: Synthesis of Reversible Circuits Based on Products of Exclusive Or Sum. ISMVL, 35–40 (2013)

[9] F. Luccio and L. Pagli: On a new Boolean function with applications. IEEE Trans. Comput., vol. 48, no. 3, 296-310 (1999)

[10] A. Mishchenko, M. Perkowski: Logic Synthesis of Reversible Wave Cascades. IWLS, 197–202 (2002)

[11] A. Mishchenko, M. Perkowski: Fast Heuristic Minimization of Exclusive Sum-of-Products. Proc. 5th RM, 242-250 (2001)

[12] RevLib - An Online Resource for Reversible Functions and Circuits. http://revlib.org/

[13] D. Maslov: Reversible Logic Synthesis Benchmarks Page. http://webhome.cs.uvic.ca/~dmaslov/

[14] A. Barenco, C.H. Bennett, R. Cleve, D.P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J.A. Smolin, H. Weinfurter: Elementary Gates for Quantum Computation. Phys Review A, 52, 3457–3467 (1995)

[15] L. Tran, B. Schaeffer, A. Gronquist, M. Perkowski, P. Kerntopf: Synthesis of Reversible Circuits Based on EXORs of Products of EXORs In: Trans. on Comput. Sci. XXIV, LNCS 8911, pp. 111–128 (2014)

[16] M. Perkowski , T. Ross, D. Gadd , J. A. Goldman, N. Song: Application of ESOP Minimization in Machine Learning And Knowledge Discovery, RM 1995, Japan.

[17] M.A. Nielsen, and I.L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, January 2011.

[18] M. Lukac, M. Kameyama, M. Perkowski, *Quantum Finite State Machines - A Circuit Based Approach,* International Journal of Unconventional Computing, Vol. 9, Num. 3-4, 2013, pp. 267-301

[19] H. Thapliyal, and N. Ranganathan, Design of reversible sequential circuits optimizing quantum cost, delay and garbage outputs, ACM Journal on Emerging Technologies, 2010.

[20] K. Fazel, M.Thornton, and J. E. Rice,"ESOP based Toffoli gate cascade generation", PACRIM, pp. 206–209, 2007

[21] Y. Sanaee, G. W. Dueck, ESOP-based Toffoli network generation with transformations, in: ISMVL, IEEE Computer Society, 2010, pp. 276–281.

[22] J. E. Rice and N. M. Nayeem, "Ordering techniques for ESOP-based toffoli cascade generation", PacRim2011, pp. 274 – 279, 2011.

[23] C. Bandyopadhyay, D. Roy, D. K. Kole, K. Dutta, H. Rahaman: ESOP-based synthesis of reversible circuit using improved cube list, Proc. ISED, pp 26-30 (2013)

[24] C. Bandyopadhyay, H. Rahaman, R. Drechesler "A Cube Pairing Approach for Synthesis of ESOP-based Reversible Circuit", ISMVL-2014, pp. 109-114.

TABLE 10. EXPERIMENTAL RESULTS FOR QUANTUM COST COMPARISON

*****SYMBOL REPRESENT THOSE CASES WHERE WE GOT BETTER RESULTS WITH RESPECT TO OTHER APPROACHES

Function	In	Out	Quantum Cost					
			EPOEM-MO-1	FTR-07 [20]	SD-10 [21]	[22]	[23]	[24]
3_17_6	3	3	25*			26	27	
ex1_82	3	3	15*					
4mod7	4	3	80*			167	110	
f2	4	4	77*	255		246	160	116
4_49_7	4	4	90*			201	174	
aj-e1_81	4	4	50*			201	167	
hwb4	4	4	62					
wim	4	7	141*	217	211	218	172	150
dc1	4	7	185*	416		454	241	
cm42a	4	10	152*	377				
pm1	4	10	152*	377		270		290
c17	5	2	87	99		81	85	
cm82a	5	3	46*	154		143	103	69
rd53	5	3	103*	265	219	269	200	136
hwb5	5	5	346					
squar5	5	8	235*	442		465	393	
c7552_119	5	16	319*	1728		942	992	
decod	5	16	545*	1728		673		976
sqr6	6	12	711*	1033	744			
hwb6	6	6	994					
con1	7	2	175	206		206	171	
rd73	7	3	280*	1143	836	1150	1022	
sqn	7	3	401*	2122				1183
z4	7	4	118*	642		642	448	260
z4ml	7	4	118*	642	577	642	448	260
ham7	7	7	70					
hwb7	7	7	2740					
inc	7	9	1700*	2140				
5xp1	7	10	1144	1430	998	1165		
rd84	8	4	1686*	2749		2558	2477	
sqr8	8	4	704	622		616	604	
radd	8	5	247*	676		669	618	349
adr4	8	5	240*	727		618	618	
dist	8	5	5075*	7601				
root	8	5	2601*	3443				
dc2	8	7	2031*	1886				
misex1	8	7	550*	982		1012	725	
hwb8	8	8	3744					
mlp4	8	8	2932*	3753				
urf2	8	8	10363					

Semi-supervised Bayesian Classification by vector features with continuous and discrete components

Vasily O. Vasilyev
Moscow Institute of Physics and Technology
(State University)
Moscow, Russian Federation, 117303
Email: vasily.o.vasilyev@gmail.com

Alexander V. Dobrovidov
V.A. Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences
Moscow, Russian Federation, 117997
Email: dobrovidov@gmail.com

***Abstract*—A solution for the traditional Bayesian classification problem in non-traditional conditions is proposed, when the distributions and a priori probabilities of classes are unknown, but a trained sample from the zero class (labeled positive) and mixed sample (unlabeled) are available. Mixed sample will be employed in the learning to restore mixed distribution and as a test sample for constructed classifier. The case with vector features containing continuous and discrete components is considered. To restore unknown distributions nonparametric kernel techniques with data-driven bandwidth are used. A new algorithm for estimating the prior probability of zero class is given using positive labeled and unlabeled samples. This allows to find a good approximation of optimal threshold for the modified Bayesian classification algorithm. Numerical verification confirms the effectiveness of the proposed classification technique even in cases of strong overlapping of class distributions.**

I. Introduction

This paper is devoted to solution of traditional problem of Bayesian classification under unconventional assumptions. We will consider the case of a dichotomy because a generalization to several classes is not particularly difficult (see [1]). Let random variable $Y \in \{0,1\}$ denotes the class number and it takes its values with probability $p = \mathbf{P}\{Y = 0\}$ and $1-p = \mathbf{P}\{Y = 1\}$. In some practical situations the observable feature vector V may be composed from two vector-valued variables $V = (X, Z)$, where $X = (X_1, \ldots, X_{n_c}) \in \mathbb{R}^{n_c}$ is a continuous random component, and $Z \in \{z = (z_1, \ldots, z_{n_d}) : z_i \in \{z_{i,j} \in \mathbb{R}\}_{j=1}^{M_i}, M_i \in \mathbb{N}, i = 1, \ldots, n_d\}$ is a discrete random component. Joint distribution of vectors V and variable Y can be written as

$$\begin{aligned}\mathbf{P}_{(V,Y)} &= \mathbf{P}_{(X,Z,Y)} = \mathbf{P}_{(Y)}\mathbf{P}_{(X,Z|Y)} \\ &= \mathbf{P}\{Y = 0\}\mathbf{P}_{(X,Z|Y)}\{*, * \mid Y = 0\} \\ &\quad + \mathbf{P}\{Y = 1\}\mathbf{P}_{(X,Z|Y)}\{*, * \mid Y = 1\},\end{aligned}$$

where $*$ means an event in corresponding probability space. Here by definition $\mathbf{P}_{(X,Z|Y)}\{*, * \mid Y = 0\} = \mathbf{P}\{B, Z = z \mid Y = 0\}$, where B is a Borel set on the probability space $(\mathbb{R}^{n_c}, \mathcal{B}, \mathbf{P})$ with σ-algebra $\mathcal{B}$ on $\mathbb{R}^{n_c}$ and an event $\{Z = z\}$ is defined above. If $B = \{X(\omega) \le x\}, x \in \mathbb{R}^{n_c}$, then

$$\begin{aligned}\nabla_x \mathbf{P}\{X \le x, Z = z \mid Y = 0\} \\ = \mathbf{P}\{Z = z \mid Y = 0\} f_0(x \mid z),\end{aligned}$$

where notation $f_0(x \mid z) = f_X(x \mid Z = z, Y = 0)$ is a conditional density of continuous observation x given the events $\{Z = z\}$ and $\{Y = 0\}$. Analogous formula for $f_1(x \mid z)$ is valid only with respect to the events $\{Z = z\}$ and $\{Y = 1\}$. Let $y \in \{0, 1\}$ be a realization of r.v. Y. For brevity, introduce functions

$$\begin{aligned}g_0(x, z) &= \mathbf{P}\{Z = z \mid Y = 0\} f_0(x \mid z), \\ g_1(x, z) &= \mathbf{P}\{Z = z \mid Y = 0\} f_1(x \mid z).\end{aligned}$$

Then the optimal Bayesian decision-making procedure is constructed as

$$\hat{y} = \begin{cases} 1, & \text{if } (1-p)g_1(x, z) \ge p g_0(x, z), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

One can see that this optimal decision can work only when the joint distribution $\mathbf{P}_{(X,Z,Y)}$ of all r.v. is completely known. But it is very rare case and usually in applications this distribution is unknown. Therefore, more than 50 years ago, first works appeared concerning the problems of classification learning (pattern recognition) under incomplete probabilistic information. Training was mainly conducted on the basis of statistical methods using marked samples from a classes (labeled) and/or samples without indicating a class (unlabeled). Different types of incompleteness of the initial statistical information generate different approaches for solving these problems. Many of these approaches are not statistical at all and sequence of observable data can be quite arbitrary (e.g. texts or gene). In this case, separation of classes will not be based on statistical criteria, but, for example, on clustering analysis based on the logical and structural properties of the observation objects. However, as noted in [2], these methods are largely heuristic and their characteristics are strongly dependent on the way of determination of distance between objects in the feature space.

The Bayesian learning assumes the existence of a mixed unlabeled sample

$$v^u = (x_i^u, z_i^u)_{i=1}^{n_u}, \quad (2)$$

where $n_u \in \mathbb{N}$, drawn from mixed distribution density

$$g(x, z) = p g_0(x, z) + (1-p) g_1(x, z) \quad (3)$$

The peculiarity of this sample is that it provides the proportions of observations from each class that correspond to the a

priori probabilities of the classes. The corresponding proportions, and therefore the estimates of the a priori probabilities of classes, are usually constructed from a labeled sample of two classes, if the latter is available. Such information is the basis for learning procedures in the most classification works in the eighties and nineties [3], [4].

Currently, more attention is paid to classifying with an unknown sample from an alternative class, i.e. when the conditional density $g_1(x,z)$ and a priory probability $(1-p)$ are unknown and a sample from this class is unavailable. This situation arises in practice (for example, in sonar problems, in econometrics), when a labeled mixed sample (v,y) is not observed and it is impossible to restore its joint a distribution. In the statistical approach, the following types of uncertainty arise:

1) *Unsupervised approach*, when different statistical information may be absent (from unknown parameters to unknown class distributions), but only a mixed unlabeled sample is available for training [5].
2) *Semi-supervised approach* (e.g. PU learning), when all densities and a priori probabilities are unknown, but a labeled sample of the zero class (corresponding to unknown density $g_0(x,z)$) and a mixed unlabeled sample v^u are available (see [6], [7], [8], [9], [10]).
3) *Supervised approach*, when all probabilistic characteristics are unknown, but a marked mixed sample from both classes is available [4].

The algorithm proposed in this paper refers to the third type of uncertainty. In a number of works using the semi-supervised approach [11], two difficulties were noted: this is a lack of information for constructing an a priori probability and the absence of a labeled sample of the negative class. Therefore, attempts were made to select observations of the zero class from the mixed sample. This was done sometimes by clustering methods that do not follow the strong Wald's theory of solutions. We give a non-heuristic, but rather accurate statistical estimate of the a priori probability of the zero class. In contrast to the well-known papers in this area [1], the present algorithm sets and solves the optimization problem to search for an estimator of this a priori probability. Such an estimate allows us to choose a threshold close to the optimum in the modified Bayesian decision function. In addition to observations (2) let us define labeled positive sample (drawn from zero class) as

$$v^p = (x_i^p, z_i^p)_{i=1}^{n_p} \tag{4}$$

where $n_p \in \mathbb{N}$ is a size of sample.

II. Developed method

A. Estimator $\hat{y}$

Let us rewrite (3) as

$$(1-p)g_1(x,z) = g(x,z) - pg_0(x,z)$$

and then modify (1) like

$$\hat{y} = \begin{cases} 1, & \text{if } g(x,z) - 2pg_0(x,z) \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

which will be the base of the proposed method. In the next sections estimation of functions $g(x,z), g_0(x,z)$ and a priori probability p are considered.

B. Estimators of $g(x,z)$ and $g_0(x,z)$

Remind that $n_d \geq 0, n_c \geq 0$ and $n_d + n_c > 0$. There are 3 possible cases: 1. if $n_d = 0$, then

$$g(x,z) = f_X(x),$$
$$g_0(x,z) = f_X(x \mid Y = 0);$$

2. if $n_c = 0$, then

$$g(x,z) = \mathbf{P}\{Z = z\},$$
$$g_0(x,z) = \mathbf{P}\{Z = z \mid Y = 0\};$$

3. if $n_d \neq 0, n_c \neq 0$, then

$$g(x,z) = \mathbf{P}\{Z = z\} f_X(x \mid Z = z),$$
$$g_0(x,z) = \mathbf{P}\{Z = z \mid Y = 0\} f_0(x \mid z).$$

Let introduce following sets

$$I_u(z) = \{i \in \{1, \cdots, n_u\} : z_i^u = z\},$$
$$I_p(z) = \{i \in \{1, \cdots, n_p\} : z_i^p = z\}.$$

For estimating probabilities:

$$\mathbf{P}\{Z = z\} \approx \frac{|I_u(z)|}{n_u},$$
$$\mathbf{P}\{Z = z \mid Y = 0\} \approx \frac{|I_p(z)|}{n_p}$$

are proposed, where $|\bullet|$ is the cardinality of set $\bullet$. For estimating probability density functions:

$$f_X(x) \approx f\left(x \mid (x_i^u)_{i=1}^{n_u}\right),$$
$$f_X(x \mid Y = 0) \approx f\left(x \mid (x_i^p)_{i=1}^{n_p}\right),$$
$$f_X(x \mid Z = z) \approx f\left(x \mid (x_i^u)_{i \in I_u(z)}\right),$$
$$f_0(x \mid z) \approx f\left(x \mid (x_i^p)_{i \in I_p(z)}\right)$$

are used, where notation $f(x \mid (x_i)_{i=1}^n)$ is the multivariate kernel density estimator (MKDE) in the point x constructed by training set $(x_i)_{i=1}^n$. The next section is devoted to MKDE.

C. Estimator $f(x \mid (x_i)_{i=1}^n)$

There are a lot of kernel density estimators and approaches to configure them. In this section is a one of the possible combinations of them, which includes two steps: pilot and subtle estimators. Firstly, for density $f(x \mid (x_i)_{i=1}^n)$ next fixed kernel estimator

$$\tilde{f}(x) = \tilde{f}(x \mid (x_i)_{i=1}^n) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

is applied, where $K(\cdot)$ is the kernel function, usually some probability density function with zero mean; h is the bandwidth (tuning parameter), $h > 0$; $x, x_i \in \mathbb{R}^{1\times d}, d \in \mathbb{N}$. Probability density function of multivariate normal distribution with zero mean and identity covariance matrix

$$\phi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{xx^\top}{2}\right)$$

as the kernel function $K(\cdot)$. One of methods of calculating the bandwidth h is applying of unbiased cross-validation (UCV)(see [12], [13]). Procedure UCV leads to an estimator

$$\hat{h} = \underset{h>0}{\operatorname{argmin}}\, \mathrm{UCV(h)},$$

with minimization function

$$\begin{aligned}&\mathrm{UCV}(h)\\ &= \frac{1}{n(n-1)h^d} \sum_{i=1}^{n} \sum_{\substack{j=1,\\ j\neq i}}^{n} \frac{1}{2^{d/2}} \phi\left(\frac{x_i - x_j}{\sqrt{2}h}\right) - 2\phi\left(\frac{x_i - x_j}{h}\right)\\ &\qquad + \frac{1}{nh^d},\end{aligned}$$

Computing minima analytically is a challenge, so a numerical calculation is popular. The function $\mathrm{UCV}(h)$ often has multiple local minima, therefore more correct way is to use brute-force search to find $\hat{h}$, however it is a very slow algorithm. In [14] it was shown that spurios local minima are more likely at too small values of h, so we propose to use golden section search between 0 and h^+, where

$$h^+ = \left(\frac{4}{n(d+2)}\right)^{\frac{1}{d+4}} \max_{i,j\in\{1,\ldots,d\}} \sqrt{|\hat{\mathbf{S}}_{i,j}|},$$

where $\hat{\mathbf{S}}$ is the sample covariance matrix of set $(x_i)_{i=1}^n$. To improve accuracy of estimator of (6) in the second step using more flexible approach is considered. Constructing estimator with fixed kernel is the first step in the methods with 'adaptive' kernel like *balloon* estimators (see [15]) and *sample point* estimators (see [16], [17]). Silverman in [18] explored Abramson's implementation and proposed following estimator

$$f(x \mid (x_i)_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(h\lambda_i)^d} K\left(\frac{x - x_i}{h\lambda_i}\right) \tag{7}$$

with local bandwidth factors

$$\lambda_i = \left\{\frac{\tilde{f}(x_i)}{g}\right\}^{-1/2},$$

where g is the geometric mean of the $\tilde{f}(x_i)$

$$g = \left\{\prod_{i=1}^{n} \tilde{f}(x_i)\right\}^{1/n}.$$

Silverman noted that using similar bandwidth h for (6) and (7) gives good results.

D. Estimator $\hat{p}$

Rewrite (3) as

$$\frac{g(x,z)}{g_0(x,z)} = p + \frac{(1-p)g_1(x,z)}{g_0(x,z)}.$$

If functions $g(x,z)$ and $g_0(x,z)$ were known then simple estimator of p

$$\bar{p} = \min_{x,z} \frac{g(x,z)}{g_0(x,z)}. \tag{8}$$

would give good results: if there is some $(x,z)_0$ such that $\frac{(1-p)g_1(x,z)}{g_0(x,z)} \approx 0$, then estimator $\bar{p}$ will be close to p. Since $g(x,z)$ and $g_0(x,z)$ are unknown then estimators are substituted instead of them. In this case estimator of $\bar{p}$ is unreasonable, because if only for one point value of $\frac{\hat{g}(x,z)}{\hat{g}_0(x,z)} \approx 0$, then the estimator $\bar{p}$ will be too undervalued in comparison with p. Therefore, it is necessary to introduce some cumulative characteristics which will have less influence of particular (x,z). As such characteristics we propose

$$S(\tilde{p}) = \sum_z \int_x \left(\tilde{p}g_0(x,z) - g(x,z)\right)^+ dx, \tag{9}$$

$$\hat{S}(\tilde{p}) = \sum_z \int_x \left(\tilde{p}\hat{g}_0(x,z) - \hat{g}(x,z)\right)^+ dx, \tag{10}$$

where $(a)^+ = \max(0,a)$ and $\tilde{p} \in [0,1]$ is some parameter. The variable $S(\tilde{p})$ is an area between two functions $g_0(x,z)\tilde{p}$ and $g(x,z)$, where $g_0(x,z)\tilde{p}$ exceeds $g(x,z)$. The meaning of the estimator $\hat{S}(\tilde{p})$ is the same. It is easy to show that $S(\tilde{p}) = 0$ for $\tilde{p} \in [0,\bar{p}]$ and $0 < S(\tilde{p}) \leq \tilde{p}$ for $\tilde{p} \in (\bar{p},1]$, where $\bar{p}$ is defined in (8). It means that $S(\tilde{p})$ changes its first and second derivatives at the point $\bar{p}$. We expect the similar changes in derivatives of the estimator $\hat{S}(\tilde{p})$. Therefore, we propose to use point $\tilde{p}$, where the curvature of the function $\hat{S}(\tilde{p})$

$$\kappa(\tilde{p}) = \frac{|\hat{S}''(\tilde{p})|}{(1 + (\hat{S}'(\tilde{p}))^2)^{3/2}}$$

reaches maximum, as a final estimator of p:

$$\hat{p} = \underset{0\leq\tilde{p}\leq 1}{\operatorname{argmax}}\, \kappa(\tilde{p}).$$

Properties of the estimator $\hat{p}$ are investigated by computer simulation. Note, that method Monte-Carlo (e.g. [19]) can help to calculate (10) as

$$\begin{aligned}\hat{S}(\tilde{p}) &= \sum_z \int_x \left(\tilde{p}\hat{g}_0(x,z) - \hat{g}(x,z)\right)^+ dx\\ &= \sum_z \int_x \left(\tilde{p} - \frac{\hat{g}(x,z)}{\hat{g}_0(x,z)}\right)^+ \hat{g}_0(x,z)dx\\ &\approx \frac{1}{n_p} \sum_{i=1}^{n_p} \left(\tilde{p} - \frac{\hat{g}(x_i^p, z_i^p)}{\hat{g}_0(x_i^p, z_i^p)}\right)^+.\end{aligned}$$

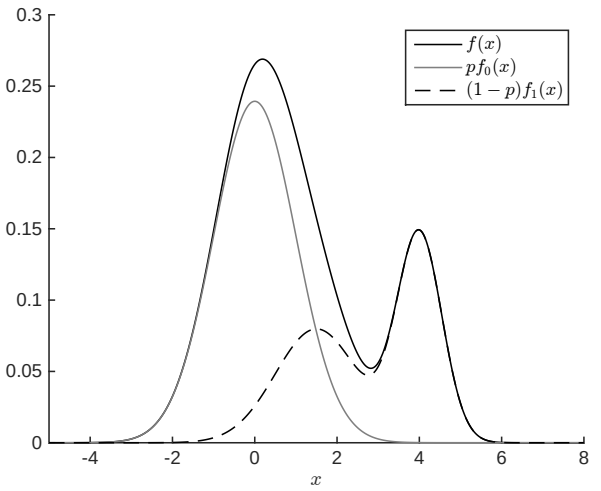


Fig. 1. Black line is $f(x)$; gray line is $pf_0(x)$; dashed line is $(1-p)f_1(x)$.

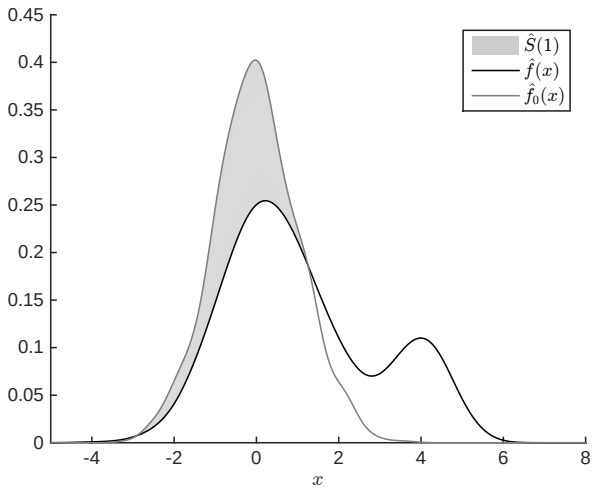


Fig. 2. Black line is density $\hat{f}(x)$; gray line is density $\hat{f}_0(x)$; gray area between two functions $\hat{f}_0(x)$ and $\hat{f}(x)$, where $\hat{f}_0(x)$ exceeds $\hat{f}(x)$.

TABLE I
SIMULATION RESULTS IN EXAMPLE 1 AFTER 50 LAUNCHES

Probabilities		Classification errors, %		
p	$\hat{p}$	Optimal method	Proposed method	Difference
0.6	0.6055	14.43	14.74	0.31

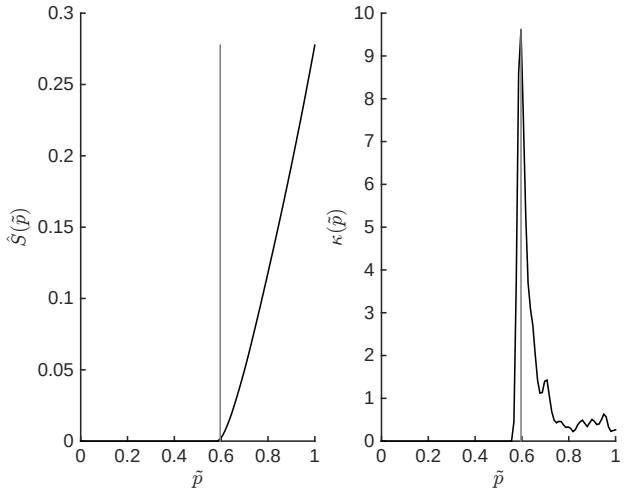


Fig. 3. Left plot: black line is $\hat{S}(\tilde{p})$. Right plot: black line is function $\kappa(\tilde{p})$; grey line shows that estimator $\hat{p} = 0.596$.

III. EXAMPLES

A. Example 1 (the feature is univariate continuous r.v.)

Let true a priori probability $p = 0.6$, $n_c = 1$ and $n_d = 0$, i.e. discrete component of feature z is absent then

$$g_0(x,z) = f_0(x) = \varphi(x;0,1),$$
$$g_1(x,z) = f_1(x) = 0.5\varphi(x;1.5,1) + 0.5\varphi(x;4,0.3),$$
$$g(x,z) = f(x) = pf_0(x) + (1-p)f_1(x),$$

where $\varphi(x;\mu,\sigma)$ is pdf of normal distribution with mean μ and covariance matrix σ. Functions $f(x), pf_0(x)$ and $(1-p)f_1(x)$ are presented on Fig. 1. Positive sample size $n_p = 1000$ and unlabeled sample size $n_u = 1000$. See estimators $\hat{f}(x)$ and $\hat{f}_0(x)$ on Fig. 2. From Fig. 3 it follows that estimator $\hat{p}$ is very close to real value of a priori probability $p = 0.6$. This experiment shows high quality of the proposed method in this one-dimensional case (see Table I). 50 repeated simulations of the experiment took 71.7 seconds[1].

B. Example 2 (the feature is two-dimensional continuous r.v.)

Let a priori probability $p = 0.7, n_c = 2$ and $n_d = 0$, as in the first example, then

$$g_0(x,z) = f_0(x) = \varphi(x;\mu_0,\sigma_0),$$
$$\mu_0 = \begin{pmatrix} 0 & 0 \end{pmatrix}, \quad \sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$
$$g_1(x,z) = f_1(x)$$
$$= 0.5\varphi(x;\mu_{1,1},\sigma_{1,1}) + 0.5\varphi(x;\mu_{1,2},\sigma_{1,2}),$$
$$\mu_{1,1} = \begin{pmatrix} 1.5 & 2 \end{pmatrix}, \quad \sigma_{1,1} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix},$$
$$\mu_{1,2} = \begin{pmatrix} -1.5 & -1.5 \end{pmatrix}, \quad \sigma_{1,2} = \begin{pmatrix} 0.3 & -0.2 \\ -0.2 & 4 \end{pmatrix},$$
$$g(x,z) = f(x) = pf_0(x) + (1-p)f_1(x),$$

where $x \in \mathbb{R}^2$. Functions $f(x)$ is presented on Fig. 4. Positive sample size $n_p = 2000$ and unlabeled sample size $n_u = 2000$. The optimal and proposed classifications are shown on Fig. 5 and Fig. 6. Despite of mismatch of optimal and developed classification boundaries their mean errors are very close to each other and differs less than 1% (see Table II).

C. Example 3

Let true a priori probability $p = 0.3, n_d = 1$ and $n_c = 1$ then

[1]Software: MATLAB R2014b, macOS Sierra 10.12.4. Hardware: MacBook Pro Model A1502 (Retina, 13-inch, Mid 2014).

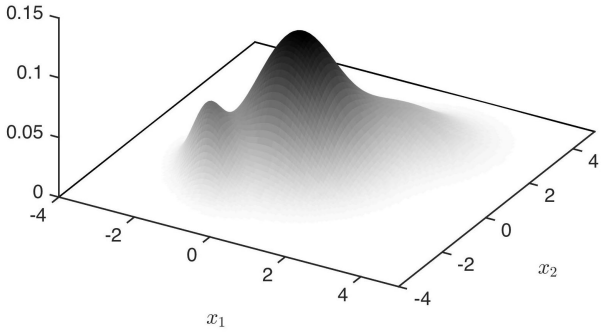


Fig. 4. Two-dimensional true mixed density $f(x) = f(x_1, x_2)$, where the left pick and the right hill describe the first class and the center pick corresponds to the zero class.

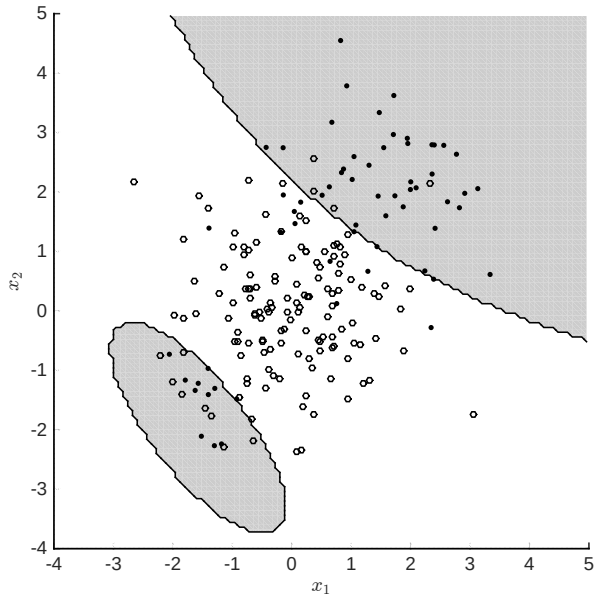


Fig. 5. Optimal classification with known characteristics. Black line is a boundary between zero class (white area) and the first class (grey area); circles belong to the zero class and dots belong to the first class. The classification error is 13.0%.

TABLE II
SIMULATION RESULTS IN EXAMPLE 2 AFTER 50 LAUNCHES

Probabilities		Classification errors, %		
p	$\hat{p}$	Optimal method	Proposed method	Difference
0.7	0.6826	12.14	12.83	0.69

$$
\begin{aligned}
g_0(x,1) &= 0.4\varphi(x;0,1.4),\\
g_0(x,2) &= 0.2\varphi(x;4,1.5),\\
g_0(x,3) &= 0.4\varphi(x;3,0.7),\\
g_1(x,1) &= 0.2\varphi(x;5,0.6),\\
g_1(x,2) &= 0.5\varphi(x;1,1),\\
g_1(x,3) &= 0.3\varphi(x;3,2).
\end{aligned}
$$

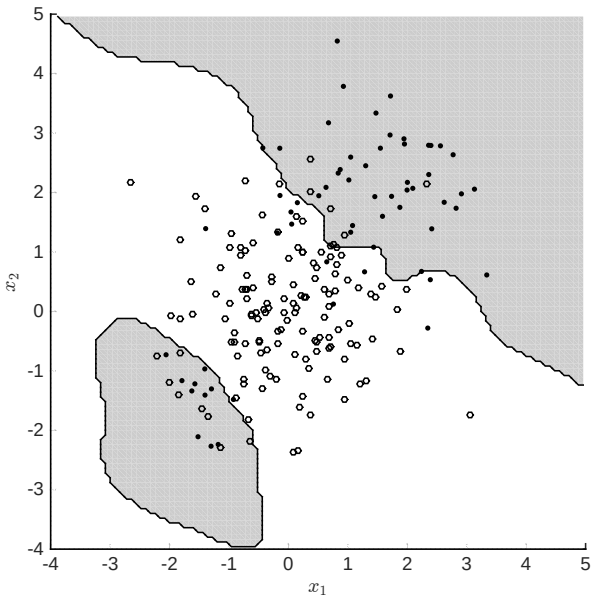


Fig. 6. Proposed classification with unknown characteristics. Black line is a boundary between zero class (white area) and the first class (grey area); circles belong to the zero class and dots belong to the first class. The classification error is 14.0%.

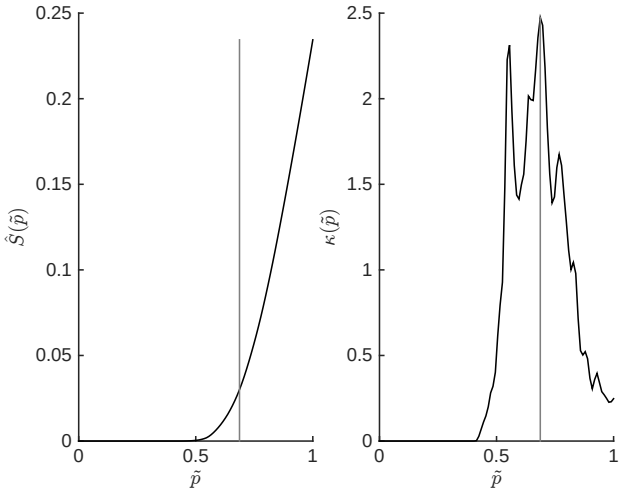


Fig. 7. Left plot: black line is $\hat{S}(\tilde{p})$. Right plot: black line is function $\kappa(\tilde{p})$; grey line shows that estimator $\hat{p} = 0.687$.

Functions $g_0(x,z)$ and $g_1(x,z)$ are presented on Fig. 8. Size of positive sample $n_p = 2000$ and unlabeled sample $n_u = 2000$.

Experiment results with multivariate signal features demonstrate high precision of the proposed classification method including fairly well estimation of a class prior probability. Developed program allows in principle to construct classification algorithm under any finite features dimensionality with continuous and discrete components.

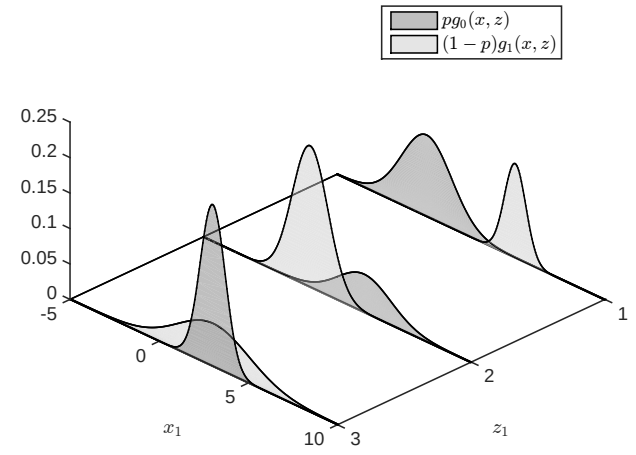


Fig. 8. Dark gray area is $pg_0(x,z)$ and light gray is $(1-p)g_1(x,z)$.

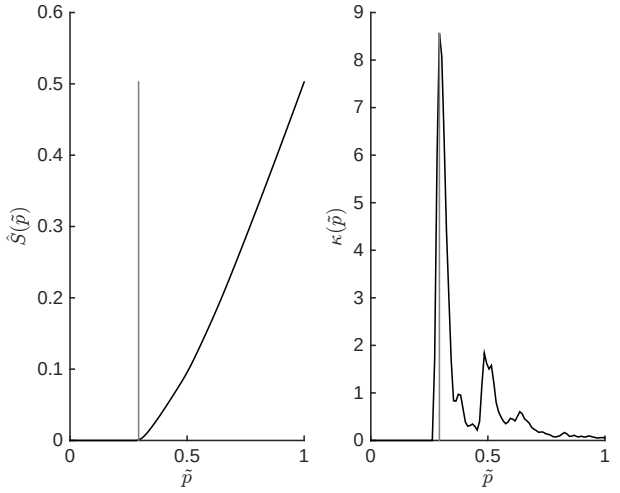


Fig. 9. Left plot: black line is $\hat{S}(\tilde{p})$. Right plot: black line is function $\kappa(\tilde{p})$; grey line shows that estimator $\hat{p} = 0.293$.

TABLE III
SIMULATION RESULTS IN EXAMPLE 3 AFTER 50 LAUNCHES

Probabilities		Classification errors, %		
p	$\hat{p}$	Optimal method	Proposed method	Difference
0.3	0.3073	11.91	13.52	1.61

IV. CONCLUSION

A multidimensional Bayesian version of semi-supervised classification problem was considered in this paper. A distinctive peculiarity of the work is 1) the use of observable vector features containing continuous and discrete components and 2) the new method of constructing fairly accurate estimators of the a priori class probabilities based on PU samples using adaptive kernel estimates of multidimensional densities with tuning parameters. This approach to solve the classification problems under nonparametric uncertainties may be very useful in practice especially since a learning mode with unknown useful signal is not reachable. It should be noted that the quality of the proposed decision rule is very close to the quality of optimal decision except the cases when the optimum itself cannot well separate a one class of signals from another one.

REFERENCES

[1] K. Fukunaga, *Statistical pattern recognition*. Academic Press, 1990.
[2] E. Sansone, F. G. B. D. Natale, and Z.-H. Zhou, "Efficient training for positive unlabeled learning," 2016. [Online]. Available: https://arxiv.org/abs/1608.06807
[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
[4] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[5] D. T. Pham and G. A. Ruz, "Unsupervised training of bayesian networks for data clustering," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 465, no. 2109. The Royal Society, 2009, pp. 2927–2948.
[6] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.
[7] Z.-H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.
[8] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," *ICML*, vol. 3, pp. 448–455, 2003.
[9] P. Yang, X.-L. Li, J.-P. Mei, C.-K. Kwoh, and S.-K. Ng, "Positive-unlabeled learning for disease gene identification," *Bioinformatics*, vol. 28, no. 20, pp. 2640–2647, 2012.
[10] J. T. Zhou, S. J. Pan, Q. Mao, and I. W. Tsang, "Multi-view positive and unlabeled learning." in *ACML*, 2012, pp. 555–570.
[11] J. He, Y. Zhang, X. Li, and Y. Wang, "Naive bayes classifier for positive unlabeled learning with uncertainty," *SIAM SDM SIAM Conference on Data Mining*, pp. 361–372, 2010.
[12] M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, vol. 9, pp. 65–78, 1982.
[13] A. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 7, pp. 353–360, 1984.
[14] P. Hall and J. Marron, "Local minima in cross-validation functions," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 53, pp. 245–252, 1991.
[15] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1049–1051, 06 1965.
[16] L. Breiman, W. Meisel, and E. Purcell, "Variable kernel estimates of multivariate densities," *Technometrics*, vol. 19, no. 2, pp. 135–144, 1977.
[17] I. S. Abramson, "On bandwidth variation in kernel estimates-a square root law," *Ann. Statist.*, vol. 10, no. 4, pp. 1217–1223, 12 1982.
[18] B. Silverman, *Density estimation for statistics and data analysis*, ser. Monographs on statistics and applied probability. London: Chapman and Hall Boca Raton, 1986.
[19] J. Hammersley, "Monte carlo methods for solving multivariable problems," *Ann. New York Acad. Sci.*, pp. 844–874, 1960.

Agent-Based Power Equilibrium in a Smart Grid with XBOOLE

Eric MSP Veith and Bernd Steinbach
Institute of Computer Science
TU Bergakademie Freiberg, Freiberg, Germany
Email: {veith, steinb}@informatik.tu-freiberg.de

***Abstract*—The share of power generated from renewable energy sources is steadily increasing. For many countries, this means an increase in power from wind or photovoltaics, which are often highly volatile. The development of the power grid to the smart grid introduces a vast amount of new sensory data that is available to plan power generation and consumption in the face of volatile generation and consumption. The vast amount of data available and new grid control paradigms such as micro grids suggests to enrich power grid control with de-centralized aspects. We propose a decentralized smart grid control architecture in which each node keeps its local power balance at an equilibrium through communication with and offers from other nodes. The system's heart piece, the power balance, uses a Boolean model to describe the combinatorial problem of picking the most efficient set of offers to solve a power disequilibrium. It employs the XBOOLE system for efficient construction and solving of the required formulae.**

***Index Terms*—Smart grid, XBOOLE, power grid control, combinatorial problem solver**

I. Introduction and Related Work

The future power grid will rely strongly on renewable energy sources [12]. Since wind power and solar energy are the most widely available sources of renewable energy, they will contribute the greatest share in most countries. However, both primary energies rely on a phenomenon that is not controllable by humanity: the weather. Both renewable energy sources exhibit a volatile power generation behavior [14]. The site considerations of wind farms and solar power plants lead to a vastly distributed generation.

Therefore, local forecasting will increase to cope with the weather conditions [8, 9, 13, 14]. Together with a tighter integration of the customer, this will lead to an increased amount of information processing in order to maintain a stable power grid in the face of both variable power consumption *and* variable power generation.

Bush [3] describes that the power grid and communication networks have much in common, and proposes to bridge the two worlds by introducing formulae that describe the influence of the communication network on the power grid, and vice versa. He does not introduce a concrete metric, though.

Optimization methods for the power grid—or parts thereof—that take different and changing load situations and power generation into account have been proposed, e.g., by Inoue et al. [7]. These approaches focus on a centralized controller for optimization and need to transfer load forecast data to the controller in order for it to work. Inoue et al. use Binary Decision Diagrams (BDDs) [2] to model the power grid. Edge-Valued Multi-valued Decision Diagrams (EVMDDs) [10] exist as a data structure akin to BDDs that can represent any p-valued function and therefore find consideration in this particular case.

In order to ease the information pressure on the control centers, we proposed to equip each node in the power grid with a software agent. The distributed software agents will act according to the protocol specified by us in the Lightweight Power Exchange Protocol (LPEP) [18, 19]; distributed generation and micro grids are supported by the semantics of the protocol. All agents are pari passu and work together to retain the power equilibrium, without a-priori knowledge of the respective power grid node types of the other agents. Since the agents are equal in terms of duty, function, and purpose, i.e., there is no discrimination of different types of agents, this particular agent is called *Universal Smart Grid Agent*.

An agent forecasting a power disequilibrium sends out a request for power or to consume the predicted surplus of power and subsequently receives offers from other agents. How the agent models its view on the grid and selects the appropriate offers to return to a power equilibrium has been shortly described in [17].

In this paper, we propose an efficient modelling of the agents' power balances through XBOOLE [1, 15, 16]. The rest of the paper is structured as follows: Section II introduces the concept of the per-agent power balance. Section III formally states the nature of the problem the proposed algorithm solves. In Section IV, we show how the power balance, the power disequilibrium, and responses by other agents are modeled in the Boolean domain and how XBOOLE is used to create a solution to the combinatorial problem of choosing the correct set of responses to return to a power equilibrium. Section V evaluates the efficiency of the approach, compares it to values from literature, and shows how we further optimize our approach. It presents the final solver algorithm. In Section VI, we summarize the results of this paper.

II. Agent-Local Power Balance

The Universal Agent models its environment in two forms: Its neighborhood that is constituted of its fellow agents it communicates with, along with their messages, and the current state and future state of its own locality, the power balance.

Each agent keeps account of the power generation and consumption of its local node. It can be very much seen

as a 'power ledger.' However, the goal of the agent is to maintain a power equilibrium at all times and insofar as the term 'power ledger' is similar to the financial term, 'balance' is more appropriately used here. The agent's power balance stores mappings of time intervals to power values, i.e.,

$$[t_1; t_2) \mapsto P \quad \Leftrightarrow \quad \tilde{t} \mapsto P. \tag{1}$$

With regards to the desired equilibrium, an agent must fundamentally consider each entry in the power balance as a *requirement*: Each mapping individually constitutes a power imbalance, i.e., a disequilibrium, and therefore requires the agent to act. The power balance is indeed then balanced when individual requirements are matched in such a way that they equalize each other; thus demand and supply are counterbalanced. The power balance is, therefore, the keystone of the agent's fundamental goal. By denoting the requirement from the i-th agent as r_i, we can write:

$$r_i : \tilde{t}_i \mapsto P_i \ . \tag{2}$$

Internally, the agent represents the power balance as an interval map, extending the notion of an interval set with mappings, using the relevant mathematical operations on sets as well as defining operations on interval overlap:

$$\{[t_1; t_2) \mapsto P_1\} \cup \{[t_1; t_2) \mapsto P_2\} = \{[t_1; t_2) \mapsto P_1 + P_2\} \ , \tag{3}$$

$$\{[t_1; t_2) \mapsto P_1\} \setminus \{[t_1; t_2) \mapsto P_2\} = \{[t_1; t_2) \mapsto P_1 - P_2\} \ . \tag{4}$$

Partial overlaps are calculated analogously. Intervals of four points in time, $t_1 < t_2 < t_3 < t_4$, result in:

$$\{[t_1; t_3) \mapsto P_1\} \cup \{[t_2; t_4) \mapsto P_2\} = \\ \{[t_1; t_2) \mapsto P_1, [t_2; t_3) \mapsto P_1 + P_2, [t_3; t_4) \mapsto P_2\} \ , \tag{5}$$

$$\{[t_1; t_3) \mapsto P_1\} \setminus \{[t_2; t_4) \mapsto P_2\} = \\ \{[t_1; t_2) \mapsto P_1, [t_2; t_3) \mapsto P_1 - P_2, [t_3; t_4) \mapsto -P_2\} \ . \tag{6}$$

Defining the 'subset' and 'subset-or-equal' relationships of individual mappings is equally feasible:

$$[t_1; t_2) \mapsto P_1 \ \subset \ [t_3; t_4) \mapsto P_2 \\ \Leftrightarrow \quad t_1 > t_3 \ \wedge \ t_2 < t_4 \ \wedge \ P_1 < P_2 \ , \tag{7}$$

$$[t_1; t_2) \mapsto P_1 \ \subseteq \ [t_3; t_4) \mapsto P_2 \\ \Leftrightarrow \quad t_1 \geq t_3 \ \wedge \ t_2 \leq t_4 \ \wedge \ P_1 \leq P_2 \ . \tag{8}$$

The Universal Agent software wraps one timespan-to-power mapping in a `Requirement` object, as depicted in Fig. 1. This class connects power messages to forecasts and contains factory methods to create an LPEP message object from a forecast and vice versa. `Requirement` objects also allow us to keep track of which requirement originated at the local node and which was sent by another agent, which is crucial for solving a disequilibrium.

III. The Combinatorial Demand-Supply Problem

In order to solve the (unbalanced) power balance, it makes use of a `PowerBalanceSolverStrategy`. Classes implementing this interface represent an actual algorithm for solving the power balance's disequilibrium in a given interval.

This disequilibrium, denoted by the power value P_0 in the time interval $\tilde{t}_0$, has been forecasted by an agent:

$$r_0 : [t_{0,1}; t_{0,2}) \mapsto P_0 \quad \Leftrightarrow \quad r_0 : \tilde{t}_0 \mapsto P_0 \ . \tag{9}$$

This agent formulates a request as an LPEP message, which, through selective broadcasting, reaches other agents. These agents then check—and, in fact, must check—whether they can help to solve the disequilibrium, and if so, answer with individual LPEP messages:

$$r_1 : \tilde{t}_1 \mapsto P_1, r_2 : \tilde{t}_2 \mapsto P_2, \ldots, r_i : \tilde{t}_i \mapsto P_i \ , \tag{10}$$

$$\forall P_i, i \neq 0 : |P_i| \leq |P_0| \ , \tag{11}$$

$$\forall \tilde{t}_i, i \neq 0 : t_{i,1} \geq t_{0,1} \wedge t_{i,2} \leq t_{0,2} \ . \tag{12}$$

The first and foremost task of the solver is now to find any combination of mappings, $\tilde{t}_1 \mapsto P_1, \ldots, \tilde{t}_i \mapsto P_i$, such that the initial disequilibrium P_0 is within the whole time interval $\tilde{t}_0$. We can therefore define the main goal of the solver as:

$$\sum_i b_i r_i \subseteq r_0 \ , \ i \neq 0, b_i \in \{0, 1\} \ . \tag{13}$$

In reality, the nature of the power grid leaves a certain margin for over- or undersupply, which depends on the size of the grid. Each agent features a *constraints module*, i.e., a software module that—among other things—supplies this margin. For the definition of the solver, it is denoted by P_C.

Furthermore, each LPEP message contains an accumulated distance value. This distance value is the sum of all impedances of all lines the message, and thus, the potential power transmission, travels. The LPEP models the power grid as an overlay network in the communications networks; hence, a communications-connection between two agents also corresponds to a power line. The higher the (accumulated) impedance, the higher the line losses of the power transmission. The second task of the solver is thusly: If more than one solution to a power disequilibrium exists, it must choose that with the lowest overall line loss. Let

$$\mathrm{d}(r_i) : r_i \mapsto \mathbb{R} \tag{14}$$

be a function the returns the accumulated distance of the requirement r_i. Using the function, we can define the secondary goal of the solver as:

$$\min \sum_i b_i \mathrm{d}(r_i), \ i \neq 0, b_i \in \{0, 1\} \ . \tag{15}$$

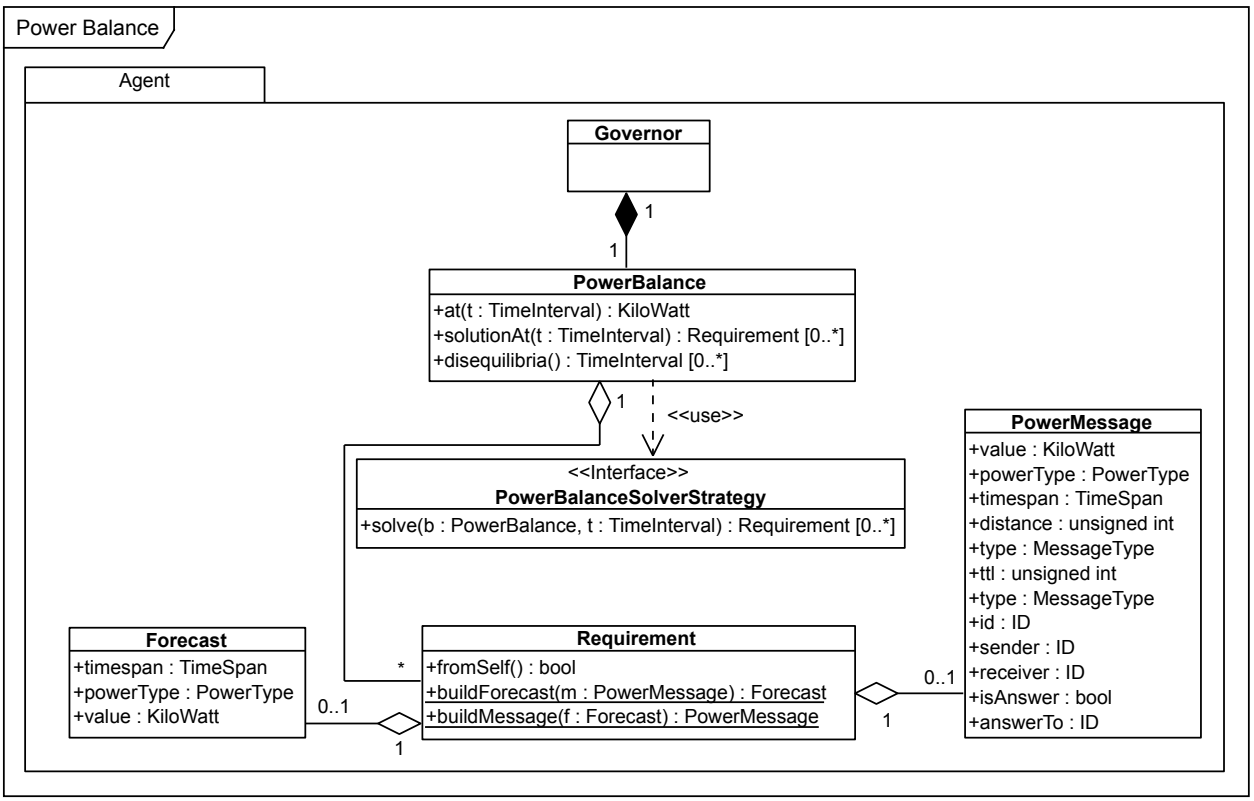


Figure 1. The Universal Smart Grid Agent's internal power balance.

The combinatorial problem the solver must find a solution to is therefore strongly reminiscent of the 0-1 knapsack problem [5].

The concept of the Universal Agent is aimed at managing the power grid at a large scale; therefore, it aggregates sub-kilovolt units at the respective transformer. For this reason, the definition of the LPEP defines the smallest amount of power that is transmitted to be $1\,\mathrm{kW}$—or $1\,\mathrm{kVAr}$, respectively—and further defines this value to be an unsigned integer. Additionally, agents must communicate the subdividability of the power values in their requests and responses. For example, many types of power plants have certain minimum load, stemming from their construction, and must therefore offer power in fixed-size chunks. Therefore, for both power and time values, the complete real numbers domain, $\mathbb{R}$, is not necessary; thus, $\mathbb{N} \cup \{0\}$ suffices. Stringently, modeling requests in terms of Boolean equations and solving the power disequilibrium in the Boolean domain provides an efficient approach to the local part of the demand-supply calculation.

IV. Modeling Demand and Supply in the Boolean Domain

A. Structure and Operation

The modelling of the local demand-supply calculation expresses the key problem in the calculation: A solution to the problem can be reduced to choosing the right combination of requirements—or, if possible, partial requirements—from other agents. We can therefore express the acceptance or rejection of a requirement by modeling it in the Boolean domain.

However, we initially face a multi-valued problem: A requirement is comprised of any power value, P_i, as well as the corresponding time interval, $\tilde{t}_i = [t_{i,1}; t_{i,2})$. Therefore, at this point, we cannot use a single variable to represent it in the demand-supply calculation.

To break down the multi-valued problem into its discrete parts, we need to split each offer into its atoms. The size of each atom is determined by the overall set of requirements the specific calculation is made of. This also includes the original request that is part of the power balance as well and which is denoted by the index 0, i.e., $[t_{0,1}; t_{0,2}) \mapsto P_0$. The size of the atoms is calculated from the vector of all power values,

$$\boldsymbol{P} = (|P_0|, |P_1|, \ldots, |P_i|, |P_C|) \ , \tag{16}$$

where P_C is the allowable power deviation, i.e., the solution must match $P_0 \pm P_C$.

Further, all timespan sizes,

$$\boldsymbol{t} = (t_{0,2} - t_{0,1}, t_{1,2} - t_{1,1}, \ldots, t_{i,2} - t_{i,1}) \ , \tag{17}$$

naturally also influence the size of the atoms, which is their respective Greatest Common Divisor (GCD):

$$\Delta P = \gcd(\boldsymbol{P}) \ , \tag{18}$$

$$\Delta t = \gcd(\boldsymbol{t}) \ . \tag{19}$$

The GCD can be calculated requirement-by-requirement due to the associative law:

$$\gcd(x, \gcd(y, z)) = \gcd(\gcd(x, y), z) \ . \tag{20}$$

The application of the GCD creates a raster of $n \cdot \Delta P \times m \cdot \Delta t$ atoms in which every requirement can be located. Thus, each agent's contribution to the power disequilibrium, i.e., its requirement, is deconstructed into a number of atoms. Each atom denotes a part of a requirement and references a time subinterval and a power subinterval in the power balance. The size of the time subinterval is Δt for all requirements; the size of each power subinterval is ΔP for all requirements. Each atom is described by a Boolean variable that expresses the origin of the requirement and the time and power interval in which it is located:

$$x_{i,\tilde{t},\tilde{P}} = \begin{cases} 1 & \text{if the agent } i \text{ influences the power grid} \\ & \text{in the time subinterval } \tilde{t} \text{ with power} \\ & \text{from the power subinterval } \tilde{P}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

An example of this subdivision is shown in Fig. 2. It contains five requirements—one forms the initial power disequilibrium, and four are (overlapping) responses from other agents—and illustrates how the $x_{i,\tilde{t},\tilde{P}}$ variables reference the parts of the respective requirement.

Using these atoms, we can express the semantics of a `Requirement` object: A requirement expresses a power delta within a certain time interval. Specifically, it expresses a number of power deltas that are available for the solver to choose from in the time interval. Agents responding to a request can do so with a single power value, an interval of power values, or even offer multiple power intervals. This information is obviously important to the solver. The *acceptance function* expresses this particular information of a requirement:

$$\mathrm{r}_i(\boldsymbol{x_{i,\tilde{t},\tilde{P}}}) = \begin{cases} 1 & \text{if } \boldsymbol{x_{i,\tilde{t},\tilde{P}}} \text{ denotes a valid inter-} \\ & \text{val for accepting the require-} \\ & \text{ment from agent } i, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

The atoms defined in Eq. (21) are now used by the solver to create the characteristic function of each requirement's acceptance function. A requirement is not simply based on its atoms: They possess a certain coherence. An agent typically indicates, through its demand notification or offer notification, what divisions into shares are possible. From this, the Boolean power balance solver creates conjunctions that specify the *acceptance function*. Each conjunction corresponds to a response and describes a possible acceptance of a requirement.

Several LPEP messages can make up one response: Through the characteristic function, the solver collapses these distinct messages back into one unified response by creating a conjunction for every possible valid interval as defined in Eq. (22). This is also required by the semantics of the protocol. Thus, for all time subintervals $\tilde{t}$ of size Δt within which the offer is valid, a conjunction for each allowable quantity of power, expressed as one or more power subintervals $\tilde{P}$ of size ΔP, is created. Thus, we can now generally express a requirement's acceptance function:

$$\begin{aligned} \mathrm{r}_i(\boldsymbol{x_{i,\tilde{t},\tilde{P}}}) = & \bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} \bar{x}_{i,\tilde{t},\tilde{P}} \\ & \vee \left(\bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} x_{i,1,1} \wedge \bar{x}_{i,1,2} \wedge \bar{x}_{i,1,3} \wedge \cdots \wedge \bar{x}_{i,\tilde{t},\tilde{P}} \right) \\ & \vee \left(\bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} x_{i,\tilde{t},1} \wedge x_{i,\tilde{t},2} \wedge \bar{x}_{i,\tilde{t},3} \wedge \cdots \wedge \bar{x}_{i,\tilde{t},\tilde{P}} \right) \\ & \vee \cdots \vee \\ & \bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} x_{i,\tilde{t},\tilde{P}} \ . \end{aligned} \tag{23}$$

All conjunctions in Eq. (23), except for the first or the last one, depend on what choices the responding agent offers. The solver needs to create at least a number of conjunctions equal to $2 \cdot \frac{P_i}{\Delta P} \cdot \frac{t_{i,2}-t_{i,1}}{\Delta t}$ since in the simplest case, a requirement is either completely accepted or not accepted at all:

$$\mathrm{r}_i(\boldsymbol{x_{i,\tilde{t},\tilde{P}}}) = \bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} x_{i,\tilde{t},\tilde{P}} \vee \bigwedge_{\tilde{t}\in r_i, \tilde{P}\in r_i} \bar{x}_{i,\tilde{t},\tilde{P}} \ . \tag{24}$$

The Binary Vector (BV) in $\boldsymbol{x_{i,\tilde{t},\tilde{P}}}$ can be represented in a compact manner by a Ternary Vector List (TVL) [1, 15, 16]. Remember Fig. 2 with its five responses; most could actually offer more than two choices to the solver—accept completely, accept partially, or not accept at all—and such a possible concrete acceptance function is shown in Fig. 3. This example illustrates the second requirement in the exemplary power balance of Fig. 1. This acceptance function, $\mathrm{r}_2(\boldsymbol{x_{i,\tilde{t},\tilde{P}}})$ uses only the atoms created from the requirement that $\mathrm{r}_2(\boldsymbol{x_{i,\tilde{t},\tilde{P}}})$ describes, i.e., all Boolean variables $x_{2,m,n}$. The agent that sent the corresponding LPEP message communicated an response comprised of $100\,\mathrm{kW}$ with the following options: $0\,\mathrm{kW}$, $50\,\mathrm{kW}$, or $100\,\mathrm{kW}$. These options are expressed through the corresponding conjunctions: $0\,\mathrm{kW}$ is described by $\bar{x}_{2,3,1} \wedge \bar{x}_{2,3,2} \wedge \bar{x}_{2,4,1} \wedge \bar{x}_{2,4,2}$, the $50\,\mathrm{kW}$ option by $x_{2,3,1} \wedge x_{2,3,2} \wedge \bar{x}_{2,4,1} \wedge \bar{x}_{2,4,2}$, and the acceptance of the whole $100\,\mathrm{kW}$ by $x_{2,3,1} \wedge x_{2,3,2} \wedge x_{2,4,1} \wedge x_{2,4,2}$.

In general, the manner in which these acceptance functions are created leads to the *characteristic function* for each acceptance function. Obviously, for any function, given a number of Boolean variables, there are many ways to formulate equivalent functions, e.g., simply by applying the axioms known to the Boolean algebra. The form we propose here in Eqs. (23) and (24) constitutes the respective characteristic function of an acceptance function, which was defined in a abstract way in Eq. (22).

Now that each non-initial requirement, i.e., all responses is converted into a number of conjunctions of its atoms, the solver

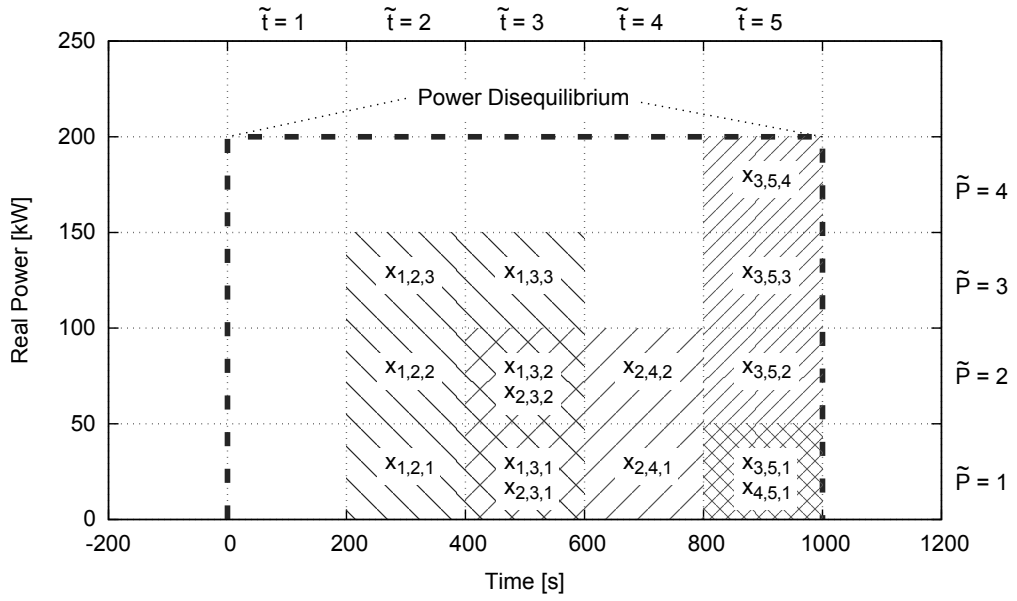


Figure 2. An example of a power balance state after the discretization.

must model the actual request, i.e., the power disequilibrium. Since the order in which the requirements are accepted is not important, but the cover is, the solver models this using a symmetric function. A symmetric function's value at any n-tuple is the same at every permutation of that n-tuple; the function does not depend on the order of its variables, but only on the number of set or unset variables. A symmetric function $S^n(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}})$ is equal to 1 iff exactly n of its variables are equal to 1, for every permutation of the assignment of its argument vector.

The solver creates m—where m denotes the amount of time subintervals that have been created through the application of the GCD—symmetric functions, one for each time subinterval $\tilde{t}$ of size Δt. Thus, each symmetric function describes $\frac{1}{m}$ of the disequilibrium. The argument of the respective symmetric function are those parts of each requirement's characteristic function for the time subinterval the symmetric function is constructed for. This is indicated by the subscript k of the symmetric function; each symmetric function describes the k-th power subinterval. Stringently, the function's argument vector is written as $\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}$ to express that only the atoms of the k-th power subinterval are a part of the argument vector.

The number of set bits corresponds to the power disequilibrium: If the power disequilibrium amounts to $n \cdot \Delta P$ kW, with $n \leq |\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}|$, the symmetric function has n 1-bits and $|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}| - n$ 0-bits. This implies that the responses the requesting agent received can actually solve the power disequilibrium. If $|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}| < n$ is true for any k, no symmetric function can be created and the disequilibrium cannot be solved. The definition of the symmetric function is:

$$\mathrm{S}_k^n(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) = \begin{cases} 1 & \text{if } n \text{ variables in } \boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}} \text{ are 1,} \\ 0 & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, m \ . \tag{25}$$

Relating to Fig. 2, the symmetric functions represent 'column-wise' slices of the respective requirements.

The solution set to the disequilibrium in $\tilde{t}$ consists of the exact covers of the symmetric function and the acceptance functions at $\tilde{t}$:

$$\mathrm{C}_k(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) = \mathrm{S}_k^n(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) \wedge \bigwedge_{i \in I'} \mathrm{r}_i(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) \ . \tag{26}$$

Here, the set I' is the set of all agents that have sent a response to the local agent's request. The complete solution set therefore consists of the complete cover, which is:

$$\mathrm{C}(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}}) = \bigwedge_k \mathrm{C}_k(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) \ . \tag{27}$$

XBOOLE can represent these equations in software and calculate the solution set efficiently. I.e., the Universal Agent uses XBOOLE as underlying software to implement the solver. The solver represents the acceptance functions as TVL in Orthogonal Disjunctive/Antivalent (ODA) form, such as the example in Fig. 3.

The ODA form of a TVL is one of the encodings for a TVL. In ODA form, the elements of the TVs are specified to be conjunctions; the TVs themselves are linked as disjunctions—hence the 'D' in 'ODA.' The 'O' denotes that the TVL is stored in orthogonal form. The 'A' specifies that the TVL can also be interpreted as a function in antivalent form. Refer to

$$\mathrm{r}_2(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}}) = \bar{x}_{2,3,1} \wedge \bar{x}_{2,3,2} \wedge \bar{x}_{2,4,1} \wedge \bar{x}_{2,4,2} \vee x_{2,3,1} \wedge x_{2,3,2} \wedge \bar{x}_{2,4,1} \wedge \bar{x}_{2,4,2} \vee x_{2,3,1} \wedge x_{2,3,2} \wedge x_{2,4,1} \wedge x_{2,4,2}$$

$x_{2,3,1}$	$x_{2,3,2}$	$x_{2,4,1}$	$x_{2,4,2}$	$\mathrm{r}_4(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}})$
0	**0**	**0**	**0**	**1**
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
1	**1**	**0**	**0**	**1**
1	0	1	0	0
1	0	0	1	0
0	1	1	0	0
0	1	0	1	0
0	0	1	1	0
1	1	1	0	0
1	1	0	1	0
1	0	1	1	0
0	1	1	1	0
1	**1**	**1**	**1**	**1**

Figure 3. An example of the acceptance function.

[1] for more details and an in-depth explanation on how the different operators are executed on a TVL in ODA form.

Fig. 3 displays the function $\mathrm{r}_2(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}})$ as well as the corresponding (expanded) TVL that would be used to represent it. Note that the lines printed in bold font not only denote those variable assignments for which $\mathrm{r}_2(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}}) = 1$ holds, but also represent the only vectors that are typically stored in a TVL.

The symmetric functions are represented in the same manner and are even easier to generate: Their TVL consists of all permutations $n \times 1$ and $(|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}| - n) \times 0$ for the k-th symmetric function. XBOOLE's actual work consists of a number of intersections of the TVLs. First, XBOOLE creates a TVL of all acceptance functions:

$$R = \bigcap_{i \in I', \tilde{t}, \tilde{P}} r_i(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}},\tilde{\boldsymbol{P}}}) \ , \tag{28}$$

as well as one TVL representing all symmetric functions:

$$S = \bigcap_{k=1}^{m} \mathrm{S}_k^n(\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}) \ . \tag{29}$$

The intersection of two disjoint TVLs—i.e., two TVLs whose variable sets are disjoint—yields their Cartesian product [1].

The complete cover then constitutes the final intersection:

$$C = S \cap R \ . \tag{30}$$

The TVL C that denotes the complete cover contains the complete solution set to the demand-supply calculation. If it contains more than one solution, i.e., $|C| > 1$, the agent must still choose among them. Ideally, if it is not restricted by contract or other constraints given by the agent's constraints module, it will try to minimize the line loss, and, therefore, prefer offers with a lower distance value. The distance value is the accumulated impedance the respective LPEP message contains; lower impedances mean physically lower line losses.

Remember Eq. (14), $\mathrm{d}(r_i) : r_i \mapsto \mathbb{R}$, which returns the distance value of a requirement r_i. This allows the ordering of all (accepted) requirements:

$$r_i \leq r_{i'} \quad \Leftrightarrow \quad \mathrm{d}(r_i) \leq \mathrm{d}(r_{i'}) \ . \tag{31}$$

C contains all variants that solve the power disequilibrium; from the TVL the solver can also extract the respective requirements that form the possible solutions. Through Eq. (31), we can calculate the sum of all distances of all requirements forming a solution. Thus, the Ternary Vectors (TVs) in C can be sorted by distance and the one with the lowest distance is returned. This is then the optimal solution to the power disequilibrium.

B. Complexity Analysis and Comparison

In order to return the result TVL, the XBOOLE solver creates a number of intermediate TVLs: One for each response, containing the individual acceptance function, one TVL that combines all acceptance functions, one for each symmetric function corresponding to a time subinterval, and finally one that combines all symmetric functions. However, since the TVLs for the individual acceptance functions as well as for the individual symmetric functions are only needed to create the Cartesian product, i.e.,

$$R_1 \times \cdots \times R_n \ \Leftrightarrow \ \mathtt{ISC}(R_1, \mathtt{ISC}(\ldots, R_n))) \ , \tag{32}$$

only at most 4 TVLs are used by the solver at a given time.

The size of the individual TVLs that represent the symmetric functions is a combinatorial classic, specifically, how often one can choose k elements from a set of the size n, disregarding the order. Here, the size of the set is the size of the argument vector of the symmetric function; the number of elements that should be chosen is n, i.e., the number of set bits in the argument vector. Therefore, we can express the size of the TVL through a binomial coefficient. Thus, each of the m individual symmetric functions that correspond to one time subinterval contains a number of TVs equal to:

$$|S_k| = \binom{|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}|}{n} = \frac{|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}|!}{n! \cdot (|\boldsymbol{x}_{\boldsymbol{i},\tilde{\boldsymbol{t}}=\boldsymbol{k},\tilde{\boldsymbol{P}}}| - n)!} \ . \tag{33}$$

We can verify this formula with the example shown in Table I that shows the TVL representing $S_5^4(\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{5},\tilde{\boldsymbol{P}}})$ the solver would create for the example power balance depicted in Fig. 2. Here, we can calculate $|S_5^4| = \binom{5}{4} = 5$.

The size of each TVL representing the individual offer's acceptance function depends on the number of power values offered, but contains at least one row for 'accept completely' and one for 'accept not at all':

Table I
TERNARY VECTOR LIST REPRESENTING S_5^4 OF THE EXAMPLE POWER BALANCE

$x_{3,5,1}$	$x_{3,5,2}$	$x_{3,5,3}$	$x_{3,5,4}$	$x_{4,5,1}$
0	1	1	1	1
1	0	1	1	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	0

$$|R_i| \geq 2 \ . \tag{34}$$

In comparison to other structures representing binary functions, such as BDDs, the TVL approach of XBOOLE proves to be more efficient in terms of space complexity. The space complexity of a BDD depends on the ordering of variables. The set of variables naturally also influences the size of any TVL. We can therefore use the $\texttt{SV\_SIZE}(P)$ operation to initially determine the number of variables:

$$nv = \texttt{SV\_SIZE}(R) \ . \tag{35}$$

Ideally, with good variable ordering, a BDD with $nv = 2k+2$ variables can be compressed to be minimal. Then, the BDD will store

$$|\boldsymbol{v}| = 2k + 2 \tag{36}$$

vertices. Under ideal circumstances, a TVL can represent an entire function with only one TV, requiring a number of ternary elements equal to:

$$nte = nv \ . \tag{37}$$

This might seem like an academic consideration, however, we will later see that through optimization, we can indeed arrive at this space requirement formula for many functions the solver considers.

Symmetric functions force us to evaluate the worst-case space complexity since they resist a data structure's approach to a compact representation. In general, using k variables, we can describe 2^{2^k} functions. Each vertex of a BDD is also the root of a subfunction; a BDD therefore describes a set of functions. In order to construct the worst case, a number of variables equal to $nv = k + 2^k$ must be given. Otherwise, the BDD can be compressed. Thus, the worst-case number of vertices of a BDD with $nv = k + 2^k$ variables is:

$$|\boldsymbol{v}| = 2 \cdot 2^{2^k} - 1 \ . \tag{38}$$

The maximum number of ternary values (i.e., elements) in a TVL with nv variables then corresponds to:

$$nte = nv \cdot 2^{nv-1} \ . \tag{39}$$

To compare the space complexity of BDDs and TVLs, we must consider their actual memory requirements. A BDD is made of data type definitions in the form of a C `struct`. Following [2], each vertex takes up 2 words to represent the edges of the tree, additional 2 words to store the `index` and `id` attributes, and, finally, 3 bits for `value` and `mark`. If we ignore padding, we can express the storage requirement of a BDD as:

$$D_{BDD} = |\boldsymbol{v}| \cdot (4D_W + 3) \text{ bits.} \tag{40}$$

On a recent x86_64 architecture, this would mean $D_W = 64\,\text{bits/word}$. In contrast, a ternary value uses only 2 bits [1] regardless of the processor architecture, sizing the TVL at:

$$D_{TVL} = 2nte \text{ bits.} \tag{41}$$

The smallest BDD that consists of one root and two terminal nodes will require:

$$\min(D_{BDD}) = 3 \cdot (4D_W + 3) = 777\,\text{bits} \ , \tag{42}$$

enough space for 388 ternary elements. Note that this difference weights heavy in the best case in favor of the TVL, and even more if the space compression ability of the ternary values can be used. Considering the worst case, the BDD becomes the better choice only for $nv \geq 521$ variables; at this point, $D \approx 8 \times 10^{146}$ GB will be required—too much by all means. Again, we will later see that we can avoid the worst case altogether, taking full advantage of the compact storage structure of a TVL and the superior computational complexity of operations on TVLs, as we will assert below.

The value of the GCD obviously directly influences the number of variables and thus the size of the individual TVLs, as the occurrences of the variables n and m indicate. One might therefore assume that a model of the given problem that works on integers would greatly reduce the memory footprint and complexity of the solver. Such a structure is an Edge-Valued Multi-valued Decision Diagram (EVMDD) that can represent any p-valued function. Let us define an offer as a p-valued function of two variables, $\tilde{t}$ and $\tilde{P}$. The function arguments denote the time and power interval for the requirement from agent i. The function remodels the acceptance equation of a requirement, described in Eq. (22) as a pure Boolean function, now as a p-valued function:

$$\mathrm{r}_i(\tilde{t}, \tilde{P}) = \begin{cases} \tilde{P} \cdot \Delta P & \text{if } t_{1,i} \leq \tilde{t}\Delta t \leq t_{2,i} \ , \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

According to [10], the memory size of an EVMDD is:

$$D_{EVMDD} = D_W \sum_{k=1}^{u} 2p_k \cdot w_k \text{ bits,} \tag{44}$$

where u denotes the number of multi-valued variables and w_k the number of non-terminal nodes for a variable. Comparing the worst-case memory size of a TVL in Eq. (41) with that of a EVMDD in Eq. (44), one can easily conclude that the EVMDD is the more efficient data structure once the rasterization leads

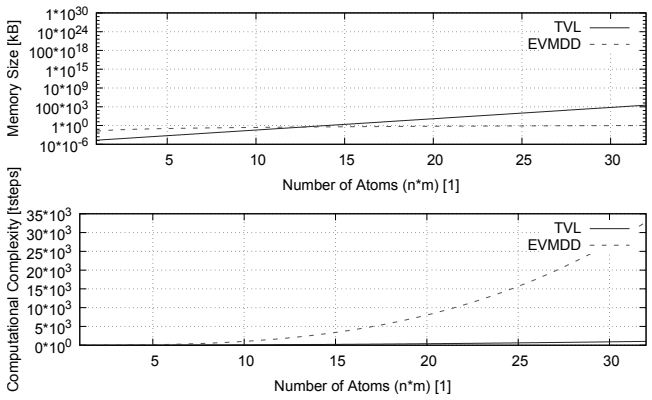


Figure 4. Memory size and computational complexity of XBOOLE-based and Edge-Valued Multi-valued Decision Diagram-based demand-supply solvers.

to a certain number of variables for the acceptance functions and, correspondingly, the symmetric function.

However, the memory of an embedded appliance running the Universal Agent software should be well suited to store a number of kilobytes to accommodate the data structures. More interestingly, the runtime behavior of the XBOOLE-based solver is more favorable. We can assume that the solution to a demand-supply calculation is achieved by a combination of two EVMDD subtrees—one for the acceptance function and one for the symmetric functions—, followed by a *satisfy-one* operation [2]. If we consider the runtime complexity of BDD and, therefore, EVMDDs, we conclude that the computational complexity of an EVMDD approach would be:

$$\mathcal{O}\left(|R|^2 \cdot |S|\right) + \mathcal{O}\left(|R| + |S|\right) \ , \tag{45}$$

whereas the computational complexity of the intersection operation—$\texttt{ISC}(P, Q)$—of XBOOLE is [1]:

$$\mathcal{O}\left(|R| \cdot |S|\right) \ . \tag{46}$$

Fig. 4 depicts the influence of n and m with regards to the TVL and EVMDD memory size and computational complexity. In comparison, while the EVMDD approach is the more space efficient data structure, the TVL-based modelling approach is by far the faster one.

This comparison serves as a transparent side-by-side analysis of a TVL-based versus an EVMDD-based approach to the solver. However, if an implementing algorithm followed Eqs. (25) to (30) by the letter, the solution would not be the optimal one. Every intersection except for the last one creates the Cartesian product of the two TVLs, since they are disjoint. No two TVLs representing their respective acceptance functions share a variable, since an offer is unambiguously identified by the variable set that makes up its atoms per Eqs. (18) and (21). The elementary symmetric functions that serve to model the initiating request add another burden: They are, by their nature, immune to simplification using the ternary value. I.e., at no time will any TVL representing a symmetric function contain a '$-$.'

Table II
BINARY VECTOR LIST FOR AN EXAMPLE SYMMETRIC FUNCTION, $\mathrm{S}^3(\boldsymbol{x})$

x_1	x_2	x_3	x_4	x_5
0	0	1	1	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	0
1	0	0	1	1
1	0	1	0	1
1	0	1	1	0
1	1	0	0	1
1	1	0	1	0
1	1	1	0	0

However, optimization is possible by exploiting two properties of the model. First, the intersection is commutative, and so is XBOOLE's $\texttt{ISC}(P, Q)$ operator. Instead of creating the complete acceptance function, R, directly for the intersection with the set of all symmetric functions, as Eqs. (28) and (30) suggest, we can compute the intersection iteratively for each acceptance function:

$$R_i \cap (\ldots \cap (R_2 \cap (R_1 \cap S))) \ . \tag{47}$$

This alleviates the solver from the task of creating the complete set R, but yields the same result as the naïve operation $R \cap S$.

The second property of the solver's model helps us to reduce the size of the set for which all symmetric functions, S, are equal to 1. It is also not necessary to create S completely. Since each row in a TVL representing a particular $S_{\tilde{t}}$ is a permutation of the first one, we can deduce the next TV in $S_{\tilde{t}}$ from the current one. Table II illustrates that one can easily establish an order within the function Binary Vector List (BVL) for any symmetric function. We can thus define and easily implement the following operations for a symmetric function's function BVL [4]:

FIRST($\boldsymbol{sbv}$) creates the first BV in the BVL $\boldsymbol{sbv}$ is a part of.
NEXTPERMUTATION($\boldsymbol{sbv}$) derives the next permutation from $\boldsymbol{sbv}$, i.e., it creates the next BV of the BVL for the symmetric function $\boldsymbol{sbv}$ is also a part of.
LAST($\boldsymbol{sbv}$) creates the last BV in the BVL $\boldsymbol{sbv}$ is a part of.

We can now permute a BV and, thus, do not need to store the complete TVL for a symmetric function. This allows us to compress the function TVLs for the symmetric functions for each time-subinterval, $S_{\tilde{t}}$, to the size of one—the current—BV. I.e., instead of working with the whole TVL for each time-subinterval's symmetric function, $S_{\tilde{t}}$, we now only require the current BV of the TVL. The current BV for the symmetric function for all time-subintervals, S, is then simply a concatenation of the respective BVs. The permutation must still obey the boundaries that are introduced through the individual time-subintervals, i.e., it must generate a permutation for each current BV in the respective $S_{\tilde{t}}$ and not over the whole S. Therefore, we must construct a function that, given the current TVL of each time-subinterval's symmetric function,

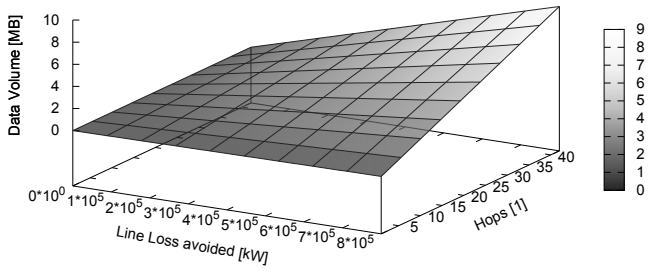


Figure 5. Data volume used compared to line loss avoided.

i.e., a vector of TVLs, creates the next valid permutation and returns the next valid BV in S. This method is called NEXTSYMMETRICFBV($\boldsymbol{S_{\tilde{t}}}$). Notice its argument type: A vector of TVLs. Algorithm 1 outlines the function's modus operandi.

We can thus construct a function that returns the next permutation of the TVs in S without having to create S completely. Therefore, we can note the solver's final form in Algorithm 2.

This final version starts with the calculation of the GCD and the construction of all relevant TVLs: First, R contains a TVL for each response—i.e., it is a collection of TVLs—, then $\boldsymbol{S_{\tilde{t}}}$ is introduced to hold a TVL for each time subinterval. However, each respective TVL, $S_{\tilde{t}}$, only contains one BV, which is the first valid permutation of the respective symmetric function's TV.

Its main part is a loop in which all permutations in S are generated, using Algorithm 1. The resulting TVL S of each permutation is then subsequently intersected with each response TVL, R_i, and the result appended to the final solution TVL. The loop ends when all valid permutations have been generated.

The best solution in the resulting TVL—if it is not empty, in which case no solution exists—is determined by sorting its TV. The basis for the sorting operation is Eq. (31), from which we can deduce that each TV in the solution TVL has an accumulated distance value. Thus, the first TV indicates the best solution. The solver now needs to map the TV back to the respective responses and to return the set of solution responses.

Fig. 5 compares the data volume required by the Universal Agent against the line loss that is avoided by its operation. The plot assumes that the node on the last hop answers and that the GCD of the Δt and ΔP atoms is 1, which means the worst-case size of the TVLs' set of variables. Through the optimized version of the solver in Algorithm 2, it is possible to keep the total amount of data required to arrive at a solution—including both, the volume of all messages transmitted and the size of the TVLs at the requesting node—below 10 MB while scaling up to quantities of 800 MW, which would include the whole capacity of a classic power plant such as one that uses bituminous coal, or even a nuclear power plant.

Fig. 5 is the foundation from which we can arrive at a general metric to evaluate the performance of the Universal

Algorithm 1 Calculation of the next valid permutation for the Binary Vector of the symmetric function

procedure NEXTSYMMETRICFBV($\boldsymbol{S_{\tilde{t}}}$)
  for $k = |\boldsymbol{S_{\tilde{t}}}|$, 1 **do** ▷ Iterates over the function TVLs.
    $S_{\tilde{t}} = \boldsymbol{S_{\tilde{t},k}}$
    $\boldsymbol{sbv} \leftarrow S_{\tilde{t},1}$ ▷ First BV in the respective TVL
    if $\boldsymbol{sbv} \neq$ LAST($\boldsymbol{sbv}$) **then**
      $\boldsymbol{sbv} \leftarrow$ NEXTPERMUTATION($\boldsymbol{sbv}$)
      break
    else if $k = 1$ **then**
      return $\emptyset$ ▷ No more permutations possible
    else
      $\boldsymbol{sbv} \leftarrow$ FIRST($\boldsymbol{sbv}$)
    end if
  end for
  $S \leftarrow S_{\tilde{t},1}$
  for $k = 2$, $|\boldsymbol{S_{\tilde{t}}}|$ **do**
    $S \leftarrow S \cap S_{\tilde{t},k}$
  end for
  return S
end procedure

Algorithm 2 The Universal Smart Grid Agent's central solver procedure

procedure SOLVE($request$, $responses$)
  $(\Delta P, \Delta t) \leftarrow$ GCD($request$, $responses$)
  $R \leftarrow$ CREATERESPONSETVLS($responses$, ΔP, Δt)
  $\boldsymbol{S_{\tilde{t}}} \leftarrow$ CREATEREQUESTTVLS($request$, $responses$)
  $S \leftarrow \emptyset$
  $C \leftarrow \emptyset$
  repeat
    $S =$ NEXTSYMMETRICFBV($\boldsymbol{S_{\tilde{t}}}$)
    $subSolution \leftarrow S$
    for all $R_i \in R$ **do**
      $subSolution \leftarrow subSolution \cap R_i$
    end for
    if $C = \emptyset$ **then**
      $C \leftarrow subSolution$
    else
      $C \leftarrow$ CON(C, $subSolution$)
    end if
  until $S = \emptyset$
  if $|C| > 0$ **then**
    $C \leftarrow$ SORT(C) ▷ Per Eq. (31)
    return SELECTRESPONSES(C)
  else
    return $\emptyset$
  end if
end procedure

Table III
COMPARISON OF THE UNIVERSAL SMART GRID AGENT AND BINARY DECISION DIAGRAM APPROACH ON THE FUKUI-TEPCO POWER GRID DURING HIGH LOAD

	BDD	Universal Agent
Line loss avoided (ΔP)	17 208 kW	17 208 kW
Compute Time	> 16 min	1 min 34 s (run time); < 11 min (simulated time)
Data Volume	100 MB	28.9 MB
Data Efficiency	0.168 kW/kB	0.581 kW/kB

Agent, as well as other approaches.

V. EVALUATION OF EFFICIENCY

Any proposal must show, at least in theory, that it improves the present situation. The Universal Smart Grid Agent makes no difference in this regard. In order to quantify the impact of the agent approach, we need to define a metric. Since one of the pillars of the solution proposed in this paper is to model the power grid in communication networks, comparing bytes and power transmitted suggests itself.

We define the *effect* of a distributed demand-supply calculation in terms of bytes it requires to transport a certain amount of power as the *data effect*:

$$\kappa = \frac{W}{D} \left[\frac{\text{kWh}}{\text{kB}}\right] . \tag{48}$$

This metric is, in fact, applicable for every solution and not specific to the Universal Agent. The data effect metric can be used to compare protocols with each other. The more Watts per Bit are transmitted, the more a effective a protocol is.

However, this does not describe the efficiency of an approach. Managing volatile power generation and consumption does not only mean to act on or react to this volatility as, e.g., curtailment is always possible, but also entails an increase in efficiency due to local generation and consumption: Power that does not need to be transmitted via longer distances suffers lower line losses than power that travels many kilometers of cable. Of course, this assumes that local generation as a reaction to increased demand or higher consumption due to temporally higher power production are possible.

We can therefore define the *data efficiency* in terms of line loss avoided per bytes transmitted:

$$\xi = \frac{\Delta P}{D} \left[\frac{\text{kW}}{\text{kB}}\right] . \tag{49}$$

The two metrics introduced in Eqs. (48) and (49) can serve to compare protocols, the quality of solvers, and approaches in general. They allow us to compare approaches that solve the problem from different angles, since all these approaches in the context of the smart grid naturally require a computer, and hence a certain volume of data.

Let us compare the approach outlined in this paper, the Universal Smart Grid Agent, to that proposed by Inoue et al. [7], which uses BDDs. These two are solutions to the same problem, but different in their characteristics: The Universal Agent is distributed and de-centralized, whereas the BDD-using solution requires only one node and can eschew network communication.

Both use the same grid, namely, an electric grid developed by Fukui University and Tokyo Electric Power Company (TEPCO). Inoue et al. briefly describe the network whose raw data is available through [6]. It models a typical Japanese distribution network. The grid is fed from 72 feeders and contains 468 switches. Including electrical and topological constraints, it allows for 1.5×10^{70} feasible configurations. Given pre-defined loads, the goal consists in finding a configuration with minimal line loss. Additionally, this paper offers the data efficiency metric, which we will apply, too.

Inoue et al. examine the Fukui-TEPCO grid at a low-load situation at 2 A.M., as well as at a high-load situation at 4 P.M. During high load, the BDDs that model the grid, as well as the constraints, require $100\,\text{MB}$ of data volume. Modelling the grid in the network simulator, we can observe how the Universal Agent solves the same load situation. Here, each node and each section as well is represented by an agent. The sections with active and reactive power loads broadcast their requests until they reach the feeder nodes, which, in turn, reply with their offers. Those sections that represent switches choose the state of the switch according to whether they have relayed an acceptance acknowledgement notification or not: If such a message was forwarded, the switch must be closed; otherwise, it will be opened. Since the LPEP features direct routing for all responses, no power will flow via these sections if no acceptance acknowledgement notification was relayed by these agents.

The results listed in Table III[1] show that, for the high-load case, the decentralized agent concept requires less data volume than BDD approach and also takes less time to compute. The simulation run in which all simulated agents share a single thread finishes in less than $2\,\text{min}$, while the simulated time within the environment amounts to a little over $10\,\text{min}$, including network latency modelled as $X \sim \mathcal{U}[20\,\text{ms}; 250\,\text{ms}]$. This ten-minute-interval stems directly from the design of the LPEP. In the high-load case, the Universal Agent therefore features a higher data efficiency than the approach by Inoue et al. based on BDDs.

However, the decentralized concept comes with a cost, which is the base data volume required to initially set up each agent instance. Fig. 6 shows the total data volume required by all Universal Agents. The graph shows not only the base data volume required by the Universal Agent approach, but also indicates that the LPEP contributes the biggest share to the overall sum. We can also note that the XBOOLE-based power balance solver requires the least amount of data, although an instance runs on each requesting agent. This stems from the space efficiency of Algorithm 2, but is also due to the

[1]The line loss avoided is calculated against the theoretical line loss in the grid, which is, in turn, calculated according to the formulae presented by Nara et al. [11].

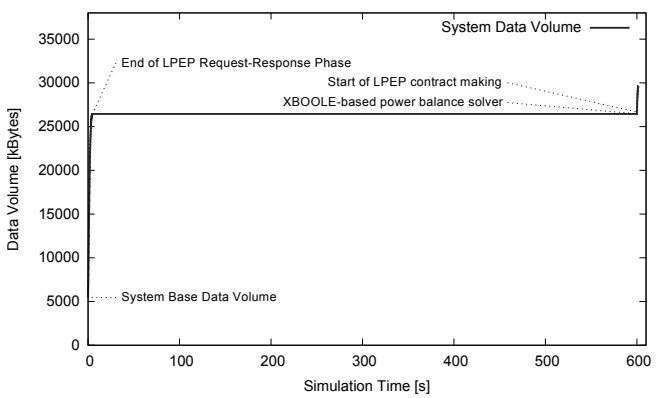


Figure 6. Total data volume by time required by the Universal Smart Grid Agent in the Fukui-TEPCO power grid during high load.

72 feeders being able to offer exactly the amount of power required; thus, the size of the individual TVLs remains low.

The nature of the Universal Agent makes it unfitting for the low-load situation. Here, the BDD approach requires only 100 kB of data to run, whereas the Universal Agent's base data volume cannot be lowered: Each node and each section, regardless of their current load situation, is always represented by an instance of the Universal Agent.

VI. Conclusion

In this paper, we presented an efficient approach to modeling demand and supply in a smart grid. We have briefly covered the underlying agent concept and discussed its heart-piece, the agent-internal power balance and its solver algorithm. We have then compared our approach to that of Inoue et al., utilizing metrics introduced in this paper.

This comparison thus served to show two results. It demonstrated that the metrics defined in Eqs. (48) and (49) can be applied to different approaches that try to improve efficiency in the smart grid. It also emphasized that the distributed, decentralized concept that is the nature of the Universal Agent is suited best for complex situations with many actors, possibly of vastly different types, but falls short in simpler environments. For the future, we anticipate the smart grid to become an ever more complex environment due to distributed generation and volatile load and generation patterns. The operation of such a grid can be governed by the Universal Smart Grid Agent.

References

[1] D. Bochmann and B. Steinbach. *Logikentwurf mit XBOOLE*. 1st ed. Berlin, Germany: Verlag Technik, 1991.

[2] R. E. Bryant. "Graph-Based Algorithms for Boolean Function Manipulation". In: *IEEE Transactions on Computers* C-35.8 (1986), pp. 677–691. ISSN: 0018-9340. DOI: 10.1109/TC.1986.1676819.

[3] S. F. Bush. *Smart Grid — Communication-enabled Intelligence for the Electric Power Grid*. 1st ed. Wiley IEEE Series. Chichester, United Kingdom: John Wiley & Sons, 2014. ISBN: 978-1-119-97580-9.

[4] cppreference.com Contributors. *std::next_permutation*. Online. http://en.cppreference.com/w/cpp/algorithm/next_permutation [Retrieved: 2016-11-02]. Oct. 2015.

[5] T. Dantzig, J. Mazur, and B. Mazur. *Number: The Language of Science*. A Plume book. Plume, 2007. ISBN: 978-0-4522-8811-9.

[6] Y. Fujimoto. *Distribution test feeders*. Online. http://www.hayashilab.sci.waseda.ac.jp/RIANT/riant_test_feeder.html [Retrieved: 2016-11-02]. 2006.

[7] T. Inoue et al. "Distribution Loss Minimization With Guaranteed Error Bound". In: *IEEE Transactions on Smart Grid* 5.1 (Jan. 2014), pp. 102–111. ISSN: 1949-3053. DOI: 10.1109/TSG.2013.2288976.

[8] S. Jafarzadeh et al. "Hour-ahead wind power prediction for power systems using Hidden Markov Models and Viterbi Algorithm". In: *IEEE PES General Meeting* (July 2010), pp. 1–6. DOI: 10.1109/PES.2010.5589844.

[9] I. Maqsood, M. Khan, and A. Abraham. "An ensemble of neural networks for weather forecasting". In: *Neural Computing and Applications* 13.2 (May 2004), pp. 112–122. ISSN: 0941-0643. DOI: 10.1007/s00521-004-0413-4.

[10] S. Nagayama and T. Sasao. "Representations of elementary functions using edge-valued MDDs". In: *Proceedings of The 37th International Symposium on Multiple-Valued Logic (ISMVL 2007)*. Oslo, Norway: IEEE, 2007. DOI: 10.1109/ISMVL.2007.49.

[11] K. Nara et al. "Implementation of Genetic Algorithm for Distribution Systems Loss Minimum Re-Configuration". In: *IEEE Transactions on Power Systems* 7.3 (1992), pp. 1044–1051. ISSN: 15580679. DOI: 10.1109/59.207317.

[12] "Fifteenth Inventory". In: *Electricity production from renewable energy sources: details by region and by country*. Ed. by Observ'ER. 2013th ed. Paris: Observ'ER, 2013. Chap. 3, pp. 3–7.

[13] C. Potter, A. Archambault, and K. Westrick. "Building a smarter smart grid through better renewable energy information". In: *Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES*. Mar. 2009, pp. 1–5. DOI: 10.1109/PSCE.2009.4840110.

[14] M. Ruppert, E. M. Veith, and B. Steinbach. "An Evolutionary Training Algorithm for Artificial Neural Networks with Dynamic Offspring Spread and Implicit Gradient Information". In: *Proceedings of the Sixth International Conference on Emerging Network Intelligence (EMERGING 2014)*. Rome, Italy: International Academy, Research, and Industry Association, 2014.

[15] B. Steinbach. "XBOOLE—A toolbox for modelling, simulation, and analysis of large digital systems". In: *System Analysis and Modelling Simulation* 9.4 (1992), pp. 297–312.

[16] B. Steinbach and M. Werner. "XBOOLE-CUDA Fast Calculations of Large Boolean Problems on the GPU". In: *Problems and New Solutions in the Boolean Domain*. Ed. by B. Steinbach. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 2016, pp. 117–149. ISBN: 978-1-4438-8947-6.

[17] E. M. Veith and B. Steinbach. "Modeling Demand and Supply in a Smart Grid". In: *Proceedings of the 24th International Workshop on Post-Binary ULSI Systems*. May. University of Waterloo. Waterloo, Canada, 2015, pp. 1–2. DOI: 10.13140/RG.2.1.1813.1689.

[18] E. M. Veith, B. Steinbach, and J. Windeln. "A lightweight messaging protocol for Smart Grids". In: *Proceedings of the Fifth International Conference on Emerging Network Intelligence (EMERGING 2013)*. Porto, Portugal: IARIA XPS Press, 2013, pp. 6–12.

[19] E. Veith, B. Steinbach, and J. Windeln. "A Lightweight Distributed Software Agent for Automatic Demand—Supply Calculation in Smart Grids". In: *International Journal On Advances in Internet Technology* 7.1 and 2 (2014), pp. 97–113.

Development of unidirectional data diode system in the secure environment

A. G. Vorontsov
Information Technology
FSUE VNIIA
Moscow, Russia
voroncov-ag@narod.ru

S. A. Petunin
Information Technology
FSUE VNIIA
Moscow, Russia
petunin@vniia.ru

The paper describes features of the development of unidirectional network (DataDiode devices) in the secure environment (the development of the unidirectional data transfer from the wide area network (WAN) to the secured enterprise network). A brief overview of the existing devices is given and their characteristics are described. The method to address limitation issues, such as lack of feedback channel related to the implementation and integration of one-way channel is described.

The main part of the research is focused on enhancing accessing process to the restricted network. The model of the unidirectional data transfer system, including DataDiode device, has been introduced.

The paper reflects:

- ***General principles of unidirectional data transfer;***
- ***Topological features of the unidirectional data transfer model;***
- ***Interaction with information security system;***
- ***Providing load balancing and interactivity under high-load conditions.***

Keywords: data diode; unidirectional; security; highload; fiber; one-way gateway

I. INTRODUCTION

A unidirectional gateway is a network appliance allowing data to be transferred only in one direction [1]. It doesn't allow data to pass in the opposite direction and connects different segments of the network with various privacy levels of the data processing and storage [2]. In the subject area under consideration, such one-way network solution must ensure no data transmission capacity in the opposite direction at the hardware level to preclude reconfiguration of security policy or firewall rules.

The class of such devices that provides isolation of network segments to prevent unauthorized access to the network information assets is referred to as Data Diode. Thus, it ensures the required data input to the closed network, and at the same time, it prevents unauthorized outputting of restricted information or any other external accessing to the closed network [3].

Currently, the Data Diode devices may be topologically different unidirectional data gateway. Some types of such systems are discussed below.

We introduce the notion of a private network segment and public network segment. The private network segment is a protected automated network system. The public network is an unsecured wide area network.

Schematic [4] of the unidirectional data gateway is shown in Fig. 1.

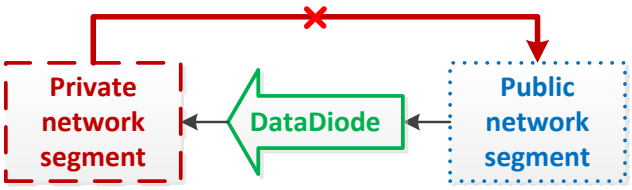


Figure 1. Data transmission to the downstream network

This Data Diode system is based on physically isolated fiber-optic communication and data feedback principle (no network autonegotiation). The functional diagram of the system kernel is shown in Fig. 2, which demonstrates the absence of a data feedback.

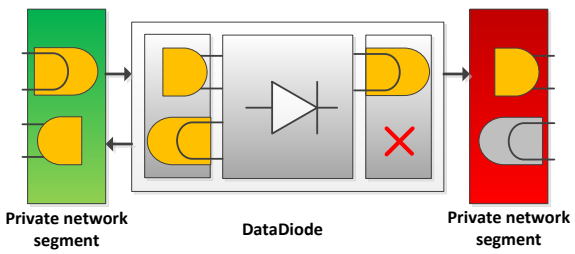


Figure 2. Functional diagram of transmitting system kernel

The devices based on the given functional diagrams have a number of advantages and disadvantages. Using Data Diodes allows resolving the problem of open data input to the downstream network, and implementing asynchronous management (without data feedback) of services within downstream network by means of control commands sent from the public network segment. It allows moving away from the

previously used input methods (e.g. specialized input points where all the data is controlled by an operator) and to partially automating data input process.

However, the absence of data feedback brings forth a number of possible technical challenges. Implementation of the unidirectional channel causes the problem of transferred data verification which in the most cases is critical for data integrity:

- TCP/IP protocol requires handshaking;
- Data feedback is required for determination and adjustment of the communication rate;
- It is impossible to run application software since it requires a data feedback or data received acknowledgement.
- Most of web-services are not functional.

There are different approaches to resolve such problems [5]. For instance, the USA patent No. 5703562 Method for transferring data from an unsecured computer to a secured computer proposes a verification mechanism that uses a warning device coupled to a secured computer and emits a warning signal if an error was introduced during data transmission. The patent suggests using a single long duration tone [6]. It is obvious that the suggestion doesn't allow transmitting the checksum calculations from a Send Node to a Receive Node to allow the latter match the results, and thus define received data integrity.

Also, there are simplified approaches that deploy commercially-available hardware devices for establishment of a topologically adjusted network, configured only to one-way data transmission. However, these methods and approaches don't provide the sufficient protection level and it doesn't protect from attacks aimed to reverse data channel flow.

Out of the existing and functional devices, the most interesting configurations are built on the basis of the diagram shown in Fig. 2.

A. Unidirectional gateways with two module and repository

It consists of two components – the transmitter of the public network and the receiver of the private network coupled without any real data feedback. Every module contains an SSD drive with XFS file system. Both components are controlled by OS Linux special build.

The following operation algorithm is used [7]:

1. The public network data flows to the receiver, and then is written to an SSD drive.
2. As data transmission to the receiver is over, the data is synchronized with the private network receiver's SSD drive via UDP protocol that doesn't require acknowledgement, thus partially eliminates the necessity of data feedback. During synchronization process, the transmitted data is splitting into data packets, where each packet entered the checksum.
3. The data stored on the receiver, becomes available to a recipient, while the data on the public network SSD drive receiver is automatically deleted after synchronization is over.

Schematic diagram of two-module unidirectional gateways is shown in Fig. 3.

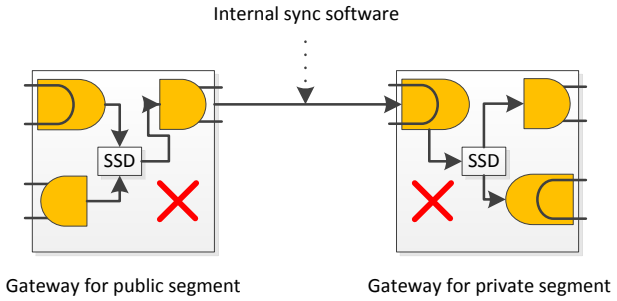


Figure 3. Functional diagram for two-module unidirectional gateways

Such implementation allows for one-way data transfer without data feedback required to confirm the integrity of the transmitted data and dynamic bandwidth adjustment of transfer rate. The data integrity is confirmed by the internal software. As the channel bandwidth and specification for the transmitter and the receiver is beforehand known, this provides an opportunity to determine synchronization speed in advance between SSDs drives.

The transmission speed has a great relevance because if the asynchronous behavior occurs, the receiver would not be able to keep up handling the input data stream, and the sender (the upstream network transmitter) would not know about it. This situation may lead to transmission failure.

The implementation disadvantages:

- XFS file system instability;
- Poor performance (about 30 Mb/s)
- The limited number of users is supported (100 users in total, 20 active users)
- The volume of SSD drives within the system is limited.

The most important measures of the device performance entered as the parameters listed in Table 1.

TABLE I. DATA TRANSFER RESULTS

Quantity*volume / Measures	1x8000 Mb	3x8000 Mb	1030x7 Mb	3030x10 Mb
Average send speed, Mb/s	20	20	15	14
Average sync speed, Mb/s	30	30	25	25
Average receive speed, Mb/s	25	25	23	22
Average transmit time, min	18	53	20	82
Average number of transmission errors	0	0	15	43

Average number of sync errors	0	0	20	1315
Number of transmissions	5	5	5	5

B. End-to-end unidirectional gateway with one module

Unlike the previously two-module architecture, it is a single device. It is also a transmission device based on the same principles as the two-module unidirectional gateways but it doesn't have any internal storage and embedded software to perform data synchronization.

Operating principle is also build on the transmitting system kernel as the two-module device, though it has its specific differences [8]:

- Absence of the internal storage;
- Supports large number of active users (511);
- High performance (up to 90 Mb/sec).

Since there is no need to store the data within the device and no sync role – the given architecture is more fault tolerant and much better to implement. The software which responsible for data receiving and transmitting is located within the device on separate servers. The diagram of the device is shown in Fig. 4.

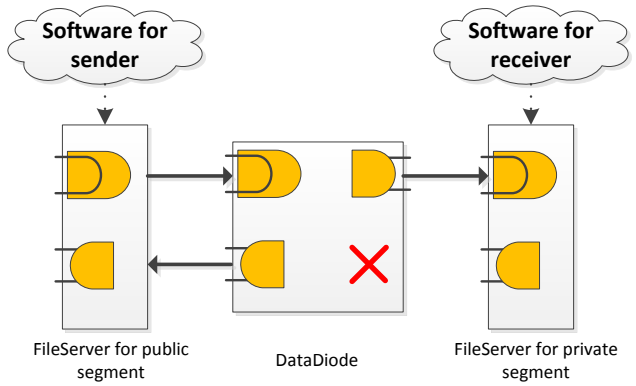


Figure 4. Scheme of the end-to-end unidirectional gateway

The most important measures of the device performance entered as the parameters listed in Table 2.

TABLE II. DATA TRANSFER RESULTS

Quantity * volume / Measures	1x8000 Mb	3x8000 Mb	1030x 7 Mb	3030x 10 Mb
Average send speed, Mb/s	87	92	52	50
Average receive speed, Mb/s	87	92	52	50
Average transmit time, min	95	265	140	610
Average number of sync errors	0	0	7	0
Average number of redundancy data packets	1	1	1	2
Number of transmissions	5	5	5	5

II. UNIDIRECTIONAL DATA TRANSFER SYSTEM REQUIREMENTS

Necessity of development methods and a number of means to implement the automated unidirectional network ensure:

- guaranteed unidirectional data transfer;
- absence of transmitted data loss;
- communication channel performance (no less than one Gb/s);
- dynamic load balancing between output points from the public network;
- dynamic load balancing between input point to the private network;
- compatibility with information protection software;
- private network topology hiding;
- integration with network domain structure.

The proposed methods and means are automated, integrated with public and private network and has centralized management. The necessary condition to be able to interact with private network is the appropriate certification of Data Diode device.

III. SYSTEM MODEL

To develop the system model, it is necessary to define the objects.

Objects of the public network:

- A transmitting file server;
- Domain controller;
- User workstations.

Objects of the private network:

- A receiving file server;
- General file server;
- Domain controller;
- Network firewall;
- Data Diode;
- User workstations.

The proposed model implements network firewall in compliance with different network classes interaction requirements.

The functional diagram is shown in Fig. 5.

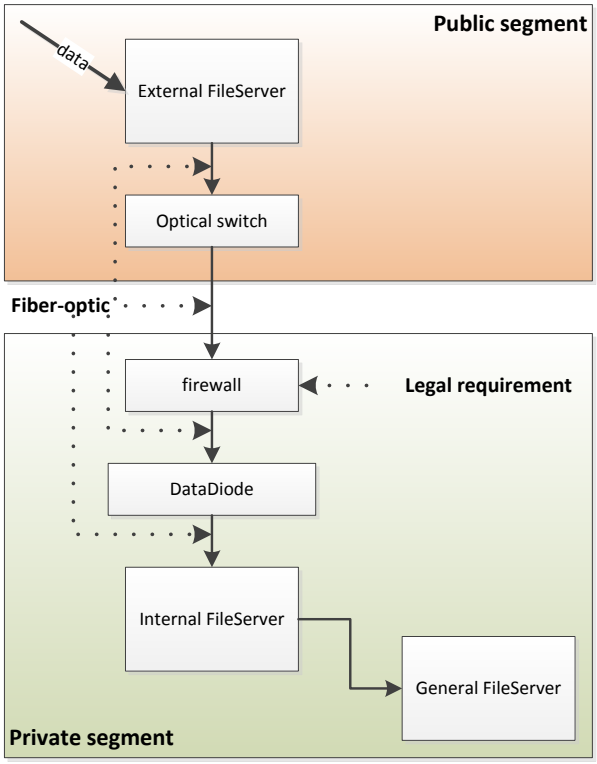


Figure 5. Example of a figure caption. *(figure caption)*

IV. STRUCTURAL IMPLEMENTATION METHODS

Architecture of the proposed solution is shown above. Besides customizing of communication equipment and sync software for the transmitting and receiving file servers, a number of problems need to be solved, such as:

- The incoming file queue problem;
- Overload of the devices that may become a reason of out of sync behavior;
- Interaction with information security system;
- Load balance;
- Interactive management and automation of the research system.

The problem of the incoming file queue is the selection of file processing priority. Before transmitting the current file in queue, the analysis of the whole ready-to-transmit queue is required, arranged by the file size, for example. Then, it is necessary to rebuild transmitting queue. This action solves such problems as large packet transmission and concentration of a great number of small-sized files in file queue. A queue manager is implemented into the proposed model. If the large-sized file is being in the transmitting process while the system has only one unidirectional channel, it is necessary to wait until the file transfer is over. To resolve this problem file queue manager should compress large-sized files into equally sized archive volumes as part of the file transfer preparation procedure. It will enhance opportunities for queue management.

As it was mentioned above, absence of data feedback and transmission error that may occur due to asynchronous behavior of data transfer speed has two potential options:

- Application of sync software;
- Application of additional control facility for received files.

For the purpose of reduction such risk it is necessary to have identical technical characteristics for transmitting and receiving servers, and also contain high-performance disk subsystems, which preferably should be based on SSD drives.

The above mentioned method for file integrity check (uses application responsible for data transfer) needs optimization. While using this method, a possible problem may occur when it comes to huge number of files transfer that becomes more complicated regarding to inability to know the size, quantity and list of transmitted files due to a lack of data feedback.

The additional check method involves full data compressing and creation of archive volumes. It guarantees visual check for complete transfer of every archive file. If error occurs, the archive would not be upload to the receiving file server, thus a user would have to retransmit the archive volume.

Interaction with information protection system when end-to-end unidirectional gateway is used becomes a trivial task, aimed to ensure integration of transmitting and receiving file servers into the existing systems [9]. The main task is to provide error-free interaction with synchronizing software without any delay of network protocol or files reading and writing operations.

In case of using one unidirectional gateway device load balance is applied if exceeding or insufficient computational resources. To resolve such problem the scaling method is applied. It involves increasing the number of transmitting and receiving file servers and unidirectional channels. It is also necessary to apply balancing gateway that redirects the user to one of the receiving servers of the public network [10]. This architecture is shown in Fig. 6.

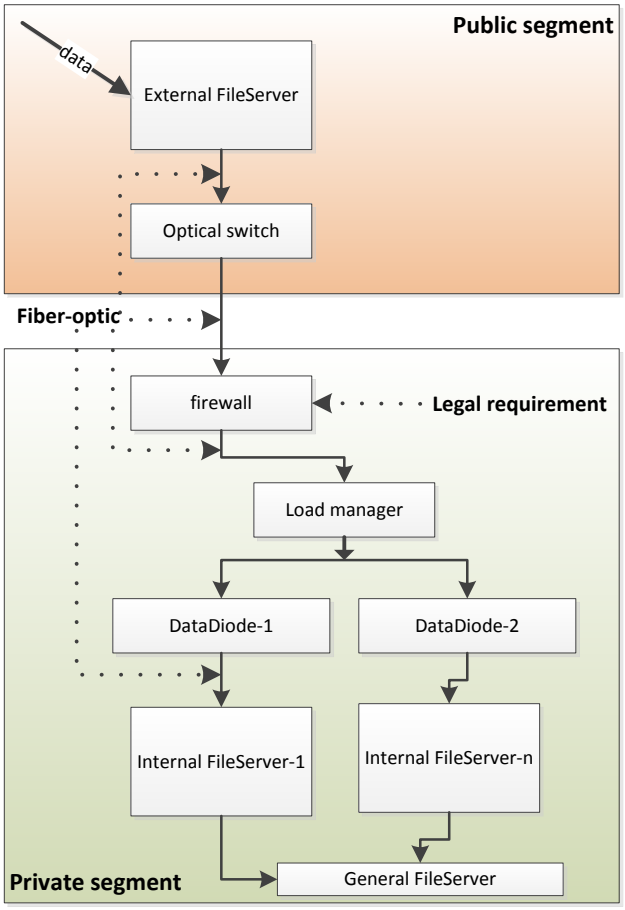


Figure 6. Example of a figure caption. *(figure caption)*

Development of a general control structure to ensure interactive control of the system is also required.

It provides:

- Addition of new network users;
- Automated control for transferred and received files that include archiving, transmission preparation, priority queuing, receiving files allocation.

The structure functions in cooperation with synchronizing software of the unidirectional gateway within public and private network. The conditions of no data feedback are observed.

Work algorithm of control structure and internal synchronization software within public network is shown in Fig. 7.

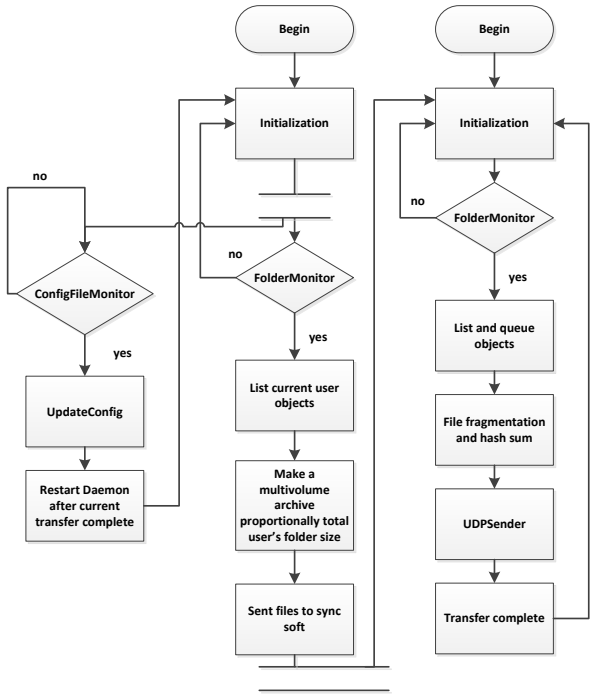


Figure 7. Sync software and control structure of the upstream network flowchart

Work algorithm of receiving data within private network is shown in Fig. 8.

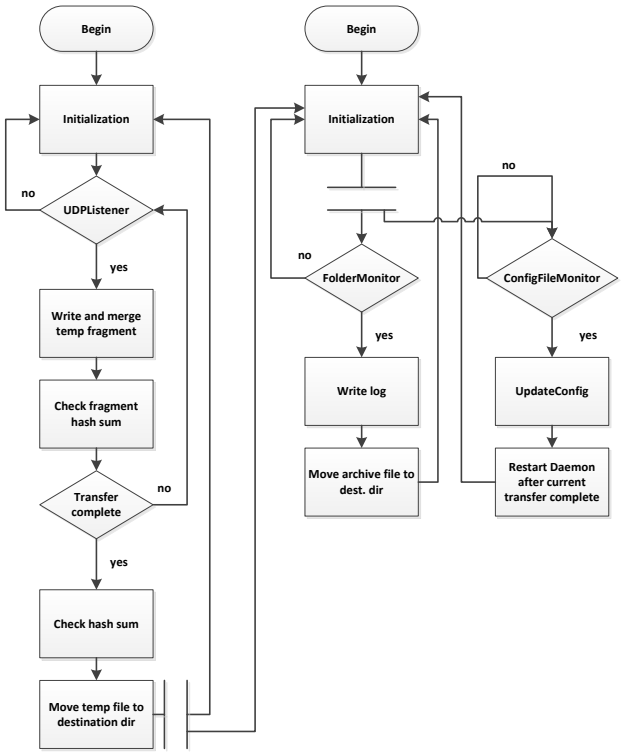


Figure 8. Sync software and control structure of the private network flowchart

V. CONCLUSION

This paper introduces and describes the development of hardware and software solutions for unidirectional gateway implementation in the conditions of secured environment. The actual implementation provided a unidirectional data flow from Wide Area Network to Local Area Network under the given conditions.

A primary result of this paper is theoretical development and implementation of unidirectional network system integrated into corporative network ensuring interaction with security systems. The developed control structure permits to scale the system considering basic load factors such as transmitted data volume and the number of users in system.

REFERENCES

[1] Unidirectional network: https://en.wikipedia.org/wiki/Unidirectional_network

[2] Australian Government Information Management Office 2003, Securing systems with Starlight, Department of Finance and Administration http://www.agimo.gov.au/archive/publications_noie/2003/06/transform/defence.html

[3] “CanSec” company – “What is unidirectional gateway”: http://cansec.ru/21/unidirectional-gateway.html

[4] AMT Group – “InfoDiode - Unidirectional gateway system”: http://www.amt.ru/rubr.aspx?rubr_id=237&art_id=1154

[5] Douglas W. Jones and Tom C. Bowersox “Secure Data Export and Auditing using Data Diodes” // 2006 USENIX/ACCURATE Electronic Voting Technology Workshop, 1 August 2006, Vancouver, https://www.usenix.org/legacy/events/evt06/tech/full_papers/jones/jones_html/

[6] OKB SAPR – “Organization of a unidirectional data transmission channel on the basis of a protected service information carrier”: http://www.okbsapr.ru/lydin_tezisy2013_1.html

[7] “CanSec” company – “Integrated file and mail unidirectional gateway Strom-File”: http://cansec.ru/products/strom_file.html

[8] “CanSec” company – “High speed unidirectional gateway Strom-1000”: http://cansec.ru/products/strom-1000.html

[9] Vorontsov A. G., Petunin S. A., Konyshev A. V. “Empowering information security systems in the conditions of domain environment” // Workshop on computer science and information technologies CSIT’2015, Rome, Italy, 2015.

[10] Somerdata – “AROW Data Diode”: http://somerdata.com/?page_id=1766

Predicting clinical status of patients after an acute ischemic stroke using random forests

Adéla Vrtková
Department of Applied Mathematics
VŠB - Technical University of Ostrava
Ostrava, Czech Republic
adela.vrtkova.st@vsb.cz

***Abstract*—According to the World Health Organization, a stroke has been the second most common cause of death in the world in the last 15 years. An ischemic stroke accounts for almost 80 % of all cases.**

The University Hospital Ostrava in the Czech Republic collects various information about patients who were transported there after suffering from an acute ischemic stroke, such as the affected brain hemisphere, duration of medical procedure or presence of hypertension. The objective of this paper was finding a model which would be able to predict patient's clinical outcome three months after an ischemic stroke based on the collected data. It was also desirable to analyse importance of the considered variables.

For this purpose, the random forests algorithm was used. To avoid biased variable importance, we used an alternative approach to the random forests which uses the conditional inference trees. Firstly, the commonly used modified Rankin Scale was used for describing the patient's outcome three months after a stroke. Secondly, only two values for the clinical status were considered, by meaning they correspond with the values 0-3 and 4-6 of modified Rankin Scale. The best performance was achieved with the second approach to description of the clinical outcome with the calculated classification accuracy 86 %.

I. Introduction

A stroke, together with ischemic heart disease, has been the world's biggest killer in the last 15 years [1]. An ischemic stroke is described as a sudden loss of blood flow to the brain which could result in permanent brain damage or death [2]. Intuitively, the long-term prediction of patient's condition after suffering from an ischemic stroke is of interest. The primary purpose of this article is the analysis of medical data of more than 600 patients with an acute ischemic stroke who were transported to the University Hospital Ostrava in the Czech Republic. This paper focuses on those variables which are available within 24 hours after admission to the hospital and the goal is using these variables to predict the patient's clinical outcome which is evaluated with the modified Rankin Scale three months after an ischemic stroke [3].

Recently, prediction of patient's clinical status has been in demand. Various studies [4]– [7] used the artificial neural networks or the logistic regression for this purpose. Eftekhar et al. [4] published a study about prediction of mortality in head trauma, Ottenbacher et al. [6] aimed to predict rehospitalization in patients with a stroke. The authors of these papers also discussed limitations of these methods and provided a comparison of the predictive abilities. Since the artificial neural networks and the logistic regression are a frequent choice when classification in medical data is of interest, the authors in [5], [7] provided methodology review where advantages, disadvantages and limitations of these methods were summarized. According to the available literature, the random forests algorithm was applied especially to microarray data or in the field of ecology. Statnikov et al. [8] published an analysis of microarray gene expression data with a comparison of the random forests algorithm and the support vector machines. The usage of the random forests algorithm for classification problems in ecology was presented in [9], [10]. Studies for complex genetic diseases using the random forests were published in [11].

For the purpose of this article, the random forests algorithm is used. We also tried the logistic regression and the artificial neural networks but since these methods achieved very poor performance, it was decided not to include them in the article. Our aim is to work with the original mixed data without any intentional transformation of the exploratory variables, e.g. dichotomization, and to reach high classification accuracy. Therefore, we decide to use the random forests with the conditional inference trees which is able to provide unbiased variable importance when data with different types of variables are of interest. In the analysis, firstly, we consider the original values of the modified Rankin Scale with values from 0 (no symptoms) to 6 (death) as the dependent variable. Secondly, we consider only two groups of patients by the clinical outcome - the first group with no symptoms to moderate disability and the second group with moderately severe disability to death. In both cases, the algorithm is performed for the purpose of classification.

We are motivated by providing a useful tool which will be used by the hospital on a daily basis. The model will be integrated into the hospital's patient database and only 24 hours after the patient's admission, the prediction of patient's clinical status after three months will be available. This long-term prediction will be used to prevent the complications after the medical procedure or for defining an optimal rehabilitation strategy. In this article, we aim to find a model with the best possible predictive abilities to achieve the highest possible accuracy in the prediction of the patient's clinical outcome.

II. Random forests

A. The basic concept of Decision Trees

Generally, decision tree methods (described in [12]) use the set of splitting rules to build the model that predicts the value of the target variable. They provide a non-parametric modelling approach with high flexibility and they are able to detect interactions between two or more exploratory variables. Since the decision trees can be used for regression or classification problems, regression and classification trees have been distinguished. Considering our aim, the basic theory of the classification trees is summarized, for detailed information about the regression trees see [12], [13].

The classification tree model (described in [13]) is appropriate when belonging of observations to a class based on independent variables is of interest, wherein the independent variables could be both qualitative and quantitative. The classification trees use a process known as recursive binary splitting. This process begins at the top of the tree (the root node) and with consideration of all exploratory variables the best split is made at each node until the terminal node is reached. The best split is chosen in the sense of minimizing classification error rate which is defined as

$$E = 1 - \max_k(\hat{p}_{mk}), \tag{1}$$

where $\hat{p}_{mk}$ is the proportion of training observations in the mth terminal node region which are from the kth class. Also, other measures (available in [13]) could be used, such as measure of total variance across all K classes called the Gini index which is defined as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{2}$$

The crucial step from the decision trees to the random forests is in building large collection of trees and combining them.

B. Random forests algorithm

The random forests algorithm is able to model complex interactions among exploratory variables [10]. Moreover, the algorithm is robust to overfitting, relatively robust to outliers and provides useful estimates of variable importance [14]. As well as the decision trees, the random forests can also be used for regression and classification problems. Since there is a difference in using the algorithm for regression and classification, we continue focusing only on the classification problem.

The first step in the random forests algorithm (introduced in [14]) is the selection of many bootstrap samples from the original dataset. The number of bootstrap samples determines the number of the trees in the model. Then, for each bootstrap sample, a classification tree is grown with a slight modification. At each node, a random sample of exploratory variables is made and the algorithm chooses the best split among these randomly chosen predictors. Typically, the number of randomly selected predictors is approximately equal to the square root of the total number of exploratory variables but in most cases, the best setting of the number of trees and the number of randomly selected predictors are dependent on the data [15]. In the next step, the predictions from all trees are aggregated within the meaning of the vote for the most popular class, in other words, the overall prediction is the most frequently occurring class [14].

In the first step of the algorithm, none of the trees uses the whole dataset. The remaining data, which are not used for building the tree, are called the *out-of-bag* observations (OOB). At each bootstrap iteration, the OOB observations are predicted and the aggregation of the OOB predictions is made. From these overall OOB predictions, we can compute the OOB error estimation which is an accurate estimation of the error rate (under the assumption of the sufficient number of trees) [16]. Moreover, thanks to the way of estimating the OOB error, there is no need to perform cross-validation or validation through a test set [13].

From the random forests algorithm, information about the variable importance can be obtained. One option how to estimate the importance of a variable is by looking at the increase of prediction error (or the decrease in prediction accuracy) when the OOB data for that variable are permuted while other variables are unchanged [16], this measure is called the permutation accuracy importance. The other option is measuring the variable importance as the mean decrease in the Gini index, see (2).

C. Bias in the random forests algorithm

Despite the promising features of this algorithm, in some cases, the biased variable importance measures can be obtained [17]. This can happen when the quantitative exploratory variables have different scale of measurement or when the categorical exploratory variables vary in the number of categories. The bias towards predictors with many categories and continuous predictors is noticeable especially in the mean decrease in the Gini index. This is caused by the process of the variable selection where continuous variables or variables with many categories are artificially preferred, for details we refer to [18].

The solution, which was proposed in [17], suggests using the conditional inference trees and the bootstrap sampling without replacement in the random forests algorithm. The main difference in the algorithm is that a significance test procedure is used for selecting variables to split instead of minimizing measures like the Gini index. In other words, the conditional inference tree chooses only those predictors which have significant relationship with the target variable [18]. This alternative way prevents overfitting as well as the original random forests algorithm described earlier. According to Strobl et al. [17], only this approach provides the unbiased variable selection in the classification trees and therefore, desired unbiased variable importance.

Since the prediction of the clinical outcome is our goal, the choice fell on the random forests algorithm mainly due to its high classification accuracy, flexibility and the ability to catch complicated interactions. Because of the different

types of predictors, the usage of the alternative approach to the random forests algorithm is necessary. For the analysis, the R package `party` is used. The usage of `party` package was demonstrated in [19].

III. Analysis

A. Data description

The original dataset consists of 649 patients who suffered from an acute ischemic stroke and who were transported for treatment to the University Hospital Ostrava in the Czech Republic. The data have been collected since June of 2005 and the database has been constantly updated. It was decided to exclude patients with missing information from the analysis, therefore, the analysed dataset contains information about 449 patients. The data come from the hospital's database where all collected information about the patients is saved. Of course, due to the confidential information, the database is fully available only to the hospital's medical staff and cannot be provided to third parties.

For our purpose, 17 exploratory variables (see Table I) are selected - 12 categorical and 5 quantitative variables. These variables are chosen from a total of 62 variables because they are available within 24 hours of admission to the hospital. The selected variables are used for the prediction of patient's clinical outcome which is evaluated with the modified Rankin Scale (mRS) three months after a stroke. The scale runs from 0 to 6 with following interpretation [3]:

- 0 - no symptoms,
- 1 - no significant disability despite symptoms,
- 2 - slight disability, unable to carry out all previous activities,
- 3 - moderate disability, require some help, but able to walk without assistance,
- 4 - moderately severe disability, unable to walk without assistance and unable to attend to own bodily needs,
- 5 - severe disability, require constant nursing care and attention,
- 6 - death.

In the analysis, we work with the original exploratory variables without any transformation. Firstly, we consider the original dependent variable which contains 7 categories. Then, we dichotomize the target variable where the patients with mRS up to 3, i.e. patients who do not require assistance with activities of daily living, are in the first category and the patients with mRS more than 3, i.e. patients who do require assistance with activities of daily living, are in the second category. The aim is to compare how well the random forests algorithm can predict these two interpretations of the patient's clinical outcome.

Since the random forests algorithm is a non-parametric approach which doesn't have assumptions for independence of the exploratory variables [17], further analysis of independence is not required. Also, we decided not to exclude possible outliers from the analysis.

B. Results

We performed the introduced alternative of the random forests algorithm using the `cforest` function in R and the variable importance was computed with the `varimp` function. It is necessary to mention, that in the `cforest` function we set only the number of trees (`ntree`) and the number of randomly selected variables (`mtry`), all other parameters were set to default, see [19] for more information about the `cforest` function. Different combinations of the number of trees and the number of randomly selected variables were tried and for presenting the results we choose the setting with the highest overall accuracy. For the estimation of other accuracy measures (sensitivity, specificity, etc.) OOB observations are used.

Firstly, we present the results of the analysis when the original mRS with 7 categories was considered. Unfortunately, the algorithm didn't provided sufficient prediction model. We reached the maximum overall accuracy of 56 % with 260 trees and 4 randomly sampled variables at each node which isn't satisfactory result. Low sensitivity and high specificity in all classes (see Table II) suggest that the classification model could cause many "false negatives". Low PPV means low probability that the patient will reach the particular outcome, if the model predicts it. The overall accuracy indicates that only 56 % patients were assigned the correct clinical status. Low classification accuracy could be caused by the fact that there aren't distinctly recognizable differences in the patients with the similar mRS, e.g. the differences in the patients with mRS equal to 3 and 4 were so small that they couldn't be properly detected by the algorithm. Additionally, in the variable importance plot (see Figure 1), we can see that the model suggests that the patient's outcome 24 hours after a stroke is the most important variable. The first top three also includes `TICI` and `DSA.inception`. That means that the outcome after three months mostly depends on how fast patient gets to the hospital (`DSA.inception`), whether blood flow to the brain tissue was restored (`TICI`) and mainly it depends on the patient's condition 24 hours after a stroke (`NIHSS.24h`).

In the other approach to the clinical outcome, we hope for better performance of the algorithm, similarly, we look for the model with high accuracy. Hereafter, only two values are used for describing the clinical outcome after three months, by meaning they correspond with 0-3 and 4-6 of mRS and can be interpreted as do not requiring assistance and do requiring assistance with activities of daily living.

The algorithm reached maximum overall accuracy of 86 % with 40 trees and 4 randomly selected variables, the accuracy measures are available in Table III. We can see the significant improvement in the performance of the algorithm. The high sensitivity and specificity suggest better predictive ability of the model than in the previous case. Despite the fact that more general evaluation of patient's condition was used, the prediction still provides essential information about the possible future outcome. Figure 2 shows the variable importance plot where it can be seen that again, the clinical status after 24

Table I: EXPLORATORY VARIABLES USED FOR LONG-TERM PREDICTION OF PATIENT'S OUTCOME

Abbreviation	Variable description	Categories	Mean (SD) or n (%)
AR	occurrence of arrhythmia	No	284 (63 %)
		Yes	165 (37 %)
CMP.before	occurrence of previous stroke	No	396 (88 %)
		Yes	53 (12 %)
DM	presence of diabetes mellitus	No	338 (75 %)
		Yes	111 (25 %)
hemisphere	affected brain area	left hemisphere	214 (48 %)
		right hemisphere	170 (38 %)
		posterior circulation	65 (14 %)
HT	presence of hypertension	No	101 (22 %)
		Yes	348 (78 %)
IVT	usage of intravenous	No	178 (40 %)
	thrombolysis	Yes	271 (60 %)
NIHSS.entry	NIH Stroke Scale at the time	minor	22 (5 %)
	of arrival to the hospital	moderate	194 (43 %)
		moderate to severe	152 (34 %)
		severe	81 (18 %)
NIHSS.24h	NIH Stroke Scale 24 hours	minor	135 (30 %)
	after ischemic stroke	moderate	190 (42 %)
		moderate to severe	57 (13 %)
		severe	67 (15 %)
non.SICH	occurrence of asymptomatic	No	402 (90 %)
	intracerebral hemorrhage	Yes	47 (10 %)
sex	sex of the patient	man	264 (58 %)
		woman	185 (42 %)
smoking	indicating whether patient smokes	No	364 (81 %)
		Yes	85 (19 %)
TICI	the thrombolysis in cerebral	no perfusion	37 (8 %)
	infarction perfusion scale grade	minimal perfusion	23 (5 %)
		partial perfusion	150 (33 %)
		complete perfusion	239 (54 %)
age	age at which patient		65 (13)
	suffered from ischemic stroke		
BMI	the body mass index		29 (12)
DSA.inception	time between ischemic stroke		290 (450)
	onset and medical procedure (min)		
DSA.arrival	time between arrival to the hospital		80 (130)
	and medical procedure (min)		
proc.dur	the length of medical procedure (min)		52 (29)

hours after stroke onset (NIHSS.24h) and the restoration of blood flow (TICI) are considered as the two most important variables. This time, age is included in the first top three.

Overall, the results showed the importance of the successful restoration of blood flow to the brain tissue which is strongly related to the patient's outcome after 24 hours. Therefore, these results suggest that if the medical procedure after patient's admission is successful, better clinical status after three months can be expected.

IV. DISCUSSION AND CONCLUSION

The aim of this paper was finding a model for the prediction of the patient's clinical status three months after an acute ischemic stroke. After stroke onset, the patient is admitted to the hospital and various information is recorded. We used the patient's data which are available at most 24 hours after stroke

Table II: ACCURACY MEASURES (IN %) FOR PREDICTION OF THE PARTICULAR OUTCOME AGAINST ALL OTHERS

mRS	Specificity	Sensitivity	PPV*	NPV*
0	95	69	64	96
1	92	54	69	87
2	94	44	64	87
3	91	62	27	98
4	92	73	31	98
5	90	60	21	98
6	97	57	87	84

* PPV = Positive Predictive Value,
NPV = Negative Predictive Value

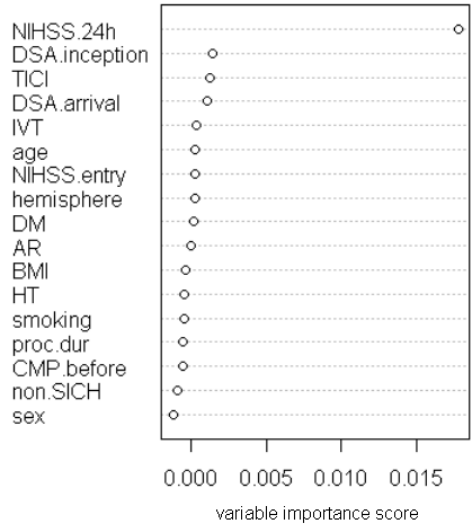


Figure 1: The variable importance plot for the model with the original clinical outcome described with mRS

Table III: ACCURACY MEASURES (IN %) FOR PREDICTION OF THE DICHOTOMIZED OUTCOME

Accuracy	Specificity	Sensitivity	PPV*	NPV*
86	92	82	95	74

* PPV = Positive Predictive Value,
NPV = Negative Predictive Value

onset. Our goal was to find a way of using the information to the long-term prediction.

For this purpose, we used the random forests algorithm with the conditional inference trees which outperformed the original random forests due to the nature of our used dataset. The dataset contains continuous variables with different scale measurements (`age`, `BMI`, `DSA.inception`, etc.) and categorical variables with different amount of categories (`IVT`, `TICI`, `hemisphere`, etc.).

The random forests algorithm also provides the suitable es-

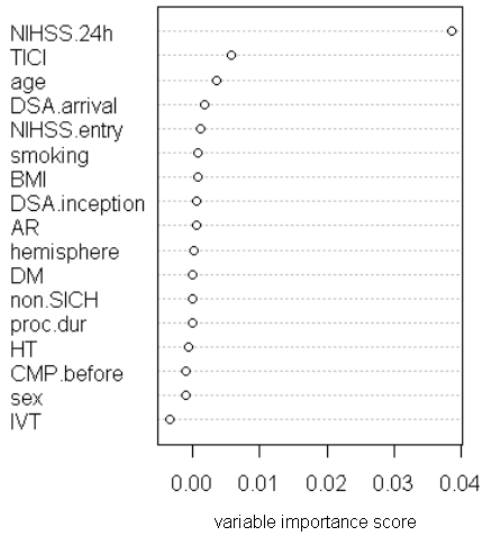


Figure 2: The variable importance plot for the model with the dichotomized clinical outcome

timation of the variable importance, so it is also possible to use the algorithm for the selection of several important variables for further analysis. The model suggested that the outcome mostly depends on the patient's condition 24 hours after stroke onset (`NIHSS.24`) and on the success in restoration of blood flow to brain tissue (`TICI`).

When considering the clinical outcome in terms of requiring (not requiring) assistance with daily activities three months after an acute ischemic stroke, we managed to build a satisfactory model for the prediction of patient's condition with overall 86 % accuracy. This means that the predictive model would be correct at least in 86 % of all cases. This measure does not fully describe the model predictive performance so it is necessary to use other accuracy measurements, such as sensitivity, specificity, positive predictive value and negative predictive value.

Sensitivity indicates that the model correctly predicted worse clinical outcome in 82 % of all patients who were actually in worse condition. Specificity suggests that patients with better clinical outcome were correctly classified in 92 % of all patients who were actually in better condition. The positive predictive value implies that 95 % of the patients classified with worse clinical status actually had worse outcome. Finally, the negative predictive value indicates that 74 % of the patients classified with better outcome actually had better clinical status.

However, the problem of the prediction of patient's outcome described with the modified Rankin Scale still lacks adequate solution. For future work, we would like to find the decent model which would be able to predict the clinical status evaluated with the modified Rankin Scale. Possible solution may be found in including more variables which can describe patient's condition more precisely, for example, the usage of the evaluation of inflammatory response or blood coagulation

parameters might be considered. Currently, new data describing patient's condition in more detail are being collected (e.g. the level of Von Willebrand Factor, the level of ADAMTS13, the number of monocytes etc.) for the purpose of the extension of the introduced model.

Acknowledgment

Special thanks go to the University Hospital Ostrava for providing the data and for the kind permission to publish this article.

This work was supported by the VŠB-Technical University of Ostrava (Project No. SP2017/56).

References

[1] (2017, Apr.) The top 10 causes of death. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310/en/index1.html

[2] H. B. van der Worp and J. van Gijn, "Acute ischemic stroke," *New England Journal of Medicine*, vol. 357, no. 6, pp. 572–579, 2007.

[3] G. Sulter, C. Steen, and J. De Keyser, "Use of the barthel index and modified rankin scale in acute stroke trials," *Stroke*, vol. 30, no. 8, pp. 1538–1541, 1999.

[4] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Medical Informatics and Decision Making*, vol. 5, no. 1, p. 3, 2005.

[5] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5, pp. 352–359, 2002.

[6] K. J. Ottenbacher, P. M. Smith, S. B. Illig, R. T. Linn, R. C. Fiedler, and C. V. Granger, "Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke," *Journal of clinical epidemiology*, vol. 54, no. 11, pp. 1159–1165, 2001.

[7] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[8] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC bioinformatics*, vol. 9, no. 1, p. 319, 2008.

[9] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[10] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[11] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh, "Identifying snps predictive of phenotype using random forests," *Genetic epidemiology*, vol. 28, no. 2, pp. 171–182, 2005.

[12] G. Tutz, *Regression for categorical data*. Cambridge University Press, 2011, vol. 34.

[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 6.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[17] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[18] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[19] C. Strobl, T. Hothorn, and A. Zeileis, "Party on!" 2009.

Exploratory ECG Analysis of Driving Events Using Wavelet Band Metrics

Renata Wachowiak-Smolíková (Member, IEEE)
Mark P. Wachowiak (Member, IEEE)
Computer Science and Mathematics
Nipissing University
North Bay, ON Canada P1B 8L7
Email: renatas@nipissingu.ca

Michel J. Johnson
École de kinésiologie et de loisir
Université de Moncton
Moncton, NB Canada E1A 3E9

***Abstract*—Electrocardiogram (ECG) signals obtained during driving simulations provide an indication of physiological response during complex driving tasks. The present study contributes to the literature on this response through the analysis of features in time-frequency representations, specifically, continuous wavelet transforms (CWTs). The main frequency band in the CWT of ECG signals, which corresponds to the heart rate, was analyzed using dispersion measures and complexity metrics (entropy and sample entropy). Results indicate that different driving conditions produce physiological responses that can be distinguished through features obtained from the CWT, which can be subsequently quantified and subjected to further statistical analysis, and used for exploring physiological responses to experimental conditions.**

1. Introduction

Automobile driving is a complex activity with high psychological and physiological demands [1]. Because the livelihood of many people either directly or indirectly depends on driving, the physiological and cognitive aspects of driving have medical, social, and economic importance. Driving simulators are useful in studying physiological responses during this activity [2], and have also shown some success in monitoring and training reaction time, visual attention, working memory, and other cognitive aspects of driving (see [2] and the bibliography cited there). This study contributes to the growing literature on these physiological responses through time-frequency analysis of electrocardiogram (ECG) signals acquired during driving simulations. Because driving is a dynamic activity, time-frequency representations, specifically continuous wavelet transforms (CWTs), facilitate gaining new insights into physiological responses during driving simulations. CWT features are visualized as a time-frequency map of spectral power at relevant frequency bands over the entire temporal duration of a signal. Although the CWT is usually assessed by visually inspecting this map, in the current paper, specific CWT features are extracted and statistically analyzed, in addition to the standard qualitative analysis. This work is part of a larger research project into rehabilitation of injured vehicular operators.

The CWT has clear benefits over traditional frequency-domain techniques. Because standard Fourier transforms contain only globally averaged information, short duration transient components cannot be easily or adequately resolved. The loss of time-based transient information can be reduced by employing a fixed-length time window, as is the case with the short-time Fourier transform (STFT). While the STFT provides temporal resolution, more flexible, alternative time-frequency methods have become readily adopted, wavelet transforms being among the most prominent for ECG analysis [3]. The CWT allows periodicities, transient events, non-stationary behaviour, and other interesting features and anomalies to be identified, localized, and qualitatively studied. Research focused on patient-specific diagnosis and analysis is one emerging application of wavelet transforms and other time-frequency techniques [4], [5], [6].

As the CWT is commonly represented as an image of spectral power, features can be isolated through pattern recognition methods, as had previously been applied to the CWT of electromyography (EMG) for the assessment of muscular effort and physiological adaptation [7] and to ECG for lower body negative pressure response [8], [9]. In this paper, individual, subject-specific response to various driving stressors, as well differences between experimental and control driving conditions, are analyzed. Both qualitative (visual) and statistical analysis of features extracted from the CWT are presented.

2. Methods

2.1. Participants

The study, conducted in July 2012 at the Université de Moncton in Moncton, Canada, collected ECG and other physiological data from $n = 17$ (12 male, 5 female) participants. The anthropometrics are: mean age = 34.69, age standard deviation = 9.05, minimum age = 22, maximum age = 51. Before starting the experiment, the procedures of the study were discussed with all participants. They read an information form, were informed of their rights, and signed a consent form approved by the Research Ethics Committee of the Université de Moncton.

2.2. Apparatus

The open car driving simulator (VS500M, Virage Simulation, Canada) resembles a General Motors compact cab interior. The simulator consists of a driver's seat,

steering column, pedals, automatic transmission, and a dashboard mounted on a three-axis motion/vibration platform to provide force feedback and vibrations. Two side screens located behind the driver provide additional visual feedback for the left and right blind spots. Rear view and side view mirrors are simulated through the screens. Three 52″ LCD displays provide a 180° front view. The resolution is 1920 × 1080 pixels per front display [10].

The participants were guided through a ten-minute practice scenario to become familiar with the driving simulator controls, followed by five minutes of rest before the final evaluation. This simulated drive occurred on a clear day through a city while encountering different levels of traffic density with other road users. Participants were guided through the simulation using a pre-programmed voice, similar to a GPS navigator, instructing them to turn left or right at various intersections. All participants underwent the same urban driving scenario (approximately 14 minutes for 7.5 kilometers).

2.3. Data Acquisition

A three lead ECG (MLA2340), was used to collect, condition (i.e. amplification, filtering, converting), and record heart signals using the Bio Amp unit (FE132) and an eight channel PowerLab unit (PL3508) (AdInstruments, Colorado Springs, CO, USA). LabChart software (version 7, AdInstruments) was used for data collection, post-processing, and data analysis. The sampling frequency was 200 Hz.

2.4. Driving Events

The simulator software allowed events to be programmed into the simulation. During the stressed driving (experimental) simulation, the following events were encountered: (1) expressway entry; (2) construction zone; (3) encountering a pedestrian; (4) seeing a cyclist in front of the driver; (5) seeing a disabled vehicle; (6) driving on a pedestrian right-of-way; (7) seeing a cyclist on the right; (8) entering a T-intersection with traffic; (9) entering a pedestrian crossing; (10) turning left in front of a truck; (11) crash 1; (12) crash 2. Each participant's ECG was recorded during the entire simulated drive. Not all participants completed the simulation, and therefore the collected data only reflect events that were actually encountered by the participant. The start and end times for each event were recorded so that the ECG could be analyzed on the basis of the entire duration of the simulation and for each individual event.

2.5. Continuous Wavelet Transform

The CWT $C_{\psi}^{x}(a,\tau)$ of a time signal $x(t)$ is the convolution of $x(t)$ with a scaled (by a) and translated (by τ) version of a nonorthogonal basis function. Varying a and τ yields a 2D matrix representation of the wavelet coefficients, which denote the amplitude of signal characteristics by frequency versus time (see [11] and the literature cited therein for a full development of wavelet theory). The CWT is computed as:

$$C_{\psi}^{x}(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^{*}\left(\frac{t-\tau}{a}\right) dt \qquad (1)$$

The CWT is flexible in that many basis functions $\psi(t)$ are available, and are usually selected to match signal characteristics. The Morlet wavelet is commonly used, as its scale parameter a has a straightforward relationship to the Fourier period $1/f$ [11].

The CWT implemented in this study, named the *sCWT* for sinusoidal CWT, is based on the selective discrete Fourier transform, which selects, for each time and frequency component, the shortest required window length, from which the corresponding spectral component is computed by the discrete Fourier transform [12]. Although in earlier studies, the Morlet and sCWT representations yielded approximately the same results and detected the same signal features [9], the sCWT was used in this work because frequencies can be directly analyzed through complex sinusoidal basis functions at that frequency [13].

For a time series $x(t)$, the goal is to determine the spectral component at frequency f at time τ, which is the centre of a time window of duration k/f, where k is a user-selectable parameter. The sCWT is given as [13]:

$$sCWT(f,\tau) = \sqrt{\frac{f}{k}} \int_{t-\frac{k}{2f}}^{t+\frac{k}{2f}} x(t) \exp\left(2\pi i f\left(t-\tau\right)\right) dt \qquad (2)$$

which is implemented as:

$$sCWT(f,nT_s) = \sum_{-\frac{k}{2fT_s}}^{\frac{k}{2fT_s}} x(m+n) \exp\left(2\pi i f T_s\right) \qquad (3)$$

The resulting complex sCWT is displayed as the squared modulus of the sCWT.

2.6. Band Metrics

The entire wavelet transform representation contains important information about frequency content in all analyzed bands. However, features that directly correspond to physiological phenomena hold the most interest. In the current context, the focus is on the ECG frequency band with the highest power, around 1 Hz, which corresponds to the heart beat, and the differences in frequency correspond to the heart rate variability (HRV). At that band, the HRV signal and CWT ridge around the main band appear to be similar. However, while HRV signals can be considered as frequency fluctuations as a function of time, the CWT is also a source of phase information, the magnitude (power) of the ridge components, and the ability to obtain coherence between ECG and other physiological obtained simultaneously [9].

The representation of the CWT can be analyzed with a variety of approaches, and features corresponding to specific physiological phenomena can be extracted [7], [8], [9]. The main feature of interest is therefore the band around 1 Hz, which also consistently contains the highest spectral power. In previous work, roughness, entropy, and approximate entropy were used to analyze this band in the context of lower-body negative-pressure conditions [8],

[9]. The current study considers longer signals over the entire course of a driving simulation, as well as short, transient events, such as experimental stressors. As in the previous studies just cited, the goal is to quantify the variability, complexity, and unpredictability of the main frequency band. The standard variability and complexity measures, mean, standard deviation, interquartile range (IQR), and entropy, were determined for the ridge of the main frequency band of each CWT.

Entropy (H), or information content, is an important measure of uncertainty, or disorder, in a random variable [14], and was used in previous work to quantify "unpredictability", or "randomness" of the main CWT frequency band ridge [9]. Entropy is computed as:

$$H(x) = -\sum_{i=1}^{N} p(x_i)\ln(p(x_i)), \tag{4}$$

where x_i is a discrete sample of $x(t)$, and p denotes an estimate of the probability density of $x(t)$, which was estimated with a Gaussian kernel at 50 equally spaced points between 0.75 and 1.40 Hz. This range was sufficient to include all the frequencies present in all ridges [9].

In addition to these standard measures, CWT bands were analyzed with sample entropy and the mean absolute difference, which are discussed below.

2.7. Dispersion Measures

Statistical dispersion quantifies the variability, scatter, or spread of a distribution. Variance and standard deviation are canonical dispersion measures, along with range and interquartile range. A relatively new dispersion measure is the mean absolute difference (MAD), also known as the Gini mean difference [15]. The MAD differs from the standard deviation in that the former is twice the second L-moment (a statistic for quantifying population asymmetry and tailedness [16]), while the latter is defined in terms of central tendency. For a sequence of sample values x_i, $i = 1, ..., N$, MAD is given as:

$$MAD = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|. \tag{5}$$

Therefore, MAD can be considered as the mean of the distance matrix of x.

A related measure is the relative mean absolute difference (RMAD), which provides a measure of dispersion when the probability distribution has a non-zero mean. It is computed from MAD as:

$$RMAD = \frac{MAD}{\bar{x}}. \tag{6}$$

Other dispersion measures computed for the CWT band are the standard deviation and the interquartile range. The latter is a robust measure of scale that is relatively unaffected by a small number of outliers.

2.8. Sample Entropy

Many metrics have been proposed to analyze dynamical nonlinear physiological signals. One such measure, approximate entropy (ApEn), was developed to quantify complexity based on Kolmogorov entropy [17]. ApEn measures regularity of fluctuations in a time series: higher frequency and similarity of such fluctuations yield low ApEn values. For some types of analysis, ApEn is a good indicator of variability and complexity in wavelet features [8], [9]. ApEn is parameterized by N, the number of points in a sequence, and m, the number of points that remain similar at the next point for sequences of size N, within a tolerance value $r > 0$ [17], [18]. ApEn(m, r, N) then approximates the likelihood that similar patterns of observations will not be followed by additional similar observations.

It has been found, however, that ApEn is a biased estimator, and is sensitive to data length (see [18] and the bibliography cited there). To improve the measure, sample entropy (SampEn) was introduced. SampEn is different than ApEn in that it produces more consistent measurements over a range of varying conditions, is relatively independent of record length, is simpler and more efficient to compute than ApEn, and, unlike ApEn, self-matches are not counted. Like ApEn, SampEn is parameterized with N, m, and r, and lower values indicate more self-similarity in the time series. SampEn has been found to be a useful indicator of disease states [18]. Because of these advantages, and especially because the driving events analyzed in this study were of variable (and often short) data length, SampEn was chosen as an indicator of complexity over ApEn. The following discussion of sample entropy follows the development presented in [18]:

Given the designations of N, m, and r above, let $B_i^m(r)$ represent $(N-m-1)^{-1}$ multiplied by the number of vectors such that $\|\mathbf{x}_m(j) - \mathbf{x}_m(i)\| < r$, and let $A_i^m(r)$ represent $(N-m-1)^{-1}$ multiplied by the number of vectors such that $\|\mathbf{x}_{m+1}(j) - \mathbf{x}_{m+1}(i)\| < r, j = 1..N-m, j \neq i$. The probability that two sequences will match for m points as is denoted as $B^m(r)$, and the probability that two sequences will match for $m+1$ points is denoted as $A^m(r)$:

$$B^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m(r), \tag{7}$$

and

$$A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r). \tag{8}$$

Then,

$$SampEn(m, r, N) = -\ln\frac{A^m(r)}{B^m(r)}. \tag{9}$$

2.9. Ridge Detection

Ridge analysis in CWT representations provides interesting insights into the underlying time series [19]. Although sophisticated ridge detection algorithms exist [20], [21], the approach taken in the current study used the maximum power value along the frequency axis in a pre-specified range of frequencies (between 0.75 and 1.40 Hz) for each time instant. The ridge was then minimally smoothed with the locally weighted scatter plot smooth (LOWESS) method with a span of 2 seconds.

2.10. Implementation and Statistical Analysis

The sCWT was implemented according to Eq. 3 [13]. It was found empirically that $k = 14$ resulted in the most pronounced high-power frequency band. Custom software was written in the Matlab environment (The Mathworks, Natick, MA, USA). Because of their computational complexity, both MAD and SampEn were calculated with custom Matlab executables. The SampEn was adapted from code written by N. Hammerla[1], which was based on the original Matlab code by K. Lee[2].

For sample entropy, the embedding dimension m was set to 0.5, 1, 2, and 4, and r was set to 0.2 times the standard deviation of $x(t)$, following the literature [18].

To assess the differences in the band features between resting and driving conditions within the same participant, a paired-samples t-test was used, with $\alpha = 0.05$. Because of the dependence of the two samples (the same individual in control and experimental conditions) and due to the relatively small sample size ($n = 17$), the non-parametric Wilcoxon signed rank test was also performed. IBM SPSS was used for statistical analyses.

3. Results

3.1. Visualization

Although statistical analysis is necessary to assess variability measures in different conditions, the focus is increasingly on patient-specific measures in biomedicine [22]. Consequently, the visual CWT representation retains importance.

An sCWT representation is shown for Subject 4 for one minute of test driving (rest) in Figure 1, and for over 13 minutes of driving under stressor conditions in Figure 2. The ridges for the main frequency band are highlighted in white. The sCWT images for a different subject (Subject 6) during resting and driving are shown in Figures 3 and 4, respectively. For the stressed driving conditions, the start times for each event are also delineated.

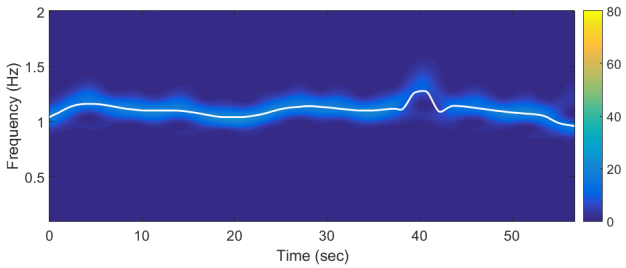


Figure 1. sCWT for test drive/resting condition for Subject 4.

3.2. Statistical Analysis

The p-values for the paired t-test and Wilcoxon signed rank test over the entire duration of the simulation are shown in Table 1. It is seen that the standard metrics (mean, standard deviation, and IQR) were all significant at $\alpha = 0.05$. The MAD and RMAD measures were also significant. However, entropy (H), which denotes complexity or unpredictability, was not significant.

To further assess the overall differences between the two conditions, entropy (Eq. 4), MAD (Eq. 5), and RMAD (Eq. 6) were computed for each frequency band in the sCWT. Using all subjects, the 95% confidence interval was determined for both conditions. The differences at a specific frequency can be considered significant if the confidence bands around the mean do not overlap at that frequency. In the wavelet domain, low entropies occur

1. https://www.mathworks.com/matlabcentral/fileexchange/53591-fast-sample-entropy
2. https://www.mathworks.com/matlabcentral/fileexchange/35784-sample-entropy

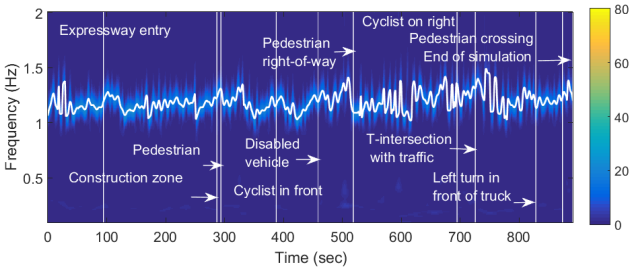


Figure 2. sCWT for driving during different events for Subject 4.

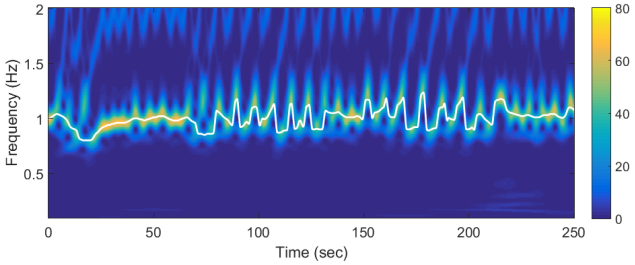


Figure 3. sCWT for test drive/resting condition for Subject 6.

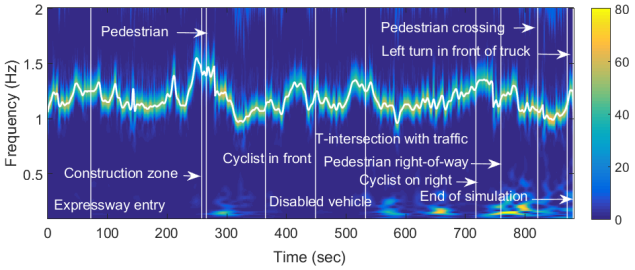


Figure 4. sCWT for driving during different events for Subject 6.

TABLE 1. VARIABILITY MEASURES FOR RESTING-DRIVING, DF = 16

Metric	Rest - Drive	T-test p	Wilcoxon p
Mean	-0.1208	0.000	0.000
Stdev	-0.0252	0.008	0.011
IQR	-0.0349	0.009	0.011
MAD	-0.0286	0.009	0.011
RMAD	-0.0166	0.036	0.049
Entropy	0.0739	0.251	0.210

when the larger coefficient energies are concentrated at only a few discrete locations [3]. Entropy was calculated using a Parzen window kernel with 100 equally spaced points. The entropy confidence intervals are shown in Figure 5, and in Figure 6 for RMAD. MAD showed no large region of significant differences in the two conditions.

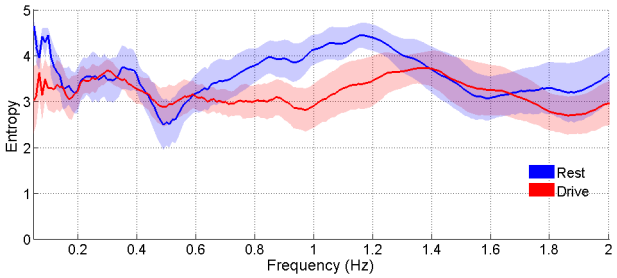


Figure 5. Entropy of frequencies from CWT for resting and driving conditions, with area of significant differences around 1 Hz.

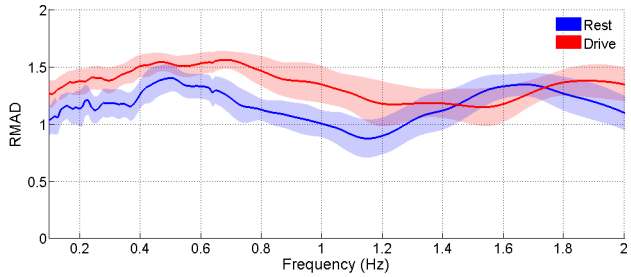


Figure 6. Relative mean absolute difference of frequencies from CWT for resting and driving conditions, with area of significant differences around 1 Hz.

From the paired t-tests of individual events, significant differences ($p < 0.05$) for SampEn (for all m) and the MAD and RMAD dispersion metrics (computed from the main frequency band of each CWT) are shown in Figure 7. As can be seen in this figure, not every event yielded significant differences between the control and experimental conditions, as measured by sample entropy and the dispersion metrics. Furthermore, the different events exhibited differing significance of the metrics.

3.3. Significance for Complexity Measures

Statistics were not computed for events 10 – 12 because of the low degree of freedom for these events. SampEn for the control was significantly higher for *expressway entry*, *pedestrian*, and *cyclist* (but only for $m = 4$). SampEn for the experimental condition was significantly higher for *pedestrian right-of-way* and *pedestrian crossing* for embeddings $m = 0.5$ and $m = 1$. The dispersion measures were significantly higher for resting in *construction zone* and *pedestrian crossing*, and lower in *pedestrian*. Interestingly, the metrics were significant, but with different polarity (sign of Rest – Drive), for *pedestrian* and *pedestrian crossing*, reflecting the different polarities shown in Figures 5 and 6.

4. Discussion

The main frequency band in the sCWT of the control (resting/test drive) condition is either smoother (Figure 1) or more regular (Figure 3, where some periodicity is exhibited) than in the stressed driving conditions (Figures 2 and 4). This result is consistent with previous studies showing more disorder and complexity in the main CWT band for experimental conditions [8], [9]. This finding is confirmed statistically in Table 1. However, aggregating over the entire time duration by frequency, the main frequency band in the control had significantly higher entropy, as evidenced by the gap between the control and experimental conditions in Figure 5 for frequencies of 0.75 to 1.25 Hz. Powers for this band are more dispersed in the experimental condition for the same range of frequencies, as seen in Figure 6. The lower entropy in the experimental condition suggests that the high power wavelet coefficients are more concentrated than in the control condition [3], but the overall dispersion of coefficients of all powers is higher in the experimental condition.

An interesting result is that over the entire duration of the driving simulation, all complexity measures (mean, standard deviation, IQR, MAD, RMAD) were higher for the experimental conditions than for the control, but the ridges for individual events more closely followed the expected behaviour of HRV under stressor conditions (i.e. lower variability). Although further study is needed to understand these findings, it can be speculated that certain events (e.g. concerning pedestrians) the autonomic nervous system causes a greater stress response, and therefore more regular (and faster) heart rate (measured by SampEn), but with a higher spread of spectral power in relevant frequency bands (measured by dispersion).

Ridge metrics should also be used with CWT visualizations. For instance, low frequency features, seen especially in the full record driving ECGs (Figure 4) and quantified with entropy and RMAD (Figures 5 and 6) indicate important physiological events requiring further study. It is notable that in heartbeat band ($\sim$ 1 Hz), wavelet entropy is significantly higher in the rest condition (Figure 5), but RMAD is significantly lower while resting (Figure 6). This may indicate that the heart rate in stressed driving has greater dispersion of spectral power in the 1 Hz band (as measured by RMAD), but less complexity and more predictability of spectral components (as measured by entropy). In this case, time-frequency analysis uncovered interesting features, the physiological implications of which require further study,

5. Conclusion

Although no clinical conclusions can be drawn or hypothesized at this preliminary stage, the results presented in this paper suggest interesting frequency features in physiological time series are obtained from both the CWT of entire duration of the signal, as well as from from individual events within the signal. Many of the ridge features contributing to the higher complexity (as measured by the metrics used in this paper) may occur during the transition from one event to another. Events happening in quick succession may also have a cumulative effect that are manifested in the entire CWT.

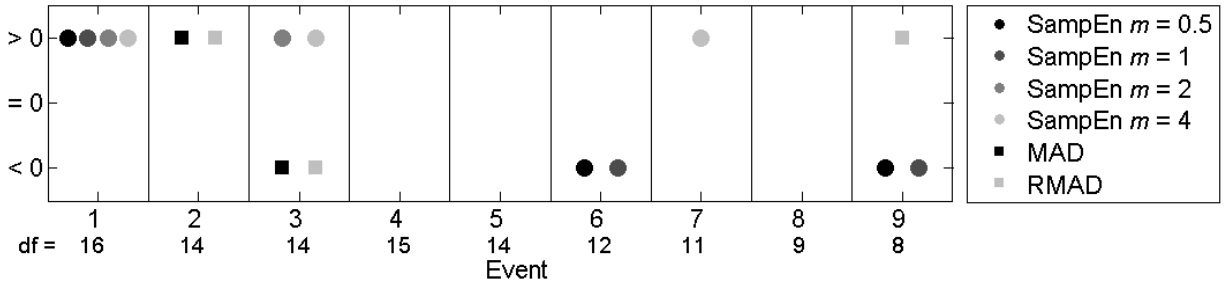


Figure 7. Significant driving events for SampEn and dispersion measures, Rest - Event. Degrees of freedom for each event is shown below the event number.

This paper proposes that, in addition to visual assessments, band metrics and other features obtained through time-frequency representations can be used for exploratory analysis of physiological signals. It is also suggested that a variety of complexity measures be used, as these metrics measure different phenomena that could potentially be useful to build new insights. For instance, in this paper, entropy/SampEn and dispersion, when significant, often had different polarities, thereby pointing to areas requiring further investigation. The results also highlight the challenges of physiological signal analysis in real, dynamic situations. Therefore, in the context of the very individual task of driving under stressor conditions, the visualizations and metrics presented here may be most useful for patient-specific analysis, rather than on broad population trends.

Acknowledgments

R.W.S. and M.J.J. are supported by the Atlantic Innovation Fund. M.P.W. is supported by NSERC grant 386586-2011. The authors thank Dr. S. Nekary for assisting with the data collection, Mr. D.J. DuVal for programming assistance, and Dr. D.C. Hay for useful critique.

References

[1] Canadian Medical Association *et al.*, "Determining medical fitness to operate motor vehicles," *CMA drivers guide. 7th ed.*, 2006.

[2] M. J. Johnson, T. Chahal, A. Stinchcombe, N. Mullen, B. Weaver, and M. Bedard, "Physiological responses to simulated and on-road driving," *International Journal of Psychophysiology*, vol. 81, no. 3, pp. 203–208, 2011.

[3] P. S. Addison, "Wavelet transforms and the ECG: A review," *Physiological measurement*, vol. 26, no. 5, p. R155, 2005.

[4] T. Ince, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ECG signals," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1415–1426, 2009.

[5] M. K. Das and S. Ari, "Patient-specific ECG beat classification technique," *Healthcare Technology Letters*, vol. 1, no. 3, pp. 98–103, 2014.

[6] A. Bhattacharyya and R. B. Pachori, "A multivariate approach for patient specific eeg seizure detection using empirical wavelet transform," *IEEE Transactions on Biomedical Engineering*, vol. PP, no. 99, pp. 1–1, 2017.

[7] R. B. Graham, M. P. Wachowiak, and B. J. Gurd, "The assessment of muscular effort, fatigue, and physiological adaptation using EMG and wavelet analysis," *PLOS ONE*, vol. 10, no. 8, p. e0135069, 2015.

[8] M. Wachowiak, D. Hay, and M. Johnson, "Quantification of wavelet band metrics for assessing heart rate variability," in *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer, 2015, pp. 1026–1029.

[9] M. P. Wachowiak, D. C. Hay, and M. J. Johnson, "Assessing heart rate variability through wavelet-based statistical measures," *Computers in Biology and Medicine*, vol. 77, pp. 222–230, 2016.

[10] M. Tremblay, F. Gallant, M. Lavallière, M. Chiasson, D. Silvey, D. Behm, W. J. Albert, and M. J. Johnson, "Driving performance on the descending limb of blood alcohol concentration (bac) in undergraduate students: a pilot study," *PLOS ONE*, vol. 10, no. 2, p. e0118348, 2015.

[11] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.

[12] L. Keselbrener and S. Akselrod, "Selective discrete fourier transform algorithm for time-frequency analysis: Method and application on simulated and cardiovascular signals," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 8, pp. 789–802, 1996.

[13] E. Toledo, O. Gurevitz, H. Hod, M. Eldar, and S. Akselrod, "Wavelet analysis of instantaneous heart rate: A study of autonomic control during thrombolysis," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 284, no. 4, pp. R1079–R1091, 2003.

[14] T. M. Cover and J. A. Thomas, "Entropy, relative entropy and mutual information," *Elements of Information Theory*, vol. 2, pp. 1–55, 1991.

[15] S. Yitzhaki *et al.*, "Ginis mean difference: A superior measure of variability for non-normal distributions," *Metron*, vol. 61, no. 2, pp. 285–316, 2003.

[16] J. R. Hosking, "L-moments: analysis and estimation of distributions using linear combinations of order statistics," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 105–124, 1990.

[17] S. Pincus, "Approximate entropy (apen) as a complexity measure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 110–117, 1995.

[18] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[19] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[20] R. A. Carmona, W. L. Hwang, and B. Torrésani, "Characterization of signals by the ridges of their wavelet transforms," *IEEE Transactions on Signal Processing*, vol. 45, no. 10, pp. 2586–2590, 1997.

[21] P. Tomassini, A. Giulietti, L. A. Gizzi, M. Galimberti, D. Giulietti, M. Borghesi, and O. Willi, "Analyzing laser plasma interferograms with a continuous wavelet transform ridge extraction technique: The method," *Applied Optics*, vol. 40, no. 35, pp. 6561–6568, 2001.

[22] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

A New Simulation Approach of the Electromagnetic Fields in Electrical Machines

D. Yarymbash
Electrical Machine Department
Zaporozhye National Technical University
Zaporozhye, Ukraine
e-mail: yarymbash@gmail.com

M. Kotsur
Department of the Electrical and Electronic Apparatuses
Zaporozhye National Technical University
Zaporozhye, Ukraine
e-mail: kotsur8@gmail.com

S. Subbotin
Software Tools Department
Zaporozhye National Technical University
Zaporozhye, Ukraine
e-mail: subbotin@zntu.edu.ua

A. Oliinyk
Software Tools Department
Zaporozhye National Technical University
Zaporozhye, Ukraine
e-mail: olejnikaa@gmail.com

***Abstract*— Theoretical researches and simulation results, which based on numerical realization of the finite element method of three-dimensional mathematical model of the induction motor are obtained. The regularities of the distribution of the induction and magnetic field energy in the short-circuit mode and their quantitative relation for active zone and the area of the coil ends of the stator windings of the low-power asynchronous motors are defined. A new approach for three-dimensional simulation of the electromagnetic process in the induction motor, which consists in differentiating the size of the finite elements and use of approximation functions of Lagrange polynomials, witch based on finite element method are realized. It provides high convergence of numerical realization of transient processes short-circuit mode, reducing the computation time, the requirements for computing resources and high simulation accuracy. Comparison of the energy values of magnetic field of the induction motor in short-circuit mode shows, that for Lagrange polynomials approximating the first degree, the relative error did not exceed 3,8% as compared with approximating polynomials of the third degree, while reducing the calculation time in 389 times and requirements for the computational resources - up to 10 times.**

Keywords*— induction motor, electromagnetic field, finite element method, approximating functions.

I. INTRODUCTION

For today, many problems associated with the peculiarities of the electrical machines in its different modes, the processes of electromagnetic and electro-thermal energy conversion and their relative effect on the operation modes are not understood [1]-[3]. Recently there is a tendency of displacement of the experimental methods of research of the electric machines prototypes with help numerical experiment, using the most precise mathematical models, which based on field theory [4] and electric circuits [5]. The most common in this sense, are three-dimensional modeling, which based on the finite element method. The advantage of such models is the possibility of considering the features of complex non-linear structure of the electrical machine, the influence of leakage fields from the end coils of the stator windings, and edge effects in the rotor and the stator of the electrical machine. This will increase the accuracy of calculations, to ensure a substantial reduction in the cost of development and production of prototypes, satisfying the modern requirements of energy efficiency and reliability of its work.

II. ANALYSIS OF LAST RESEARCHES

As practice shows, the classical methods of calculating of the electrical machines and the analysis of their work in different modes, based on a number of assumptions, that can be lead to significant errors of calculation, which does not satisfy modern requirements for energy efficiency. Specifically, it is inherent to electrical machines for the middle and high powerful, which have non-traditional design [6]-[13].

With the development of computer technology and specialized software systems, it became possible to use electromagnetic field simulation using the finite element method to calculate the parameters of the electrical machine, based on the analysis of electromagnetic processes occurring in them, providing high accuracy [14]-[15].These problems discussed in works, such as [16]-[25].

However, a significant difference of geometric parameters of the individual elements of the electrical machines, a significant non-linearity of the electrical and magnetic properties of materials, lead to substantial difficulties in the implementation of the electromagnetic field models of the electrical machines and significant computing resources and simulation time [20],[21]. Therefore, in most researches [16], [18], [22]-[24] electromagnetic processes in electrical machines considered in plane-parallel approximation, without the

influence of the leakage fields from the end coils of the stator windings, edge effects in the rotor and stator (distortion of the magnetic field in the edge regions), and other effects that can lead to increase the accuracy of calculations [25]-[27]. Also, in the assumption of linearity of the magnetic properties of ferromagnetic materials have been applied [28]-[30].

The application of these assumptions for the calculation of the electrical machines excludes possible to specify its characteristics and takes adequate estimates of the parameters of the processes in them. This leads to the relevance of developing new approaches that expend the application domain of the existing methods for calculation of spatial electromagnetic fields, considering constructive features of the electrical machine, the nonlinearity of the electrical and magnetic properties of the active and structural materials, operating modes, which providing computational efficiency and accuracy of the calculation.

III. FORMULATION OF THE WORK PURPOSE

The aim is to the development of the new effective approach for the realization of three-dimensional mathematical model of transient electric and magnetic fields in induction motors, which based on their design features, the nonlinearity of the electrical and magnetic properties of the active and structural materials, which provides reliability and high simulation accuracy.

IV. EXPOUNDING THE MAIN MATERIAL

Research carried out for the short-circuit dynamic mode of the induction motor for type 4A50A2 (0.09 kW). The geometric model and computational domains is shown in Fig. 1 and Fig.2

The three-dimensional model of AD includes: 3D geometric domains of the of the stator core – 1 (j = 1); 3D geometric domain of the stator windings – 2 (j = 2); 3D geometric domain of the stator insulation – 3 (j = 3); 3D geometric domain of the rotor core – 4 (j = 4); 3D geometric domain of the rotor winding bars – 5 (j = 5); 3D geometric domain of the rotor squirrel cage rings – 6 (j = 6); 3D geometric domain of the shaft – 7 (j = 7); 3D geometric domain of the case induction motor – 8 (j = 8). The geometric domains of the case and shaft of the induction motor are not displayed.

At the mathematical description of electromagnetic processes, the assumption of isotropy of the electrical and electromagnetic properties of materials, lack of bias currents and free of charges are accepted [7]. In this case, the transient electromagnetic processes in induction motor in the short-circuit mode can be represented by coupling system of the nonlinear partial differential equations [4],[22],[31]:

$$\begin{cases} \sigma_j \partial \mathbf{A}_j / \partial \tau + \sigma_j(\theta_j) \nabla V_j + \\ + \nabla \times \left[\left(\mu_0 \mu_j(B) \right)^{-1} \nabla \times \mathbf{A}_j \right] = \mathbf{J}_{ej}; \\ - \nabla \cdot \partial \left(\varepsilon_0 \varepsilon_r \nabla V_j \right) / \partial \tau - \\ - \nabla \cdot \left(\sigma_j(\theta_j) \cdot \nabla V_j - \mathbf{J}_{ej} \right) = 0, \end{cases} \quad (1)$$

where $\mathbf{A}$ – vector magnetic potential; V – electric potential; $\sigma(\theta)$ – specific conductivity; B – magnetic field induction; μ – relative magnetic permeability; ε_r - relative permittivity; θ – temperature; $\mathbf{J}_e$ – external current source density; indices *j*– corresponds to subareas of the computational domain (fig.1).

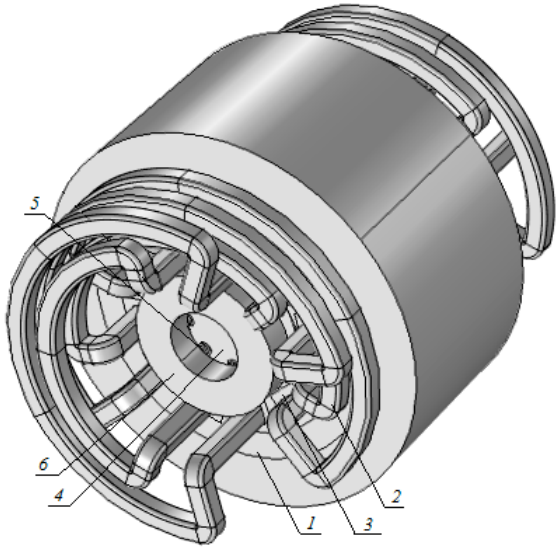


Figure 1. Geometric model of the induction motor for type 4A50A2

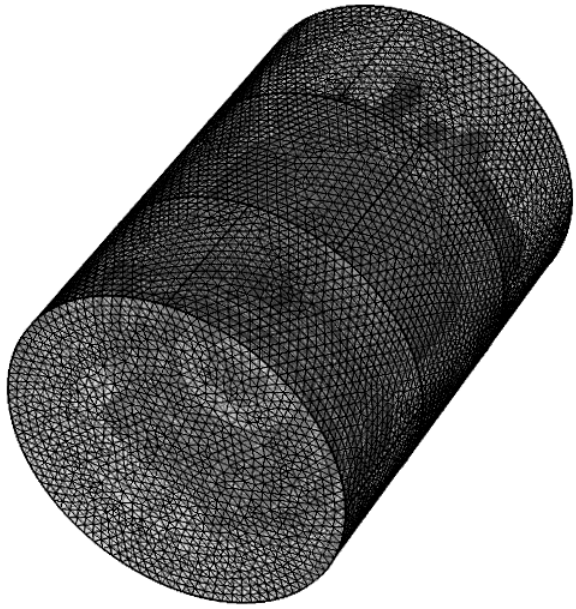


Figure 2. Computational domains of the induction motor for type 4A50A2

In accordance with [20]-[21],[31] the system of equations (1) is supplemented by the condition of Coulomb calibration. Conditions for coupling of the magnetic and electric fields can be formulated as:

$$\begin{cases} \mathbf{n}_{i,k} \times (\mathbf{H}_i - \mathbf{H}_k) = 0\big|_{\forall i,k \in (1,4), i \neq k}, \\ \mathbf{H} = (\mu_0 \mu)^{-1} \nabla \times \mathbf{A}, \\ \mathbf{n}_{i,k} \cdot (\mathbf{J}_i - \mathbf{J}_k) = 0\big|_{\forall i,k \in (1,4), i \neq k}, \\ \mathbf{J} = \sigma(\theta) \cdot \mathbf{E}, \ \mathbf{E} = -\nabla V - j\omega \mathbf{A}, \end{cases} \quad (2)$$

where H – magnetic field intensity; E – electric field intensity; ω – angular frequency.

On the external borders of the computational domain, the boundary conditions are defined as [20],[21]:

$$\begin{cases} \mathbf{A}_j = 0\big|_{\forall j \in (1,4)}, A_j = \overline{k} \cdot A_y(x,z)|_{j=4}, \\ V_j = \varphi_j\big|_{\forall j \in (1,3)}, \\ \mathbf{n}_j \cdot (\mathbf{J}_j) = 0\big|_{j=4} \end{cases} \quad (3)$$

The temperature mode of the induction motor is estimated in accordance by [9].

The initial conditions associated with the first commutation law [4],[26]:

$$\begin{cases} i_A|_{0-} = i_A|_{0+} = 0; \\ i_B|_{0-} = i_B|_{0+} = 0; \\ i_C|_{0-} = i_C|_{0+} = 0, \\ u_A|_{0+} = 0; \\ u_B|_{0+} = \sqrt{2} U_\phi \sin(2\pi/3); \\ u_c|_{0+} = \sqrt{2} U_\phi \sin(4\pi/3). \end{cases} \quad (4)$$

The three-dimensional model (1) with the conditions Coulomb calibration, the boundary conditions (2), (3) and the initial conditions (4), is realized the finite element method [35] in the structure of the software systems Comsol Multiphysics.

Evaluation of the magnetic field and its energy average values for each j-th zone computational domain is performed by next expression [20],[21]:

$$\begin{aligned} W\big|_j &= \frac{1}{2} \iiint_{V_j} (\mathbf{B} \cdot \mathbf{H}) dxdydz; \\ w\big|_j &= W\big|_j / V_j; V_j = \iiint_{V_j} dxdydz. \end{aligned} \quad (5)$$

For visualization results of the numerical simulation, the magnetic field localization area is displayed, for point in time, when short circuit dynamic process approaching stationary mode at τ = 0,04 (sec). This area corresponds to the changes in the value range module of the magnetic vector potential, which limited values |Amin| = $0{,}74 \cdot 10^{-3}$ (Wb/m) and |Amax| = $3{,}74 \cdot 10^{-3}$ (Wb/m) (Fig. 3).

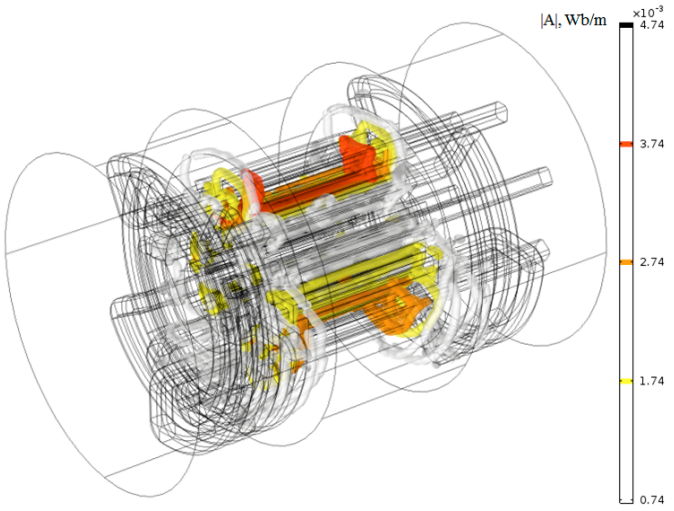


Figure 3. Equipotential lines of the magnetic vector potential |A|.

Equipotential lines of the magnetic vector potential, which equal |Amin| = $0{,}74 \cdot 10^{-3}$ (Wb/m), located along the slots, the crowns of the teeth of the stator and rotor of the induction motor. They are become isolated through the end surfaces of the stator and the active part of the rotor with squirrel cage rings and partially - through the frontal part of the stator windings (Fig.3).

Location equipotential lines for values of the vector magnetic potential module, which equal $1{,}74 \cdot 10^{-3}$, $2{,}74 \cdot 10^{-3}$, $3{,}74 \cdot 10^{-3}$ (Wb/m), has an identical character. They extend along the crown of the stator teeth and through of the slots among end surface of the stator at a distance of 15 ... 20% of the length of the active part of the stator are become isolated (Fig. 3).

Equipotential lines corresponding to the maximum value of the magnetic vector potential, concentrated along the slots with conductors sections of the stator winding with the highest current load at a given time.

Thus the boundary surface of the localization domain of the magnetic field of the induction motor extend along the stator and rotor slots, extending at their end surfaces and cover crowns of teeth, rotor squirrel cage rings and partially frontal part of the stator windings (Fig. 3).

These features of the localization of the magnetic field in the stator and rotor ends are caused by their final axial dimensions, that enhance the effect of the mutual influence of their own fields in the active zones of the stator and rotor of the induction motor, as well as the from stray fields caused by currents in the frontal parts of the stator windings [32],[36].

In addition, these factors determine the character of the distribution in the computational domain specific magnetic field energy (Fig. 4), (Fig. 5).

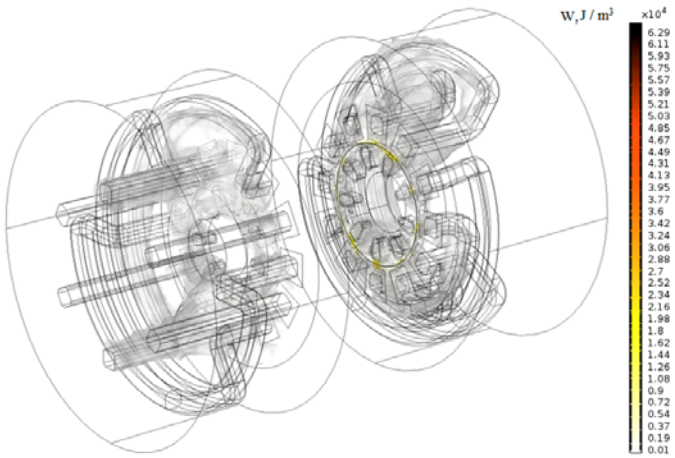


Figure 4. Equienergy surface of the magnetic field energy (W) in the frontal areas of the computational domain

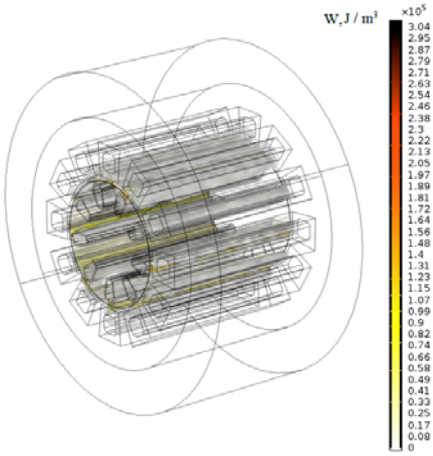


Figure 5. Equienergy surface of the magnetic field energy (W) in the center areas of the computational domain

The frontal subareas computational domain of the induction motors, the equienergy surface with an initial specific energy of the magnetic field, equal to 10 (J / m^3), covers the conductors of the phase of the stator winding with the greatest current load at a given time, as well as the rotor winding (Fig. 4). The increase of the specific energy of the magnetic field in the range from $1{,}4\cdot10^4$ (J / m^3) to $6\cdot10^4$ (J / m^3) in the air gap and slots with the sections of the phase stator windings, at the largest at a given time current loadings, in their subarea of the induction motor is observed (Fig. 4).

Equienergy surface with an initial specific energy of the magnetic field equal to 100 (J / m^3) in the subarea of the active part of the stator and rotor induction motor covers the air gap and partly stator winding conductors of the induction motor with maximum load current at a given time. Next, they expand near ends subarea of the stator and rotor induction motor. In this subarea localization, values of the specific energy of the magnetic field are in the range ($0{,}74\ldots3\cdot10^5$ J / m^3) and observed in the air gap of the induction motor. The maximum density of the magnetic field energy is observed in the area near the conductor with the highest current load (Fig. 5).

Based on the results follows, that the value of the magnetic field energy in the frontal subareas computational domain of the induction motor (Fig. 4), (Fig. 5) is to 12,5% of the magnetic field energy value for the total estimated area of the induction motor. Moreover, in 6,7 times less than the magnetic field energy value for the central subarea of the induction motor (cores of the stator and the rotor with only slot part stator and rotor windings). Geometric features of the distribution of the magnetic field energy in the computational domain of the induction motor (Fig. 4), (Fig. 5) determinate the quantitative estimation of the effects of the self and mutual inductance of the windings, the surface and the edge effects in the stator and rotor of the induction motor [36]. Therefore, the ratio of the values of the magnetic field energy in the frontal and central subareas should be taken into account in the calculation of the parameters of the low powers induction motor and their modes of operation researches.

For reduce the computational resources and expenses of time of the implementation of the three-dimensional model of electromagnetic processes (1), differentiation of size of the finite element computational domain was applied (Fig. 2). The density of the finite elements in the conductive elements and at the boundaries of the conjugation elements was increased. To the outer boundaries progresses of the computational domain, the finite element density was decreased. Also, to effectively reduce the expenses of time and computational resources, improving the convergence of the numerical model implementation (1), Lagrange polynomials of the first, second and third degrees for approximating functions of finite element method are considered [37]. Higher order Lagrange polynomials are usually used in greater fragmentation of the computational mesh, especially for complex nonlinear surfaces and three-dimensional formulation of the problem. However, the use of higher order shape functions complicates the transformation method, and therefore provides increased the dimension of equations and their coefficients.

To evaluate the degree of influence of the Lagrange polynomials on the accuracy of the realization three-dimensional dynamic model (1) by finite element method, the ratio for the relative discrepancies of the magnetic field energy values was used (Table 1):

$$\delta\tilde{W}_i\big|_\tau = \left[\frac{\left|W_i\big|_\tau - W_{\text{base}}\big|_\tau\right|}{\max\left(W_{\text{base}}\right)}\right]\cdot 100\%$$

For the average values of magnetic field induction in the air gap:

$$\delta\tilde{B}_{\delta i}\big|_\tau = \left[\frac{\left|B_{\delta i}\big|_\tau - B_{\delta\text{base}}\big|_\tau\right|}{\max\left(B_{\delta\text{base}}\right)}\right]\cdot 100\%$$

as well as their average values for the magnetic energy and magnetic field induction in the air gap for the time interval $0 \le \tau \le 2 \cdot T$

$$\tilde{W}_{av\_i}\Big|_{\tau} = \left[\frac{\sum_{\tau=0}^{\tau=\tau_{max}} \left| W_i\big|_{\tau} - W_{base}\big|_{\tau} \right|}{\sum_{\tau=0}^{\tau=\tau_{max}} \left| W_{base}\big|_{\tau} \right|} \right] \cdot 100\%$$

$$\tilde{B}_{av\_\delta i}\Big|_{\tau} = \left[\frac{\sum_{\tau=0}^{\tau=\tau_{max}} \left| B_{\delta i}\big|_{\tau} - B_{\delta base}\big|_{\tau} \right|}{\sum_{\tau=0}^{\tau=\tau_{max}} \left| B_{\delta base}\big|_{\tau} \right|} \right] \cdot 100\%$$

where W_{base} – value of the magnetic field energy of the according to the numerical realization of the system of equations (1) using the approximate Lagrange polynomial of the third degree; i = 1,2 –Lagrange polynomial degree.

TABLE I. TOLERANCES, REQUIREMENTS TO COMPUTING RESOURCES AND COMPUTATION TIME OF THE DYNAMIC MODEL (1) BY FINITE ELEMENT METHOD

Approximation of the shape function	Relative discrepancies, %	Weighted average discrepancies, %
	W, (J / m^3)	
linear	0,23…3,78	3,635
Quadratic	0,09…0,89	0,883
Cubic	-	-
	B_δ, (Wb/m^2)	
linear	0,02…1,12	0,781
Quadratic	0,02…0.98	0,778
Cubic	-	-
	Computation time, r.u	Memory (RAM), r.u.
linear	1	1
Quadratic	17,4	2,51
Cubic	389,02	9,65

The accuracy, expenses of time and computing resources requirements (the amount of RAM) for the numerically-field simulation using the finite element method and the approximate Lagrange functions with the first polynomials, second and third degree, realized in the structure of the software systems funds Comsol Multiphysics in the simulation was estimated (Table. 1).

According to simulation, the discrepancies of the relative values of the magnetic field energy in the computational domain for the linear approximation function varies from 0.23% to 3.78% and the weighted average discrepancies is equal to 3,635%. For approximating the Lagrange function with second-degree polynomial, the interval changes is narrowed to 0.09 ... 0.89% and the weighted average discrepancies value decreases to 0.885%. The relative values discrepancies for the magnetic field induction in the air gap at linear approximation function changes in the range of 0.02 ... 1.12%, at the value of the discrepancies weighted average - 0.781%, and for approximating quadratic polynomial function of the second order - 0.02% and 0.98 ... 0.778%, respectively. If you use quadratic approximation for the finite element method, the computational expense increased in 2.51 times, while for the approximation of the cubic shape function - increased in some orders of magnitude, and duration of the numerical- field simulation increases in 17.4 and 389 times, respectively.

Thus, if the problem of mathematical simulation of the electromagnetic processes in induction motor is included in the system limitations, that set the interrelations between the vectors of the dependent and independent variable of structural parameters for optimization problem in electric machines, it is more preferred to use a finite element method for the linear approximation of the shape function. Calculations by the finite element method and approximation of the shape functions by Lagrange polynomials of the second order provide high accuracy for verification calculations of electric machines with optimal values of structural parameters. For induction motor low and medium powers, the numerical realization of the 3D model (1) by the finite element method with Lagrange approximating third-degree polynomials practically does not improve accuracy in comparison with second-degree polynomials, but increases simulation time to two or more orders, which is not advisable.

V. CONCLUSION

The three-dimensional mathematical model, which describes electromagnetic processes in induction motor in dynamic short-circuit modes, and taking into account the structural features, nonlinearity of the electro physical and magnetic properties of active and structural materials is proposed.

The features of the distribution of the vector magnetic potential and the magnetic field energy in the calculated domain of the induction motor, which have a significant effect on the parameters of the induction motor in the short-circuit mode are identified.

An effective approach for the realization of the three-dimensional model of electromagnetic processes in induction motor by the finite element method is proposed. It involves the use of differentiation of the density of finite elements in the space of the computational domain and the approximation of the shape function by Lagrange polynomials of the first degree. This will reduce the requirements for computational resources up to 17 times, and the time simulation up to 400 times at the relative discrepancies of these calculations less than 3.8%, compared with the use of approximating Lagrange polynomials of the third degree.

REFERENCES

[1] Kotsur M. Synchronization methods of the induction motors rotation in energy-efficient electric drive system. Fundamental and Applied Studies in the Modern World: papers and commentaries / The University of Oxford. – Oxford, 2016. – Volume XV. –P. 384-389.

[2] Kotsur M. Features of the impact of thermal effects on the asynchronous motor with the modified pulse control system in under conditions of

frequent starts. Elektrotehnika ta elektroenergetika, 2014, Issue 1, pp. 32 – 36.
[3] Kotsur M. Kotsur I, Bliznyakov A. The efficiency increase of the reversible braking mode of the wound-rotor induction motor. Eastern-European Journal of Enterprise Technologies, 2015, Volume 73, Issue 1/8, pp. 27-30.
[4] Wilow V. Electromagnetical model of an induction motor in COMSOL Multiphysics, Master's thesis, KTH University, Sweden, 2014, 41 p.
[5] Tykhovod S., Kornus T. Accelerated Calculation of Steady States Processes in Complex Nonlinear Circuits. Proceedings 2015 16th International Conference on Computational Problems of Electrical Engineering (CPEE), 2-5 September, 2015, Lviv, pp. 225-228.
[6] Kopulov I. P. Mathematical modeling of electrical machines, Moscow, Vysshaya shkola, 2001, 327 p.
[7] Kopulov I. P., Klokov V. K., Morozkin V. P. Design of electrical machines, Moscow, Vysshaya shkola, 2005, 767 p.
[8] Tolochko, O. I. Rozkaryaka P. I, Zhurov I. O. Simulation of asynchronous motor with stator phase failure. Electrotechnical and computer systems. Issue 15 (91), 2014 – pp. 262 – 266.
[9] Jannati M.; Idris N.R.N., and Aziz M.J.A. A new Method for RFOC of Induction Motor under open-phase Fault, Industrial Electronics Society, IECON 2013 – 39th Annual Conference of the IEEE, 2013, pp. 2530 – 2535.
[10] Jannati M., Idris N.R.N., Salam Z. A New Method for Modeling and Vector Control of Unbalanced Induction Motors, Energy Conversion Congress and Exposition (ECCE), IEEE, 2012, pp. 3625 – 3632.
[11] Jannati M., and Fallah E. Modeling and Vector Control of Unbalanced Induction Motors (Faulty Three Phase or Single Phase Induction Motors), 1st. Conference on Power Electronic & Drive Systems & Technologies (PEDSTC), May 2010, pp. 208 – 211.
[12] Danyal Mohammadi. Dynamic modeling of single-phase induction motor loads. Master's thesis, Boise State University, Idaho, USA, 2012, 80 p.
[13] Islam, R., Iqbal, H. (2010). Analytical Model for Predicting Noise and Vibration in Permanent-Magnet Synchronous Motors. IEEE Transactions on Industry Applications, 2010, Volume 46(6), pp. 2346-2354.
[14] Persova M. G., Soloveychik Yu. G., Temlyakova Z. S. A new approach to the design of electrical machines based on numerical simulation, Elektrotehnika. Issue 9, 2007 – pp. 15 – 21.
[15] Tarımer İ, Güven, M. E, Arslan, S. Investigation for Losses of M19 and Amorphous Core Materials Asynchronous Motor by Finite Elements Methods. Elektronika i Elektrotechnika, 2012, Volume 18, pp. 15-18.
[16] Vaskovskiy, Yu. V., Geraskin A. A. Mathematical modeling of electromagnetic fields in the squirrel cage induction motor with damaged rotor winding, Tehnicheskaya elektrodinamika. Issue 2, 2012 – pp. 56 – 61.
[17] Zamchalkin, A. S., Tyukov V. A. Numerical simulation of the process of starting an induction motor, Dokladyi TUSURa, Issue 1(25), 2012 – pp. 171 – 177.
[18] Milyih, V. I., Polyakova N. V. An analysis of harmonic composition the AC magnetic field associated with a rotating rotor turbine generator, at idle speed and short circuit modes, Elektrotehnika i elektroenergetika. Issue 2, 2013 – pp. 5 – 12.
[19] Plyugin, V. E. Numerical simulation of the electromagnetic field of the induction motor with the external massive rotor, Vestnik NTU «KHPI»,Issue 51(1024), 2013 – pp. 66 – 75.
[20] Yarymbash D. S. The research of electromagnetic and thermoelectric processes in the AC and DC graphitization furnaces. Naukoviy visnyik NGU, 2015, Issue 3., pp. 95–102.
[21] Yarymbash D.S., Oleinikov A.M. On specific features of modeling electromagnetic field in the connection area of side busbar packages to graphitization furnace current leads. Russian Electrical Engineering, 2015, Volume 86, Issue 2, pp. 86 – 92.
[22] Subbotin S., Oliinyk A., Levashenko V., Zaitseva E. Diagnostic rule mining based on artificial immune systems for a case of uneven distribution of classes in sample. Communications. Issue 3, 2016. pp. 4–12.
[23] Subbotin S., Oliinyk A., Skrupsky S. Individual prediction of the hypertensive patient condition based on computational intelligence. Proceedings of the International Conference on Information and Digital Technologies, 2015, pp. 336–344.
[24] Oliinyk A., Subbotin S.A. The decision tree construction based on a stochastic search for the neuro-fuzzy network synthesis. Optical Memory and Neural Networks (Information Optics), 2015, Vol. 24, ¹ 1, pp. 18–27.
[25] Oliinyk A.O., Skrupsky S.Yu., Subbotin, S.A. Experimental Investigation with Analyzing the Training Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search. Automatic Control and Computer Sciences, 2015, Vol. 49, Issue 1, pp. 11–20.
[26] Nguyen M. K. Predicting Electromagnetic Noise in Induction Motors, Master's thesis, KTH University, Sweden, 2014, 54 p.
[27] G. He, Z. Huang, D. Chen. Two-Dimensional Field Analysis on Electromagnetic Vibration-and-Noise, 2011, Volume. 47, Issue 4, pp. 787–794.
[28] Pellerey P., Lanfranchi V., Friedrich G. Coupled Numerical Simulation between Electromagnetic and Structural Models. Influence of the Supply Harmonics for Synchronous Machine Vibrations in Magnetics, IEEE Transactions, 2012, Volume 48, Issue 2, pp.983-986.
[29] Zamchalkin, A. S., Tyukov V. A. Numerical simulation of the process of starting an induction motor, Dokladyi TUSURa, Issue 1(25), 2012 – pp. 171 – 177.
[30] Plyugin, V. E. Numerical simulation of the electromagnetic field of the induction motor with the external massive rotor, Vestnik NTU «KHPI», 2013, Volume 1024, Issue 51, pp. 66 – 75.
[31] Yetgin, A.G., Turan, M. Efficiency Optimization of Slitted–Core Induction Motor. Journal of Electrical Engineering, 2014, Volume 65, Issue 1, pp. 60-64.
[32] Serdal Arslan, Sibel Akkaya Oy, İlhan Tarimer. Investigation of Stator and Rotor Slits Effects to the Torque and Efficiency of an Induction Motor. TEM Journal. 2017, Volume 6, Issue 1, pp. 117-125.
[33] Yetgin, A.G., Turan, M. Efficiency Optimization of Slitted–Core Induction Motor. Journal of Electrical Engineering, 2014, Volume 65, Issue 1, pp. 60-64.
[34] P. Zhou, S. Stanton, Z. J. Cendes, Dynamic modeling of three phase and single phase induction motors, Proc. IEEE Int. Conf. Electric Machines and Drives, 1999-May, pp. 556-558.
[35] Yarymbash D. S., Kilimnik I. M. Simulation features of the electromagnetic processes in the inductor sizing die of the press, // Visnik kremenchutskogo derzhavnogo politehnichnogo universitetu, Volume 1, Issue 4(45), 2007, pp. 53–55.
[36] Mogilnikov B. C., Oleynikov A. M., Strelnikov A. N. Induction motors with two-layer rotor and their application, Moscow, Energoatomizdat, 1983, 120 p.
[37] Zenkevich, O., Morgan K Finite Elements and Approximation, Moscow, Mir, 1986, 318 p.

Vehicle recognition and its trajectory registration on the image sequence using deep convolutional neural network

Dmitry Yudin, Alexander Knysh
Department of Technical Cybernetics
Belgorod State Technological University named after V.G. Shukhov (BSTU n.a. V.G. Shukhov)
Belgorod, Russian Federation
yuddim@yandex.ru

Abstract – **The article shows the methods of vehicle recognition on the image sequence and its trajectory registration. As a recognition algorithm authors used Viola-Jones method with optical flow filter and the deep convolutional neural network in combination with sliding window technique for vehicle detection task. Also authors analyze approaches to registration of detected vehicle trajectories on image sequence based on its linear and angular velocities and Kalman filter. The efficiency of vehicle detection is shown in terms of the precision and recall of recognition. Quality of vehicle registration on the image sequence is estimated by the standard deviation of results from sample values. The article also shows usage prospects of proposed algorithms as a part of driver assistance system and unmanned vehicle control system.**

Keywords – recognition, detection, vehicle, image sequence, registration, trajectory, Viola-Jones method, deep learning, convolutional neural network, driver assistance system.

I. INTRODUCTION

Machine vision methods are actively researched in direction of development of driver assistance systems and unmanned vehicles to ensure trouble-free driving and road safety. Vehicle recognition and tracking its trajectory on image sequence is one of such methods. On the first hand, its implementation involves the solution of the vehicle classification and detection problem and, on another hand, registration and prediction of their trajectory on the basis of the video stream obtained from moving cameras. A number of approaches to solve both problems have been developed in the world.

Some methods of machine learning for vehicle detection are constructed on the principle of multi scale sliding window: the method reduces the detection problem to binary classification for each point of the image over a certain rectangular neighborhood. For each rectangular area of the image taken with all possible shifts and scales, the hypothesis of the desired object's presence in the region is checked using a pre-trained classifier.

Good results are shown using cascade schemes to recognize complex objects in images, when the detector is a sequence of cascades of the so-called strong classifiers. In a turn, a strong classifier is built using an algorithm of machine learning, for example, AdaBoost (or some other version of the booster), as a linear combination of weak classifiers. Cascading of strong classifiers allows to achieve high productivity due to fast (at the cascade's first-second level) failure for the overwhelming number of regions that do not contain the desired object. Amount of such "empty" windows is several times higher than windows containing the object. So processing time of the "empty" sub-window differs from processing time of sub-window with the object several times (in proportion to the cascade length). This approach is the basis of Viola-Jones method, which uses Haar features [1-3], but this approach is poorly understood in case of vehicle recognition in different angles and lighting conditions. Classifiers based on the support vector method (SVM), which use histograms of oriented gradients (HOG), are also effective in similar problems [4], but they require a large amount of computation and are not effective in real-time systems.

Modern computing devices (GPU, FPGA) made it possible to solve in real time the recognition objects task on stereo images according to the results of depth maps evaluation [5]. This approach allows you to separate partially overlapped objects, and also effectively limit the areas of interest in the image, for example, according to characteristic dimensions in the physical world. An additional advantage is the possibility of obtaining data for estimating the coarse three-dimensional profile of the road on which the movement is being made. But as a rule, special stereo camera and parallel computing devices are expensive. So it is important to solve this vehicle detection task using monocular camera. For these algorithms information about observable objects movement is obtained as a result of the optical flow analysis, which will allow us to identify areas of interest and / or supplement the results of the operation of the vehicle detectors [6, 7]. However, estimation quality of the optical flow depends on the nature of image textures, speed of movement, information about the motion, and therefore research in this area continues.

Practical works show a significant influence of lighting conditions (time of the year, time of day, sun position relative to the camera, etc.) on the recognition quality (7 basic types of lighting are allocated and marked in the iRoads dataset image base [8]). Compensation of this condition is made by

This article is sponsored by grant of the President of the Russian Federation for state support of young Russian scientists № MK-3130.2017.9 (contract №14.Z56.17.3130-MK) on the theme "Recognition of road conditions on images using deep learning".

configuration of the recognition algorithm parameters. In [9] it is proposed to make "samples" in the "sky" and "road" zones in order to clarify the illumination parameters.

Currently, there are a number of developments and publications related to usage of deep convolutional neural networks in vision systems for the vehicle recognition. The achievements and perspectives of convolutional and recurrent neural networks in the recognition of images, speech, texts and other multidimensional signals are briefly described in the article by Y. LeCun, I. Bengio and G. Hinton in "Nature" journal [10]. They note the effectiveness and significant versatility of deep learning and high interest in applying it in the field of driver assistance systems and unmanned vehicles.

Work [11] considers the usage of convolutional neural networks for the detection of vehicles and road markings on images. Results showed that with increasing distance, the method used by the authors shows a significant decrease in the recognition recall.

The approach to recognition and detection of vehicles on satellite image is proposed in [12]. Results of this study show high accuracy in the detection of vehicles, but they are obtained for a fairly limited training sample for one particular city. At the same time, this approach with some modifications can be used to detect vehicles not just the top view but in other view angles.

The construction of a multiclass classifier for detecting various objects on images is described in the article by authors from Google [13], which showed the need to improve the quality of object recognition.

A solution of image classification problem to 1,000 classes (including cars) of the ImageNet dataset was proposed in [14] on base of convolutional neural networks. It notes the need to further improve applied architectures and training methods of the network.

When an object (vehicle) is found, it is necessary to track its trajectory, on the first hand, to prevent a collision with the observer, and on another hand, not to lose this object in case of error of the second kind during the recognition algorithm execution.

For the vehicle trajectory restoration in the dissertation [15] a method is suggested that it involves the integration of object keypoints' trajectories to construct the whole object's trajectory.

Development of driver assistance system to evaluating and predicting the vehicle trajectory according to the information obtained from the three-dimensional machine vision system is presented in [16]. During estimates calculation authors have used the model of the car's movement on which the camera was installed and the detecting and tracking method of wheels centers of the vehicle hidden from the driver.

The application of an extended Kalman filter to prevent collisions of an unmanned aerial vehicle with moving objects is described in [17]. In this case, the optimal trajectory of these objects is calculated. Then it is used in the aircraft's motion model. Results and effectiveness of the method are confirmed with the simulation. In addition, it is possible to use that method for a ground robotic vehicle.

In [18] a probabilistic prediction model of road incidents is proposed using three-dimensional tracking of the vehicles trajectory according to their motion patterns. The fuzzy self-organizing neural network is trained to identify patterns of vehicle behavior on sections of its trajectories. Each section of the trajectory is associated with the vehicle behavior pattern, on the basis of which the probability of an accident (collision) on the road is calculated. Usage of motion patterns allowed the authors of [19] to predict the vehicle behavior for a few seconds in advance. The probability of a new vehicle position is calculated using the history of traffic patterns of this vehicle. However, it is noted that developed approach allows to predict not a new position itself but the probability distribution of possible positions of the observed vehicle. In a similar paper [20] aimed at the development of advanced driver assistance systems a method is proposed that combines vehicle trajectory prediction based on the existing motion model with constant yaw and acceleration and vehicle maneuver recognition. However, to consider all possible conditions and a lot of types of maneuvers, further researches in this direction required.

II. Task Formulation

In this article we consider an algorithm that provides vehicle detection and its trajectory registration. It involves the following stages:

1) reading an image from a sequence of images or capturing a new image from a monocular camera;

2) detection of the object (vehicle, participant of the road traffic) by one of methods or algorithms, using multi scale sliding window technique and providing acceptable recognition precision and recall, for example, by Viola-Jones method or deep convolutional neural network. Objects can be vehicles of various types, for example cars, trucks. Under the detection we mean finding coordinates of the object bounding box. Overlapping of the vehicle bounding box found by algorithm and real vehicle bounding box should exceed 50%;

3) finding estimates of the new position and size of the object bounding box by one of effective methods;

4) usage of the found estimates as the predicted position of the object on the next image and registration of refined coordinates and sizes of the object bounding box in the current frame. If on the next few images the object disappears and then appears again, the algorithm still remains functional, because for these images coordinates and sizes of the object bounding box are used as the measured coordinates and sizes until the disappearance.

If the object is not found on more than n following images, this indicates a loss of the object from the scope and algorithm termination.

III. Vehicle Recognition on Image Sequence

For the vehicle detection task, statistical classifiers based on Viola-Jones method and deep convolutional neural network are used in a combination with sliding window technique.

The quality of the vehicle model depends on the power of the training sample for statistical classifiers. The model is

reliably trained only with a sufficient number of images corresponding to the possible vehicle appearance and lighting conditions. When the model hits the environment with vehicles, an appearance of which differs significantly from the training sample, the model needs to be retrained. When creating a vehicle model we were considered vehicles in an arbitrary perspective (the training sample contained 12,000 positive images (vehicles) and 20,000 negative images (non-vehicles), the testing sample contained 7,000 positive images (vehicles) and 10,000 negative images (non-vehicles)). A fragment of the training sample is shown in Fig. 1. Testing image dataset contains 7,000 vehicles and 7,000 negative images.

Viola-Jones method uses concepts of integral image, Haar features and their cascading classification using AdaBoost [1]. In this paper, classical Viola-Jones method of image classification in sliding window is supplemented by using the Lucas-Kanade-Tomashi method for optical flow calculation [21]. The construction of such filter for our task is described in [7]. Usage of an optical flow in this case increases the image recognition recall by detecting missing vehicle bounding rectangles (boxes) on image sequence. However, at the same time, it leads to false positives and lowers the recognition precision, and also increases the processing time per frame.

To eliminate false positives of such modified method of Viola-Jones, an attempt was made to apply an additional geometric filter for vehicles bounding rectangles, based on determining the horizon line and estimating the distance to objects [3]. It led to an increase in recognition accuracy on the training and test samples, but a significant decrease in recall, as shown in Table 2.

Also for image classification in sliding window we have used deep convolutional neural network which architecture is shown in Table 1. As loss function we have used binary cross entropy. Implementation of this network was made on python 3.5 using Keras [22] and Tensor Flow [23] libraries.

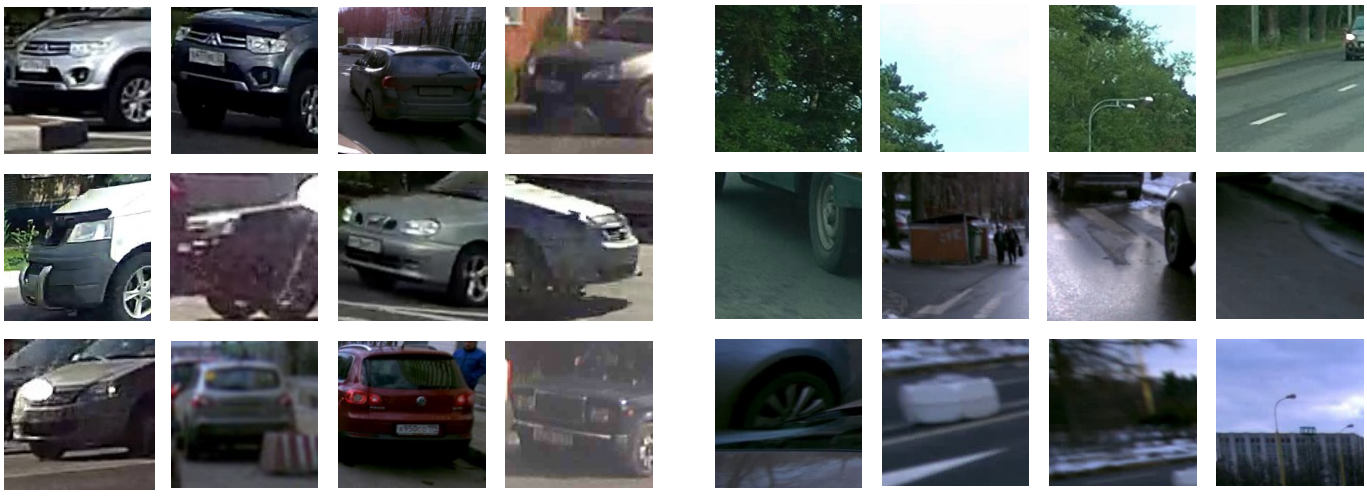
TABLE I. DEEP CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

No.	Type of Layer	Details
1.	Input Layer	24×24 neurons
2.	2D Convolutional Layer with ReLu Activation Function	3×3 core, 1×1 strides, 128 output maps with sizes 22×22
3.	2D Max Pooling	Pool size 2×2, stride 2×2, 128 output maps with sizes 11×11
4.	2D Convolutional Layer with ReLu Activation Function	3×3 core, 1×1 strides, 128 output maps with sizes 8×8
5.	2D Max Pooling	Pool size 2×2, stride 2×2, 128 output maps with sizes 4×4
6.	2D Convolutional Layer with ReLu Activation Function	3×3 core, 1×1 strides, 256 output maps with sizes 2×2
7.	2D Max Pooling	Pool size 2×2, stride 2×2, 256 output maps with sizes 2×2
8.	Dense (fully connected) Layer with ReLu Activation Function and Dropout	100 neurons, dropout with probability 0.5
9.	Dense (fully connected) Layer with Sigmoid Activation Function	1 neuron

Results of image classification for different methods are given in Table. 2.

Results of vehicle detection are estimated using three measures of evaluation: Precision, Recall, F-score [24].

By analyzing results of the vehicle classification method we can conclude that the recognition quality is at an acceptable level only for the deep convolutional neural network which gave precision of 90.16% and recall of 88.93% on testing set. So results of the deep convolutional neural network allow use it as inputs to algorithms of vehicle trajectory registration and prediction. But it also indicates the need for additional research for improving recognition quality.



a b

Figure 1. Fragment of training set with vehicles in an arbitrary perspective: a – positive images, b – negative images

TABLE II. VEHICLE CLASSIFICATION RESULTS

Type of classifier	Training set			Testing set		
	Precision	Recall	F-score	Precision	Recall	F-score
The usage of Viola-Jones method and optical flow filter without the filter of false positives	0,8005	0,6971	0,7453	0,7720	0,7309	0,7509
The usage of Viola-Jones method and optical flow filter with the filter of false positives	0,8570	0,6910	0,7650	0,8530	0,6560	0,7420
Proposed deep convolutional neural network	0,8971	0,8853	0,8912	0,9016	0,8893	0,8954

An example of vehicle detection using Viola-Jones method and optical flow is shown in Fig. 2.

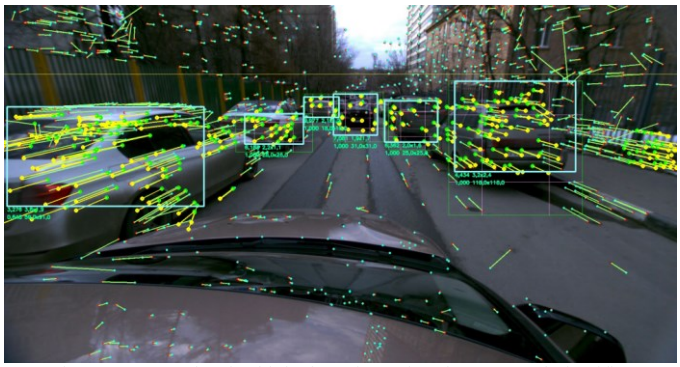


Figure 2. Example of vehicle detection using the proposed classifier

IV. TRAJECTORY REGISTRATION OF DETECTED VEHICLES

The vehicle movement trajectory is defined as a set of their positions at times t_i, $i = 1, 2,\ldots, n$, n – number of frames available in the video sequence. An object position detected on the image is determined by the bounding rectangle (box) $r = (x, y, w, h)$, where x and y – coordinates of the rectangle center, w and h – its width and height in pixels.

The registration and prediction model of the vehicle position should provide a return of complex information about the current (x_{ip}, y_{ip}, w_{ip}, h_{ip}) position of vehicle found in t_i-th moment of time and thereby ensure its trajectory registration.

To predict the trajectory of an object recognized on the image, an algorithm based on an estimation of its linear and angular velocities can be used. A simplified diagram of the observed object motion is shown in Fig. 3.

Here we use the following notations: (x_i, y_i) – coordinates of the center of the object found on i-th frame, v_i – linear object velocity on i-th frame, ω_i – angular object velocity on i-th frame, h_i u w_i – sizes (width and height) of the vehicle bounding box.

A distance between centers of the object on the current i-th frame and previous (i-1)-th one is determined on the basis of Euclidean distance on the basis of formula

$$L_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \quad (1)$$

An angle of the object movement direction on the i-th frame is calculated as

$$\varphi_i = \arccos\left(\frac{x_i - x_{i-1}}{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}\right) \cdot sign(y_i - y_{i-1}) \quad (2)$$

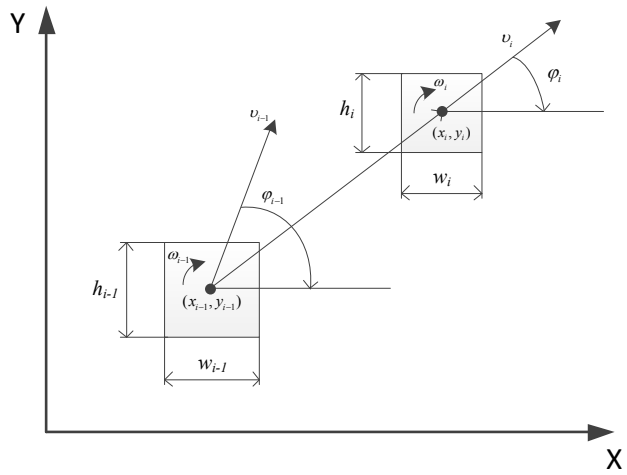


Figure 3. Simplified diagram of the observed object motion taking into account it's linear and angular velocities

To register object's position we propose the model a simplified diagram of which is shown in Fig. 4. It uses as input data the object position, found on the basis of the detection (recognition) algorithm, on the current frame (x_i, y_i, w_i, h_i) and previous three frames (x_{i-1}, y_{i-1}, w_{i-1}, h_{i-1}), (x_{i-2}, y_{i-2}, w_{i-2}, h_{i-2}), (x_{i-3}, y_{i-3}, w_{i-3}, h_{i-3}) forming an input vector V_1.

At the output of the model the registered (x_{ip}, y_{ip}, w_{ip}, h_{ip}) vehicle position in t_i–th moment is formed. This position is the output vector O_{1i}.

Thus the mathematical model is a functional relationship expressed by the formula $O_{1i} = f_1(V_{1i})$. An evaluation of coordinates of the object is calculated from formulas

$$\begin{aligned} \hat{x}_i &= x_{i-1} + L_{i-1} \cdot \sin(2 \cdot \varphi_{i-1} - sign(x_{i-1} - x_{i-2}) \cdot \varphi_{i-2}), \\ \hat{y}_i &= y_{i-1} + L_{i-1} \cdot \sin(2 \cdot \varphi_{i-1} - sign(y_{i-1} - y_{i-2}) \cdot \varphi_{i-2}), \end{aligned} \quad (3)$$

where $(\hat{x}_i, \hat{y}_i)$ – estimation (prediction result) of coordinates of the object at the i-th step (frame), computed using three previous points (i–1), (i–2) and (i–3), L_{i-1} – Euclidean distance between points (i–1) and (i–2), calculated by the formula (1), φ_{i-1} and φ_{i-2} – angles of the object movement direction respectively at (i–1)-th and (i–2)-th frames, calculated by the formula (2).

Registered coordinate values (x_{ip}, y_{ip}) are calculated as the average value between the forecast and the value found on the i-th frame

$$x_{ip} = \frac{(\hat{x}_i + x_i)}{2}, \; y_{ip} = \frac{(\hat{y}_i + y_i)}{2}. \quad (4)$$

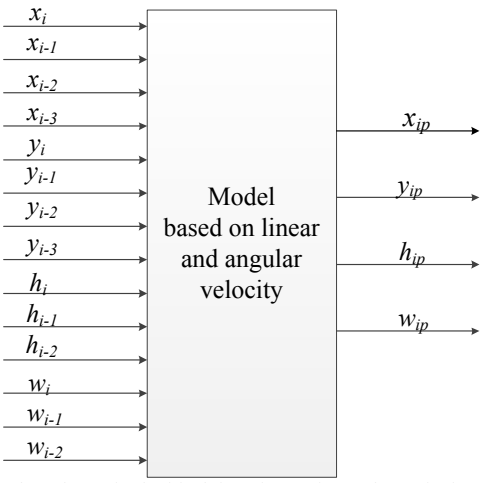


Figure 4. The scheme in the black box form of a mathematical model for object position registration on the basis of its linear and angular velocities

Estimation (forecasting) of the size of the object ($\hat{w}$, $\hat{h}_i$) is carried out as its linear extrapolation:

$$\hat{w}_i = 2 \cdot w_{i-1} - w_{i-2}, \; \hat{h}_i = 2 \cdot h_{i-1} - h_{i-2}. \tag{5}$$

Object size values (w_{ip}, h_{ip}) which are subject to registration are calculated as average values between the forecast and the value found on the i-th frame

$$w_{ip} = \frac{(\hat{w}_i + w_i)}{2}, \; h_{ip} = \frac{(\hat{h}_i + h_i)}{2}. \tag{6}$$

Found position (x_{ip}, y_{ip}, w_{ip}, h_{ip}) clarifies vehicle coordinates and dimensions obtained as a result of the image recognition algorithm. So it reduces noise and improves the quality of object detection.

To register the detected vehicles position we can also use the model using advanced Kalman filter [25], which is schematically shown in Fig. 5.

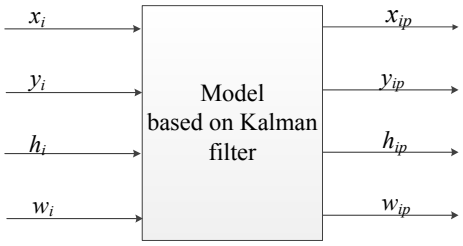


Figure 5. The scheme in the black box form of a mathematical model for object position registration on the basis of the advanced Kalman filter

As the input data the object position (x_i, y_i, w_i, h_i) found by the detection algorithm on the current i-th frame is used. It forms an input vector V_{2i} =[x_i, y_i, w_i, h_i]^T. At the model output the registered vehicle position (x_{ip}, y_{ip}, w_{ip}, h_{ip}) is formed in the t_i–th moment. It forms the output vector O_{2i} = [x_{ip}, y_{ip}, w_{ip}, h_{ip}]^T.

Thus the mathematical model is a functional relationship expressed by the formula $O_{2i} = f_2(V_{2i})$. At the next i-th step before the arrival of measurement results Y_i (Y_i = (x_i, y_i, h_i, w_i)^T) in Kalman filter, new object position is evaluated (prediction of the state vector) in accordance with an expression

$$\hat{X}_{i|i-1} = \hat{X}_{i-1|i-1}. \tag{7}$$

This kind of estimation is due to the fact that the motion law of found object is unknown beforehand, therefore transition matrix F and control matrix B are accepted as single ones, and control vector U is zero [25]. Here, the matrix of a priori estimation $\hat{X}_{i|i-1}$ is composed of predicted values of the object coordinates and sizes $\hat{X}_{i|i-1} = (\hat{x}_i, \hat{y}_i, \hat{h}_i, \hat{w}_i)^T$. Matrix $\hat{X}_{i-1|i-1}$ is composed of estimates of the object coordinates and sizes of in the previous step $\hat{X}_{i-1|i-1} = (\hat{x}_{i-1}, \hat{y}_{i-1}, \hat{h}_{i-1}, \hat{w}_{i-1})^T$.

New covariance matrix (a priori error estimation) is calculated as

$$P_{i|i-1} = F \cdot P_{i-1|i-1} \cdot F^T + Q. \tag{8}$$

Since during object detection the probability of imposing random systematic errors is small, then elements of the covariance matrix-column Q are assigned a value equal to one pixel.

By an a priori estimate of the state $\hat{X}_{i|i-1}$ from (7) we can calculate the measurement forecast:

$$\hat{Y}_i = H \cdot \hat{X}_{i|i-1}. \tag{9}$$

The measurement matrix H is selected as a unit.

After the next measurement is received $Y_i = (x_i, y_i, h_i, w_i)^T$ a forecast error of the i-th measurement is calculated by the formula

$$E_i = Y_i - H \cdot \hat{X}_{i|i-1}. \tag{10}$$

Then the state evaluation is corrected by selecting a point lying somewhere between the initial estimation $\hat{X}_{i|i-1}$ and a point corresponding to a new dimension Y_i:

$$\hat{X}_{i|i} = \hat{X}_{i|i-1} + G_i \cdot E_i, \tag{11}$$

where G_i – matrix of filter coefficients. This adjusted estimation is the registered position of the object $\hat{X}_{i|i} = (x_{ip}, y_{ip}, h_{ip}, w_{ip})^T$.

Finally the estimation of the covariance matrix of the state estimation error is corrected:

$$P_{i|i} = (I - G_i \cdot H) \cdot P_{i|i-1}, \tag{12}$$

where I – unit matrix. Covariance matrix of measurement forecast error E_i is calculated by the formula:

$$S_i = H \cdot P_{i|i-1} \cdot H^T + R, \tag{13}$$

and the matrix of filter coefficients at which the minimum error of the state estimation is reached is calculated as

$$G_i = P_{i|i-1} \cdot H^T S_i^{-1}. \tag{14}$$

Elements of the covariance matrix-column of measurements R are also set equal to one pixel because of the low probability of measurement errors.

Actions are repeated for each new value of the vector Y_i. At the initial time $\hat{X}_{0|0} = (x_0, y_0, h_0, w_0)^T$, $P_{1|0} = 0$.

During Kalman filter operation the center coordinates and sizes of the vehicle bounding box on the new frame are calculated on the basis of all previous information about the object's movement (by analogy with the integration operation over the entire time interval). It increases the registration accuracy in comparison with the method based on linear and angular velocity estimates.

Both models of object's trajectory registration have limitations on the case when the movement direction of observed vehicle changes sharply on the new frame. It is rarely observed in the case of high frame processing speed, for example, more than 20 frames per second.

The proposed mathematical models based on linear and angular object velocities and Kalman filter were implemented in Matlab environment. These implementations were tested in six cases of vehicle behavior (width and height for each vehicle bounding box was assumed to be the same, so we considers square boxes):

1) overtaking the observer on the left by the vehicle,

2) appearance and approach on the right of a slowly moving vehicle in the transverse traffic direction,

3) overtaking the observer on the right by the vehicle,

4) following the observed vehicle,

5) beginning of the vehicle overtaking on the left by the observer,

6) overtaking the vehicle on the right by the observer (shown in Figure 6).

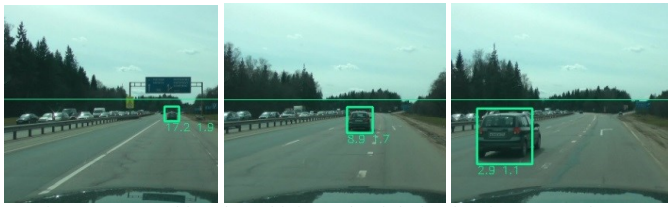


Figure 6. Overtaking the vehicle on the right by the observer (6-th case)

Registration results of the center coordinates and sizes of the vehicle bounding box for case 6 are shown in Fig. 7-8. The solid line shows coordinates or sizes of the region obtained by the vehicle detection algorithm on image sequence. The dashed line is registered coordinates or sizes obtained by the algorithm on the basis of the linear and angular object velocity. Bar-dashed line – registered coordinates or sizes obtained using an algorithm based on Kalman filter.

An application of both proposed models yielded adequate results, and the model in which Kalman filter is applied provides smaller "emissions" on graphs. This feature is required in real systems of trajectory registration and prediction of the observed objects, when a sharp change in position is usually the result of inaccurate vehicle detection.

As a measure of comparison of developed models, the following are applied:

- standard deviation (σ_1) of registered object coordinates (x, y) from the coordinates obtained as a result of image recognition, and

- standard deviation (σ_2) of the registered object sizes (w, h) from sizes obtained as a result of image recognition.

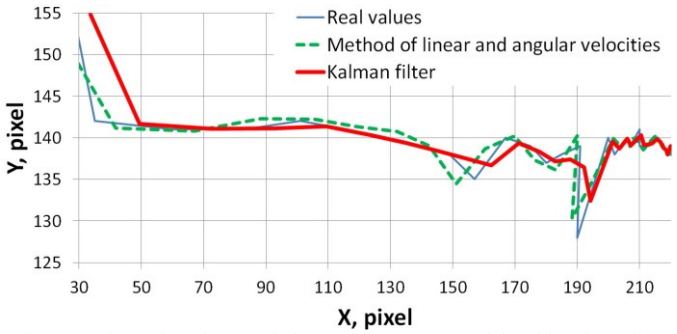


Figure 7. The registration result for coordinates (x, y) of the object bounding box center for the 6th case

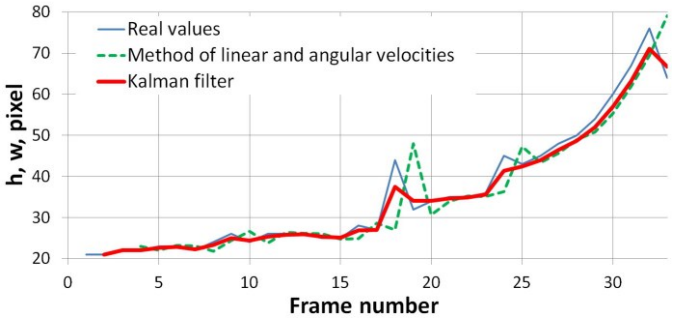


Figure 8. The registration result of the object bounding box sizes (w, h) for the 6th case

Results of calculation measures σ_1 and σ_2 for both models for all 6 cases are presented in Table 3.

TABLE 3. RESULTS OF CALCULATION OF STANDARD DEVIATIONS σ_1 AND σ_2 FOR BOTH MODELS FOR ALL 6 CASES

Algorithm of trajectory registration	Registered value	Standard deviation (σ_1 or σ_2), pixels.						
		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Average value
Algorithm for object trajectory registration based on its linear and angular velocities	The coordinates of the center (x, y) of the object bounding box	4,79	32,04	4,20	12,37	10,96	4,33	10,74
	Sizes (w, h) of the object bounding box	5,79	20,71	2,24	5,8	4,24	4,13	6,71
Algorithm for object trajectory registration based on the Kalman filter	The coordinates of the center (x, y) of the object bounding box	7,91	20,04	4,82	8,26	9,32	5,84	8,75
	Sizes (w, h) of the object bounding box	3,39	7,38	1,92	2,28	3,14	2,10	3,85

On the basis of this table we can conclude that both standard deviations σ_1 and σ_2 are lower on average for the algorithm of object trajectory registration based on Kalman filter and are at an acceptable level for use in driver assistance systems.

V. CONCLUSION

In this research two algorithms for the vehicle detection (recognition) on image sequence are proposed. The first one represents a combination of Viola-Jones method and Lucas-Kanade-Tomashi method for optical flow calculation. Second one is a deep convolutional neural network with multi scale sliding window technique. An implementation of the first method gave precision of 85.30% and recall of 65.6% on testing set. Second algorithm gave precision of 90.16% and

recall of 88.93% on testing set. So results of deep convolutional neural network are better and allow us use it as inputs to algorithms of the vehicle trajectory registration.

We have analyzed effectiveness of the proposed algorithms for the vehicle trajectory registration based on vehicles linear and angular velocities and Kalman filter. We have implemented it in Matlab environment. The algorithm using Kalman filter, on average, showed the best performance in the sense of the standard deviation from the sample values. If the object disappears on the next few images and then appears again, the method still works, because for these images coordinates and sizes of the vehicle bounding boxes are used as measured coordinates and sizes until the disappearance.

Received test results note to prospects of the developed algorithms for usage as a part of information support for the driver assistance system, navigation and control system for a robotic vehicle [26] or an unmanned vehicle.

ACKNOWLEDGMENT

This article is written in a course of the grant of the President of the Russian Federation for state support of young Russian scientists № MK-3130.2017.9 (contract № 14.Z56.17.3130-MK) on the theme “Recognition of road conditions on images using deep learning”.

REFERENCES

[1] P. Viola and M. J. Jones, “Rapid object detection using a boosted cascade of simple features”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’01), vol. 1, pp. I-511–I-518, 2001.

[2] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance”, Proceedings of the 9 th International Conference on Computer Vision (ICCV’03), vol. 1, pp. 734–741, 2003.

[3] D. A. Yudin, A. S. Knysh, and E. O. Kapustina, “Detection of cars on static images using the Viola-Jones method”, Collected materials of the III International Scientific and Practical Conference “Innovative development of automation, information and energy-saving technologies, metallurgy and metal science. Current state, problems and prospects”, Moscow, MISA, pp. 280–287, 2015.

[4] N. Dalal, B. Triggs, and D. Europe, “Histograms of Oriented Gradients for Human Detection”, CVPR 2005, vol. 1, pp. 886 – 893, 2005.

[5] C. Rabe, T. Muller, A. Wedel, and U. Franke, “Dense, Robust, and Accurate Motion Field Estimation from Stereo Image Sequences in Real-Time”, Computer Vision – ECCV 2010, pp. 582–595, 2010.

[6] A. Ramirez, E. Ohn-Bar, and M. M. Trivedi, “Go with the flow: Improving Multi-View vehicle detection with motion cues”, IEEE International Conference on Pattern Recognition (ICPR), 2014.

[7] K. A. Rudakov, A. S. Pikalov, and D. A. Yudin. “Analysis of methods of tracking objects on a sequence of images”, International conference “Actual problems of robotics and automation”, Belgorod, pp. 154–158, 2015.

[8] M. Rezaei and M. Terauchi, “Vehicle Detection Based on Multi-feature Clues and Dempster-Shafer Fusion Theory”, Image and Video Technology, Springer, vol. 8333, pp. 60–72, 2014

[9] M. Rezaei, M. Terauchi, R. Klette, N. Zealand, and N. Zealand, “Robust Vehicle Detection and Distance Estimation Under Challenging Lighting Conditions”, IEEE Transactions on Intelligent Transportation Systems (T-ITS), pp. 2723–2743, 2015.

[10] Y. LeCun, Y. Bengio, and G.Hinton, “Deep learning”, Nature, vol. 521, pp. 436–444, 2015.

[11] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, F. Mujica, A. Coates, A.Y. Ng, “An Empirical Evaluation of Deep Learning on Highway Driving”, ArXiv: 1504.01716 [cs.RO], 2015.

[12] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, “Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks”, IEEE geoscience and remote sensing letters, vol. 11, no. 10, pp. 1797–1801, 2014.

[13] C. Szegedy, A. Toshev, and D. Erhan, “Deep Neural Networks for Object Detection”, Advances in Neural Information Processing Systems 26 (NIPS 2013), pp. 2553–2561, 2013.

[14] A. Krizhevsky, I. Sutskever, G. Hinton, “ImageNet classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

[15] V. D. Kustikova, “Methods and algorithms for analysis of motion trajectories in solving the problem of video detection of vehicles: dis. ... cand. Tech. Sciences”, The Nizhny Novgorod state. University of. N.I. Lobachevsky, Nizhny Novgorod, 145 p., 2015.

[16] W. Hu, X. Xiao, D. Xie, and T. Tan, “Traffic accident prediction using 3-D model-based vehicle tracking”, IEEE Transactions on Vehicular Technology, vol. 53, no. 3, pp. 677–694, 2004.

[17] C.G. Prevost, A. Desbiens, and E. Gagnon, “Extended Kalman Filter for State Estimation and Trajectory Prediction of a Moving Object Detected by an Unmanned Aerial Vehicle”, American Control Conference. ACC '07, pp. 1805–1810, 2007.

[18] J. Wiest, M. Hoffken, U. Kresel, and K. Dietmayer, “Probabilistic trajectory prediction with Gaussian mixture models”, IEEE Intelligent Vehicles Symposium (IV), pp. 141–146, 2012.

[19] C. Oh and T. Kim, “Estimation of rear-end crash potential using vehicle trajectory data”, Accident Analysis & Prevention, vol. 42, no. 6, pp. 1888–1893, 2010.

[20] A. Houenou, P. Bonnifait, V. Cherfaoui, and W. Yao. “Vehicle trajectory prediction based on motion model and maneuver recognition”. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4363–4369, 2013.

[21] C. Tomasi and T. Kanade, “Detection and Tracking of Point Features”, Carnegie Mellon University Technical Report CMU-CS-91-132, 1991.

[22] “Keras: Deep Learning library for Theano and TensorFlow”, URL: https://keras.io/.

[23] “TensorFlow. An open-source software library for Machine Intelligence”, URL: https://www.tensorflow.org/.

[24] D. L. Olson and D. Delen, “Advanced Data Mining Techniques”, Springer, 1st edition, p. 138, 2008.

[25] I. N. Sinitsyn, “Filters of Kalman and Pugachev”, Moscow, University Book, Logos, 640 p., 2006.

[26] D. A. Yudin, G. G. Postolsky, A. S. Kizhuk, and V. Z. Magergut, “Mobile Robot Navigation Based on Artificial Landmarks with Machine Vision System”, World Applied Sciences Journal, 24 (11), pp. 1467–1472, 2013.

Healthcare System Reliability Analysis Addressing Uncertain and Ambiguous Data

Elena Zaitseva, Vitaly Levashenko, Miroslav Kvassay
University of Zilina
Department of Informatics
Zilina, Slovakia
e-mail: {elena.zaitseva, vitaly.levashenko, miroslav.kvassay}@fri.uniza.sk

Paul Barach
Wayne State University
School of Medicine
Detroit, MI, USA
e-mail: pbarach@gmail.com

***Abstract*—Healthcare systems in point are complex, inhomogeneous, highly variable, and require special mathematical representations. The structure function is offered for the mathematical representation of healthcare systems and its reliability evaluation is considered in this paper. Methods of system reliability evaluation based on the structure function are well established but deterministic. This restricts its use for uncertain or incomplete data. The structure function is constructed by a new method that is based on application of Fuzzy Decision Trees (FDTs), where input and output attributes are interpreted as component states and values of the structure function, respectively. For illustration, we consider the system of laparoscopic abdominal surgery. The analysis of these system components is implemented based on the structure function constructed with the FDT use. This structure function is formed based on incomplete data according to the new method. Laparoscopic surgery allows us to conclude that FDT-based reliability analysis is applicable for incomplete data and improved predictions in healthcare systems.**

Keywords—reliability; healthcare system; multi-state system; structure function; importance measures; fuzzy decision tree

I. INTRODUCTION

Reliability is one of important properties of any system function and of any reliable healthcare system [1]. The reliability analysis of a system includes typical steps using the construction of mathematical representation of the investigated system. This mathematical representation correlates with a mathematical method used for the system evaluation and algorithms for calculation of reliability indices and measures. A healthcare system includes components of different types, such as technical components (equipment/devices), software and human factors [2]. Therefore, this system is non-homogenous and complex [1, 3, 4], and the construction of mathematical representation including all types of components is challenging [1, 2, 4]. The simplest decision is the evaluation of stationary states, where the time dependence of system behavior can be ignored. One of the possible representations for this condition is a structure function that allows describing a system with any structural complexity. The structure function defines the correlation of system components states and system performance levels. This is a deterministic model and all possible component states and performance levels must be indicated and reflected in the structure function.

However, complete information about healthcare systems structure and behavior cannot be obtained, since the observation of all situations is impossible: some of them agree with the hazard to a patient's health. For example, [1] consider reliability and safety of pacemaker application that is interpreted as a system with uncertainty and some aspects of surgical patient care evaluated [5]. Furthermore, parts of the information are deducted from the expert's experiences. The investigation about collection of knowledge experts about healthcare system is considered [6, 7]. This information is ambiguous and unequal and can be evaluated with just some reliability or confidence. Therefore, the structure function of a healthcare system needs to be constructed based on uncertain data.

A new method for the structure function construction based on uncertain and incompletely specified data has been introduced [8, 9]. The problem of the structure function construction in which the system is classified into some classes that agree with the system performance levels. The problem in this interpretation is a typical problem in Data Mining and can be decided with application of well-known tools. We propose to use Fuzzy Decision Trees (FDTs) for this classification. FDTs are widely used in data mining for analysis of uncertain data and decision making in ambiguities [10, 11]. FDTs can be used naturally to analyze fuzzy data. In addition, FDTs allow taking into account uncertainties caused by incompletely specified data. This is still possible if there is only sparse data, or if acquired data is incomplete due to poor documentation. An FDT allows reconstructing data with different levels of the confidence [12]. The use of FDTs for the design of the structure function assumes the induction of a tree that is based on the data (fuzzy and/or crisp). The values of the structure function are then defined by the FDT for all combinations of the component states.

The structure function is constructed based on uncertain data that is analyzed by methods of reliability engineering. In particular, in this paper we argue that the importance analysis of healthcare system is considered [13, 14]. The influence of every system component impacts the changes in the system

This work was partly supported by the grants of VEGA 1/0038/16 and VEGA 1/0354/17

performance and can be evaluated under the importance analysis.

In this paper, the examination of laparoscopic surgery is considered. The system consists of human factors (doctors and nurses) and a medical device [4]. This mathematical system representation (structure function) is constructed based on uncertain and incompletely specified data. The importance analysis indicates the components (doctors, nurses, device) that have maximum influence on medical error, that may have health consequences to the patient [15].

This paper is structured as follows. Section II discusses the background of a proposed approach for the health system evaluation. The conception of structure function and methods for importance measures calculation are presented in this section. These measures are calculated with the use of Direct Partial Logical Derivatives [13]. These derivatives are developed under the Logical Calculus to investigate the changes of the structure function values depending on changes in this function variables values. Principal steps of the proposed method for structure function construction based on uncertain and incompletely specified data are considered in Section III. The results of method evaluation are presented in Section IV.

II. BACKGROUND

A. The Structure Function in Healthcare Analytics

The structure function is one of the mathematical models in reliability analysis. Typically, this mathematical model is used for representation of a system based on deterministic data: the system is represented as a mapping that assigns a system state to every possible profile of component states. There are two possibilities for system components states and system performance levels interpretation. The first of them allows determining two states for every of system components and two system performance levels. This interpretation is known in reliability analysis as the Binary-State System (BSS). However, in practice, many systems can exhibit different performance levels between these two extremes of full functioning and fatal failure [16, 17]. Therefore the Multi-State System (MSS) has been introduced for system mathematical representation in reliability analysis [18]. MSS is a mathematical model that is used to describe a system with several (more than two) levels of performance [16, 19, 20].

The concept of the structure function can be used to represent BSS and MSS, and associates the space of component states and system performance levels. It is obvious that BSS is a particular case of MSS. Therefore, in this paper, the structure function of MSS is used and considered for healthcare system representation.

In this case the system components states agree with the structure function variables x_i (i = 1, …, n, n is number of system components) and $\boldsymbol{x} = (x_1,\ldots, x_n)$ is a vector of system component's states (state vector). The state $x_i = 0$ represents the total failure of the component while state $x_i = m_i - 1$ corresponds to perfect functioning of the i-th component. The system performance level (system state) is defined depending on components states. This correlation is represented by the structure function value $\phi(\boldsymbol{x})$ that depends on variables values. The structure function values are from 0 to $M - 1$ (from failure ($\phi(\boldsymbol{x}) = 0$) to perfect functioning ($\phi(\boldsymbol{x}) = M - 1$) [21]:

$$\begin{gathered}\phi(x_1,\ldots,x_n) = \phi(\boldsymbol{x}): \\ \{0,\ldots,m_1 - 1\}\times\ldots\times\{0,\ldots,m_n - 1\} \to \{0,\ldots,M-1\}.\end{gathered} \tag{1}$$

Furthermore, each system component is characterized by the probabilities of its' states:

$$p_{i,s} = \Pr\{x_i = s\}, \quad s \in \{0,\ldots,m_i - 1\}. \tag{2}$$

As is shown in [2], that structure function represents a system in a stationary state or in fixed time. The typical measure in reliability analysis for such mathematical representation is system availability and probabilities of the system performance levels [2, 19]. The probabilities of the system performance levels, the system availability and unavailability based on structure function (1) are calculated as [2, 19]:

$$A_j(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) = j\}, j = 0, \ldots, M\text{-}1, \tag{3}$$

$$A^{\geq j}(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) \geq j\}, \tag{4}$$

$$U^{\geq j}(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) < j\} = 1 - A^{\geq j}(\boldsymbol{p}), \tag{5}$$

where j = 0, …, M-1; $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2,\ldots, \boldsymbol{p}_n)$ is a vector defining state probabilities of all system components, and $\boldsymbol{p}_i = (p_{i,0}, p_{i,1},\ldots, p_{i,mi\text{-}1})$ is a vector of probabilities of individual states of the i-th system component.

Let us to consider a simple example of healthcare system [9]. This system consists of two components: a doctor (x_1) and diagnostic device (x_2). The goal of this system investigation is definition of the conditions that are needed to create a medical error. The doctor's work in this system is represented by three states: 0 is interpreted as fatal error, 1 is incorrect work (without fatal result) and 2 is perfect work. The devices can have two states only: 0 is failure and 1 is proper function. The system has three performance levels: 0 is incorrect diagnosis with fatal consequence for the patient (or medical error), 1 is incorrect diagnosis without fatal consequence for the patient, and 2 is correct diagnostics. According to expert opinion, this system can be represented by structure function in Table I.

The structure function in Table I allows us to compute the probability of the system failure (possibility of medical error) as:

$$A_0(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) = 0\} = p_{1,0}\cdot p_{2,0}.$$

TABLE I. STRUCTURE FUNCTION OF SIMPLE SYSTEM

The system components		$\phi(\boldsymbol{x})$
x_1	x_2	
0	0	0
0	1	1
1	0	1
1	1	2
2	0	2
2	1	2

The probability of the incorrect diagnosis with fatal consequence for the patient is computed as the probability of the system performance level 1:

$$A_1(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) = 1\} = p_{1,0} \cdot p_{2,1} + p_{1,1} \cdot p_{2,0},$$

and the probability of the correct diagnosis agrees with the system performance level 2:

$$A_2(\boldsymbol{p}) = \Pr\{\phi(\boldsymbol{x}) = 2\} = p_{1,1} \cdot p_{2,1} + p_{1,2} \cdot (p_{2,0} + p_{2,1}).$$

The probabilities of system performance levels and system availability allow to evaluate the system as a whole and not discover the influence of system components on its availability. This examination is possible by importance measures of the system [13, 14].

B. Importance Analysis

Importance analysis is part of reliability analysis for the evaluation of a fixed component influence into the system reliability, availability, maintainability and safety. Different aspects of this influence are evaluated by importance measures that evaluate the relative importance of individual components or groups of components to system performance. The authors provide a review of importance measures and specifics of their applications for BSS and MSS [14]. Various importance measures have been proposed with different motivations. Some of the most commonly used measures are the *Structural Importance* (SI), *Birnbaum's Importance* (BI), *Criticality Importance* (CI) and *Fussell-Vesely's Importance* (FVI) and whose definitions are in Table II. These importance measures are actually related to each other in some aspects. In case of MSSs, importance measures can be used to investigate several dependencies in the system. More precisely, they permit identifying [13]:

- influence of a specific component state on a specific system state;
- total influence of a given component state on the whole system;
- total influence of a component on a specific system state; and,
- total influence of a given component on the whole system.

Several importance measures can be defined. In this paper, we focus on SI and BI measures. There are different approaches and algorithms to compute individual importance measures and transformations between them. In paper [13] the importance measures calculation based on structure function of MSS is considered. The algorithms in are used for calculation of the measures in Table II by structure function are based on the Direct Partial Logic Derivative [13].

TABLE II. IMPORTANCE MEASURES

Importance measure	Meaning
SI	The SI analyzes system topology, and it corresponds to the relative number of situations in which a given component is critical, i.e. in which its degradation results in system degradation.
BI	The BI takes into account not only system topology but also state probabilities of the system components. This measure agrees with the probability that a component is critical, i.e. the probability that system degradation results from degradation of the component.
CI	The CI of a given component is computed as the probability that the system failure has been caused by the component failure given that the system is failed.
FVI	The FVI of a given component agrees with the probability that the component contributes to system failure.

C. Direct Partial Logic Derivatives

A Direct Partial Logic Derivative (DPLD) is one of the principal techniques in Logical Differential Calculus that is part of Algebra Logic [22]. These derivatives have been used for analysis of the structure function $\phi(\boldsymbol{x})$ and DPLD with respect to variable x_i has been defined as[21, 13]:

$$\frac{\partial \phi(j \to h)}{\partial x_i (s \to r)} = \begin{cases} 1, & \text{if } \phi(s_i, \boldsymbol{x}) = j \text{ and } \phi(r_i, \boldsymbol{x}) = h \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where $s, r \in \{0,1, \ldots, m_i - 1\}$, $s \neq r$; $j, h \in \{0,1, \ldots, M - 1\}$, $j \neq h$, and $\phi(a_i, \boldsymbol{x}) = \phi(x_1, x_2, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n)$ for $a \in \{s, r\}$.

DPLD (6) defines circumstances under which change of state of component i from value s to r results in change of the system from state j to h.

The computer implementation of DPLD is very simple and can be considered as a comparison of structure function values. For example, let us consider DPLD in Fig. 1. This derivative $\partial\phi(1\to0)/\partial x_1(1\to0)$ provides the comparison of the structure function values for state vectors $(1, x_2)$ and $(0, x_2)$, $x_2 \in \{0, 1\}$: if $\phi(1, x_2) = 1$ and $\phi(0, x_2) = 0$, then the DPLD's values equal 1, else the DPLD's values are equal to 0. Therefore, this derivative has one nonzero value for state vector (1, 0). This DPLD does not depend on variable x_1, therefore, it is defined for the two values of the second variable.

From the point of view of reliability analysis derivative $\partial\phi(1\to0)/\partial x_1(1\to0)$ indicates critical state of system where the fault of the first component (x_1 change value from 1 to 0) causes the system failure (structure function value $\phi(\boldsymbol{x})$ changes from 1 to 0). In context of the considered task, it means the indication of the conditions required for a medical error

depends on a mistake by the doctor provided his/her incorrect work: the medical error in diagnosis depending on doctor's mistake can be if the device does not work ($x_2 = 0$). Other DPLDs allow us to investigate all possible changes in system state depending on components states changes. For example, in Fig. 2 DPLD $\partial\phi(2\to0)/\partial x_1(2\to0)$ allows us to consider other aspects of medical error in diagnosis depending on doctor mistake, where doctor's mistake occurrences in case of his/her perfect work. According to DPLD in Fig. 2 it should be possible if the diagnostic device is not functioning.

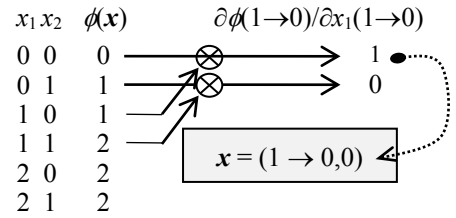


Figure 1. Example of DPLD calculation to define condition of medical error if doctor's work has some incorrection.

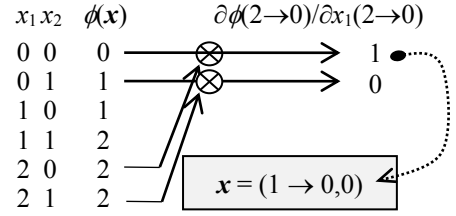


Figure 2. Example of DPLD calculation to define condition of medical error if doctor's work is perfect.

Similarly other critical system states based on other DPLDs can be calculated. In Table III all critical states of the considered system are shown.

The analysis in Table III is a qualitative analysis of the system that considers the condition of critical system states. In context of considered example, it is a medical error that can be caused by degradation/mistake of doctor's work or failure of the diagnostics devices. DPLD allows investigating all possible changes of system state (correct diagnosis) depending on the doctor's mistake or the diagnostic devices failure.

DPLDs permit calculating critical system states for every system performance level depending on components states changes. This property of DPLD is used in calculation of importance measures that is considered in detail [13]. But DPLD can be used to analyze the completely specified structure function that means the function values must be known for all possible state vectors $\boldsymbol{x}$. In real world of healthcare initial data for the system mathematical representation is uncertain and incompletely specified [1, 9, 23]. Therefore, special methods and algorithms have to be developed to construct the structure function based on uncertain data. There are different approaches to solve this problem, e.g. Bayesian networks [24, 25] or Fuzzy Decision Trees [8]. In this paper, we consider the application of Fuzzy Decision Tree for construction of the structure function for the healthcare system.

TABLE III. CRITICAL STATES OF SYSTEM

DPLD	Critical state	Descriptions
$\frac{\partial\phi(1\to0)}{\partial x_1(1\to0)}$	(1→0, 0)	The medical error occurs if doctor errs in case of his/her incorrect work (without fatal result) and the diagnostic devices is not function
$\frac{\partial\phi(1\to0)}{\partial x_1(2\to0)}$	—	
$\frac{\partial\phi(1\to0)}{\partial x_1(2\to1)}$	—	
$\frac{\partial\phi(2\to0)}{\partial x_1(1\to0)}$	—	
$\frac{\partial\phi(2\to0)}{\partial x_1(2\to0)}$	(2→0, 0)	The doctor's mistake in case of his/her perfect work causes medical error if diagnostic devices is not function
$\frac{\partial\phi(2\to0)}{\partial x_1(2\to1)}$	—	
$\frac{\partial\phi(2\to1)}{\partial x_1(1\to0)}$	(1→0, 1)	The doctor's mistake in case of his/her incorrect work (without fatal result) causes incorrect diagnosis without fatal consequence for the patient if the diagnostic devices is function
$\frac{\partial\phi(2\to1)}{\partial x_1(2\to0)}$	(2→0, 1)	The degradation of diagnosis without fatal consequence for the patient is caused by doctor's mistake in case of his/her incorrect work (without fatal result) and the diagnostic devices is function
$\frac{\partial\phi(2\to1)}{\partial x_1(2\to1)}$	(2→1, 0)	The degradation of diagnosis without fatal consequence for the patient is caused by the doctor's work degradation from perfect work into incorrect work without fatal result if the diagnostic devices is function
$\frac{\partial\phi(1\to0)}{\partial x_2(1\to0)}$	(0, 1→0)	The medical error is caused by diagnostic devices failure if the doctor's mistake is fatal
$\frac{\partial\phi(2\to0)}{\partial x_2(1\to0)}$	—	
$\frac{\partial\phi(2\to1)}{\partial x_2(1\to0)}$	(1, 1→0)	The failure of the diagnostic devices occurs the degradation of diagnosis without fatal consequence for the patient if the doctor's work is incorrect without fatal result

D. Importance Measures

Two types of importance measures considered in this paper are SI and BI. The SI measures focus on the topological properties of the system. They allow finding components (or specific states) with the greatest influence on the system performance (or on a specific system state) from a topological point of view. The basic SI is denoted as $SI_{i,s}^{j\downarrow}$, and it is computed as a relative number of situations in which a minor degradation of state s of component i (i.e., change of the component from state s to s -1) results in degradation of system state j. These situations can be identified as the nonzero elements of all DPLDs of the form of $\partial\phi(j\to h)/\partial x_i(s\to s-1)$ for $h\in\{0,1,\ldots,j\text{-}1\}$. This implies that j DPLDs have to be calculated to compute $SI_{i,s}^{j\downarrow}$. However, definition (6) of DPLD implies that at most one derivative from

all these DPLDs can be nonzero for a state vector (s_i, $\boldsymbol{x}$), what indicates that DPLDs carry quite little information from reliability point of view. Because of that, new types of DPLDs have been defined in [13]. Those DPLDs were named as Integrated DPLDs (IDPLDs) because they combine several DPLDs together. For calculation of $\mathrm{SI}_{i,s}^{j\downarrow}$, the most important one is IDPLD of type I that is defined as follows:

$$\begin{aligned}\frac{\partial\phi(j\downarrow)}{\partial x_i(s\to r)} &= \bigcup_{h=0}^{j-1}\frac{\partial\phi(j\to h)}{\partial x_i(s\to r)}\\ &= \begin{cases}1, & \text{if } \phi(s_i,\boldsymbol{x})=j \text{ and } \phi(r_i,\boldsymbol{x})<j\\ 0, & \text{other}\end{cases},\end{aligned} \tag{7}$$

where $s, r \in \{0,1,\ldots,m_i\text{-}1\}$, $s \neq r$; $j, h \in \{0,1,\ldots,M\text{-}1\}$, $j \neq h$, and symbol $\cup$ represents logical disjunction. Clearly, this derivative identifies situations in which change of component i from state s to r results in degradation of system state j. For $r = s$ -1, this IDPLD allows finding situations based on which $\mathrm{SI}_{i,s}^{j\downarrow}$ can be computed and, therefore, we can write [13]:

$$\mathrm{SI}_{i,s}^{j\downarrow} = \mathrm{TD}\left(\partial\phi(j\downarrow)/\partial x_i(s\to s-1)\right), \tag{8}$$

where TD(.) denotes truth density of the argument interpreted as a function with Boolean-valued output (it agrees with the relative number of vectors for which the argument takes non-zero values).

SI (8) quantifies topological importance of a specific component state for a specific system state assuming that the component degrades state by state. Using this measure, we can also compute the total importance of a given component state on the whole system [13]:

$$\mathrm{SI}_{i,s}^{\downarrow} = \sum_{j=1}^{M-1}\mathrm{SI}_{i,s}^{j\downarrow}, \tag{9}$$

or the total topological importance of the component on a specific system state [13]:

$$\mathrm{SI}_{i}^{j\downarrow} = \frac{1}{m_i-1}\sum_{s=1}^{m_i-1}\mathrm{SI}_{i,s}^{j\downarrow}. \tag{10}$$

SI (10) agrees with the relative number of situations in which a minor degradation of state s of component i results in system degradation while SI (10) corresponds to the relative number of situations in which a degradation of component i causes degradation of system state j. If we want to compute the total topological importance of a given component, then we can calculate the average from all $\mathrm{SI}_{i,s}^{\downarrow}$, i.e. for s = 1,2,…, m_i -1, or the sum of the $\mathrm{SI}_{i}^{j\downarrow}$ measures through j = 1,2,…, M [13]:

$$\mathrm{SI}_{i}^{\downarrow} = \frac{1}{m_i-1}\sum_{s=1}^{m_i-1}\mathrm{SI}_{i,s}^{\downarrow} = \sum_{j=1}^{M-1}\mathrm{SI}_{i}^{j\downarrow}. \tag{11}$$

This measure agrees with the relative number of situations in which a degradation of component i results in a decrease in system state.

According to [13], relationships between SI measures (8) – (11) can be expressed in the form of Table IV. The central part of the table contains SI measures (8). SI measures (9) and (10) are located in the bottom row and the right column respectively, and they are computed based on the data from the central part. The total importance of the component measured by SI (11) is located in the lower right cell. Such a table can be constructed for all system components, and allows to perform a detailed topological analysis of the system.

TABLE IV. STRUCTURAL IMPORTANCE MEASURES

		Component state				Average
		1	2	…	m_i -1	
System state	1	$\mathrm{SI}_{i,1}^{1\downarrow}$	$\mathrm{SI}_{i,2}^{1\downarrow}$	…	$\mathrm{SI}_{i,m_i-1}^{1\downarrow}$	$\mathrm{SI}_{i}^{1\downarrow}$
	2	$\mathrm{SI}_{i,1}^{2\downarrow}$	$\mathrm{SI}_{i,2}^{2\downarrow}$	…	$\mathrm{SI}_{i,m_i-1}^{2\downarrow}$	$\mathrm{SI}_{i}^{2\downarrow}$
	⋮	⋮	⋮	⋱	⋮	⋮
	M -1	$\mathrm{SI}_{i,1}^{(M-1)\downarrow}$	$\mathrm{SI}_{i,2}^{(M-1)\downarrow}$	…	$\mathrm{SI}_{i,m_i-1}^{(M-1)\downarrow}$	$\mathrm{SI}_{i}^{(M-1)\downarrow}$
Sum		$\mathrm{SI}_{i,1}^{\downarrow}$	$\mathrm{SI}_{i,2}^{\downarrow}$	…	$\mathrm{SI}_{i,m_i-1}^{\downarrow}$	$\mathrm{SI}_{i}^{\downarrow}$

SI measures (8) – (11) take into account only system topology and, therefore, these measures are primarily used in a phase of system design when state probabilities of the components are not known yet. If the probabilities are known, another set of measures can be computed. These measures are known as BI, and they can be computed using the similar formulae as SI measures. The only difference is in computation of $\mathrm{BI}_{i,s}^{j\downarrow}$, which agrees with the probability that a minor degradation of component i results in the degradation of system state j, and which can be computed using IDPLDs in the following way [13]:

$$\mathrm{BI}_{i,s}^{j\downarrow} = \Pr\left\{\partial\phi(j\downarrow)/\partial x_i(s\to s-1)=1\right\}. \tag{12}$$

The importance measures SI (8) – (11) and BI (12) allow evaluation of system represented by structure function. Let us consider this function construction for mathematical representation of healthcare system based on uncertain data with the use of FDT.

III. Structure Function Construction Based on Uncertain Data with Application of Fuzzy Decision Tree

A. General Representation of the Method

The analysis of the healthcare system is based on experts' experience and judgment. This data is uncertain as a rule. This uncertainty can be caused by a lot of factors, but we have considered two of them. The first factor is incompletely specified data, because some values of system states or performance levels cannot be obtained. For example, it might need an unacceptable long time to get the data. The second factor is ambiguity and vagueness of collected data values. This type of ambiguity can be caused, for instance, by the subjectivity of the expert's evaluations.

Therefore, the construction of the structure function must consider two aspects. The first is a mapping that assigns the system performance level to each possible profile of the component states (for example, see Table I). The second is addressed by interpreting it as a classification problem for uncertain data. Therefore the structure function construction based on uncertain data can be interpreted as classification a problem for uncertain data which is typical for data mining. One of the possible options is applying decision trees or FDT [10, 11]. Such methods have been proposed for structure function construction in [8] firstly and are based on the method of FDT induction with application of cumulative information estimations introduced in [9]. This method for structure function construction includes the following steps (Fig. 3):

- Collection of data in the repository according to requests of FDT induction;
- Representation of the system model in the form of an FDT that classifies component states according to the system performance levels;
- Construction of the structure function as a decision table that is created by the inducted FDT.

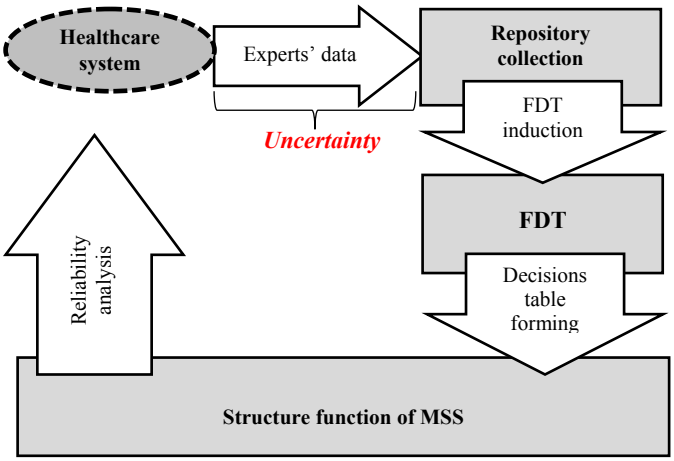


Figure 3. The structure function construction with application of FDT.

According to the method in [8] the structure function is constructed as a decision table that classifies the system performance level to each possible profile of the component states. The decision table is formed based on a FDT that provides the mapping of all possible component states (input data) in M performance levels.

B. Data Repository

Initial data for the structure function construction is collected in a form of the repository by the monitoring or expert evaluations of values of system component states and system performance levels. The repository for the FDT induction is formed as a table. The number of columns is $n+1$. The first n columns are used for the representation of states of the components. The system performance level is indicated in the last column. The i-th column in the first n columns includes m_i sub-columns ($i = 1,2,\ldots, n$). The last column for the system performance level has M sub-columns. The sub-column is linked with one of the values of component states or performance levels. Every row of the table represents one monitoring situation or one expert's evaluation. The number of rows agrees with the number of samples for FDT induction.

The table's cell contains a number (from interval from 0 to 1) that is interpreted as the possibility (likelihood) of the value (from 0 to m_i -1 or M -1) indicated in the sub-column. Note that the sum of these possibilities in cells of one column and one row equals to 1. It means that the sum is 1 for the possibilities of all states of every system component or system performance levels in the sample. Such data can be obtained from expert evaluations or by possibilistic fuzzy clustering [26]. These possibilities correspond to a membership function of fuzzy data [27]. This demand for initial data representation is caused by the method of FDT induction.

For example, let us assume that data from an expert about the system for investigation about a medical error considered above is presented in Table V. This data is incompletely specified. Only four situations are obtained by experts' evaluation. All values are defined with possibility (likelihood).

TABLE V. The Collected Data for Analysis of Successful Laparoscopic Surgery

No.	x_1			x_2		$\phi(x)$		
	0	1	2	0	1	0	1	2
1	0.9	0.1	0.0	0.9	0.0	0.9	0.1	0.0
2	1.0	0.0	0.0	0.2	0.8	0.1	0.9	0.0
3	0.0	1.0	0.0	0.3	0.7	0.1	0.3	0.6
4	0.1	0.1	0.8	0.9	0.0	0.0	0.3	0.7

C. FDT Induction

There are different types of FDT. The induction of an FDT of any type is performed by the determination of the correlation between n input attributes $\{A_1, A_2,\ldots, A_n\}$ and one output attribute B. The construction of the system structure function supposes that the system performance level is the output attribute and state vectors are input attributes [8]. The correlation between the terminologies and basic concepts of FDTs and reliability analysis was introduced in [8]. Let us now consider this correlation succinctly.

Input attribute (component state) A_i (i = 1,2,…, n) corresponds to the i-th column of repository and is measured by a group of discrete values ranging from 0 to $m_i - 1$. These values $\{A_{i,0},..,A_{i,j},..,A_{i,m_i-1}\}$ agree with the values of states of the i-th component and correspond to sub-columns of the i-th column in the repository. An FDT assumes that the input set A = $\{A_1, A_2,\ldots, A_n\}$ is classified into one of the values of output attribute B. Output attribute B corresponds to the last (the n +1-th) column of the repository. Value B_w of output attribute B agrees with one of the system performance levels and is defined by M values ranging from 0 to M -1 (w = 0,1,…, M -1).

The FDT induction is based on the determination of expanded attributes A_i. The type of FDT is caused by the criteria for the choice of this attribute. The selection criterion of expanded attributes A_i for induction of FDT is defined by cumulative information estimations considered in [9]. There are two tuning thresholds α and β in this method of FDT induction. A tree branch stops to expand when either the frequency f of the branch is below α or when more than β percent of instances left in the branch has the same class label. These values are thus key parameters needed to decide whether we have already arrived at a leaf node or whether the branch should be expanded further. Decreasing the parameter α and increasing the parameter β allow us to build large FDTs. On one hand, large FDTs describe datasets in more detail. On the other hand, these FDTs are very sensitive to noise in the dataset. We selected empirically parameters α from 0.10 to 0.30 and β from 0.70 to 0.90. We estimated that a confidence degree of more than 0.90 would allow us to reach a decision with sufficient confidence. Moreover, the threshold frequency 0.10 eliminates the variants of no-principal decisions. Notably, increasing the size of the FDT has no influence on the FDT root or the higher FDT nodes. It only adds new nodes and leaves to the bottom part of the FDT. These new nodes and leaves have a low bearing on decision making.

The FDT for the considered system for investigation of medical error (Fig. 4) has 2 levels and includes all input attributes. This implies that all input attributes are considered to be significant for this system. This FDT has been inducted with thresholds α = 0.20 and β = 0.70.

The attribute A_2 corresponds to the 2-nd variable of the structure function that is 2-nd system's component. This component is a diagnostic device and has two states. Therefore, the attribute A_2 has two values that are associated with two branches of the FDT in Fig. 4. Each branch agrees with a block of the output attribute values with the confidences of every value. This block is a leaf if one value of the output attribute has a sufficient level of confidence (it is greater than or equal to the value of the threshold β) or frequency of the decision is less than or equal to threshold α. In the other case, the FDT provides the analysis of the next input attribute. For this example the value $A_{2,0}$ of attribute A_2 has the leaf value B_0 for the output attribute. It is interpreted as the medical error in diagnosis with confidence 0.767 if the diagnostic device is not functioning. In this example, the value $A_{2,1}$ of the attribute A_2 implies the analysis of the attribute A_1. The attribute A_1 agrees with the doctor's work. This attribute can have the values $A_{1,0}$, $A_{1,1}$, and $A_{1,2}$, which are associated with branches of the FDT. Every value agrees with value of the first variable of the structure function that represents the first component states. The value $A_{1,0}$ of the first attribute agrees with value B_1 of the output attribute and values $A_{1,1}$, and $A_{1,2}$ cause value B_2 of the output attribute.

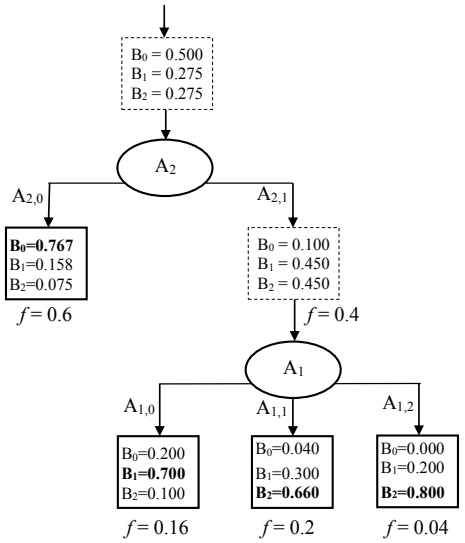


Figure 4. The FDT for the construction of structure function.

So, the FDT allows us to represent all system states, and it is inducted based on incomplete specified data in which values can be defined.

D. Construction of Structure Function based on FDT

According to [8], FDTs allow developing fuzzy decision rules or a decision table for decision support system. A decision table contains all possible values of input attributes and the corresponding values of the output attribute that is calculated using the FDT. Because there is a unique correlation between FDT's attributes and values of variables and structure function, the decision table can be described in term of structure function (reliability analysis). In this case the inducted FDT allows forming a table that agrees with the structure function. This implies that all possible combinations of values of the component states (all state vectors) have to be analyzed by the FDT to classify state vectors into M classes of the system performance levels.

Let us illustrate the structure function construction for the introduced example based on FDT in Fig. 4. Assume that the state vector is $\boldsymbol{x}$ = (0 0), $\boldsymbol{x}$ = (1 0) or $\boldsymbol{x}$ = (2 0). According to the FDT (Fig. 4), the system output attribute for these state vectors has value B_0, because they have value $x_2 = 0$ that corresponds value $A_{2,0}$ of the attribute A_2 with confidence 0.767: $\phi(0, 0) = 0$, $\phi(1, 0) = 0$ and $\phi(2, 0) = 0$.

The analysis of the first attribute A_1 is necessary if the value of the attribute A_2 is $A_{2,1}$. The output attribute has value B_1 with confidence 0.700 for value $A_{1,0}$ the attribute A_1. The structure function value for state vector $\boldsymbol{x}$ = (0 1) is $\phi(0, 1) = 1$.

The value $A_{2,1}$ of the attribute A_2 cause the value of output attribute B_2 with confidence 0.660 but the frequency of this decision is equal to threshold α (in additional, it is finished classification), therefore $\phi(1, 1) = 2$. The structure function for the state vector $\boldsymbol{x}$ = (2 1) is $\phi(2, 1) = 2$ because the output attribute value is B_2 with confidence 0.800 for values $A_{1,1}$ and $A_{2,2}$ input attributes A_1 and A_2.

All computed values of the structure function based on FDT (Fig. 4) are equal to values of the initial structure function in Table I. This example shows the principal steps in construction of a structure function based on uncertain data. The same result for the initial structure function and constructed structure function implies the correctness of the proposed method.

IV. Illustrative Example

A. Initial Data

For illustration of the approach presented above, let us consider a system representing laparoscopic surgery. This system is composed of two types of components: medical devices and human factors. The principal medical device is a laparoscopic robotic surgery machine (component 1) that can be in either functioning (state 1) or failed (state 0). The human factor is represented by the medical personnel that include [4]: an anesthetist and anesthetic nurse (components 2 and 4), which perform their activities perfectly (state 1) or with an error (state 0), and surgeon and surgical nurse (components 3 and 5), which perform their activities either perfectly (state 2) or with a small error (state 1) or with a critical error (state 0). Based on the performance of the surgery machine and the staff, the laparoscopic surgery can be successful (state 2), or adverse events can occur (state 1), or it can fail (state 0). If we want to analyze reliability of this system, we need to find its structure function. This can be done using the FDT induction considered above.

Let us assume that several laparoscopic surgeries were observed and evaluated by experts. The results of the observations are presented as fuzzy data collected in a repository (Table VI). Using the data in the repository, the FDT (Fig. 5) can be inducted. The structure function for the mathematical representation of laparoscopic surgeries is constructed by the FDT according to principle considered in section III.D. After obtaining the structure function, a reliability analysis can be performed. Firstly, we can compute the probabilities of the individual states of the system (3). For this purpose, the state probabilities of the system components are needed (Table VII). Based on the data presented in Table VII the structure function constructed by FDT in Fig. 5, we find that:

$$A_0(\boldsymbol{p}) = 0.0249, \quad A_1(\boldsymbol{p}) = 0.0351, \quad A_2(\boldsymbol{p}) = 0.9400, \quad (13)$$

which implies that the probability of unsuccessful laparoscopic surgery is about 0.0249.

TABLE VI. Repository for Construction of the Structure Function of the System Representing Laparoscopic Surgery

No.	A_1		A_2		A_3			A_4		A_5			B		
	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{2,0}$	$A_{2,1}$	$A_{5,0}$	$A_{5,1}$	$A_{5,2}$	B_0	B_1	B_2
1	0.9	0.1	1.0	0.0	0.8	0.2	0.0	0.1	0.7	0.1	0.2	0.7	0.9	0.1	0.0
2	0.7	0.3	1.0	0.0	0.0	0.9	0.1	0.8	0.2	0.0	1.0	0.0	0.8	0.1	0.1
3	0.7	0.3	0.0	1.0	0.8	0.2	0.0	0.8	0.1	0.6	0.3	0.1	0.9	0.1	0.0
4	1.0	0.0	0.1	0.9	1.0	0.0	0.0	0.1	0.9	0.0	0.1	0.9	0.8	0.2	0.0
5	0.9	0.1	0.0	1.0	0.9	0.1	0.0	0.8	0.2	0.6	0.4	0.0	1.0	0.0	0.0
6	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
7	1.0	0.0	0.0	1.0	0.7	0.3	0.0	0.0	0.3	0.0	0.3	0.7	0.1	0.8	0.1
8	0.8	0.2	0.1	0.0	0.9	0.1	0.0	0.6	0.4	0.1	0.8	0.1	0.9	0.1	0.0
9	0.9	0.1	0.1	0.9	0.1	0.1	0.8	1.0	0.0	0.7	0.2	0.1	0.8	0.1	0.1
10	0.7	0.3	0.1	0.9	0.0	0.0	1.0	0.6	0.4	0.0	0.0	1.0	0.0	0.1	0.0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
25	0.3	0.7	0.0	1.0	0.0	0.1	0.9	0.1	0.8	0.0	0.0	1.0	0.0	0.1	0.9

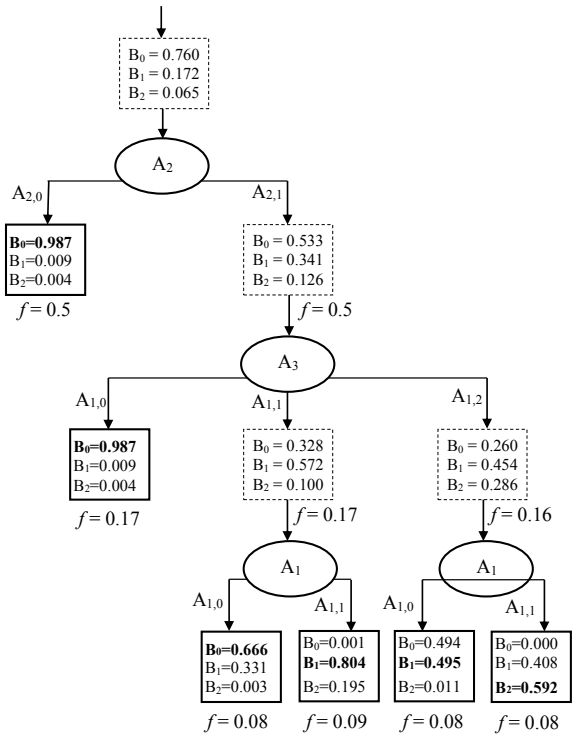


Figure 5. The FDT for the structure function of laparoscopic surgery.

TABLE VII. State Probabilities of Components of the System Representing Laparoscopic Surgery

Component	Component state		
	0	1	2
1	0.002	0.998	–
2	0.015	0.985	–
3	0.010	0.115	0.875
4	0.012	0.988	–
5	0.015	0.155	0.830

B. Structural Importance

The system state probabilities (13) are very important global characteristics of the system. However, they do not allow us to investigate importance individual elements of the system to assess their performance. This can be done using importance measures. Consider two most used importance measures as SI and BI for evaluation of the laparoscopic surgery by the structure function constructed based on FDT in Fig. 5.

Firstly, let us investigate the topological importance of the system components. Let us illustrate this for component 3 (the surgeon). The system and the component has 3 states and, therefore, measures $\mathrm{SI}_{3,1}^{1\downarrow}$, $\mathrm{SI}_{3,2}^{1\downarrow}$, $\mathrm{SI}_{3,1}^{2\downarrow}$, and $\mathrm{SI}_{3,2}^{2\downarrow}$ can be computed. If we want to compute $\mathrm{SI}_{3,1}^{1\downarrow}$, which quantifies topological importance of state 1 of the component on system state 1, the nonzero elements of derivative $\partial\phi(1\downarrow)/\partial x_3(1\to 0)$ have to be find. This can be done using formula (7) as follows:

$$\begin{aligned}\frac{\partial\phi(1\downarrow)}{\partial x_3(1\to 0)} &= \{x_1 \leftrightarrow 0\}\{x_2 \leftrightarrow 1\}\{x_5 \leftrightarrow 2\} \\ &\vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 0\} \\ &\vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 1\}\{x_5 \leftrightarrow 0\} \\ &\vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 1\}\{x_5 \leftrightarrow 1\},\end{aligned} \tag{14}$$

where symbol $\leftrightarrow$ denotes logical biconditionality. This result implies that the considered change of component 3 causes degradation of system state 1 if component 1 is in state 0 and component 2 in state 1 and component 5 in state 2 (the first row) or if component 1 is in state 1 and component 2 in state 1 and component 4 in state 0 (the second row) or etc. (the third and the fourth row). Since this derivative is a function with a Boolean-valued output, which takes a non-zero value for 7 out of the 24 state vectors for which the derivative is defined, its truth density is 7/24. Therefore, $\mathrm{SI}_{3,1}^{1\downarrow} = 7/24$. Using the same procedure, the remaining SI measures can be computed. The results are presented in the central part of Table VIII. The data in the table implies that state 1 of component 3 has the greatest topological influence on state 1 of the system, while state 2 has the biggest influence on system state 2.

TABLE VIII. STRUCTURAL IMPORTANCE MEASURES FOR COMPONENT 3

		Component state		Average
		1	2	
System state	1	0.2917	0.0417	0.1667
	2	0.0417	0.1250	0.0833
Sum		0.3333	0.1667	**0.2500**

In the next step, we can compute measures $\mathrm{SI}_{3,1}^{\downarrow}$ and $\mathrm{SI}_{3,2}^{\downarrow}$, which quantify the total topological importance of state 1 and 2 respectively of component 3. This can be done using formula (9), and the results are in the bottom part of Table VIII. This data indicates that state 1 of the component has greater influence on the system than state 2. Similarly, using formula (10), measures $\mathrm{SI}_3^{1\downarrow}$ and $\mathrm{SI}_3^{2\downarrow}$ can be computed. The results are in the right column of Table VIII, and they imply that component 3 has bigger influence on system state 1 than on system state 2.

Finally, the total topological importance of component 3, i.e. $\mathrm{SI}_3^{\downarrow}$, can be computed using formulae (11). The result of this computation is the bottom right cell of Table VIII. Value 0.25 implies that the degradation of component 3 results in system degradation in a quarter of situations in which the component is not in state 0.

In the similar way, the other components can be analyzed. The final results are in Table IX. Based on this data, we can state that the components with the greatest topological influence on the system representing laparoscopic surgery are components 1 and 2, i.e., the anesthetist and the laparoscopic robotic surgery machine.

TABLE IX. TOTAL STRUCTURAL IMPORTANCE MEASURES FOR COMPONENTS OF THE SYSTEM REPRESENTING LAPAROSCOPIC SURGERY

Component (i)	$\mathrm{SI}_i^{\downarrow}$
1	0.3056
2	0.4722
3	0.2500
4	0.0556
5	0.1458

C. Birnbaum's Importance

The BI measures corresponding to the other SI measures can be computed based on $\mathrm{BI}_{i,s}^{j\downarrow}$ using the same formulae as (9) – (11). For a specific component, all these measures can be expressed in the form of Table VII.

For example, if we want to investigate the relative importance of component 3 we must consider not only the topology of the system but also the state probabilities of the components. We can then proceed in the same way as in the case of SI. The only difference is the of calculation of $\mathrm{BI}_{3,s}^{j\downarrow}$ for $s = 1,2$ and $j = 1,2$. For illustration, let us compute $\mathrm{BI}_{3,1}^{1\downarrow}$. According to (12) and based on (14) we can write:

$$\begin{aligned}\mathrm{BI}_{3,1}^{1\downarrow} &= \Pr\left\{\begin{array}{l}\{x_1 \leftrightarrow 0\}\{x_2 \leftrightarrow 1\}\{x_5 \leftrightarrow 2\} \\ \vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 0\} \\ \vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 1\}\{x_5 \leftrightarrow 0\} \\ \vee \{x_1 \leftrightarrow 1\}\{x_2 \leftrightarrow 1\}\{x_4 \leftrightarrow 1\}\{x_5 \leftrightarrow 1\}\end{array}\right\} \\ &= p_{1,0}p_{2,1}p_{5,2} + p_{1,1}p_{2,1}p_{4,0} \\ &\quad + p_{1,1}p_{2,1}p_{4,1}p_{5,0} + p_{1,1}p_{2,1}p_{4,1}p_{5,1}.\end{aligned} \tag{15}$$

Using the state probabilities from Table VI, we obtain:

$$BI_{3,1}^{1\downarrow} = 0.1785, \quad (16)$$

which implies that degradation of state 1 of component 3 results in degradation of system state 1 with the probability 0.1785.

The results of all the calculations related to investigation of the importance of component 3 using BI measures are shown in Table X. Based on this data we can state that the most important state of component 3 is state 1 and that the component has the biggest influence on state 2 of the system. We can also say that a minor degradation of this component degrades the system with the probability 0.5736. The final results for all the components are in Table XI. The values in this table rank importance of the components similarly as the SI measures in Table IX and, therefore, after taking the state probabilities of the components into account, the most important components are components 1 and 2, while the least important is component 4 (the anesthetic nurse).

TABLE X. BIRNBAUM'S IMPORTANCE MEASURES FOR COMPONENT 3

		Component state		Average
		1	2	
System state	1	0.1785	0.0003	0.0894
	2	0.8061	0.1622	0.4841
Sum		0.9847	0.1625	**0.5736**

TABLE XI. TOTAL BIRNBAUM'S IMPORTANCE MEASURES FOR COMPONENTS THE SYSTEM REPRESENTING LAPAROSCOPIC SURGERY

Component (i)	$BI_i^{\downarrow}$
1	0.9740
2	0.9899
3	0.5736
4	0.0941
5	0.4869

Importance measures have their own specificity in evaluation of the healthcare system reliability. For laparoscopic surgery, the anesthetist and the laparoscopic robotic surgery machine are most critical components. The importance of these components is confirmed if the probabilities of medical errors of various team members and their device failures are taken into account.

V. CONCLUSION

In this paper we have presented a novel and original method for the construction of the structure function based on incompletely specified and ambiguous data for evaluation of a healthcare system. It directly supports the MSS reliability estimation and can cope with uncertain data for the analysis of system reliability/availability and calculation of importance measures. This is a atypical problem for reliability analysis of healthcare systems, where the data cannot be obtained for all possible situations. We consider the reliability analysis of laparoscopic surgery and compute a set of importance measures that allows to examine the influence mistakes in the work of physicians and/or nurses and the errors in the performance of medical devices.

REFERENCES

[1] B.S. Dhillon, Reliability Technology,Human Error, and Quality in Health Care, CRC Press, 190 p., 2008.

[2] E. Zaitseva, "Reliability analysis methods for healthcare system", in Proc. of the 3rd Int. Conf. on Human System Interaction, Rzeszow, Poland, pp.211-216, 2010, http://dx.doi.org/10.1109/HSI.2010.5514564

[3] A. Taleb-Bendiab, D. England, at al. "A principled approach to the design of healthcare systems: Autonomy vs. governance," Reliability Engineering and System Safety, vol.91, no. 12, pp.1576–1585, 2006, http://dx.doi.org/10.1016/j.ress.2006.01.011

[4] B.S. Dhillon, Human Reliability and Error in Medicine, World Scientific, 2003

[5] Surgical Patient Care. Improving Safety, Quality and Value, Sanchez, J.A., Barach, P., Johnson, J., Jacobs, J.P. (Eds.), Springer, 2017, http://dx.doi.org/10.1007/978-3-319-44010-1

[6] P.Barach "Designing high-reliability healthcare teams," in Proc of Int. Conf. on Information and Digital Technologies, Rzeszów, Poland, pp.17-23, 2016, http://dx.doi.org/10.1109/DT.2016.7557144

[7] F. Bracco, M. Masini et al., "Adaptation of non-technical skills behavioural markers for delivery room simulation," BMC Pregnancy Childbirth, vol. 17(1), pp.89-95, 2017, http://dx.doi.org/10.1186/s12884-017-1274-z

[8] V.Levashenko, E.Zaitseva, "Construction of a reliability structure function based on uncertain data," IEEE Trans. on Reliability, vol.65(4), pp. 1710-1723, 2016, http://dx.doi.org/10.1109/TR.2016.2578948

[9] V.Levashenko, E.Zaitseva et al, "Reliability estimation of healthcare systems using Fuzzy Decision Trees," in Proc. of 2016 Fed. Conf. on Computer Science and Information Systems (FedCSIS 2016), Gdansk, Poland, pp.331-340, 2016, http://dx.doi.org/10.15439/2016F150

[10] J.R.Quinlan, "Simplifying decision trees", Int. J. Man-Machine Studies, vol. 27, pp. 221-234, 1987, http://dx.doi.org/10.1016/S0020-7373(87)80053-6

[11] S.Mitra, K.M.Konwar, and S.K.Pal, "Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation". Trans. on Syst., Man Cybernetics — Part C: Applications and Reviews, vol. 32, pp.328-339, 2002, http://dx.doi.org/10.1109/TSMCC.2002.806060

[12] M.G.Tsipouras, T.P.Exarchos, and D.I.Fotiadis, "A methodology for automated fuzzy model generation," Fuzzy Sets and Systems, vol. 159, no 23, pp.3201-3220, 2008, http://dx.doi.org/10.1016/j.fss.2008.04.004

[13] M.Kvassay, E.Zaitseva, and V.Levashenko, "Importance analysis of multi-state systems based on tools of logical differential calculus", Reliability Engineering and System Safety, vol.165, pp.302-316, 2017, http://dx.doi.org/ 10.1016/j.ress.2017.03.021

[14] W. Kuo, X. Zhu, Importance Measures in Reliability, Risk, and Optimization: Principles and Applications, Wiley, 2012, http://dx.doi.org/10.1002/9781118314593

[15] A.Bognar, P. Barach, J. Johnson, R. Duncan, D. Woods, J. Holl, D. Birnbach, E. Bacha, "Errors and the Burden of Errors: Attitudes,

Perceptions and the Culture of Safety in Pediatric Cardiac Surgical Teams.", Ann Thoracic Surgery, vol.4, pp.1374-1381, 2008

[16] B.Natvig, Multistate Systems Reliability Theory with Applications, Wiley, New York, 2011, http://dx.doi.org/ 10.1002/9780470977088

[17] E.Zio, "Reliability engineering: Old problems and new challenges", Reliability Engineering and System Safety, vol. 94, pp.125–141, 2009, http://dx.doi.org/ 10.1016/j.ress.2008.06.002

[18] R.E. Barlow, F. Proschan, Mathematical theory of reliability, Wiley, New York, 1965

[19] A.Lisnianski, and G.Levitin, Multi-state System Reliability. Assessment, Optimization and Applications. Singapore, SG: World Scientific, 2003.

[20] T.Aven, and B.Heide, "On performance measures for multistate monotone system", Reliability Engineering and System Safety, vol. 41, pp.259–266, 1993, http://dx.doi.org/

[21] E.Zaitseva, and V.Levashenko, "Reliability analysis of multi-state system with application of multiple-valued logic", International Journal of Quality and Reliability Management, vol.34(6), pp.862-878, 2017, http://dx.doi.org/10.1108/IJQRM-06-2016-0081

[22] M.A.Tapia, T.A.Guima, and A. Katbab, "Calculus for a multivalued-logic algebraic system," Applied Mathematics and Computation, vol.42, pp. 255-285, 1991

[23] M. Lyons., S. Adams, M. Woloshynowych. and C. Vincent, "Human reliability analysis in healthcare: A review of techniques," Int. J. of Risk and Safety in Medicine, vol. 16(4), pp. 223–237, 2004

[24] P. Baraldi, L. Podofillini et al., "Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application," Reliability Engineering and System Safety, vol.138, pp.176-193, 2015, https://doi.org/10.1016/j.ress.2015.01.016

[25] G.L Yao, N.Novielli, S. Manaseki-Holland etc. "Evaluation of a predevelopment service delivery intervention: an application to improve clinical handovers", BMJ Qual Saf., vol.21 pp. 29-38, 2012.

[26] R.Kruse, Ch. Doring, and M.-J. Lesot, "Fundamentals of Fuzzy Clustering," in Advances in Fuzzy Clustering and its Applications, eds. J.Valente de Oliveira and W. Pedrycz, Wiley, 434 p., 2007, http://dx.doi.org/10.1002/9780470061190.ch1

[27] H. Tanaka, L.T. Fan, F.S. Lai, and K. Toguchi, "Fault-tree analysis by fuzzy probability," IEEE Trans. on Reliability, vol.32, pp.453-457, 1983, http://dx.doi.org/10.1016/S0019-9958(65)90241-X

DEVELOPMENT THE AUTOMATED INFORMATION SYSTEM OF LADLE-FURNACE PROCESS TO PREDICT THE CONTENT OF ALLOYING ELEMENTS IN BEARING STEEL

O.V. Zhadanos, I.V. Derevyanko, Y.S Proydak, M.I. Gasik
Department of electrometallurgy
National Metallurgical Academy of Ukraine
4 Gagarin Ave., Dnipro, 49600, Ukraine
Alexjad@mail.ru

O.I. Panchenko, A.S. Salnikov, O.V. Yakovitsky
Joint stock company «Dneprospetsstal»,
81 Yuzhnoe Shosse, Zaporizhzhya, 69008, Ukraine,
Salnikov.anat.@dss.com.ua

Abstract— **Regression models of chromium, silicon, manganese and carbon content behavior in metal depending on the amount of added carbonaceous materials, ferrosilicomanganese SiMn17, ferromanganese FeMn78, ferrosilicon FeSi65, ferrochromium FeCr800 are obtained as a result of analysis of experimental data for bearing electric steel ШХ15 and ШХ15СГ-В. These models enable to forecast chemical composition of steel in order to save reducing agents and alloying elements. The structural diagram of automated information system of ladle-furnace is designed according to results of investigations.**

Keywords— ladle furnace; bearing electrical steel; prediction of chemical composition of steel; regression models; carbon-containing materials; ferrosilicomanganese; ferromanganese; ferrosilicon; ferrochromium; adequacy of the model; automated information system (AIS)

I. INTRODUCTION

The most important problem at the stage of bearing electric steel treatment on the ladle-furnace is to provide stable regulated chemical composition of metal and rational charge of alloying and reduction alloys when steelmaking. According to current technology, at ladle-furnace treatment the chemical composition of steel is controlled only by mechanical sample taking and subsequent analysis in the laboratory. Therefore it is important to have data about element concentration behavior in the processed metal and to define the rational charge of alloying and reduction alloys based on these results. One of areas of this problem solution is working out of mathematical model for forecasting the final content of elements in the melt. There are two types of models that characterize the content behavior of chemical elements in metal during steel out-of-furnace treatment: physic-chemical based on thermo chemistry and thermo kinetics laws and regression models.

II. RETROSPECTIVE OF RESEARCHES AND PUBLICATIONS

The advantage of the first models is a high accuracy of forecast [1-5], but structural of such models needs rather complicated calculations. At the same time, actual values of counted magnitudes do not coincide with theoretical ones which require their subsequent correction based on obtained experimental data. Regression models are less accurate, however it is possible to obtain data meeting the requirements to forecasting steel chemical composition at their application [6]. Regression models of C, Mn and Si content change depending on the weight of added alloying and reduction alloys (carbonaceous materials, ferrosilicomanganese SiMn17, ferrosilicon FeSi65) during treatment of 100 t structural steel in ladle-furnace are obtained [7]. To raise accuracy of the regression models it is necessary to consider additionally change of melt weight, the content of leading elements in ferroalloys of each batch as well as other elements which content in ferroalloy is regulated by standards, mass exchange in the system semiproduct – slag-metal mixture. Thus, it is reasonable to develop the regression models of alloying element content change in the process of roller-bearing steel treatment on the ladle-furnace in order to save reduction alloys and alloying ferroalloys.

III. THE PRINCIPLES OF INNOVATIVE TECHNOLOGY OF BEARING ELECTRIC STEELMAKING AT JSC "DNEPROSPETSSTAL"

According to GOST (interstate standard for Commonwealth of Independent States) 801-78, chemical composition of roller-bearing steel grades ШХ15 and ШХ15СГ-В smelted at JSC "Dneprospetsstal" (Ukraine, Zaporizhzhya city) is presented in Table 1.

Roller-bearing steelmaking is carried out under through flow diagram: “electrical arc steel furnace (EAF) (metal-semiproduct), ladle-furnace (sulfur removal, deoxidization, alloying) and vacuum plant (deoxidization, degasification, correcting alloying)” (Fig. 1).

TABLE I. CHEMICAL COMPOSITION OF STEELS ШХ15 AND ШХ15-СГ ACCORDING TO GOST 801-78, SPECIFICATIONS OF JSC "DNEPROSPETSSTAL"

Steel grade	C	Si	Mn	Cr	S	P	Ni	Cu	Ni+Cu
					not more				
ШХ15	0.95-1.05	0.17-0.37	0.2-0.4	1.30-1.65	0.02	0.027	0.30	0.25	0.50
ШХ15СГ-В	0.95-1.05	0.40-0.65	0.9-1.2	1.30-1.65	0.02	0.027	0.30	0.25	0.50

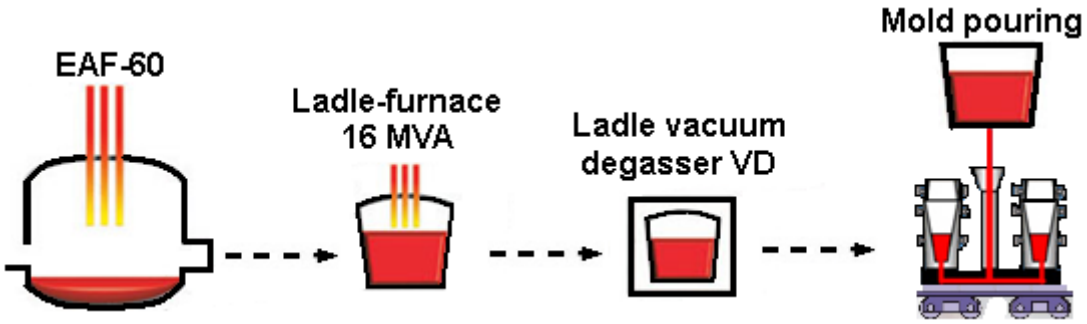


Figure 1. The flow diagram of smelting, out-of-furnace treatment and pouring of roller-bearing steel at JSC "Dneprospetsstal"

Technology of roller-bearing steel ШХ15СГ-В smelting in arc furnaces EAF-60, which was in force at JSC "Dneprospetsstal" till 2008, provided use of ferrosilicon FeSi65 for deoxidization and alloying of metal-intermediate product in EAF and ladle-furnace according to DSTU (Ukrainian state standard) 4127-2002 [8] (63-68 % Si, impurities not more, %: 0.2 C; 0.02 S; 0.05 P; 2.5 Al; 2.5 Mn; 0.5 Cr), high-carbon ferromanganese FeMn78A DSTU 3547-97 [8] (78-82 % Mn, ≤7 % C, impurities not more, %: 2 Si; 0.03 S; 0.05 P), ferrochromium FeCr800A according to GOST 4757-79 [8] (not less than 65% of Cr and not more, %: 8.0 C, Si, 0.06 S; 0.03% P). Solid slag-forming materials consisting of CaO and CaF2 are added when metal tapping from EAF. Metal is deoxidated by Al in the electric furnace-ladle and then subjected to degasification with final deoxidation by Al. Steel is poured in steel molds.

Physic-chemical analysis of processes at all technological stages of production and steel pouring showed [9, 10] that application of home produced ferrosilicon FeSi65 (DSTU 4127-2002) with not regulated content of calcium (0.3-0.6 %) is one of the key uncontrollable factors affecting the formation of globular and oxide inclusions in electric steel grades ШХ15 and ШХ15СГ-В. At the same time, using ferrosilicon FeSi75 that almost does not contain Ca and Al did not help increase the yield of rolled section from the first delivery control by nonmetallic inclusions [9, 10]. It is determined [11] that amount and type of inclusions is defined by the final content of calcium and aluminum in metal based on the results of analysis of effect of technological parameters of steelmaking and refining by flow diagram “arc furnace-ladle-furnace – vacuum vessel” on composition and dimensional groups of inclusions. Ferrosilicon with not restricted high content of calcium (0.3-0.6 %) and high-basic slag on ladle-furnace, from which calcium enters steel as a result of calcium oxide reduction by ferrosilicon and aluminum, are calcium sources in metal [9, 11].

On the basis of stated above, the innovative practice of ШХ15СГ-В steelmaking with the use of ferrosilicomanganese SiMn17A (≥65 % Mn; 15-20 Si; ≤2.5 C; ≤0.03 S; ≤0.1 % P) DSTU 3548-97 was scientifically proved and developed to maintain high yield of rolled section of all five dimensional groups from the first delivery control according to GOST 801-78 [12]. Though calcium content in manganese ferroalloys is not regulated by DSTU standards but is always stably low (less than 0.1-0.15 % of each) proceeding from conditions of ferroalloy production in ferroalloy furnaces. Preliminary deoxidization and almost complete alloying of metal-intermediate product by manganese is carried out in the arc furnace by ferrosilicomanganese.

IV. DEVELOPMENT OF REGRESSION MODELS

The following materials are used for reducing and alloying of bearing steel: ferrosilicon of grade FeSi65 (63-68% Si) DSTU (state standard of Ukraine) 4127-2002 [8], silicomanganese SiMn17 (Mn 65%, 15-20% Si) DSTU 3548-97 [8], ferrochromium FeCr800 (not less than 65% of Cr) GOST 4757-79 and carbon in the form of electrode breakage. Data of industrial smelting operations are processed by following parameters in order to construct regression models:

- weight of metal in the ladle M_{melt} = 58-66.7 tons;
- content of Si, Mn, C, Cr in metal-semiproduct prior to ladle-furnace treatment, %: $[Si]_{init}$ = 0.09-0.55%, $[Mn]_{init}$ = 0.13-1.14%, $[C]_{init}$ = 0.78-1.0%, $[Cr]_{init}$ = 0.49-1.49%;
- weight of ferrochromium FeCr800A, ferrosilicon FeSi65, ferrosilicomanganese SiMn17, high-carbon ferromanganese FeMn78A, carbon, kg:

$m_{FeCr800A}$ = 0-870, m_{FeSi65} = 0-280, m_{SiMn17} = 0-200, $m_{FeMn78A}$ = 0-160, m_C = 0-105;

- content of main elements in ferroalloys for each melting;
- content of Si, Mn, C, Cr in steel upon completion of ladle-furnace treatment, %: $[Si]_{fin}$, $[Mn]_{fin}$, $[C]_{fin}$, $[Cr]_{fin}$;
- change of Si, Mn, C, Cr content in steel according to results of ladle-furnace treatment, %: $\Delta[Si]$, $\Delta[Mn]$, $\Delta[C]$, $\Delta[Cr]$.

Specific charges of alloying and reduction alloys are computed, content of main elements in added ferroalloys is corrected. Data of 47 smelting operations of steel ШХ15 and ШХ15С-Г are approximated by linear regression equations using personal computer [13]. The following model is suggested for estimation of chromium content change [13]:

$$\Delta[Cr] = a_1 \cdot m_{FeCr800A_{sp}} \cdot \frac{[\%Cr]}{[\%Cr_{bas}]} + a_2, \qquad (1)$$

where a_1, a_2 - equation factors, $m_{FeCr800A_{sp}}$ - specific weight of added highcarbon ferrochrome (kg/t), $[\%Cr]$ - chromium content when current melt alloying, $[\%Cr_{bas}]$ - base content of chromium in FeCr800A (accepted 65 %). Estimation of effect (significance) of regression equation factors on change of chromium content $\Delta[Cr]$ by Student criterion is carried out. T-statistics values for each factor of equation are defined by the following equation [14]:

$$t_{a_j} = \left| \frac{a_j}{s_{a_j}} \right|, \qquad (2)$$

where a_j - estimation of j - regression factor, s_{a_j} - estimation of average quadratic deviation of regression factor.

Estimation of average quadratic deviation of regression factors is carried out as follows [15]:

$$s_{a_j} = \frac{s_{rem}}{\sqrt{\frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}{n}} \cdot \sqrt{n-m-1}}, \; s_{a_0} = \frac{s_{rem}}{\sqrt{n-m-1}}, \qquad (3)$$

where n – volume of sampling, m – number of input variables in equation, s^2_{rem} – estimation of remainder variance.

$$s^2_{rem} = \frac{1}{n-m-1} \sum_{i=1}^{n} [y_i - f_i]^2 \qquad (4)$$

We compared obtained T-statistics values of factors to critical value tcr which is defined depending on number of degrees of freedom k = n – m - 1 and significance value α = 0.95 under special tables or is computed on PC [14]. If $|t_{a_j}| \geq t_{cr}$, regression equation factor is considered significant.

T-statistics values of (1) factors are as follows: t_{a_1} = 39.25; t_{a_2} = 1.23. As t-statistics value of factor a_2 is less then critical t_{cr} = 2.01, this equation factor is insignificant and excluded from equation. Therefore, the final equation (1) is as follows [13]

$$\Delta[Cr] = a_1 \cdot m_{FeCr800A_{sp}} \cdot \frac{[\%Cr]}{[\%Cr_{bas}]}, \qquad (5)$$

Regression model adequacy by Fisher's ratio test is also estimated. F - statistics value is computed from equation 6 [15].

$$F_{calc} = \left(\frac{S_{regr}}{S_{rem}} \right) \cdot \left(\frac{k_2}{k_1} \right), \qquad (6)$$

where k_1 = m, k_2 = n - 2 degree of freedom.

If F_{calc} outnumbers the critical value of Fisher distribution F_{cr}, the equation is significant. As a result of calculations [13] we obtained the following values F_{calc} = 1521 and F_{cr} = 4.06 (α = 0.05), i.e. (2) is significant and numerical value of factor a_1 = 0.064 (Fig, 2). Determinacy factor r^2 of developed model is 0.97 and absolute accuracy of forecast – 0.05 %.

Regression model of manganese content change depending on specific charges of SiMn17 and FeMn78 is obtained in a similar way [13].

$$\Delta[Mn] = b_1 \cdot m_{SiMn17_{sp}} \cdot \frac{[\%Mn]_{SiMn17}}{[\%Mn]_{SiMn17_{bas}}} + \\ + b_2 \cdot m_{FeMn78A_{sp}} \cdot \frac{[\%Mn]_{FeMn78}}{[\%Mn]_{FeMn78_{sp}}} + b_3, \qquad (7)$$

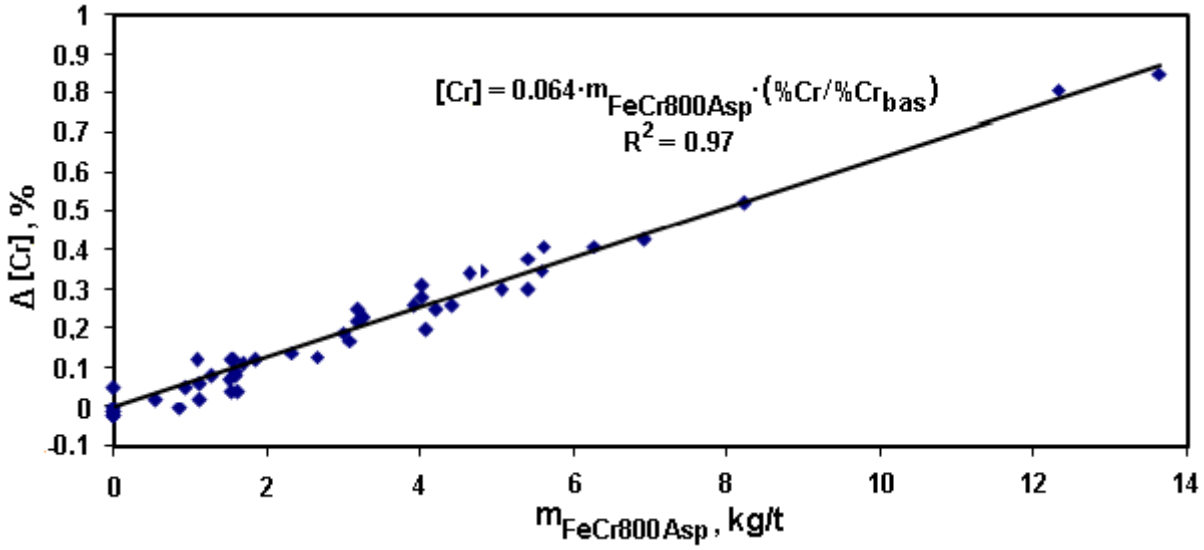


Figure 2. Dependence of Cr content change in roller-bearing steel ШХ15 and ШХ15СГ-В during ladle-furnace treatment on specific charge of high-carbon ferrochrome $m_{FeCr800A_{sp}}$

where b_1 , b_2 , b_3 - equation factors, $m_{SiMi17_{sp}}$ - specific weight of added ferrosilicomanganese (kg/t), $m_{FeMn78A_{sp}}$ - specific weight of high carbon ferromanganese (kg/t), $[\%Mn]_{SiMn17}$, and $[\%Mn]_{FeMn78}$ - manganese content in ferrosilicomanganese and high-carbon ferromanganese when alloying, $[\%Mn]_{SiMn17_{bas}}$, $[\%Mn]_{FeMn78_{bas}}$ - base content of manganese in ferrosilicomanganese and high-carbon ferromanganese. T-statistics values of (7) factors are obtained as a result of calculations: $t_{б1}$ = 13.3; $t_{б2}$ = 9.74, $t_{б3}$ = 2.13. As t-statistics value of factors is more than critical $t_{cr} = 2.01$, all equation factors are significant. Based on the results of (7) by Fisher's ratio test it is determined that F_{calc} = 95 and F_{cr} = 4.06, i.e. the equation is significant. Numerical values of equation factors are: b_1 = 0.067, b_2 = 0.071, b_3 = 0.006 (Fig. 3). Determinacy factor r^2 = 0.81 and absolute accuracy of forecast for this model is 0.06 %.

Equation (8) is obtained based on analysis of effect of addition of alloying and reduction alloys on change of silicon content.

$$\Delta[Si] = c_1 \cdot m_{SiMn17_{sp}} \cdot \frac{[\%Si]_{SiMn17}}{[\%Si]_{SiMn17_{bas}}} + c_2 \cdot m_{FeSi65_{sp}} \cdot \frac{[\%Si]_{FeSi65}}{[\%Si]_{FeSi65_{bas}}} + c_3 \cdot m_{FeCr800A_{sp}} + c_4 \cdot m_{FeMn78A_{sp}} + c_5 \quad (8)$$

where c_1 , c_2 , c_3 , c_4 , c_5 - equation factors, $m_{SiMi17_{sp}}$ - specific weight of added ferrosilicomanganese (kg/t), $m_{FeSi65_{sp}}$ - specific weight of ferrosilicon (kg/t), $m_{FeCr800A_{sp}}$ - specific weight of high-carbon ferrochrome (kg/t), $m_{FeMe78A_{sp}}$ - specific weight of high-carbon ferromanganese (kg/t), $[\%Si]_{SiMn17}$ and $[\%Si]_{FeSi65}$ - silicon content in ferrosilicomanganese and ferrosilicon during alloying, $[\%Si]_{SiMn17_{bas}}$, $[\%Si]_{FeSi65_{bas}}$ - base content of silicon in ferrosilicomanganese and ferrosilicon.

For (8): t_{c1} = 2.51; t_{c2} = 8.02, t_{c3} = 0.81, t_{c4} = 0.3, t_{c5} = 1.0. As values t-statistics for factors c_3 , c_4 , c_5 are less than critical t_{cr} = 2.01, these factors are not significant. And the final model is:

$$\Delta[Si] = c_1 \cdot m_{SiMn17_{sp}} \cdot \frac{[\%Si]_{SiMn17}}{[\%Si]_{SiMn17_{bas}}} + c_2 \cdot m_{FeSi65_{sp}} \cdot \frac{[\%Si]_{FeSi65}}{[\%Si]_{FeSi65_{bas}}} \quad (9)$$

Check by Fisher's ratio test (F_{calc} = 50 и F_{cr} = 4.06) showed that equation (9) is significant. Numerical values of equation factors are c_1 = 0.021, c_2 = 0.043 (Fig. 4). Determinacy factor of model is r^2 = 0.74 and absolute accuracy of forecast – 0.07 %.

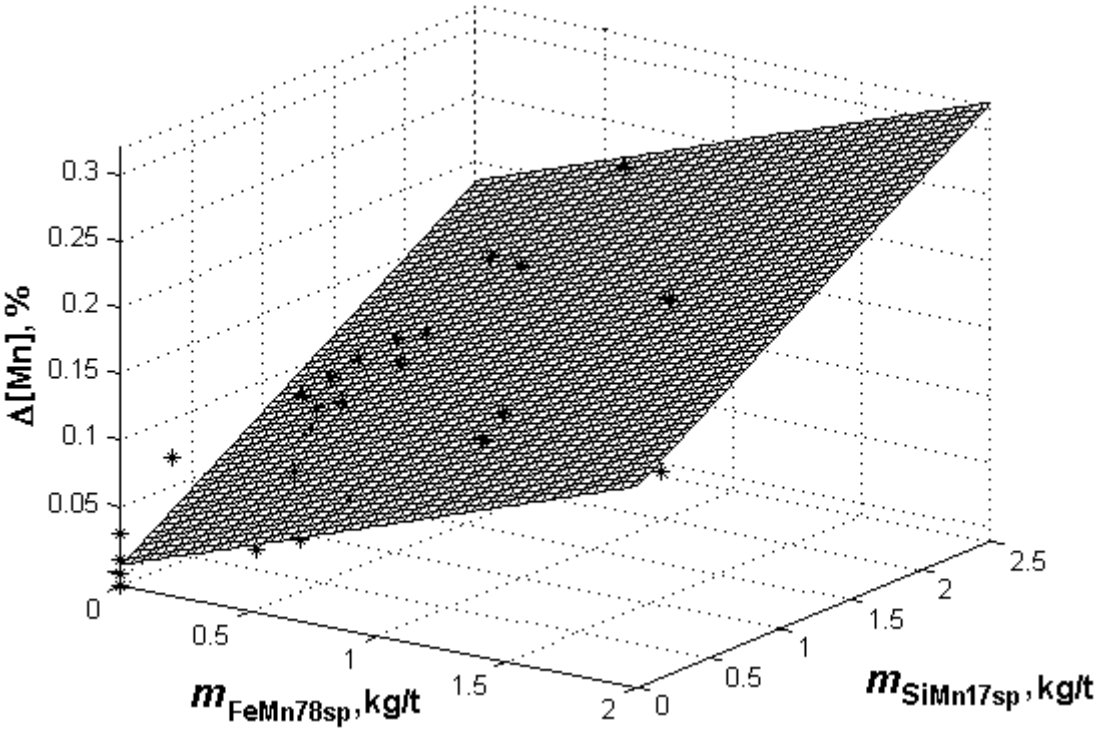


Figure 3. Dependence of Mn content change in bearing electric steel during ladle-furnace treatment on specific charge $m_{SiMi17_{sp}}$ and $m_{FeMn78A_{sp}}$: points - rated values of smelting operations, plane – obtained model

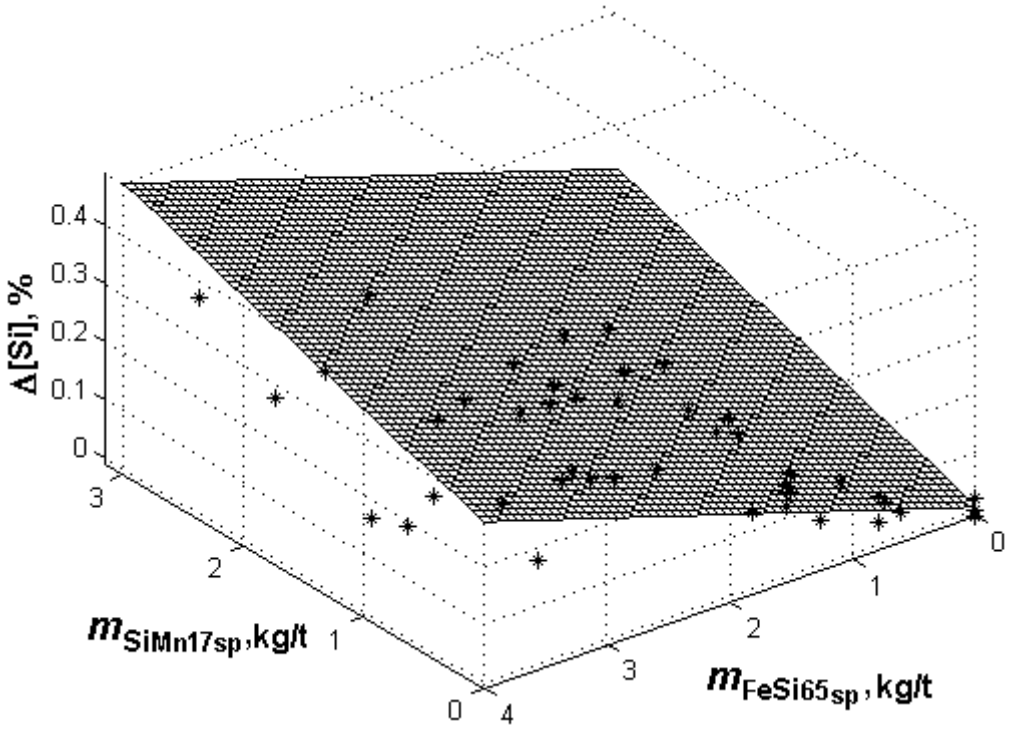


Figure 4. Dependence of Si content change in bearing electric steel during ladle-furnace treatment on specific charge $m_{SiMi17_{sp}}$ and $m_{FeSi65_{sp}}$

Equation (10) is developed for estimation of carbon content change in metal

$$\Delta[C] = d_1 \cdot [C]_{init} + d_2 \cdot m_{C_{sp}} + d_3 \cdot m_{SiMn17_{sp}} + d_4 \cdot m_{FeSi65_{sp}} + d_5 \cdot m_{FeCr800A_{sp}} + d_6 \cdot mFeMn78A_{sp} + d_7, \quad (10)$$

where d_1, d_2, d_3, d_4, d_5, d_6, d_7 - equation factors, $m_{C_{sp}}$ - specific weight of carbonaceous flux cored wire (kg/t), $m_{SiMi17_{sp}}$ - specific weight of ferrosilicomanganese (kg/t), $m_{FeSi65_{sp}}$ - specific weight of ferrosilicon (kg/t), $m_{FeCr800A_{sp}}$ - specific weight of high-carbon ferrochrome (kg/t), $m_{FeMe78A_{sp}}$ - specific weight of high-carbon ferromanganese (kg/t).

T-statistics values of equation (10) factors are defined: t_{d1} = 10.4; t_{d2} = 12.5, t_{d3} = 3.2, t_{d4} = 1.03, t_{d5} = 6.3, t_{d6} = 5.1, t_{d7} = 8.0. As t-statistics value for factor d_4 is less than critical t_{cr} = 2.01, this factor of equation is not significant. Therefore, the total equation will be as follows:

$$\Delta[C] = d_1 \cdot [C]_{init} + d_2 \cdot m_{C_{sp}} + d_3 \cdot m_{SiMi17_{sp}} + \\ + d_4 \cdot m_{FeCr800A_{sp}} + d_5 \cdot m_{FeMe78A_{sp}} + d_6 \quad , \quad (11)$$

As F_{calc} = 61 and F_{cr} = 4.06, equation (11) is significant, and factors are: d_1 = - 0.84, d_2 = 0.01, d_3 = 0.0027, d_4 = 0.009, d_5 = 0.006, d_6 = 0.82. Absolute accuracy of the forecast for this model is 0.03 %, determinacy factor r^2 = 0.78.

Obtained mathematical models are presented in Table 2.

V. Modeling and Analysis of Results

To analyze the effectiveness of mathematical models considered real industry situations:

- Metal-semiproduct with the mass of 63 tons released from electrical arc furnace into the ladle with the following content of alloying elements ($[Si]_{init}$ = 0.22%, $[Mn]_{init}$ = 0.14%, $[C]_{init}$ = 0.9%, $[Cr]_{init}$ = 1.29%);
- during ladle-furnace processing added 207 kg FeCr800A ($m_{FeCr800A_{sp}}$ = 3.29 kg/t), 107 kg SiMn17 ($m_{SiMi17_{sp}}$ = 1.7 kg /t), 80 kg FeMn78A ($m_{FeMn78A_{sp}}$ = 1.27 kg/t), 95 kg FC65 ($m_{FeSi65_{sp}}$ = 1.5 kg/t), 95 kg of C ($m_{C_{sp}}$ = 1.5 kg/t).
- After ladle-furnace processing the metal had next content of alloying elements: $[Si]_{fin}$ = 0.32%; $[Mn]_{fin}$ = 0.35%; $[C]_{fin}$ = 1.01%; $[Cr]_{fin}$ = 1.5%.

With using the developed mathematical models the consumption of alloying and deoxidizing expense was 138 kg FeCr800A ($m_{FeCr800A_{sp}}$ = 2,19 kg/t), 82 kg SiMn17 ($m_{SiMi17_{sp}}$ = 1.3 kg/t), 61 kg FeMn78A ($m_{FeMe78A_{sp}}$ = 0,97 kg/t), 48 kg FeSi65 ($m_{FeSi65_{sp}}$ = 0.76 kg/m), 95 kg of C ($m_{C_{sp}}$ = 1.5 kg/t), and the content of alloying elements was $[Si]_{fin}$ = 0.27%, $[Mn]_{fin}$ = 0.3%, $[C]_{fin}$ = 1.0%, $[Cr]_{fin}$ = 1.43% (i.e. the middle of the range). Saving of alloying and deoxidizing was 69 kg FeCr800A kg, 25 kg SiMn17, 19 kg FeMn78A, 47 kg FeSi65.

VI. The Block Diagram of Automated Information System

To implementation the developed models proposed automated information system (AIS) was proposed as a part of the observable automated control system of ladle-furnace (Fig. 5). The main purpose of AIS is to give the operator at the control panel the current content of carbon ($[C]_t$), Si ($[Si]_t$), Mn ($[Mn]_t$), Cr ($[Cr]_t$) in the metal during processing within the ladle furnace and recommendations about quantity of chemical additives, which necessary add in the melt C ($m_{C_{rec}}$) , FeSi65 ($m_{FeSi65_{rec}}$), SiMn17 ($m_{SiMi17_{rec}}$) FeMn78A ($m_{FeMn78A_{rec}}$) FCr800A ($m_{FeCr800A_{rec}}$). AIS consists of the following subsystems: "predict of $\Delta[C]$ "; "predict of $\Delta[Si]$ "; "predict of $\Delta[Mn]$,"predict of $\Delta[Cr]$; "calc. of $[C]_t$ "; "calc. of $[Si]_t$ "; "calc. of $[Mn]_t$ "; "calc. of $[Cr]_t$ "; "Recommendations about mMnC17»; 'Recommendations about mFeSi65 "," recommendations about mFeMn78A", "recommendations about mFeCr800A". The input parameters of the system are: $[Si]_{init}$, $[Mn]_{init}$, $[C]_{init}$, $[Cr]_{init}$, (the results of measurements enter to the system of mathematical models through a programmable microprocessor controller); the amount, time and kind of added to the melt chemical additives, mC_t , m_{FeSi65_t} , m_{SiMn17_t} , m_{FeMn78_t} , $m_{FeCr800_t}$, the aim values of changing content Si, Mn, C, Cr - $[Si]_{fin.aim}$, $[Mn]_{fin.aim}$, $[C]_{fin.aim}$, (setting by the operator PPC). The above mentioned input and output subsystem's parameters, together with the results of the intermediate measurements additionally transferred to the subsystem "Backup". In the case technology changes of ladle-furnace's process, the availability of subsystem "Backup" allows to perform automatic correction of model coefficients, embodied in the subsystems of automated information system.

AIC is realized by integrating into existing process control system industrial computer with developed mathematical models.

TABLE II. MATHEMATICAL MODELS OF CHANGE OF ALLOYING ELEMENTS CONTENT IN THE LADLE-FURNACE

Forecasted parameter	Mathematical model
$\Delta[Cr]$	$\Delta[Cr] = a_1 \cdot m_{FeCr800A_{sp}} \cdot \frac{[\%Cr]}{[\%Cr_{bas}]}$
$\Delta[Mn]$	$\Delta[Mn] = b_1 \cdot m_{SiMi17_{sp}} \cdot \frac{[\%Mn]_{SiMi17}}{[\%Mn]_{SiMi17_{bas}}} + b_2 \cdot m_{FeMn78A_{sp}} \cdot \frac{[\%Mn]_{FeMn78}}{[\%Mn]_{FeMe78_{bas}}} + b_3$
$\Delta[Si]$	$\Delta[Si] = c_1 \cdot m_{SiMn17_{sp}} \cdot \frac{[\%Si]_{SiMi17}}{[\%Si]_{SiMi17_{bas}}} + c_2 \cdot m_{FeSi65_{sp}} \cdot \frac{[\%Si]_{FeSi65}}{[\%Si]_{FeSi65_{bas}}}$
$\Delta[C]$	$\Delta[C] = d_1 \cdot [C]_{init} + d_2 \cdot m_{C_{sp}} d_3 \cdot m_{SiMi17_{sp}} + d_4 \cdot m_{FCr800A_{sp}} + d_5 \cdot m_{FeMe78A_{sp}} + d_6$

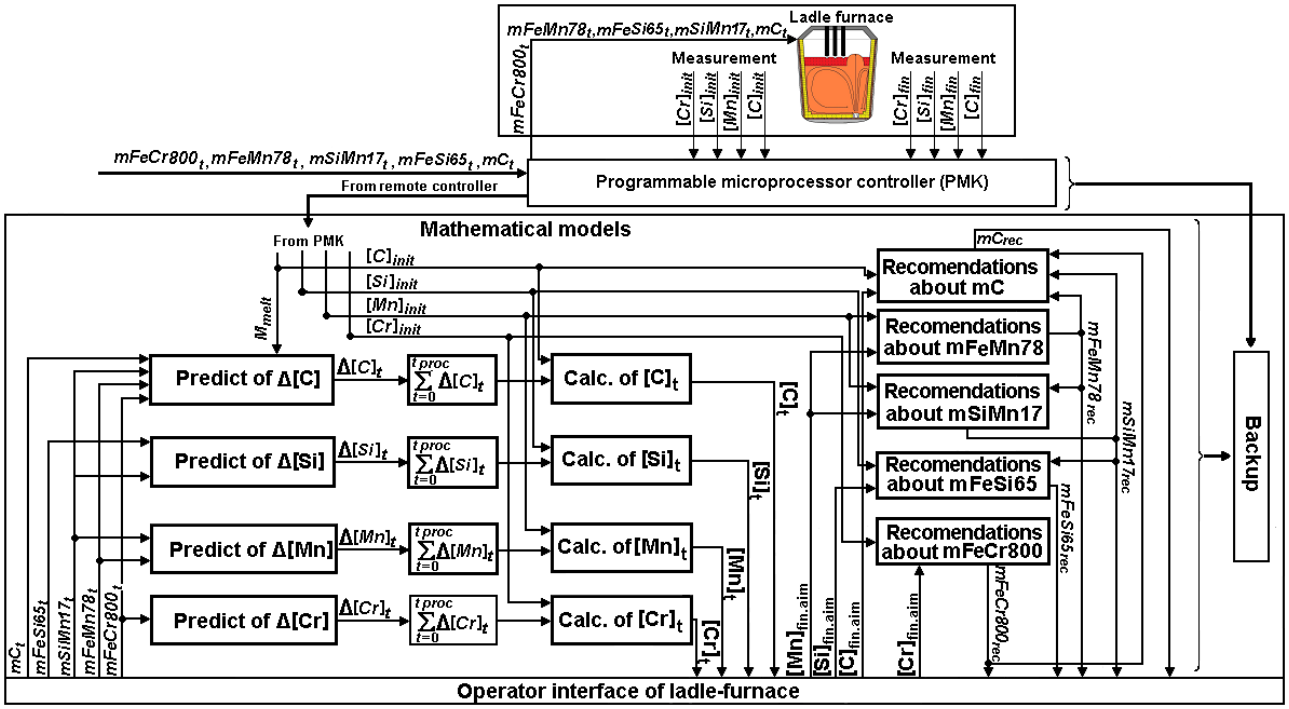


Figure 5. The block diagram of an automated information system for monitoring the chemical composition of bearing electric steel during processing in ladle-furnace based on developed models

VII. CONCLUSIONS

1. Developed regression models of alloying elements content change in the process of bearing electric steel ШХ15 and ШХ15С-Г treatment using ladle-furnace enable to forecast the concentration of Si, Mn, Cr and C in steel during treatment.

2. The structural diagram AIS for implementation of out-of-furnace treatment of bearing electric steel for the purpose of chemical composition monitoring and recommending the rational charge of alloying and reduction alloys is made.

REFERENCES

[1] A.G. Ponomarenko, M.P. Gulyaev, I.V. Derevyanchenko, S.A .Khrapko, R.N. Martynov, R.V. Sinyakov, D.A. Ponomarenko, O.L. Kucherenko, R.N. Pilchuk. Industrial development of computer control of steel smelting at BMP and MMP, based on physico-chemical model ORACLE. Proceedings of Fifth Congress of Steelmakers, Rybnitsa (October 14-17, 1998), Moscow, Chermetinformatsiya, 1999, pp. 174-177.*

[2] S.V. Kazakov. Prediction the composition of steel melts during smelting and secondary treatment. Metal and Casting of Ukraine, 2005, No. 3-4, pp. 17-20.*

[3] Snegirev, Yu.V., Tutarova, V.D. and Fedorova, A.A. Using artificial neural network to predict steel chemical composition during secondary treatment in ladle furnace. Software of systems in the industrial and social fields. 2014. - №1. p. - 41-48.*

[4] Zora JANCÍKOVÁ, Pavel ŠVEC. Prediction of chemical composition of refining slag with exploitation of artificial neural networks. Cybernetic letters: informatics, cybernetics and robotics. ISSN 1802-3525. 2008. - №2.

[5] E.V. Prikhodko, D.N. Togobitskaya, A.F. Petrov, A F. Khamkhotko, S.V. Grekov. Physic-Chemical Properties Forecasting for Manganese Ferroalloy Production Slag. Metallurgical and Mining Industry, 2010, Vol. 2, No. 3 – p.p. 186-192*.

[6] Vihlevschuk V.A., Kharakhulakh V.S., Brodsky S.S. Ladle finishing of steel. Dnepropetrovsk: SSPC "System Technology", 2000 - 190 p.

[7] Zhadanos Oleksandr, Derevyanko Ihor. Development the mathematical models for prediction the content of alloying elements in structural steel during ladle-furnace process. Central European Researchers Journal, Vol. 2, Issue. 1, 2016 – pp. 16-21.

[8] Gasik M.I., Lyakishev N.P.. Physicochemistry and technology of electric ferroalloys. Dnipropetrovsk, Systemnye Technologii, 2005, 448 p.*

[9] A.I. Panchenko, I.N. Logozinsky, A.S. Salnikov, L.N. Korol, S.S. Kazakov, M.I. Gasik, A.P. Gorobets. Comparative pilot-industrial influence researches of 65% ferrosilicon with different contents of calcium on ШХ15СГ-В steel contamination with aluminum-calcium inclusions. Advances in Electrometallurgy, 2007, No. 4, pp. 49-55.*

[10] M.I. Gasik, K.V. Grigorovich, A.I. Panchenko, A.S. Salnikov, S.S. Shibayev, A.K. Gerber, A.Y. Dalmatov. Non-metallic inclusions in sort rolling mill products of ШХ15СГ-В steel. Electrometallurgy, 2010, No. 5, pp. 2-14.*

[11] M.I. Gasik, A.P. Gorobets, A.I. Panchenko, I.N. Logozinsky, A.S. Salnikov, S.S. Kazakov, L.M. Skripka, S.A. Kasian, O.N. Sezonenko. Theoretical preconditions for the formation of oxide and globular inclusions at various residual contents of calcium and aluminum. Metallurgical and Mining Industry, 2008, No. 1, pp. 48-54.*

[12] M.I. Gasik, A.I. Panchenko, L.M. Skripka, A.S. Salnikov, S.L. Mazuruk. Technology of smelting pure ШХ15СГ-В electric steel with diversification of ferroalloys. Steel, 2009, No. 6, pp. 25-28.*

[13] A.V. Zhadanos, M.I. Gasik A.I. Panchenko, A.S. Salnikov, L.M. Skripka. Mathematical Model of Roller-Bearing Electric Steel Chemical Composition Control on the Ladle-furnace. Metallurgical and Mining Industry, 2010, Vol. 2, No. 6 – p.p. 390-396.

[14] Kukushkin O.N., Beitcun S.V., Zhadanos A.V. Statistics in Excel. Dnepropetrovsk: National metallurgical academy of Ukraine, 2002. - 64 c.

[15] William J. Orvis. Excel for Scientists and Engineers. SYBEX Inc. Alameda, CA, USA ©1995. ISBN 0782117619. 508 p.

*Published in Russian

Co-funded by the Tempus Programme of the European Union

This publication is the result of the project implementation:
TEMPUS CERES: Centers of Excellence for young RESearchers.
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES

CERES