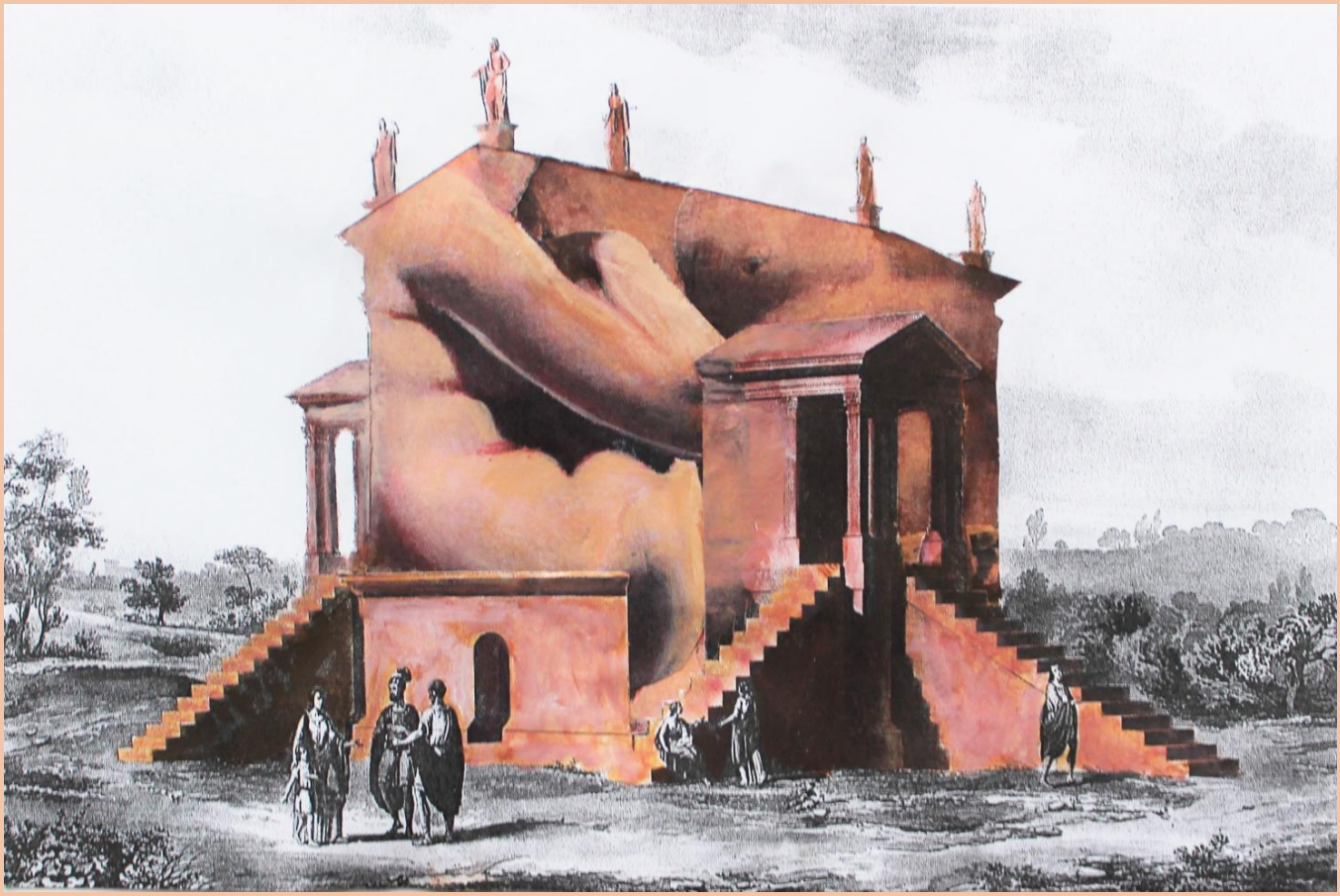
Annals of Computer Science and Information Systems
Volume 26

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems

September 2–5, 2021. Online



Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki,
Dominik Ślęzak (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 26

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Dominik Ślęzak (eds.)

2021, Warszawa, Polskie Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 26

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems

USB: ISBN 978-83-962423-0-3
WEB: ISBN 978-83-959183-9-1

ISSN 2300-5963
DOI 10.15439/978-83-959183-9-1

Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover art:
Karolina Kardas,
Elbląg, Poland

Also in this series:

Volume 25: Proceedings of the 16th Conference on Computer Science and Intelligence Systems, **ISBN Web 978-83-959183-6-0, ISBN USB 978-83-959183-7-7, ISBN ART 978-83-959183-8-4**
Volume 24: Proceedings of the International Conference on Research in Management & Technovation 2020, **ISBN WEB: 978-83-959183-5-3, ISBN USB: 978-83-959183-4-6**
Volume 23: Communication Papers of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-959183-2-2, ISBN USB: 978-83-959183-3-9**
Volume 22: Position Papers of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-959183-0-8, ISBN USB: 978-83-959183-1-5**
Volume 21: Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-955416-7-4, ISBN USB 978-83-955416-8-1, ISBN ART 978-83-955416-9-8**
Volume 20: Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-955416-3-6, ISBN USB: 978-83-955416-4-3**
Volume 19: Position Papers of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-955416-1-2, ISBN USB: 978-83-955416-2-9**
Volume 18: Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-952357-8-8, ISBN USB 978-83-952357-9-5, ISBN ART 978-83-955416-0-5**
Volume 17: Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-952357-0-2, ISBN USB: 978-83-952357-1-9**
Volume 16: Position Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-949419-8-7, ISBN USB: 978-83-949419-9-4**
Volume 15: Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-949419-5-6, ISBN USB 978-83-949419-6-3, ISBN ART 978-83-949419-7-0**

DEAR Reader, it is our pleasure to present to you Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems (FedCSIS'2021), which took place, fully remotely, on September 2-4, 2021. Conference was originally planned to take place in Sofia, Bulgaria, but the global COVID-19 pandemics forced us to adapt and organize the conference online.

Before proceeding further, let us share a very important information. In June 2021 FedCSIS conference series has been ranked B in the CORE ranking system. This constitutes a major achievement for the series. This is particularly valuable achievement since the series was not ranked before. We would like to thank prof. Paweł Sitek for leading our efforts and preparing all necessary documentation.

Position papers comprise two categories of contributions: *challenge papers* and *emerging research papers*. *Challenge papers* propose and describe research challenges in theory or practice of computer science and information systems. Papers in this category are based on deep understanding of existing research, or industrial problems. Based on such understanding and experience, they define new exciting research directions and show why these directions are crucial to the society at large. *Emerging research papers* present preliminary research results from work-in-progress, based on sound scientific approach but presenting work not completely validated as yet. They precisely describe the research problem and its rationale. They also define the intended future work including the expected benefits from solution to the tackled problem. Subsequently, they may be more conceptual than experimental.

The *communication papers* were introduced in 2017 as a separate category of contributions. They report on research topics worthy of immediate communication. They may be used to mark a hot new research territory, or to describe work in progress in order to quickly present it to scientific community. They may also contain additional information omitted from the earlier papers or may present software tools and products in a research state.

FedCSIS'2021 was chaired by prof. Stefka Fidanova, while dr. Nina Dobrinkova acted as the Chair of the Organizing Committee. This year, the FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, Warsaw University of Technology, Wrocław University of Economics and Business, and Institute of Information and Communication Technologies, Bulgarian Academy of Sciences.

FedCSIS'2021 was technically co-sponsored by: IEEE Bulgarian Section, IEEE Poland Section, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, IEEE Poland Section Control System Society Chapter, Committee of Computer Science of the Polish Academy of Sciences, Mazovia Cluster ICT, Poland, Eastern Cluster ICT, Poland, and Bulgarian Section of SIAM.

We also glad to announce that this year (and we believe that in future also) the FedCSIS conference formed strategic alliance with QED Software, a Polish software company developing AI-based products.

During FedCSIS 2021, the keynote lectures were delivered by:

- David Bader, Distinguished Professor, New Jersey Institute of Technology: "*Solving Global Grand Challenges with High Performance Data Analytics"*
- Rajkumar Buyya, Director, Cloud Computing and Distributed Systems (CLOUDS) Lab, The University of Melbourne, Australia and CEO, Manjrasoft Pvt Ltd, Melbourne, Australia: "*Neoteric Frontiers in Cloud and Edge Computing*"
- Hristo Djidjev, Los Alamos National Laboratory: "*Using quantum annealing for discrete optimization"*
- Moshe Y Vardi, Professor, Rice University: *"Lessons from COVID-19: Efficiency vs Resilience"*

FedCSIS'2021 consisted of five Tracks and one special event for Young Researchers. Within each Track, topical Technical Sessions have been organized. Some of these Technical Sessions have been associated with the FedCSIS conference series for many years, while some of them are relatively new. Their role is to focus and enrich discussions on selected areas pertinent to the general scope of each Track. Here is the list of tracks and topical Technical Sessions organized within their scope.

Track 1: Artificial Intelligence in Applications (16th Symposium AAIA'21)

- Computational Optimization (14th International Workshop WCO'21)

Track 2: Computer Science & Systems (CSS'21)

- Computer Aspects of Numerical Algorithms (14th Workshop CANA'21)
- Multimedia Applications and Processing (14th International Symposium MMAP'21)

Track 3: Network Systems and Applications (NSA'21)

- Internet of Things – Enablers, Challenges and Applications (5th Workshop IoT-ECAW'21)
- Cyber Security, Privacy, and Trust (2nd International Forum NEMESIS'21)

Track 4: Advances in Information Systems and Technology (AIST'21)

- Data Science in Health, Ecology and Commerce (3rd Special Session DSH'21)
- Information Systems Management (16th Conference ISM'21)
- Knowledge Acquisition and Management (27th Conference KAM'21)

Track 5: Software and System Engineering (S3E'21)

- Cyber-Physical Systems (8th International Workshop IWCPS-8)
- Software Engineering Workshop (41st IEEE Workshop SEW-41)

Artificial Intelligence and Cybersecurity (1st Young Researchers Workshop YRW'21)

Each paper, found in this volume, was refereed by at least two referees.

The program of FedCSIS required a dedicated effort of many people. We would like to express our warmest gratitude to all Committee members, of each Track and each Technical Session, for their hard work in attracting and later refereeing 129 submissions.

We thank the authors of papers for their great contribution into theory and practice of computing and software systems. We are grateful to the invited speakers for sharing their knowledge and wisdom with the participants.

We hope that you had an inspiring conference. We also hope to meet you again for the 17th Conference on Computer Science and Intelligence Systems (FedCSIS 2022). This time we are almost certain that we will be able to organize the conference on site and will finally reach Sofia, Bulgaria.

Co-Chairs of the FedCSIS Conference Series:

Maria Ganzha, *Warsaw University of Technology, Poland and Systems Research Institute Polish Academy of Sciences, Warsaw, Poland*

Leszek Maciaszek, *Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, Warsaw Poland and Management Academy, Warsaw, Poland*

Dominik Ślęzak, *Institute of Informatics, University of Warsaw, Poland*

Annals of Computer Science and Information Systems, Volume 26

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems

September 2–5, 2021. Online

TABLE OF CONTENTS

15th International Symposium Advances in Artificial Intelligence and Applications

THIS track is a continuation of international AAIA symposiums, which have been held since 2006. It aims at establishing the synergy between technical sessions, which encompass wide range of aspects of AI. With its longest-tradition threads, such as WCO focusing on Computational Optimization, it is also open to new initiatives categorized with respect to both, the emerging AI-related methodologies and practical usage areas. Nowadays, AI is usually perceived as closely related to the data, therefore, this track's scope includes the elements of Machine Learning, Data Quality, Big Data, etc. However, the realm of AI is far richer and our ultimate goal is to show relationships between all of its subareas, emphasizing a cross-disciplinary nature of the research branches such as XAI, HCI, and many others.

AAIA'21 brings together scientists and practitioners to discuss their latest results and ideas in all areas of Artificial Intelligence. We hope that successful applications presented at AAIA'21 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

TOPICS

Papers related to theories, methodologies, and applications in science and technology in the field of AI are especially solicited. Topics covering industrial applications and academic research are included, but not limited to:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Reinforcement Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- Business Intelligence and Online Analytics
- Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications
- AI for Combinatorial Games, Video Games and Serious Games
- Evolutionary Algorithms and Evolutionary Computation
- Computational Optimization (14th Workshop WCO'21)

TRACK CHAIRS

- **Ślęzak, Dominik,** University of Warsaw, Poland
- **Matwin, Stan,** Dalhousie University, Canada

PROGRAM CHAIRS

- **Świechowski, Maciej,** QED Software, Poland
- **Sosnowski, Łukasz,** Dituel, Poland

PROGRAM COMMITTEE

- **Agre, Gennady,** Bulgarian Academy of Sciences, Bulgaria
- **Bianchini, Monica,** University of Siena, Italy
- **Calpe Maravilla, Javier,** University of Valencia, Spain
- **Chelly, Zaineb,** Université de Versailles Saint-Quentin en Yvelines UFR des Sciences, France
- **Cyganek, Bogusław,** AGH University of Science and Technology, Poland
- **Dey, Lipika,** TCS Innovation Lab Delhi, India
- **Düntsch, Ivo,** Brock University, Canada
- **Girardi, Rosario,** UNIRIO, Brazil
- **Grabowski, Adam,** University of Bialystok, Poland
- **Ignatov, Dmitry,** National Research University Higher School of Economics, Russia
- **Jaromczyk, Jerzy,** University of Kentucky, United States
- **Jin, Xiaolong,** Chinese Academy of Sciences, China
- **Kasprzak, Włodzimierz,** Warsaw University of Technology, Poland
- **Kayakutlu, Gulgun,** Istanbul Technical University, Turkey
- **Lingras, Pawan,** Saint Mary's University, Canada
- **Loukanova, Roussanka,** Stockholm University, Sweden; and Institute of Mathematics and Informatics Bulgarian Academy of Sciences, Bulgaria
- **Markowska-Kaczmar, Urszula,** Wroclaw University of Technology, Poland
- **Matson, Eric,** Purdue University, USA
- **Matwin, Stan,** Dalhousie University, Canada
- **Menasalvas, Ernestina,** Universidad Politécnica de Madrid, Spain
- **Meneses, Claudio,** Universidad Católica del Norte, Chile
- **Moshkov, Mikhail,** King Abdullah University of Science and Technology, Saudi Arabia
- **Mozgovoy, Maxim,** University of Aizu, Japan
- **Myszkowski, Paweł,** Wroclaw University of Science and Technology, Poland

- **Pataricza, András,** Budapest University of Technology and Economics, Hungary
- **Peters, Georg,** Munich University of Applied Sciences & Australian Catholic University, Germany & Australia
- **Po, Laura,** Universitá di Modena e Reggio Emilia, Italy
- **Porta, Marco,** University of Pavia, Italy
- **Przybyła-Kasperek, Małgorzata,** University of Silesia, Poland
- **Raghavan, Vijay,** University of Louisiana at Lafayette, USA
- **Ramanna, Sheela,** University of Winnipeg, Canada
- **Rauch, Jan,** University of Economics, Prague, Czech Republic
- **Reformat, Marek,** University of Alberta, Canada
- **Schaefer, Gerald,** Loughborough University, England
- **Sikora, Marek,** Silesian University of Technology, Poland
- **Stanczyk, Urszula,** Silesian University of Technology, Poland
- **Stoean, Catalin,** University of Craiova, Romania
- **Subbotin, Sergey,** Zaporozhye National Technical University
- **Szczech, Izabela,** Poznan University of Technology, Poland
- **Unland, Rainer,** University of Duisburg-Essen, ICB, Germany
- **Weber, Richard,** University of Chile, Chile
- **Verstraete, Jörg,** Systems Research Insitute, Polish Academy of Sciences, Poland
- **Zakrzewska, Danuta,** Institute of Information Technology Technical University of Lodz, Poland
- **Zdravevski, Eftim,** Ss.Cyril and Methodius University, Macedonia
- **Zaineb Chelly,** Aberystwyth University, Wales
- **Zielosko, Beata,** University of Silesia, Poland

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 3–11
DOI: 10.15439/2021F127
ISSN 2300-5963 ACSIS, Vol. 26

An impact of tensor-based data compression on deep neural network accuracy

Jakub Grabek*† and Bogusław Cyganek*†
*Department of Electronics,
AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland
Email: *jakub.grabek@qed.pl, cyganek@agh.edu.pl*
†QED Software Sp. z o.o.,
Mazowiecka 11/49, 00-052 Warszawa, Poland

***Abstract*—The emergence of the deep neural architectures greatly influenced the contemporary big data revolution. However, requirements on large datasets even increased a necessity for efficient data storage. The storage problem is present at all stages, from the dataset creation up to the training and prediction stages. However, compression algorithms can significantly deteriorate the quality of data and in effect the classification models. In this article, an in-depth analysis of the influence of the tensor-based lossy data compression on the performance of the various deep neural architectures is presented. We show that the Tucker and the Tensor Train decomposition methods, with properly selected parameters, allow for very high compression ratios, while conveying enough information in the decompressed data to achieve only a negligible or very small drop in the accuracy. The measurements were performed on the popular deep neural architectures: AlexNet, ResNet, VGG, and MNASNet. We show that further augmentation of the tensor decompositions with the ZFP floating-point compression algorithm allows for finding optimal parameters and even higher compressions ratios at the same recognition accuracy. Our experiments show data compressions of 94%-97% that result in less than 1% accuracy drop.**

I. Introduction

The tendency of generating huge volumes of data dynamically increases. In this light data compression methods are of high importance. This is also true in the ML/AI domain, where modern deep neural architectures require larger training data sets. On the other hand, it has been shown that tensor decomposition methods allow to achieve huge compression ratios on multidimensional data [1]–[4]. Not surprisingly then that the tensor decompositions have been applied to compress weights of the deep neural networks [5], [6]. There are also works focusing on problem of learning from compact representations of images [7], [8]. However, to the best of our knowledge, there are no studies on the influence of the training data compression with tensor decompositions on the accuracy of the deep neural networks. This paper fills this gap providing an in-depth analysis of the Tucker and Tensor Train (TT) decomposition based data compression methods on the performance of the common deep network architectures such as AlexNet, ResNet, VGG, and MNASNet. The networks are trained in different scenarios, and due to the batch processing not all data need to be decompressed at the same time. Furthermore, we propose an additional step of data compression based on the ZFP floating-point lossy compression method [9]. Our experimental results show that the properly setup tensor decompositions followed by the ZFP module allow for as high as 94%-97% data compression ratios with less than 1% drop in accuracy of the deep neural networks.

The rest of the paper is organized as follows. Section II describes the related works. In Section III the two tensor decomposition methods used in our experiments are discussed. Section IV presents the ZFP floating-point compression method. In Section V neural network models utilized during experiments are described. Our proposed tensor-based training method is explained in Section VI. Experimental results with a discussion on results are described in Section VII. Finally, Section VIII concludes the paper.

II. Related Works

Compression methods gained much attention over the recent few decades. Since the seminal works of Lempel & Ziv in the lossless compression [10], much more diversed methods have been proposed in the lossy compression domain. Soon it was realized that various matrix, such as the well-known SVD one, and tensor decompositions can lead to significant data reductions [11]. However, the latter depends on many parameters, such as the tensor rank [2], [12]. In the case of multidimensional data, such as images, video, etc., the tensor-based approach offers much more possibilities, as will be discussed. In the era of deep neural networks, tensor-based methods proved to be superior in compressing their weights. However, it is also possible to compress their training and testing data - in this paper, we explore this branch.

Balle et al. proposed an image compression method consisting of nonlinear transformations for analysis and signal synthesis [13]. Three stages of convolutional linear filters with nonlinear activation functions were used to create both transforms. It achieved better rate-distortion performance than the standard JPEG and JPEG 2000 compression methods.

The tensor approach was explored by Zhang et al. [14]. The hyperspectral images were stacked into 3D tensors in which

spatial-spectral structure is preserved. The data - approximated and stored in projection matrices - achieved a high compression ratio with a low value of introduced artifacts.

Aidini et al. used the CANDECOMP/PARAFAC tensor decomposition to the compression method [15]. The multispectral image time series was expressed as a linear combination of the learned tensors and the quantization of the coefficients using the learned encoding dictionary.

Watkins and Sayeh proposed deep neural networks based method for gray image compression. The network is based on the autoencoder structured network, capable of both compressing and decompressing images [16] with a high compression ratio.

A multispectral image compression method based on the convolutional neural network was proposed by Li and Liu [5]. The processing path consists of the encoder and decoder parts. Both parts use CNN in combination with discrete cosine transform and nonnegative tensor decomposition (NDT) in pseudo-autoencoder structure. The method shows improved computational efficiency with comparable PSNR values compared to direct NDT in the wavelet domain.

Friedland et al. [17] analyze an impact of artifacts from perceptual compression on deep learning, concluding that classification accuracy is tightly connected to compression rate and data quantization. The loss of classification efficiency was mainly due to artifacts introduced during the compression process.

Similarly, in the PhD thesis on the impact of standard image compression techniques on CNN performance, Dejan concluded that network trained on JPEG encoded images partially relies on artifacts introduced by the compression [18].

In the paper, the authors do not use weight compression [6], [19], [20]. This topic is an additional option for further saving space in neural networks and can be implemented in future works. In recent years, tensors gained much attention from the ML/AI community, also in the context of data decompositions for compression [2]–[4], [12]. However, there is no work to analyze the influence of tensor-based compression of the training and testing data on the performance of the deep neural networks. We fill this gap, starting in the next section with a brief introduction to tensors.

III. Tensor Compression

Tensors are mathematical objects which can be regarded as multidimensional arrays of data, in which each separate dimension corresponds to a different degree of freedom of a measurement. Such an approach provides tools that extend the classical matrix analysis and which can take into account correlations hidden in data, to yield better results in various applications, such as compression or filtering [2].

Tensors extend the notion of vectors and matrices into higher-dimensional objects [2], [3], [21]–[23]. As discussed below, they allow for better representation and processing of the multidimensional signals and, in effect, also for higher compression ratios.

The multidimensional compression can be based on the following tensor product

$$\hat{\mathcal{T}} = \mathcal{T} \times_1 \mathbf{F}_1 \times_2 \mathbf{F}_2 ... \times_P \mathbf{F}_P \tag{1}$$

where decompressed version of input tensor $\mathcal{T}$ is denoted as $\hat{\mathcal{T}}$, whereas $\mathbf{F}_i$ is the ith factor matrix. The key idea is that the set of factor matrices and their product with $\mathcal{T}$, from the right side of (1), require much less storage space than the original tensor $\mathcal{T}$, while its recovered version $\hat{\mathcal{T}}$ is close enough in terms of the chosen norm, allowing e.g. for proper CNN training, as will be discussed. Also in the above equation, the k-th modal product $\mathcal{T} \times_k \mathbf{M}$ of a tensor $\mathcal{T} \in \Re^{N_1 \times N_2 \times ... \times N_P}$ and a matrix $\mathbf{M} \in \Re^{Q \times N_k}$ is used. The result is also a tensor $\mathcal{S} \in \Re^{N_1 \times N_2 \times ... N_{k-1} \times Q \times N_{k+1} \times ... \times N_P}$, whose elements are expressed as follows:

$$\mathcal{S}_{n_1 n_2 ... n_{k-1} q n_{k+1} ... n_P} = (\mathcal{T} \times_k \mathbf{M})_{n_1 n_2 ... n_{k-1} q n_{k+1} ... n_P} = \sum_{n_k=1}^{N_k} t_{n_1 n_2 ... n_{k-1} q n_{k+1} ... n_P} m_{q n_k} \tag{2}$$

As shown below, the compression matrices $\mathbf{F}_i$, called factors, can be obtained using the Tucker decomposition of tensors [12]. Thanks to the proper selection of the ranks of the tensor decomposition factors, decomposition usually well separates useful signal from its high-frequency components while taking multidimensional characteristics of the signal into account. The decomposition procedure of the tensor $\mathcal{T}$ is done by calculation of an approximating tensor $\hat{\mathcal{T}}$ that is close to the input tensor in terms of the Frobenius norm. Hence, a minimization function is defined as follows

$$\Theta(\hat{\mathcal{T}}) = ||\hat{\mathcal{T}} - \mathcal{T}||_F^2 \tag{3}$$

A. Tucker-based methods

The concept of the Tucker decomposition of a 3D tensor is presented in Figure 1. Assuming that the approximating tensor $\hat{\mathcal{T}}$ contains the same amount of useful information as the original tensor $\mathcal{T}$, it can be expressed as follows

$$\hat{\mathcal{T}} = \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 ... \times_P \mathbf{S}_P \tag{4}$$

where $\mathcal{Z} \in \Re^{R_1 \times R_2 \times ... \times R_P}$ is a core tensor and $\mathbf{S}_i \in \Re^{N_i \times R_i}$ are the so-called mode matrices. Using algebraic operations, from Equation (4), the formula for the core tensor is obtained:

$$\mathcal{Z} = \hat{\mathcal{T}} \times_1 \mathbf{S}_1^T \times_2 \mathbf{S}_2^T \times_2 ... \times_P \mathbf{S}_P^T \tag{5}$$

Combining Equation (5) with Equations (3) and (4) yields

$$\Theta(\hat{\mathcal{T}}) = ||\hat{\mathcal{T}} - \mathcal{T} \prod_{k=1}^{P} \times_k (\mathbf{S}_k \mathbf{S}_k^T)||_F^2 \tag{6}$$

The Tucker decomposition in Equation (6) reads that a tensor $\mathcal{T}$ is approximated by its projection onto space spanned by the matrices $\mathbf{S}_k$. To compute the series of $\mathbf{S}_k$ matrices, the alternating method can be used [2], [21], [24]–[26]. Also, let's observe that $\mathbf{S}_i \mathbf{S}_i^T$ in (6) is equivalent to the factor matrix $\mathbf{F}_i$ from (1).

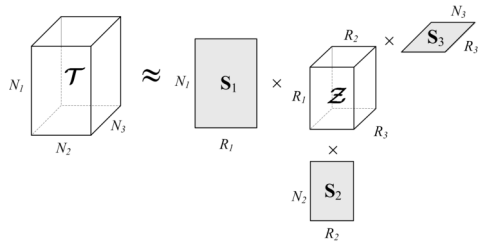


Fig. 1. Visualization of the Tucker decomposition of a 3D tensor.

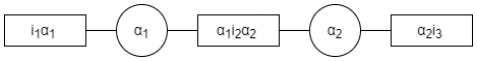


Fig. 2. Visualization of the Tensor Train network.

The approximation in Equation (4) includes only the components conveying the majority of energy available in the signal. However, to ensure the best quality, the estimation of the proper ranks R_1, R_2, and R_3 of the mode matrices S_i is necessary. Although fixed values can be used as a first approximation, the proper ranks need to be based on the signal content in real dynamic systems with unpredictable noise values. Multiple methods were presented to help solve this problem, i.e., Muti and Bourennane [22] using the minimum description length parameter (MDL) computed for each dimension separately, allowing optimal rank selection.

The Tucker format is stable, but its computational complexity grows exponentially with input tensor dimensions. It makes the method suitable only for "small" dimensions [27], [28].

B. Tensor-Train

To mitigate the computational complexity problem, a D-th order tensor $\mathcal{T} \in \Re^{N_1 \times N_2 \times \ldots \times N_D}$ can be represented as $\mathcal{T}(j_1, j_2, ..., j_d) = \mathbf{G}_1[j_1]\mathbf{G}_2[j_2]...\mathbf{G}_d[j_d]$, where $\mathbf{G}_k[j_k]$ is factor matrix with size of $r_{k-1} \times r_k$ $r_0 = r_d = 1$ and $j_k \in \{1, 2, ..., n_k\}$ $(k \in 1, 2, ..., d)$ [45].. The set of r_k is collectively called Tensor Train ranks. The $\mathbf{G}_k[j_k]$ belonging to the same mode can be stacked into 3-rd order core tensor $\mathcal{G}_k \in \Re^{N_1 \times r_{k-1} \times r_k}$ allowing the $\mathcal{T}$ be represented as follows:

$$\mathcal{T} = \mathcal{G}_1 \times^1 \mathcal{G}_2 \times^1 ... \times^1 \mathcal{G}_d \tag{7}$$

where $\times^1$ is called mode-(N,1) contracted product, presented here [29].

The decomposition can be presented graphically by the linear tensor network [30], [31], illustrated in Figure 2. There are two types of nodes: rectangular and circular. Rectangular contains spatial indices (i_k), some auxiliary indices (α_k), and a tensor with these indices associated with such nodes. On the other hand, a circular node is a link and contains only the auxiliary indices. If an auxiliary index is present in two cores, the cores are connected. To evaluate an input tensor, all tensor in rectangles need to be multiplied, and then summation is performed over all auxiliary indices.

Compared to Tucker Decomposition, the Tensor Train format has lower spatial complexity, making it more computationally efficient for tensors with larger dimensions [32].

IV. Fixed-Rate Compressed Floating-Point Arrays

The need for floating-point array compression is expressed by multiple lossy and lossless compression algorithms developed throughout the years. The most widely spread are image compression methods allowing encoding 2D and 3D arrays. For instance, PNG and JPEG-LS use linear prediction; JPEG - the block transform coding; JPEG2000 relies on the higher-order wavelets.

The Fixed-Rate Compressed Floating-Point Arrays (ZFP) compression scheme is based on ideas developed to compress 2D image data efficiently [33]. The input 3D array is divided into small, fixed-size blocks of dimensions 4 x 4 x 4, stored using a user-specified number of bits, which can be accessed independently. The method compresses the block performing the following steps:

1) Align the values in a block to a common exponent;
2) Convert the floating-point values in a block to a common exponent;
3) Convert the floating-point values to a fixed-point representation;
4) Apply an orthogonal block transform to decorrelate the values;
5) Order the transform coefficients by the expected magnitude;
6) Encode the resulting coefficients one "bit plane" at a time

The conversion to fixed-point is done by expressing each block value with respect to the largest floating-point exponent in a block, which is stored uncompressed resulting in normalized values in the range (-1, +1).

The prepared values are transformed to a basis allowing the spatially correlated values to be mostly decorrelated, as this results in many near-zero coefficients that can be compressed efficiently.

A separable orthogonal transform in d dimensions is employed to take advantage of regularly gridded data, resulting in a basis that is the tensor product of 1D basis functions. The proposed transform, due to coefficient selection, replaces divisions and multiplications into bitshifts. This choice achieves near-optimal results in terms of decorrelation efficiency and coding gain and is very efficient from a computational perspective. Further details of the ZFP method can be accessed from [33].

V. Neural Network Models

The advent of modern ML/AI methods, especially deep neural networks (DNN), resulted in a real IT revolution [34]. In a very short time, people realized real power in these systems that can be directly trained from data with marvelous results. Hence, the term "data" gained even more importance. With these came the eruption of modern deep neural network architectures, such as AlexNet [35], VGG [36], ResNet [37],

and dozens of their derivatives [38]–[41]. These were possible thanks to novel scientific achievements such as a solution to the vanishing gradient problem in very deep networks, optimization algorithms, and thanks to the availability of the efficient general-purpose graphics processing units (GGPU).

For example, in the computer vision domain, neural networks dominate in majority of tasks, such as object detection & recognition, segmentation, filtering, to name a few. However, performance of the neural networks depends heavily on availability of the high quality labeled data. However, this can be jeopardized by many factors, such as compression-decompression processes, we focus upon in this work.

To examine the impact of the proposed compression method, state-of-the-art models capable of high accuracy with reasonable low training time were used. Their architectures are shortly described below.

AlexNet

The neural network contains eight learned layers [35] - five convolutional and three fully connected. It introduces such features as ReLU nonlinearity, the capability of training on multiple GPUs, local response normalization, and overlapping pooling. The new approach combined with a high emphasis on overfitting reduction results in a highly accurate model, even today. The network is recognized as a milestone, and solutions proposed in the article are now considered as standard by AI/ML community.

ResNet

The ResNet networks implements skip connections with ReLU nonlinearity and batch normalization [41]. It allows to mitigate vanishing gradient problem and speeds up the training process, which positively impacts the feasibility of training deeper neural networks.

VGG

The model uses small receptive fields, decreasing the number of trainable parameters and increasing the ReLU unit count [36]. Such an approach makes the decision function more discriminative, which in turn increases overall network performance.

MNASNet

It is a neural architecture for mobile devices [42]. It bases on Factorized Hierarchical Search Space approach, which balances the diversity of layers and the size of total search space. The resulting network generally runs faster and uses less computational power.

VI. Tensor-based Data Processing

The main goal of our approach is to decrease data size with the lossy compression process without a significant impact on the quality of object prediction by the benchmark neural-network architectures. The proposed method requires an input in a tensor form. Based on images selected to process, the width (W) and height (H) parameters, describing tensor dimensions, are selected based on the biggest image in a set. The last parameter depth (D) depends only on the number of input images. The input tensor is assembled utilizing the method shown in Algorithm 1 and produces $W \times H \times D$ data chunk used for further processing.

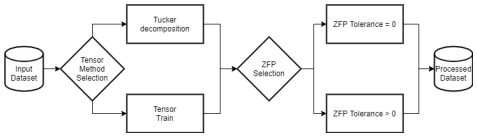


Fig. 3. Block diagram of proposed method's compression step.

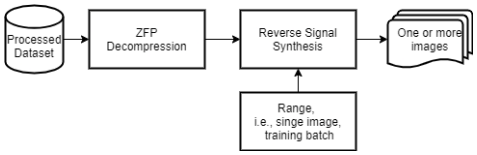


Fig. 4. Block diagram of proposed method's decompression step.

The next step is a tensor decomposition, as presented in Figure 3. Depending on the selected method, the result is a set of mode matrices (Tensor Train) or core tensor and mode matrices (Tucker decomposition). These approximated signals contain the most relevant information, and higher frequency components are removed. Smaller rank values selected for the method translate to higher compression and more significant high-frequency attenuation. In the next step, obtained results are compressed with the ZFP algorithm using a tolerance argument ZFP_t that controls the compression quality. Such compressed data objects can be stored on a disk for further use or be utilized as an intermediate step in real-time processing.

Before the next step, processed data needs to be decompressed, as presented in the Figure 4. First, using the ZFP decompression method, then in the reverse signal synthesis process, the aforementioned matrices are merged into the result tensor. However, the reverse signal synthesis can be calculated for the entire tensor, a single image, or a set of consecutive images. Such flexibility allows decrease computational requirements and allows the method to be used as part of modern neural-network training architecture.

VII. Experimental Results

The presented method was implemented in Python, using the NumPy, SciPy, and scikit-image packages. TensorLy library was used for tensor decomposition [43], and ZFP library for the additional compression of multidimensional floating-point arrays [33]. As the benchmark, the following neural network architectures were selected: AlexNet, ResNet-18, ResNet-34, VGG-11, VGG-13, and MNASNet0.5.

Presented Experiments were performed on a server computer, equipped with 256 GB of RAM, 64-core processor AMD Ryzen Threadripper 3990X with the 2.9 GHz base clock, and 64-bit Ubuntu 20.04.2 LTS OS.

The quantitative results were measured in terms of compression ratio (Cr) and object detection accuracy of neural

Algorithm 1 Tensor assembler.

```
T = ∅
while I ≠ ∅ do
    load original image I_i and prepare container for resized image I_p
    calculate x and y offsets needed to place I_i in the center of I_p
    resize selected image from I to dimensions specified in W and H, keeping aspect ratio of original data and save it to I_p
    append I_p to T
    remove I_i from I
end while
return T
```

network (Acc). The compression ratio is defined as follows:

$$Cr = \frac{\text{Uncompressed data size}}{\text{Compressed data size}} \tag{8}$$

Furthermore, an object detection accuracy of a neural network is described as the proportion of correct predictions over the total examined cases:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP, TN, FP, FN represents true positive, true negative, false positive, and false negative values achieved by the network model, on the tested dataset, respectively.

For the quantitative evaluation of the proposed method, Imagenette [44] was selected. The dataset is a subset of 10 easily classified classes from the Imagenet dataset and was selected to decrease the time needed for development and tests. The original structure contains train and validation sets with 9469 and 3925 images, respectively. For better quality estimation, the presented accuracy of tested neural networks is measured using a test set containing 1500 images. The before-mentioned set was separated from the original subsets using 900 images from the train set and 600 images from the validation set. The final image count for used image sets is as follows: training 8569 images, validation 3325 images, and test 1500.

The input data was processed using the tensor-based compression methods described in Section VI. The collection of test datasets was created depending on selected parameters and methods, as shown in Table I.

To cover a broad spectrum of possible utilization of the proposed method, it was decided to check an impact of tensor based data compression/decompression on networks depending on the learning process. Both random weight initialization and the transfer learning technique were used. For the latter, models were pre-trained on the Imagenet dataset.

The training process was conducted for each of the prepared datasets and each of the selected benchmark neural networks, and results were measured. All prepared data plots from Figure 5 to Figure 19 are presented below.

Denoting dimensions of the input tensor as $W \times H \times D$, values of the corresponding ranks for tensor methods were calculated as $[0.5W, 0.5H, D]$ for Tucker decomposition. Similarly, in the case of the tensor train method, dimensions of the factor matrices were set to $[1, 0.5min(W, H), 0.5min(W, H), 1]$. Obtained compression ratios for different tensor methods and values of ZFP tolerance are presented in Table II. Results of the neural network accuracy are presented in Table III for the random weight initialization and in Table IV for the transfer learning, respectively. The results are discussed below.

For the transfer learning technique, it can be observed that TT achieved better accuracy for the selected ranks. Depending on the tested neural network, it achieves 0.6 - 2.5% worse results compared to the original dataset. At the same time, the TT method provides users with a lower compression ratio, between 14.92 and 16.86 for lossless and lossy ZFP compression, respectively. On the other hand, the Best rank method achieves higher compression 20.41 - 21.39, with lower network accuracy however. In this case the difference between original data and tested datasets is higher, from 2.1% to 6.3%. Additionally, for nearly all cases, methods combining tensor-based algorithm with lossy ZFP yield better results than ZFP lossless ones. Depending on the tested neural network, the lossy ZFP version achieves 0.6 - 1.1% and 2.1 - 4.1% difference between original for the TT and Best rank methods, respectively.

In the random weight training case, again TT achieves better accuracy, with an accuracy loss between 3.3% - 7.2% depending on the tested neural network. The Best rank method yields 6.6% - 13.5% drop in accuracy. In the considered context, for nearly all tested cases, lossy ZFP compression does not change the accuracy achieved by tested neural networks.

Hence, the proposed method combines high compression capabilities with good retention of important signals for the detection process. For the most common neural network training technique used today - transfer learning - the method achieves results very close to the values obtained on the original dataset.

Higher compression rate impacts all tested networks' quality by removing high-frequency components from the images, which means lowering object detection accuracy by 0.6 - 1.1% and 2.1 - 4.1% for the Tensor train and Best rank methods, respectively. However, tensor methods inherently allow an easy change in the compression rate by selecting different ranks during the decomposition process, making it possible to find an acceptable trade-off for a given application.

VIII. Conclusion

In this paper, a novel framework and experiments on data compression/decompression in order to measure the impact on the deep neural network training and prediction are presented. The compression/decompression is based on tensor decomposition methods combined with the floating-point array compression. We show that the presented methods can achieve very high compression ratios while still preserving enough

TABLE I
DATASETS USED IN NETWORK QUALITY ASSESSMENT.

	Name	Tensor compression type	ZFP tolerance	Validation set compressed?
A	original	-	-	false
B	best_rank	Best rank	lossless	true
C	best_rank_comp	Best rank	1e-3	true
D	best_rank_orig_val	Best rank	lossless	false
E	best_rank_comp_orig_val	Best rank	1e-3	false
F	tensor_train	Tensor train	lossless	true
G	tensor_train_comp	Tensor train	1e-3	true
H	tensor_train_orig_val	Tensor train	lossless	false
I	tensor_train_comp_orig_val	Tensor train	1e-3	false

TABLE II
COMPRESSION RESULTS.

Tensor compression type	ZFP tolerance	Compression ratio (Space saving)	Complete processing time [h:mm:ss]
Best rank	lossless	20.41 (0.951)	2:46:02
Best rank	1e-3	21.39 (0.953)	2:47:09
Tensor train	lossless	14.92 (0.932)	1:27:38
Tensor train	1e-3	16.86 (0.941)	1:27:43

TABLE III
NETWORK ACCURACY VERSUS DATASET TYPE. TRAINING RESULTS FOR RANDOM WEIGHT INITIALIZATION. DATASETS: ORIGINAL (A), BEST_RANK (B), BEST_RANK_COMP (C), BEST_RANK_ORIG_VAL (D), BEST_RANK_COMP_ORIG_VAL (E), TENSOR_TRAIN (F), TENSOR_TRAIN_COMP (G), TENSOR_TRAIN_ORIG_VAL (H), TENSOR_TRAIN_COMP_ORIG_VAL (I)

	AlexNet	ResNet-18	ResNet-34	VGG-11	VGG-13	MNASNet0.5
A	0.7687	0.8025	0.7962	0.8018	0.7806	0.7389
B	0.6497	0.6683	0.6581	0.6838	0.6713	0.6033
C	0.6555	0.7292	0.7304	0.7139	0.6945	0.6454
D	0.6482	0.681	0.7083	0.6851	0.6566	0.6191
E	0.6662	0.693	0.6856	0.6823	0.6741	0.6321
F	0.7093	**0.7554**	0.7605	0.7422	0.7228	0.6652
G	0.7271	0.7521	0.7475	0.7427	**0.7338**	**0.681**
H	**0.7389**	0.7493	**0.7623**	0.7417	0.7335	0.6795
I	0.7284	0.7439	0.7475	**0.7483**	0.7264	0.6561

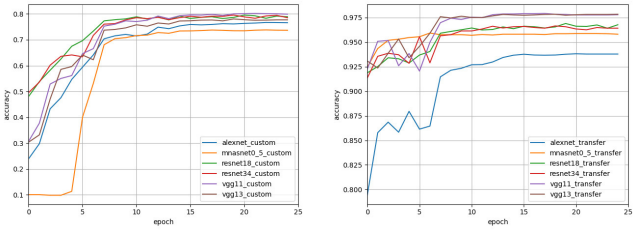
TABLE IV
NETWORK ACCURACY VERSUS DATASET TYPE. TRAINING RESULTS FOR TRANSFER LEARNING TECHNIQUE. DATASETS: ORIGINAL (A), BEST_RANK (B), BEST_RANK_COMP (C), BEST_RANK_ORIG_VAL (D), BEST_RANK_COMP_ORIG_VAL (E), TENSOR_TRAIN (F), TENSOR_TRAIN_COMP (G), TENSOR_TRAIN_ORIG_VAL (H), TENSOR_TRAIN_COMP_ORIG_VAL (I)

	AlexNet	ResNet-18	ResNet-34	VGG-11	VGG-13	MNASNet0.5
A	0.9381	0.9689	0.9664	0.9791	0.9783	0.9592
B	0.8943	0.9378	0.9434	0.9577	0.9595	0.9177
C	0.8961	0.9411	0.9434	0.9562	0.9575	0.8963
D	0.893	0.9394	0.9338	0.9572	0.9597	0.9185
E	0.8854	0.9378	0.9437	0.9572	0.9618	0.9141
F	0.9124	0.9549	0.9498	0.9654	0.9697	0.9335
G	**0.9302**	**0.961**	**0.959**	0.971	**0.972**	**0.9508**
H	0.9243	0.9575	0.9582	**0.9715**	0.9715	0.9501
I	0.9261	0.9585	0.9575	0.9674	0.9689	0.9503

significant information in data to achieve high accuracy of object detection in the neural models.

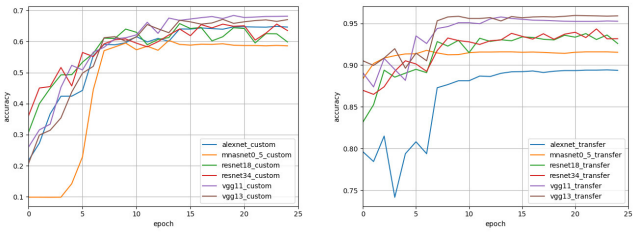
The utilized algorithms can smoothly change the achieved compression rate, which impacts network accuracy during the training process, allowing users to find parameters that are optimal for a given application.

Furthermore, storing data in the proposed form allows for a selective decompression of a single or a group of images without the need to decompress the entire tensor. The accuracy drop in respect to the original data is visible, but for the problems where storage or data transfer speed are important, it can increase performance both during training and normal operation. Alternatively, we can say that thanks to data compression a larger amount of data can be transferred and



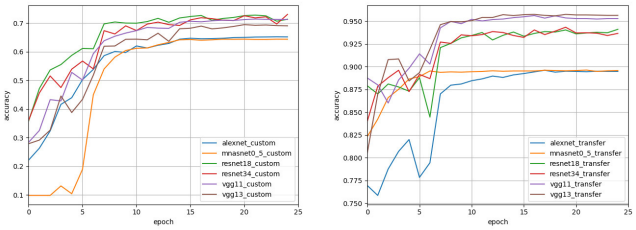
(a) Random initialization. (b) Transfer learning.

Fig. 5. Comparison between all used architectures trained on the original dataset.



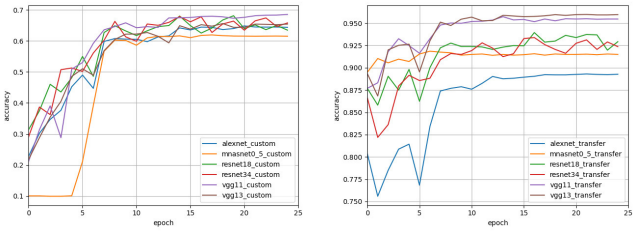
(a) Random initialization. (b) Transfer learning.

Fig. 6. Comparison between all used architectures trained on the dataset compressed with Tucker decomposition and lossless ZFP.



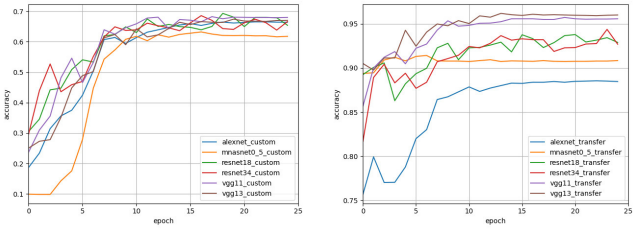
(a) Random initialization. (b) Transfer learning.

Fig. 7. Comparison between all used architectures trained on the dataset compressed with Tucker decomposition and lossy ZFP.



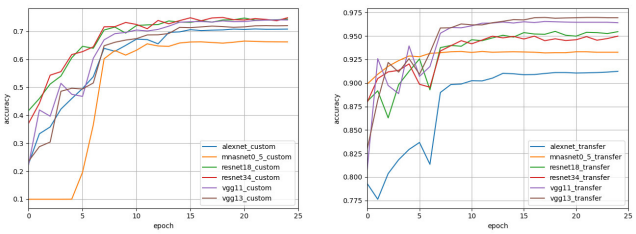
(a) Random initialization. (b) Transfer learning.

Fig. 8. Comparison between all used architectures trained on the dataset compressed with Tucker decomposition and lossless ZFP with original validation set.



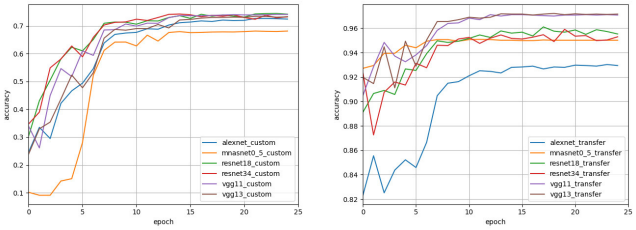
(a) Random initialization. (b) Transfer learning.

Fig. 9. Comparison between all used architectures trained on the dataset compressed with Tucker decomposition and lossy ZFP with original validation set.



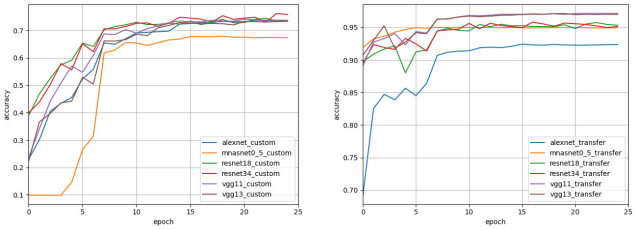
(a) Random initialization. (b) Transfer learning.

Fig. 10. Comparison between all used architectures trained on the dataset compressed with Tensor Train algorithm and lossless ZFP.



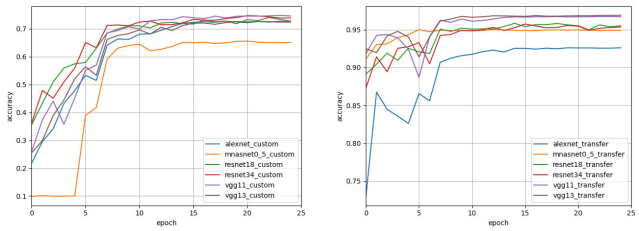
(a) Random initialization. (b) Transfer learning.

Fig. 11. Comparison between all used architectures trained on the dataset compressed with Tensor Train algorithm and lossy ZFP.



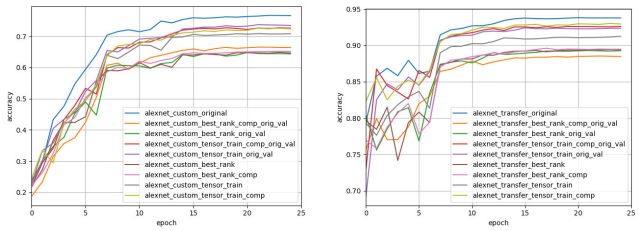
(a) Random initialization. (b) Transfer learning.

Fig. 12. Comparison between all used architectures trained on the dataset compressed with Tensor Train algorithm and lossless ZFP with original validation set.



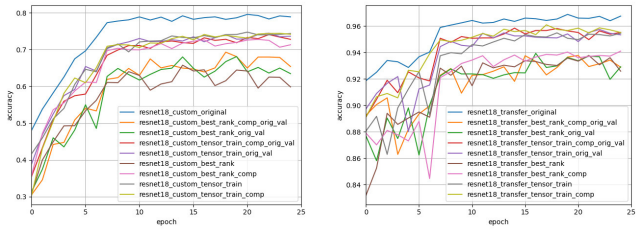
(a) Random initialization. (b) Transfer learning.

Fig. 13. Comparison between all used architectures trained on the dataset compressed with Tensor Train algorithm and lossy ZFP with original validation set.



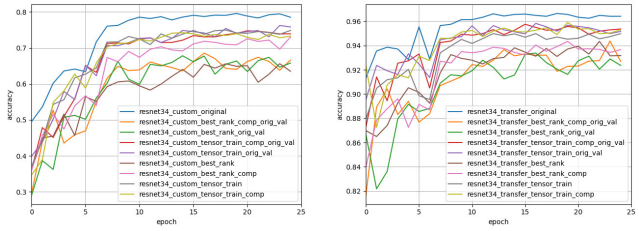
(a) Random initialization. (b) Transfer learning.

Fig. 14. AlexNet training results for each tested dataset.



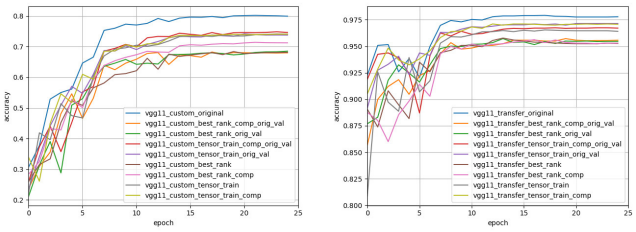
(a) Random initialization. (b) Transfer learning.

Fig. 15. ResNet-18 training results for each tested dataset.



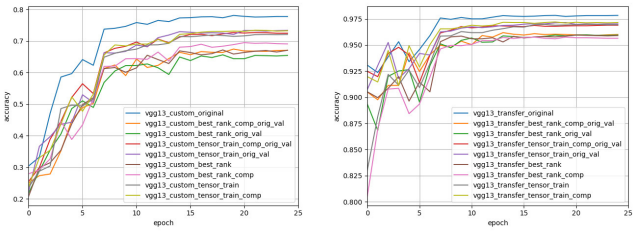
(a) Random initialization. (b) Transfer learning.

Fig. 16. ResNet-34 training results for each tested dataset.



(a) Random initialization. (b) Transfer learning.

Fig. 17. VGG-11 training results for each tested dataset.



(a) Random initialization. (b) Transfer learning.

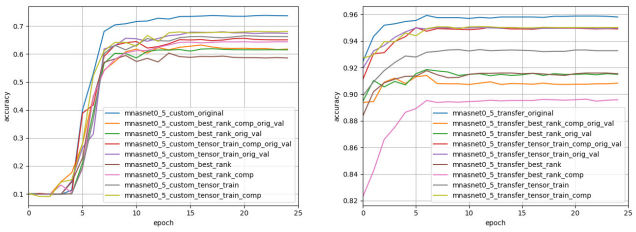
Fig. 18. VGG-13 training results for each tested dataset.

used for training. Also, although the method was developed for images, it can be useful for other data types, such as physical or industry measurements, etc.

Summarizing, the best results were achieved using the transfer learning technique, where a dataset is processed with Tensor Train decomposition paired with the lossy version of the ZFP algorithm. In this setting 0.6% - 0.7% drop in accuracy of the deep neural networks in respect to the original dataset was achieved for all tested methods. The best accuracy, both in respect to the original and processed dataset, was obtained with the VGG-11 and VGG-13 models.

ACKNOWLEDGMENT

This research was co-funded by Smart Growth Operational Programme 2014-2020, financed by European Regional Development Fund, in frame of project POIR.01.01.01-00-0570/19, operated by National Centre for Research and Development in Poland.



(a) Random initialization. (b) Transfer learning.

Fig. 19. MNASNet0.5 training results for each tested dataset.

REFERENCES

[1] COCOCCIONI, M., et al. Novel arithmetics in deep neural networks signal processing for autonomous driving: Challenges and opportunities. In *IEEE Signal Processing Magazine*, **2020**, 38.1: 97-110. doi: 10.1109/MSP.2020.2988436

[2] Cyganek, B. *Object Detection and Recognition in Digital Images: Theory and Practice*; John Wiley & Sons: New York, NY, USA, 2013. doi: 10.1002/9781118618387

[3] Kolda, T.; Bader, B. Tensor Decompositions and Applications. *SIAM Rev. 51.3* **2009**, *51*, 455–500. doi: 10.1137/07070111X

[4] Cyganek, B., Thumbnail Tensor—A Method for Multidimensional Data Streams Clustering with an Efficient Tensor Subspace Model in the Scale-Space, Sensors, 19(19), 4088, 2019, doi: 10.3390/s19194088

[5] LI, J., LIU, Z., Multispectral transforms using convolution neural networks for remote sensing multispectral image compression. In *Remote Sensing 11.7: 759*, **2019**. doi: 10.3390/rs11070759

[6] CHOI, Y., EL-KHAMY, M., LEE, J., Universal deep neural network compression. In *IEEE Journal of Selected Topics in Signal Processing, 14.4*, **2020**, pp. 715-726. doi: 10.1109/JSTSP.2020.2975903

[7] Przyborowski M., et al. Toward Machine Learning on Granulated Data – a Case of Compact Autoencoder-based Representations of Satellite Images. In *2018 IEEE International Conference on Big Data (Big Data)*, **2018**, pp. 2657-2662, doi: 10.1109/BigData.2018.8622562.

[8] WANG, N; YEUNG, D. Y., Learning a deep compact image representation for visual tracking. In *Advances in neural information processing systems*, **2013**

[9] Lindstrom, P., Fixed-Rate Compressed Floating-Point Arrays. In *IEEE Transactions on Visualization and Computer Graphics 20(12)* **2014**, pp. 2674-2683, doi:10.1109/TVCG.2014.2346458

[10] ZIV, J., LEMPEL, A., Compression of individual sequences via variable-rate coding. In *IEEE transactions on Information Theory*, **1978**, 24.5: 530-536. doi: 10.1109/TIT.1978.1055934

[11] Cyganek, B., A Framework for Data Representation, Processing, and Dimensionality Reduction with the Best-Rank Tensor Decomposition. Proceedings of the ITI 2012 34th International Conference Information Technology Interfaces, June 25-28, 2012, Cavtat, Croatia, pp. 325-330, doi:10.2498/iti.2012.0466, **2012**.

[12] De Lathauwer, L.; De Moor, B.; Vandewalle, J. On the best rank-1 and rank-(R1, R2,..., Rn) approximation of higher-order tensors. *Siam J. Matrix Anal. Appl.* **2000**, *21*, 1324–1342. doi: 10.1137/S0895479898346995

[13] BALLÉ, J., LAPARRA, V., SIMONCELLI, E. P., End-to-end optimized image compression. In *arXiv preprint arXiv:1611.01704*, **2016**.

[14] ZHANG, L., et al. Compression of hyperspectral remote sensing images by tensor approach. In *Neurocomputing, 147*, **2015**, pp. 358-363. doi: 10.1016/j.neucom.2014.06.052

[15] AIDINI, A., TSAGKATAKIS, G., TSAKALIDES, P., Compression of high-dimensional multispectral image time series using tensor decomposition learning. In: *2019 27th European Signal Processing Conference (EUSIPCO). IEEE*, **2019**. p. 1-5. doi: 10.23919/EUSIPCO.2019.8902838

[16] WATKINS, Y. Z., SAYEH, M. R., Image data compression and noisy channel error correction using deep neural network. In *Procedia Computer Science, 95*, **2016**, pp. 145-152. doi: 10.1016/j.procs.2016.09.305

[17] FRIEDLAND, G., et al. On the Impact of Perceptual Compression on Deep Learning. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, **2020**, p. 219-224. doi: 10.1109/MIPR49039.2020.00052

[18] DEJEAN-SERVIÈRES, M., et al. Study of the impact of standard image compression techniques on performance of image classification with a convolutional neural network. **2017**. PhD Thesis. INSA Rennes; Univ Rennes; IETR; Institut Pascal.

[19] ULLRICH, K., MEEDS, E., WELLING, M., Soft weight-sharing for neural network compression. In *arXiv preprint arXiv:1702.04008*, **2017**.

[20] JIN, S. et al. DeepSZ: A novel framework to compress deep neural networks by using error-bounded lossy compression. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, **2019** p. 159-170. doi: 10.1145/3307681.3326608

[21] DENG, Lei, et al. Model compression and hardware acceleration for neural networks: A comprehensive survey. In *Proceedings of the IEEE*, **2020**, 108.4: 485-532. doi: 10.1109/JPROC.2020.2976475

[22] Muti, D.; Bourennane, S. Multidimensional filtering based on a tensor approach. *Signal Process.* **2005**, *85*, 2338–2353. doi: 10.1016/j.sigpro.2004.11.029

[23] Cyganek, B.; Smołka, B. Real-time framework for tensor-based image enhancement for object classification. *Proc. SPIE* **2016**, *9897*, 98970Q. doi: 10.1117/12.2227797

[24] Cyganek, B.; Krawczyk, B.; Wozniak, M. Multidimensional Data Classification with Chordal Distance Based Kernel and Support Vector Machines. *Eng. Appl. Artif. Intell.* **2015**, *46*, 10–22. doi: 10.1016/j.engappai.2015.08.001

[25] Cyganek, B.; Wozniak, M. Tensor-Based Shot Boundary Detection in Video Streams. *New Gener. Comput.* **2017**, *35*, 311–340. doi: 10.1007/s00354-017-0024-0

[26] Marot, J.; Fossati, C.; Bourennane, S. Fast subspace-based tensor data filtering. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 3869–3872. doi: 10.1109/ICIP.2009.5414048

[27] Khoromskij, B. N., Khoromskaia, V., Multigrid accelerated tensor approximation of function related multidimensional arrays. In *SIAM J. Sci. Comput., 31*, **2009**, pp. 3002–3026. doi: 10.1137/080730408

[28] Oseledets, I. V., Savostianov, D. V., Tyrtyshnikov, E. E., Tucker dimensionality reduction of three-dimensional arrays in linear time. In *SIAM J. Matrix Anal. Appl., 30*, **2008**, pp. 939–956. doi: 10.1137/060655894

[29] Lee, N., Cichocki, A., Fundamental tensor operations for large-scale data analysis using tensor network formats. In *Multidimensional Syst. Signal Process., vol. 29, no. 3*, **2017**, pp. 921–960 doi: 10.1007/s11045-017-0481-0

[30] Hubener, R., Nebendahl, V., Dur, W., Concatenated tensor network states. In *New J. Phys., 12*, **2010**, 025004. doi: 10.1088/1367-2630/12/2/025004

[31] Van Loan, C. F., Tensor network computations in quantum chemistry Technical report, available online at www.cs.cornell.edu/cv/OtherPdf/ZeuthenCVL.pdf, **2008**.

[32] Oseledets, I., Tensor-Train Decomposition. In *SIAM J. Scientific Computing. 33.*, **2011**, pp. 2295-2317. doi: 10.1137/090752286.

[33] Lindstrom, P., Fixed-Rate Compressed Floating-Point Arrays. In *IEEE Transactions on Visualization and Computer Graphics vol. 20*; **2014**, doi :10.1109/TVCG.2014.2346458.

[34] Lemley, J., Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision. In *IEEE Consumer Electronics Magazine vol. 6, Iss. 2*; **2017** doi: 10.1109/MCE.2016.2640698

[35] Krizhevsky, A., Sutskever, I., Hinton, G. E., ImageNet classification with deep convolutional neural networks. In *Communications of the ACM. 60 (6)* pp. 84–90. doi: 10.1145/3065386

[36] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*. **2014**

[37] He, Kaiming, et al. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* **2016**, pp. 770-778 doi: 10.1109/CVPR.2016.90

[38] Krizhevsky, A., et al., ImageNet classification with deep convolutional neural networks. In *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 1.*, Red Hook, NY, USA: Curran Associates, **2012**, pp. 1097–1105. doi: 10.1145/3065386

[39] Simonyan K. and Zisserman A., Very deep convolutional networks for large-scale image recognition. In *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., **2015**, pp. 1–14.

[40] Xie S., et al., Aggregated residual transformations for deep neural networks.In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, **2017**, pp. 5987–5995. doi: 10.1109/CVPR.2017.634

[41] Szegedy, S. et al., Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, S. P. Singh and S. Markovitch, Eds., **2017**, pp. 4278–4284.

[42] Tan, M., et al. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2019**, pp. 2820-2828 doi: 10.1109/CVPR.2019.00293

[43] Kossaifi, J.; Panagakis, Y.; Kumar, A.; Pantic, M. TensorLy: Tensor Learning in Python. *arXiv preprint* **2018**, arXiv:1610.09555.

[44] Howard, J., imagenette dataset, https://github.com/fastai/imagenette/

[45] Oseledets, I. V., Tensor-train decomposition. In *SIAM J. Sci. Comput., vol. 33, no. 5*, **2011**, pp. 2295–2317 doi: 10.1137/090752286

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 13–21
DOI: 10.15439/2021F129
ISSN 2300-5963 ACSIS, Vol. 26

Impact of time series clustering on fuel sales prediction results

Joanna Henzel*, Marek Sikora‖
Department of Computer Networks and System
Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: *joanna.henzel@polsl.pl, ‖marek.sikora@polsl.pl

Jakub Bularz
FuelPrime Department
AIUT Ltd.
Wyczółkowskiego 113, 44-109 Gliwice, Poland
Email: jakub.bularz@aiut.com

***Abstract*—The purpose of the paper is to check the impact of data clustering in the process of predicting demand. We checked different ways of adding information about similar datasets to the forecasting process and we grouped the measurements in multiple ways. The experiments were executed on 50 time series describing fuels sales (gasoline and diesel sales) on 25 petrol stations from an international company. We described the data preparation process and feature extraction process. In the 9 presented experiments, we used the XGBoost algorithm and some typical time series forecasting methods (ARIMA, moving average). We showed a case study for two datasets and we discussed the practical usage of the tested solutions. The results showed that the solution which used XGBoost model utilising data gathered from all available petrol stations, in general, worked the best and it outperformed more advanced approaches as well as typical time series methods.**

I. Introduction

THE proper forecasting of sale is a crucial element in many sectors, also in the retail industry. This plays an important role in the competitiveness of any company. Accurate forecasts can help to create proper planning of demand and deliveries, which can lead to cost reductions by the company. In the paper, we focus on the petrol retail sector and the problem of forecasting fuel sales from multiple petrol stations.

A lot of research was done for forecasting sale in multiple retail industries with the use of Machine Learning methods, however not so much was done in this field regarding the fuel sales short-term predictions. Also, the usefulness of clustering similar petrol stations based on historical data and similarities of time series was not tested in this area.

The objective of this paper is to check the usefulness of data clustering in the process of predicting fuel sales. In our research, we inspected different ways of adding information about similar datasets to the forecasting process and we grouped the measurements in multiple ways. We also checked how adding the information about predicted sales for similar datasets can affect the results obtained for a single dataset. We hypothesised that using information about the historical sales from different petrol stations and the sales predictions for them can improve sales predictions results for the dataset.

The paper is organised as follows: the next section provides the review of the literature and related works, section III describes the data preparation process. Afterwards, the experiments and results are presented, followed by a more detailed case study. The paper ends with some conclusions and a discussion of the results.

II. Related Works

Traditionally, forecasting was made using statistical methods. These were exponential smoothing [1], moving average, the Auto-regressive Integrated Moving Average (ARIMA) model and the Seasonal Auto-regressive Integrated Moving Average (SARIMA) model.

More complex methods were also used for the task of sales forecasting. In [2] a comparison of various linear and non-linear models for the sales forecasting task was conducted and the results suggest that non-linear models should be highly considered when dealing with modelling retail sales. In the paper [3] a comparison of different Machine Learning Techniques was conducted regarding sales forecasting of retail stores. The authors concluded that boosting algorithms gave better results than the regular regression ones. For them, the best results were obtained for the GradientBoost algorithm and the XGBoost [4] implementation has been used in order to increase the accuracy. The successful usage of the XGBoost algorithm for sales forecasting was also presented in [5]. The authors in the paper extracted features based on the historical sales and then they trained one model using XGBoost. The proposed model performed extremely well for sales prediction with less computing time and memory resources. The paper [6] presented a framework based on Facebook's Prophet algorithm and backtesting strategy for forecasting future sales in the retail industry.

In the field of sales forecasting a lot of different neural network and Deep Learning approaches were tested e.g. evolutionary neural networks (ENN) [7], the Dynamic Artificial Neural Network [8] and the backward propagation neural network [9]. In the paper [10] a case study for using long short-term memory (LSTM) recurrent neural networks (RNN) was proposed. Deep Bi-Directional LSTM Networks was presented in [11]. The authors of the paper [12] proposed a novel forecasting method that combined the deep learning method – LSTM – and random forest (RF) for demand forecasting problems. They compared the results with neural networks, multiple regression, ARIMAX, LSTM networks, and RF

method and their proposition was statistically significantly better.

Fuel sales forecasting was not considered in many scientific articles. Most of them were focused on the marketing research area and not machine learning models strategies. The paper [13] described the problem of forecasting fuels sales but from a perspective of long-term forecasts - for a few years ahead. The authors of the paper [14] focused more on the attributes describing the petrol stations and their surroundings and the models were developed for planning purposes in order to assess potential sales of new sites. The use of judgmental forecasting for predicting fuels sales on petrol stations was presented in the paper [15]. The authors used experts' knowledge to create forecasts. They used PROLOG in order to map the knowledge into production rules, thus enabling computer processing. The paper [16] showed the solution for real-time petrol sales forecasting using dilated causal convolutional neural network (CNNs). To the best of our knowledge, the tree boosting algorithm, especially the extreme gradient boosting (XGBoost) algorithm, has not been used to forecast fuels sales.

Clustering methods are widely used for forecasting purposes. In the paper [17] a k-means clustering algorithm was used along with decision trees, artificial neural networks and support vector machines methods. The author of the paper [18] described using k-means clustering and ARIMA model for forecasting electricity load. In [19], a hybrid sales forecasting system based on clustering and decision trees was presented. Clustering of time series data was broadly discussed in [20]. In this paper the clustering algorithms were divided into six groups: Partitioning, Hierarchical, Grid-based, Model-based and Density-based. Predicting based on similarities was also mentioned in [21]. The authors applied the fuzzy network of comparators (NoC) for the problem of ovulation date prediction.

III. Data preparation

In the experiments, we used data from 25 petrol stations located in different parts of Poland. All of them belong to one international company and are part of a network of petrol stations. In the research, we focused on forecasting the sale of two types of regularly sold fuels: Gasoline and Diesel. We did not consider different types of fuels e.g. Premium Gasoline or Premium Diesel which are more expensive than the regular ones. We were making experiments for 2 types of fuels from 25 petrol stations, so finally, we were working with 50 tabular datasets, which described the time series of fuel sales. The datasets covered a 3-year period from 2017-01-01 to 2019-12-31.

The raw data derived from stations contained 5 columns:

- *meter* - ID of a fuel tank.
- *start_time* - Start of a considered period of time; timestamp from which fuel sale in a given period is counted.
- *end_time* - End of a considered period of time; timestamp to which fuel sale in a given period is counted.

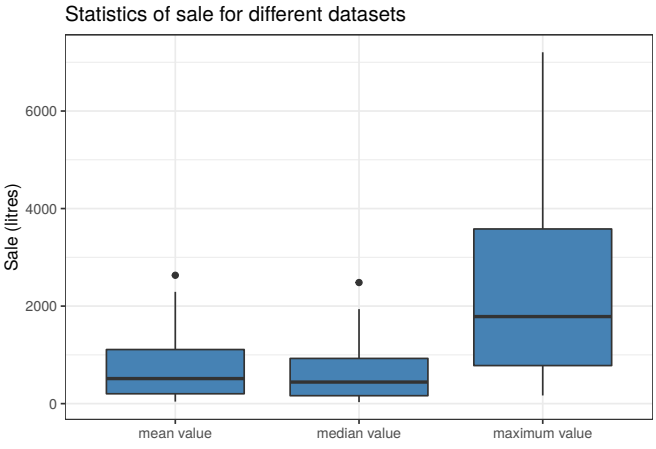


Fig. 1. A mean, median and maximum value of sales from datasets with 4-hour intervals.

- *sales* - Total fuel sales (in litres) in a given period of time.
- *temp* - Fuel temperature.

Most of the periods (*end_time* - *start_time*) last for about 15 minutes, but the periods were not exactly equal. The median time of periods (n = 5,062,511) was equal to 15.27 minutes, the mean value of periods was 16.11 minutes $\pm$ 10.42 minutes. The datasets also contained missing periods, i.e. the *end_time* and the following row *start_time* were not always equal.

In the first step of the data preparation process, we decided to prepare datasets in a way that they will contain equal periods of time. We decided to aggregate sales into 4-hour intervals, because experts gave us the information that forecasting for this amount of time is a reasonable approach. Forecasting sale for every 15 minutes does not have big practical usage but forecasting for bigger periods of time can be used e.g. for demand planning.

The sales value for a new interval was calculated as the sum of the sales from all periods that were entirely contained within a given 4-hour interval. To this was added the proportional value of sales from periods that only partially fell within the interval.

Basic statistics of sales from all of 50 datasets are presented in figure 1. They were calculated after creating modified datasets and cleaning the data. The minimum value was not presented because it is equal to 0 in each dataset.

A. Attributes

Each of the fifty time series of fuel sales was presented as a separate tabular dataset with 5 columns. In our experiments, we wanted to consider not only typical time series forecasting methods but also method used mostly for tabular datasets. Because of this, a crucial step was to create proper attributes which will describe the sale in each created interval. We decided to prepare attributes that could be divided into two categories:

- attributes based on time column and
- attributes based on previous values of sales.

1) Attributes from date: Each row of the datasets described one interval. After the first step of the data preparation process, all intervals were equal so in the next steps we used only one time column - in our case we used *end_time* column. Attributes derived from the time column were created in order to describe the time of sale – for example, different fuel sales can be between midnight and 4 am and different between 12 am and 4 pm. Also, weekday, public holidays, vacations etc. can have an impact on fuel sales at petrol stations.

The list of proposed time attributes created based on *end_time* column is as follow:

- basic attributes from time column: year number, quarter number, month number, day number, number of a day in the year, week number, hour, weekday, season, logical attribute describing the weekend;
- logical attributes describing the school holidays: information on whether any school holidays are in progress, information on whether summer holidays are in progress, information on whether winter holidays are in progress, information on whether spring holidays are in progress, information on whether Christmas brake is in progress;
- logical attributes describing the public holidays: information on whether any public holiday is in progress, information about New Year's Day, Easter, Easter Monday, All Saint's Day, All Souls Day, Christmas Eve, Christmas, New Year's Eve;
- cyclical attributes created based on basic attributes from time column - the problem of creating these attributes are described in detail in [22].
- numerical attributes describing a number of days before public holidays (any public holiday, Christmas and Easter) – these attributes have value from a range [0;7].

All considered petrol stations are located in Poland and because of this attributes for public and school holidays were prepared based on polish holidays.

2) Attributes from previous values of sales: Using forecasting methods typical for tabular datasets has some limitations and problems comparing with time series methods. In a tabular dataset, each row describes one measurement and they are not chronologically connected with each other. In a simple tabular dataset, we lose information about time series. Because of this, we proposed adding attributes describing the history of fuel sales:

- fuel sales observed in previous intervals – from one interval before to six intervals before;
- mean value from x previous intervals (x = (6, 12, 18, 24, 30, 36, 42));
- median value from x previous intervals (x = (6, 12, 18, 24, 30, 36, 42));
- fuel sales observed in previous days in the same hours as a considered measurement - from 2 days before to 7 days before;
- mean value from x previous days in the same hours as a considered measurement (x = (7, 10, 14, 21));
- fuel sales observed x weeks before and in the same hours as a considered measurement (x = (2, 3, 4));
- mean value from x previous weeks and in the same hours as a considered measurement (x = (2, 4, 6, 8));
- attributes describing trend: a preceding interval sales value compared with previous values, a preceding interval sales value compared with mean values from previous intervals, mean values of preceding intervals compared with previous mean values;

The example will be presented in order to better explain creating these attributes. Suppose we want to create attributes for the period where *end_time*= '2019-02-22 08:00:00' (Friday). Some attributes would be as follows:

- fuel sales observed in previous intervals – these attributes would describe sales from intervals '2019-02-22 04:00:00', '2019-02-22 00:00:00', ..., '2019-02-21 08:00:00';
- mean value from 6 previous intervals - this attribute would be a mean value of sales from intervals '2019-02-22 04:00:00', '2019-02-22 00:00:00', ..., '2019-02-21 08:00:00';
- fuel sales observed 5 days before in the same hours as a considered measurement – this attribute would describe sale from interval '2019-02-17 08:00:00';
- mean value from 7 previous days in the same hours as a considered measurement – this attribute would be a mean value of sale from intervals '2019-02-21 08:00:00', '2019-02-20 08:00:00', ... , '2019-02-15 08:00:00';
- fuel sale observed 2 weeks before in the same hours as a considered measurement – this attribute would describe sale from interval '2019-02-08 08:00:00';
- mean value from 2 previous weeks in the same hours as considered measurement – this attribute would be a mean value of sale from intervals '2019-02-15 08:00:00' and '2019-02-08 08:00:00'.

After the data preparation process, each dataset had 157 columns and $3 \cdot 365 \cdot 6 = 6570$ rows (3 years of data; 365 days in a year; 6 4-hours intervals). It is worth mentioning that not all rows might be used in the experiments because some of them could have missing values of sales. Also, not all columns were used when training the model. For example columns *id_meter*, *start_time*, *end_time* were not used.

IV. Experiments

The purpose of this paper is to analyse the benefit of using clustering methods and information about similar records in the process of forecasting sales. In our experiments, we used datasets representing fuel sales on petrol stations. We assumed that there might be similarities among the different sales characteristics. However, some of them may also differ significantly from each other. For example, in some datasets sales can be strongly correlated with their previous values and in other datasets, we won't see a similarly strong connection

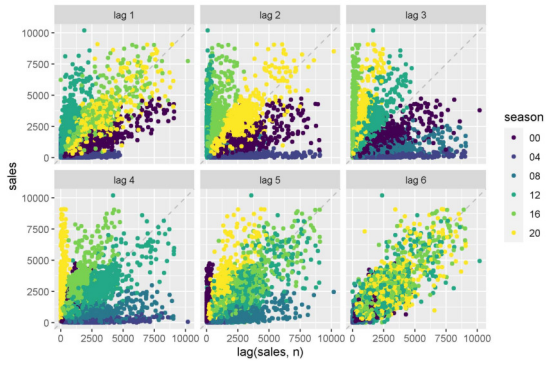


Fig. 2. Exemplary sales values plotted against lagged values of itself. It is an example of a dataset where the sales value is highly correlated with 6th lagged value.

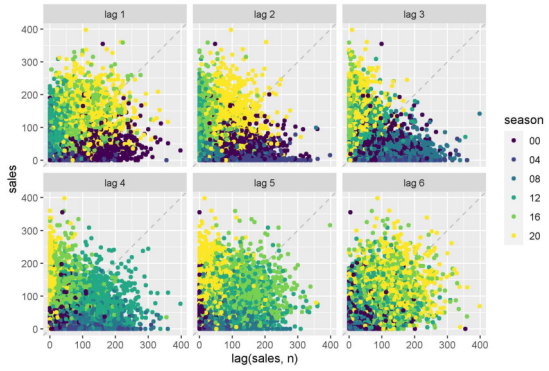


Fig. 3. Exemplary sales values plotted against lagged values of itself. It is an example of a dataset where the sales value is weakly correlated with previous sales values.

with historical data. It can be visible in the figure 2 and figure 3. Both illustrations show the time series of fuel sales against lags of itself. Each figure shows the results for a different dataset. In the figure 2 we can see that the sales value is highly correlated with lag number 6 of itself. As it was said before, the datasets described 4-hour intervals, so lag number 6 was an interval that occurred 24 hours before the interval under consideration. A different situation can be observed for the sales of the dataset presented in the figure 3 – here the correlation of historical values with the current value of the sales is not so prominent.

The differences between sales could be also visible when considering different attributes e.g. the attributes describing a time of a sale.

Taking into account the differences and similarities of sales characteristics, we wanted to find out how adding information about these similarities can improve sales forecasting.

A. Description of experiments

In the experiments, we used typical methods for time series and methods intended for tabular datasets. We forecasted one value ahead, so we always forecasted sales for the next 4 hours.

In all experiments in which we trained a predictive model using XGBoost, the first model was trained on data from 2017-01-01 to 2018-12-31, and then the model was re-trained every 7 days of data. Testing of a given model was done on data that fell between model recalculations.

We carried out following experiments:

1) Station-based forecast – experiments "*moving average*", "*arima predictions*" and "*separate groups predictions*".
2) Forecast after putting all datasets into one group – experiments "*one group – predict normalized sale*" and "*one group – add column with prediction*";
3) Forecast based on similar records – experiment "*regression knn*".
4) Forecast based on clustering time series – experiment "*partitional clustering – predict normalized sale*" and "*partitional clustering – add column with prediction*".
5) Forecast based on double clustering – experiment "*double clustering*".

Each experiment will be described below.

Method "*moving average*" (*MovAvg*) is a simple forecasting method based on moving average. The window for these calculations was 42 measurements (sales from the previous 7 days).

Experiment "*arima predictions*" (*ARIMA*) was based on ARIMA method. In this experiment, `auto.arima` function from the `forecast` R package was used. 84 previous measurements (2 weeks of data) was a training set (argument `learning_set_size = 84`).

In both experiments, "*moving average*" and "*arima predictions*", only the value of the sales was taken into consideration.

Experiment "*one group – predict normalized sale*" (*1G.PredNormSale*) is the experiment where each dataset in a first step was normalised. It means that the sales were normalised and all attributes created based on a sales history were generated again using normalised values of sale. Then all records from all datasets were connected into one dataset. The model was trained using XGBoost method on data from 2017-01-01 to 2018-12-31 and then recalculated every 7 days of data. The model forecasted the normalised value of the sales. For each recalculation predictions were made. Each prediction was then reversed from a normalised value to the regular one.

Experiment "*one group – add column with prediction*" (*1G.AddPred*) is the experiment where, as before, data from all datasets were normalised and combined into one training set. A model that predicted the normalised sales value was created. Then, for each dataset describing one time series (to the train set and test set), a column was added with the forecast from the model (forecast of the normalised sales value). A separate XGBoost model was then created for each dataset (the decision column was the actual sales value, without normalisation). The model was also recalculated every 7 days.

In the experiment "*separate groups predictions*" each dataset was considered independently. A separate XGBoost model was trained for each dataset. A model was trained only on historical data from the specific time series.

Experiment "*regression knn*" (*RegKNN*) used a typical k-NN regression method. Firstly, all datasets were normalised and connected into one dataset. Then we found an optimal number of neighbours for this dataset (parameter `k` for k-NN algorithm) – repeated cross-validation was used on records from 2017-01-01 to 2018-12-31 in order to do this. In the next step, we performed k-nearest neighbour regression that was recalculated every 7 days. Predicted normalised values were reversed to the regular values of sales. k-NN does not work with missing data so in training datasets only rows with no missing values were used. In a test set, missing values were imputed using the last observed value.

In the experiment "*partitional clustering – predict normalized sale*" (*PC.PredNormSale*) the datasets were grouped by the nature of their normalised time series describing sales. The `dtwclust` R package and the partitional (or partitioning) clustering algorithm were used for clustering. Within each group, all examples from the datasets comprising the group were combined. An XGBoost model that predicted the normalised sales value was created for each group. The final prediction was made based on these models - the output was a normalised forecast, which then had to be inverted to a regular value. For example, assume that we have five datasets describing fuels sales - datasets *a, b, c, d, e*. In the first step of the experiment, each dataset was normalised and the value of the sale was normalised. Then only time series (not any additional created attributes) were compared and grouped using `dtwclust` package - assume that datasets *a* and *c* were classified to group *X* and datasets *b, d, e* were classified to group *Y*. Then 2 models were created - the first model that was trained on connected datasets *a* and *c* (group *X*) and a second model that was trained on connected datasets *b, d, e* (group *Y*). Then predictions were made based on an appropriate model for each data set - predictions from a test set from dataset *a* were made based on a model trained for group *X*, predictions from the test set from dataset *b* was made based on a model trained for group *Y* etc.

Experiment "*partitional clustering – add column with prediction*" (*PC.AddPred*) – as before, data from all time series were normalised and grouped by the nature of the sales time series. Within each group, a training set was created and a model was trained that predicted the normalised sales value. Then, for each datasets belonging to this group (to the train and test set), a column was added with the forecasts from this model (forecast of normalised sales value). A separate model was then created for each dataset (the decision column was the actual sales value, without normalisation). Showing it on the example: again, suppose that datasets *a* and *c* were classified to group *X* and datasets *b, d, e* were classified to group *Y* based on their sales characteristics. Then models were trained for group *X* and group *Y* - these models predicted the normalised value of the sales. Then for each non-normalised datasets (*a, b, c, d, e* before normalisation) new column with the prediction of normalised sale was added. So for the non-normalised dataset *a*, column with predictions from a model created for group *X* was added. Then for each non-normalised dataset *a, b, c, d, e* separate XGBoost model was trained.

In experiment "*double clustering*" (*2Clust*) data from all time series were also normalised and grouped by the characteristic of the sales time series. Then for each group, a new logical column was added to each normalised dataset. For example, if two groups were created, two new columns would be added to each dataset. The columns indicated whether the row belongs to the group represented by the column. At this point, all records from one dataset would have the same values in these columns – if the dataset was classified to group number 2 then in the second column all records would have a value "1", and in the first column all records would have a value "0". Clustering based on time series and adding columns with information about groups was the first clustering in this experiment. Then all datasets were combined into one dataset. Then an optimal number of clusters for algorithm k-means was calculated and then k-means clustering was performed on the created dataset. This was the second clustering in this experiment. Then for each cluster obtained from k-means, we created XGBoost model. It is worth mentioning, that the rows from one dataset could be added to the different clusters from k-means clustering. It means that rows classified into one k-means cluster could have different values in columns representing results from time series clustering.

The flowcharts of the more advanced experiments – *PC.PredNormSale*, *PC.AddPred* and *2Clust* – are presented in figures 4 and 5.

In clustering procedures always normalised sales and normalised datasets were used.

The important aspect of these experiments is the utility of the created models and using the proposed solution in a real-life scenario. One interesting problem is forecasting sales for new, unknown petrol station or new fuel on an existing petrol station. We do not have a history of sales for this fuel, so models from experiments *SepGr*, *ARIMA* and *MovAvg* couldn't be used. Because of the lack of historical data, we wouldn't also be able to assign this fuel into one of the previously created clusters that were created based on similarities of time series. It means that models created in experiments *PC.AddPred*, *PC.PredNormSale* and *2Clust* also couldn't be used. The only way to use these models would be to manually assign the fuel from the station to an existing cluster. Such a task could be performed by an expert, but this could create additional errors and bias. In the discussed problem, the models from experiment *RegKNN* could be possibly used however k-NN do not work with missing data so many attributes would have to be imputed. In the problem of forecasting sale for new petrol station, the only way would be to use models from experiments *1G.PredNormSale* and *1G.AddPred*, because XBGoost method handles missing data.

B. Results

The errors metrics for each experiment were presented in the table I. These are the standard error metrics that are frequently used in papers that describe creating forecasting models. In describing the results we used *Root Mean Square*

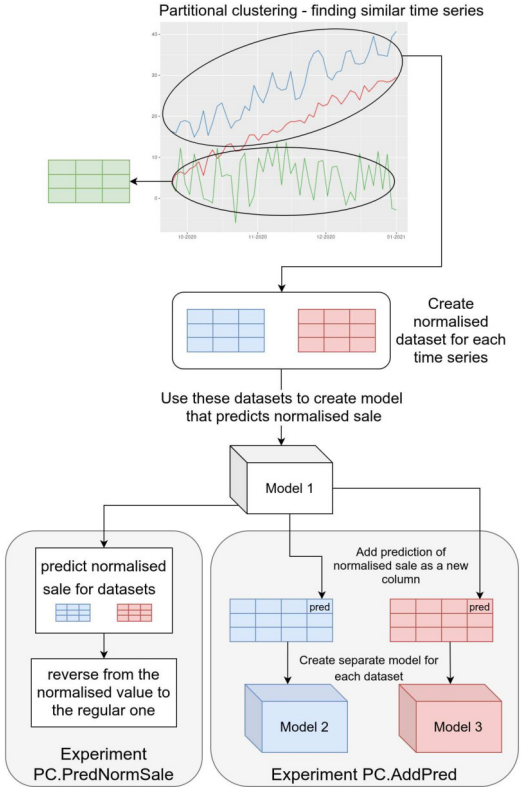


Fig. 4. Flowchart of the experiments *PC.PredNormSale* and *PC.AddPred*.

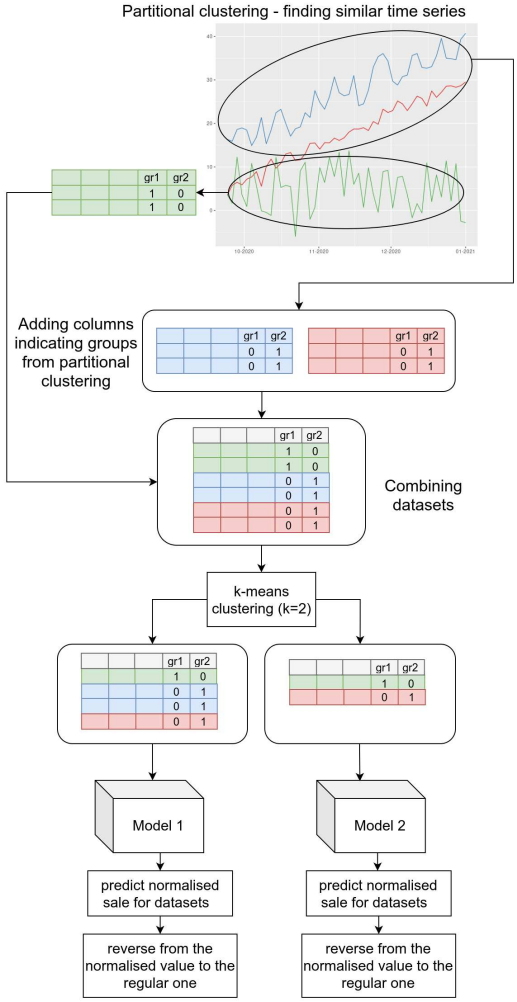


Fig. 5. Flowchart of the experiment *2Clust*.

Error (RMSE), *Mean Absolute Error (MAE)*, *Coefficient of determination (R2)* and *Weighted Mean Absolute Percentage Error (WMAPE)*. In our opinion, the most intuitive errors are *MAE* and *WMAPE* and these metrics will be briefly explained in order to understand the results correctly.

The MAE express what error, on average, the model makes. It does not explain if it was an overestimation or underestimation error. In the case of described experiments, it says what was the average error in forecasting sales for a 4-hour period.

The WMAPE is an advanced version of a metric *mean absolute percentage error (MAPE)*. The MAPE is meaningful only when the values are large. If the actual value is close to 0, the value of MAPE gives uninterpreted results. In order to bypass this problem, a similar measure – WMAPE – was developed. It is the sum of absolute errors divided by the sum of the actual values and it works well with smaller numbers. It is widely used in the retail sector. The sale of fuel on some petrol stations was often equal to 0, so it was important to use a modified version of the MAPE. The WMAPE, multiplied by 100, can be interpreted as the average percentage by which the model is wrong.

The results presented in the table I show that the best solution among the tested approaches is for experiment *1G.PredNormSale*.

In figure 6 the histogram of errors for the experiment *1G.PredNormSale* is presented, where $error = prediction - real\_value$. We can see that the distribution of errors is balanced for this approach.

TABLE I
ERROR METRICS FOR THE EXPERIMENTS.

model	RMSE	MAE	R2	WMAPE
1G.PredNormSale	347.97	207.13	0.65	0.32
PC.PredNormSale	349.55	208.95	0.64	0.33
SepGr	365.17	222.46	0.60	0.36
1G.AddPred	372.86	225.61	0.58	0.36
PC.AddPred	374.83	228.69	0.58	0.36
RegKNN	413.61	262.83	0.50	0.42
2Clust	459.69	343.93	0.17	0.54
ARIMA	505.40	351.03	0.23	0.57
MovAvg	583.09	435.50	0.07	0.67

We also wanted to compare how these errors differ in real-life situations. In order to do this, for each dataset and each experiment we compared the absolute errors of predictions with the maximum value of sale from a test dataset. The calculation of this error can be presented as follows $\frac{|pred-real|}{maxReal} \cdot 100$, where $pred$ is the predicted value, $real$ is the actual value and

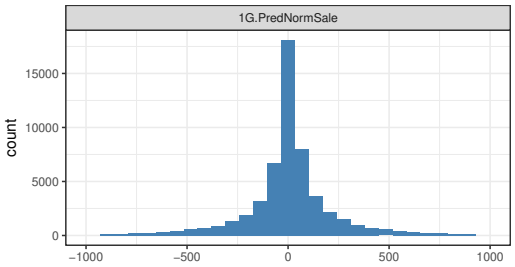


Fig. 6. Histograms of errors for the experiment *1G.PredNormSale*.

$maxReal$ is the maximum value of the actual sales in the test set. Then, for each experiment, the average for each set was calculated and then the average values from all datasets were obtained. The average real-world prediction errors are given in table II. This error allows us to understand the scale of the error we are dealing with and allows us to answer the question of how significant these errors are in a relation to the domain. Here, the *1G.PredNormSale* approach was also the best one.

TABLE II
RELATIVE ERRORS FOR THE EXPERIMENTS.

	mean_realtive_error [%]
1G.PredNormSale	6.59
PC.PredNormSale	6.65
SepGr	7.11
1G.AddPred	7.16
PC.AddPred	7.25
RegKNN	8.38
2Clust	10.99
ARIMA	11.61
MovAvg	14.44

In most industries where the demand for products needs to be forecast, overestimation and underestimation errors mostly are not equal to each other. We can imagine that in a problem of forecasting fuel demand on petrol station, underestimation errors are much more harmful to the industry. If these forecasts would be used for deliveries planning then underestimation could create a situation where there wouldn't be fuel on a petrol station. It could lead to losing clients. Because of this, we performed an analysis of underestimation and overestimation errors. We considered prediction as an overestimation when the prediction was at least 20 litres larger than the real sales during the 4-hour period. The prediction was an underestimation when it was at least 20 litres smaller than the real sales. The numbers of examples that were overestimated or underestimated in each dataset and in each experiment are presented in figure 7. We can see that for each experiment the average number of underestimation is bigger than the number of overestimated examples. We can also see from the plot that for experiments *RegKNN* and *ARIMA* the average number of examples overestimated and underestimated is higher than for other experiments. It means that these methods made more errors with an absolute value higher or equal to 20 litres, so they made more significant errors. This conclusion

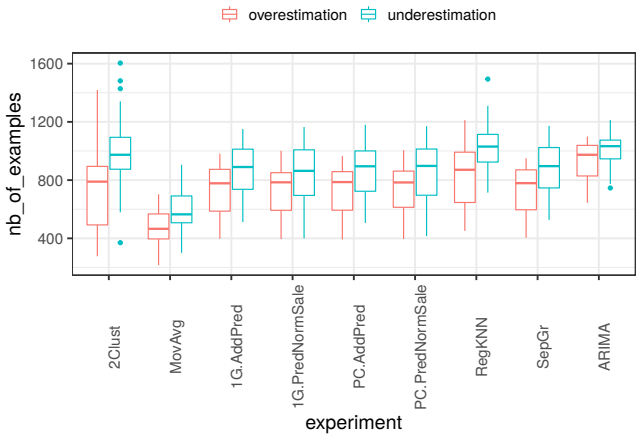


Fig. 7. Box plots representing number of examples that were overestimated or underestimated in each dataset for each experiment.

is consistent with the results obtained from the RMSE metric, which penalises large mistakes.

V. CASE STUDY

In this section, we would like to present a case study for two different datasets and discuss the utility of created methods. Both of the datasets represent the sale of gasoline on two different petrol stations.

In the previous section, we discussed what are the errors and the error metrics. Each error was calculated for every 4-hour period independently. Suppose that the errors in 6 consecutive 4-hour periods were (-250, 250, -300, 100, 300, -100). Each of these errors is significant, but they add up to 0. In this case, we can say that the model at the end of the day did not make a mistake, because ultimately as much fuel was sold in one day as the model predicted. This is an important aspect for demand forecasting.

In the following case study, we present the moving sums of errors. The window for moving sum was equal to 7 days (42 measurements). We presented what was the average moving sum of errors for the test dataset and we compared it with an average moving sum of real sales that have taken place within 7 days.

Results for the first dataset (dataset A) are presented in table III (*mean* – the average value of the moving errors, *stder* – the standard deviation of the moving errors, *%mean* - the ratio of the average moving error to the moving average of sales expressed as a percentage, *%stder* - the ratio of the standard deviation of the moving errors to the moving average of sales expressed as a percentage). The smallest average value of a moving error within 7 days was obtained for the experiment *1G.PredNormSale*. This is in line with the results presented in the table I and II where this solution had the best results for error metrics. The average value of moving error was equal to 0.3% of a moving sum of sales within 7 days for this dataset. The standard deviation of moving error was equal to 6.3% of a moving sum of sales within 7 days for this dataset.

We can observe that the average moving error for most of the experiments is greater than zero – it means that mostly we will deal with overestimation and as it was discussed before this is the better kind of error in the retail sector.

The moving sum of the actual sale in the 7-days window for dataset A was around 23,000 litres.

TABLE III
THE ANALYSIS OF THE MOVING ERROR CALCULATED IN WINDOW EQUAL TO 7 DAYS FOR THE DATASET A

experiment	mean	stder	%mean	%stder
1G.PredNormSale	67.57	1472.71	0.29	6.25
PC.PredNormSale	93.93	1440.51	0.40	6.11
1G.AddPred	193.71	1167.56	0.82	4.96
PC.AddPred	246.76	1273.60	1.05	5.41
MovAvg	303.66	1301.34	1.29	5.52
ARIMA	311.59	2312.32	1.32	9.81
SepGr	360.86	1022.90	1.53	4.34
RegKNN	-587.24	1400.77	2.49	5.95
2Clust	-2912.15	4487.68	12.36	19.05

As a contrast, we present the results for the second dataset (dataset B). The results are presented in IV. For this dataset the method *1G.PredNormSale* got worse results comparing with other methods. The value of the average moving error was also less than zero so it can lead more frequently to underestimation. The best average value of moving error was obtained for the experiment *PC.AddPred* and it was equal to 0.06% of a moving sum of sales within 7 days for this dataset.

The moving sum of the actual sale in 7-days window for dataset B was around 13,000 litres.

This case study shows that the error metrics presented in the table I can indicate the general best solution taking into account the measurements independently but it does not mean that it will always be the best solution in each case (e.g. for the demand planning). Also, the average best solution does not need to be the best for each considered dataset.

TABLE IV
THE ANALYSIS OF THE MOVING ERROR CALCULATED IN WINDOW EQUAL TO 7 DAYS FOR THE DATASET B

experiment	mean	stder	%mean	%stder
PC.AddPred	8.29	880.62	0.06	6.54
1G.AddPred	30.26	839.20	0.22	6.24
SepGr	54.63	902.84	0.41	6.71
ARIMA	-64.55	1425.59	0.48	10.59
MovAvg	73.78	1019.11	0.55	7.57
PC.PredNormSale	-87.06	794.90	0.65	5.91
1G.PredNormSale	-122.13	775.17	0.91	5.76
RegKNN	189.07	603.95	1.41	4.49
2Clust	-489.86	4107.20	3.64	30.52

VI. CONCLUSIONS

The accurate forecasts are an important aspect in all companies from a retail sector, also from the petrol retail sector. This study checked the usefulness and impact of data clustering in the process of predicting fuel sales. The research showed that the best results, in general, were obtained for the experiment where for each time-series the new features were extracted and then all obtained tabular datasets were connected into one dataset. This dataset was used to train one predictive model which predicted the normalised sale for each fuel and each petrol station. This solution outperformed the typical simple time series forecasting methods (ARIMA, moving average), but also it was better compared with the results where advanced clustering methods were used. It also performed better than creating a predictive model for each time series separately based only on its historical data with extracted features. In general, adding the predictions obtained from the model generated on a bigger number of datasets as a new column did not improve the results of creating models separately for each dataset.

The worst results were obtained for time series methods, so it proves that in some cases the use of methods typical for tabular datasets can lead to better results for time series predictions. Also, it shows that the process of developing proper features from historical data, as well as using information about the time of the measurements (information derived from the date and time and information about holidays), can lead to obtaining better results. However, this hypothesis should be checked in further studies where our results will be compared with results obtained for more advanced methods used in time series forecasting problems.

The additional asset of the presented best performing approach (*1G.PredNormSale*) is also the fact that it can be quickly and straightforward used when predicting sale for a new petrol station or a new fuel on the petrol station because the sale of this fuel wouldn't need to be compared and assigned to one of the previously created cluster based on its historical data. In this proposed solution, one model is created for all petrol stations and it can be used with new data. Also, the XGBoost method, which is used in the proposed solution, would deal with missing data. However, if we would like to have the real values for all of the extracted features then only 8 weeks of historical sales would be needed to do this. The 4-week historical data would leave only 2 from 157 attributes with empty values. The solution with the best results is also practical from the perspective of time constraints because only one model is trained for all datasets. In the proposed solution, the model was recalculated weekly, which minimises the chances of a concept drift problem.

In the presented case study, we showed a different aspect of the predictions. We presented cumulative errors for the 7-day long periods. The results showed that for each presented dataset, the average moving sum of errors obtained for the best approach from the considered case study was relatively low. However, the case study also proved that the method which is generally the best do not give the best results in every case. In practical applications, a method would need to be developed to select the appropriate model (e.g. by tracking recent errors). This is a challenge for future research.

Other aspects that we will investigate in further work will be making similar experiment with the use of Deep Learning and ensemble methods. The experiments will be also performed on different datasets from another retail sector e.g. on data describing the sale of fast-moving consumer goods (FMCG).

The forecasting of sale during the promotion in FMCG sector and the efficiency of promotions were conducted in our previous works ([22], [23]), but the analysis was not yet performed on the data describing regular sales and the clustering methods were not tested by us in this field.

In conclusion, the results showed that the use of clustering methods did not produce the best results among the considered approaches, however, our study showed that it is useful to use historical data from other stations when making predictions and it should be considered to create one predictive model for many datasets and not for each dataset separately.

ACKNOWLEDGMENT

This research was realised in co-operation with FuelPrime Department in AIUT Ltd. and was partially supported by the European Union through the European Social Fund (grant POWR.03.05.00-00-Z305).

REFERENCES

[1] E. S. Gardner Jr., "Exponential smoothing: The state of the art," vol. 4, no. October 1983, pp. 1–28, 1985.

[2] C. W. Chu and G. P. Zhang, "A comparative study of linear and nonlinear models for aggregate retail sales forecasting," *International Journal of Production Economics*, vol. 86, no. 3, pp. 217–231, dec 2003. doi: 10.1016/S0925-5273(03)00068-9

[3] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, "Sales-forecasting of retail stores using machine learning techniques," in *Sales-forecasting of Retail Stores using Machine Learning Techniques*. IEEE, 2018. doi: 10.1109/CSITSS.2018.8768765. ISBN 9781538660782 pp. 160–166.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. KDD '16. ACM, 2016. doi: 10.1145/2939672.2939785. ISBN 9781450342322 pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[5] X. Dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. Institute of Electrical and Electronics Engineers Inc., jan 2021. doi: 10.1109/ICCECE51280.2021.9342304. ISBN 9781728183190 pp. 480–483.

[6] E. Žunić, K. Korjenić, K. Hodžić, and D. Đonko, "Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data," *International Journal of Computer Science and Information Technology*, vol. 12, no. 2, pp. 23–36, apr 2020. doi: 10.5121/ijcsit.2020.12203

[7] K.-F. Au, T.-M. Choi, and Y. Yu, "Fashion retail forecasting by evolutionary neural networks," *International Journal of Production Economics*, vol. 114, no. 2, pp. 615 – 630, 2008. doi: 10.1016/j.ijpe.2007.06.013

[8] V. Adithya Ganesan, S. Divi, N. B. Moudhgalya, U. Sriharsha, and V. Vijayaraghavan, "Forecasting food sales in a multiplex using dynamic artificial neural networks," in *Advances in Intelligent Systems and Computing*, vol. 944. Springer Verlag, 2020. doi: 10.1007/978-3-030-17798-0_8. ISBN 9783030177973. ISSN 21945365 pp. 69–80.

[9] C. Giri, S. Thomassey, J. Balkow, and X. Zeng, "Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression," in *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*. Institute of Electrical and Electronics Engineers Inc., aug 2019. doi: 10.1109/ICESI.2019.8863024. ISBN 9781728121741 pp. 1–6.

[10] Q. Yu, K. Wang, J. O. Strandhagen, and Y. Wang, "Application of Long Short-Term Memory Neural Network to Sales Forecasting in Retail—A Case Study," in *Advanced Manufacturing and Automation VII*. Springer Singapore, 2018. doi: 10.1007/978-981-10-5768-7_2. ISBN 978-981-10-5768-7. ISSN 18761119 pp. 11–17.

[11] D. Ruta, L. Cen, and Q. H. Vu, "Deep Bi-Directional LSTM Networks for Device Workload Forecasting," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, 2020. doi: 10.15439/2020F213. ISBN 9788395541674 pp. 115–118.

[12] S. Punia, K. Nikolopoulos, S. P. Singh, J. K. Madaan, and K. Litsiou, "Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail," *International Journal of Production Research*, vol. 58, no. 16, pp. 4964–4979, aug 2020. doi: 10.1080/00207543.2020.1735666

[13] J. Chai, S. Wang, and S. Wang, "Demand Forecast of Petroleum Product Consumption in the Chinese Transportation Industry," *Energies*, vol. 5, no. 3, pp. 577–598, 2012. doi: 10.3390/en5030577. [Online]. Available: www.mdpi.com/journal/energiesArticle

[14] I. Themido, A. Quintino, and J. Leitao, "Modelling the Retail Sales of Gasoline in a Portuguese Metropolitan Area," *International Transactions in Operational Research*, vol. 5, no. 2, pp. 89–102, mar 1998. doi: 10.1111/j.1475-3995.1998.tb00106.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1475-3995.1998.tb00106.x

[15] K. Kalid, J. Ahmad, S. Yong, and K. H. Yew, "Petronas Petrol Station Fuel Consumption Forecast System," in *Proceedings of the Second International Conference on Artificial Intelligence in Engineering & Technology*, 2004. doi: 10.13140/2.1.4493.8568. ISBN 10.13140/2.1.4. [Online]. Available: https://www.researchgate.net/publication/271637545

[16] S. M. Rizvi, T. Syed, and J. Qureshi, "Real-time forecasting of petrol retail using dilated causal CNNs," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, p. 3, feb 2021. doi: 10.1007/s12652-021-02941-3. [Online]. Available: https://doi.org/10.1007/s12652-021-02941-3

[17] P. F. Jiménez-Pérez and L. Mora-López, "Modeling and forecasting hourly global solar radiation using clustering and classification techniques," *Solar Energy*, vol. 135, pp. 682–691, oct 2016. doi: 10.1016/j.solener.2016.06.039

[18] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Japan Architectural Review*, vol. 3, no. 1, pp. 62–76, jan 2020. doi: 10.1002/2475-8876.12135. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/2475-8876.12135

[19] S. Thomassey and A. Fiordaliso, "A hybrid sales forecasting system based on clustering and decision trees," *Decision Support Systems*, vol. 42, no. 1, pp. 408–421, oct 2006. doi: 10.1016/j.dss.2005.01.008

[20] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - A decade review," *Information Systems*, vol. 53, pp. 16–38, may 2015. doi: 10.1016/j.is.2015.04.007

[21] Ł. Sosnowski, I. Szymusik, and T. Penza, "Network of Fuzzy Comparators for Ovulation Window Prediction," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager, Eds. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-50153-2_59. ISBN 978-3-030-50153-2 pp. 800–813.

[22] M. Blachnik and J. Henzel, "Estimating the Performance Indicators of Promotion Efficiency in FMCG Retail," in *Neural Information Processing. ICONIP 2020. Lecture Notes in Computer Science*, vol. 12533. Springer, Cham, 2020. doi: 10.1007/978-3-030-63833-7_27. ISBN 9783030638320. ISSN 16113349 pp. 320–332.

[23] J. Henzel and M. Sikora, "Gradient Boosting and Deep Learning Models Approach to Forecasting Promotions Efficiency in FMCG Retail," in *Artificial Intelligence and Soft Computing. ICAISC 2020. Lecture Notes in Computer Science*, vol. 12416. Springer, Cham, 2020. doi: 10.1007/978-3-030-61534-5_30. ISBN 978-3-030-61533-8. ISSN 16113349 pp. 336–345.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 23–32
DOI: 10.15439/2021F12
ISSN 2300-5963 ACSIS, Vol. 26

State-of-the-Art Techniques in Artificial Intelligence for Continual Learning: A Review

Bukola Salami
School of Computing, University of Eastern Finland, Kuopio, Finland
bukolas@uef.fi

Keijo Haataja
School of Computing, University of Eastern Finland, Kuopio, Finland
keijo.haataja@uef.fi

Pekka Toivanen
School of Computing, University of Eastern Finland, Kuopio, Finland
pekka.toivanen@uef.fi

***Abstract*—Artificial neural networks are used in many state-of-the-art systems for perception, and they thrive at solving classification problems, but they lack the ability to transfer that learning to a new task. Human and animals both have the capability of acquiring knowledge and transfer them continually throughout their lifespan. This term is known as continual learning. Continual learning capabilities are important to ANN in the real world especially with the continuous stream of big data. However, it remains a challenge to be achieved because they are prone to a problem called catastrophic forgetting.. Fixing this problem is critical, so that ANN incrementally learn and improve when deployed to real life situations. In this paper, we did a taxonomy of continual learning in human by introducing plasticity-stability dilemma, hence the Hebbian plasticity and compensatory homeostatic plasticity process of learning and memory formation that occurs in the brain. We also did a state-of-the-art review of three different approaches to continual learning to mitigate catastrophic forgetting.**

***Index Terms*—Artificial Intelligence; Continual Learning; Catastrophic Forgetting; Artificial Neural Networks; Stability–Plasticity Dilemma**

I. Introduction

LEARNING continually has always been the grand goal of any Artificial Intelligent (AI) systems functioning in the real-world scenario because AI systems can be continuous exposed to streams of data and so are required to remember existing tasks when modelled on new stream of data. This has recently attracted much attention in the AI community, especially related to Artificial Neural Networks (ANNs) [1]. Humans and animals have an exceptional ability to learn large number of different skills and tasks but also to select the ones which are useful and relevant without negatively interfering with each other and at the same time being able to recall information when needed on such tasks that were previously learned [2, 3]. The ability to do this is called Continual Learning and can also be referred to as Lifelong Learning or Incremental Learning [1, 2]. AI agents should demonstrate a capability for continual learning [4]. The goal is to gather knowledge across tasks, particularly through model sharing and possibly having only one model that can perform well on all the learned tasks [5]. However, the existing standard for model deployment has a critical flaw: data are dynamic, and this continues to change [1, 6].

With remarkable successes accomplish over the past few years in AI, deep network applications are however restricted to sole, distinct problem. Where every single network has to be trained and re-trained from the beginning every single time a new task is fed into the network and as a result their training remains very challenging to deal with particularly in real-world settings and in situations where data are scarce and/or computation is costly [7]. Furthermore, the sequence of tasks may not be clearly labelled tasks and they may switch randomly, leading to an individual task recuring in long time intervals. Therefore, the main challenge of an AI agent to learn continually is being susceptible to catastrophic forgetting or catastrophic interference [4].

This well-known phenomenon was first recognized by McCloskey and Cohen in [8]. Catastrophic forgetting always leads to a degrade generalization performance or in the worst case, a complete loss of information on an older task that was previously performed because it was simply re-trained on the new task or dataset sequentially [1, 2]. It specifically happens when a network is trained in sequence on several tasks because the weights that are imperative for task A are modified to incorporate the goal of task B, and as a result of these changes to the network, the accuracy on task A can severely reduce after some training updates on task B [4, 9]. This is conceivably, one of the main gap between modern ANNs design and biological neural networks because of the complexity of synapses [10].

To overcome the lack of continual learning in ANNs, recently three main strategies have been proposed: Progressive/Architectural Strategy, Rehearsal Strategy, and Regularization Strategy [2, 5].

A lifelong learning architecture capable of continual learning could guide the field of AI into a period of extraordinary performance, generality, and integration. This architecture could also prevent the need for costly data collection, labelling and retraining that sets constrains on today's state-of-the-art computer systems [9]. In essence, to overcome catastrophic forgetting, an AI system should display the capability to gain new knowledge and simultaneously improve the network on existing tasks based on the continuous stream of data, thereby, preventing significant existing knowledge from being forgotten [1, 2]. This is known as the stability-plasticity dilemma. Plasticity is the ability to integrate new knowledge, while stability is preserving existing knowledge while new stream of data is processed. Although, a model will not be able to gain new knowledge from new training data if they are too stable. Likewise, a model with abundant plasticity can suffer from a great weight change and forget a previously learned task. [11, 12]. One of the effective approaches to plasticity was the one addresses by Stephen Grossberg, articulated in 1980 on the solution to the stability–plasticity dilemma which states that "a system must remain plastic enough to learn important and new information, while also maintaining stability in its memories for information that has already acquired" [9]. Such adaptation and memory formation are what can be observed in biological neurosystems. Humans have remarkable ability to preserve old knowledge and skills learned and it is mainly reliant on how often they are recollected and used. Tasks that are practiced and performed regularly, tend to be unforgettable, unlike the ones that are

This work is supported and funded by Digital Innovation HUB of Northern Savo Region – DigiCenterNS Kuopio, Finland

so old and frequently not used. Strangely, this adaptation and memory formation sometimes happens with little or no form of supervision whatsoever. This process, at the fundamental level, according to Hebbian theory, is the consolidation of neurons connected to synapses, that performs together at the same time, compared to neurons with unrelated performance behavior [5].

Most of the algorithms presented in this paper are based on the current state and advancements in both neurophysiology and computational neuroscience field that are capable of continual learning in AI. [7, 9]

In this paper, we review some of the major works in continual learning both in advanced animals such as humans, whales, dogs, dolphins etc. and AI agent: we focus on how humans and animals acquire new knowledge and memories and at the same time been able to retain the useful ones over time. We also discussed several proposed algorithms for continual learning systems to overcome catastrophic forgetting. The rest of this paper is organized as follows: Section 2 reviews continual learning in humans and animal. Section 3 also reviews few continual learning strategies and algorithms proposed in the last 4 years. section 3.1 introduced the fundamental of continual learning, its desiderata, and the three different strategies. Section 3.2 reviewed some common algorithms proposed for continual learning with their respective mathematical equations and in table 1, a summary of different recent algorithms to continual learning is given. In section 4 we proposed our novel research idea for forgetting in ANN, and in section 5 the conclusion of the paper .

II. Continual Learning In Advanced Animals

New skills and knowledge can easily be acquired and transferred across domains in advanced animals to complete tasks, while artificial neural systems are still in the early stages regarding transfer learning, which is prone to catastrophic forgetting [1]. Likewise, humans and animals can learn in a continual way, but it has been somewhat challenging for an AI system to do the same [5].

Evidence found recently suggests that the human and animal brain can avoid forgetting by shielding previously learnt knowledge and skills in the neocortical circuits. The brain significantly benefits from the integration of multisensory information, which provide the means for an effective communication. Furthermore, in conditions of sensory hesitation with respect to the predominant tendency to train ANNs on uni-sensory information, such as audio or visual information [1]. For example, when a mouse learns a new task/skill, a part of its excitatory synapse is reinforced, and this leads to an increase in the capacity of individual dendritic spines of the mouse brain neurons [5]. Afterward, these increased dendritic spines persevere in spite of learning some other skills alongside the old one, and it results to retention of such skill after a few months later. When some of these dendritic spines are selected and cleared up, the matching skill is forgotten. This gives a fundamental evidence that neural mechanisms for supporting the protection of these synapses are important to retention of task performance. The results obtained with the mouse experiment alongside with some other neurobiological models suggested that continual learning in the neocortex depends on task-specific synaptic consolidation, by which knowledge is strongly encoded by reducing the plasticity of synapses that are vital to previously learned tasks and therefore stable over a long timeframe [13].

The principal core idea is that learning is associated with persistent and experience-driven changes to the brain, as given with the mouse example, that help them in the effective performance of vital tasks, such as the acquisition of necessities like food and shelter while avoiding the unpleasantness that accompanies injury or predation [14]. This is the inspiration behind autonomous embodied agents research on multisensory features for early development and sensorimotor specialization in human brain [1].

A. Stability–Plasticity Dilemma

The human brain experiences neural plastic changes across its lifespan both in healthy conditions and also after brain lesions. The process where the brain adapts to environmental challenges and disease is referred to as plasticity [15, 16]. This process was first demonstrated by neuroanatomist Michele Vincenzo Malacarne in 1783 when he intensively trained one in each pair of two birds and two dogs from the same clutch of eggs and litter respectively [15]. The external environment surrounding animals can be considered as static for a short period of time but will become dynamic over a long time. Animals essentially learn quickly about new stimuli to adapt to such environments when it changes, so also, the plasticity that occurs at the neural pathways and continuously changes with respect to internal and external stimuli [17, 18].

Plasticity is an important part for neural malleability at the cells and circuits level in the brain. Neural plasticity can serve multiple functions, such as been *homeostatic* in nature for excitement within a network, it could also be *mnemonic* to form the basis of the memory and lastly, been *metaplasticity* [18].

One important form of plasticity is indicated across sensory modalities, however, a large part of the human brain neurons are present at birth, therefore plasticity and associated learning are expected to occur early in life. The brain needs to be plastic enough to acquire new knowledge and memories but stable enough to retain them over time. This balance is known as the plasticity-stability dilemma [19, 20]. Humans have amazing ability to adapt by efficiently gaining new skills, transforming them to new experiences, and recalling and transferring them across several areas where they are needed. It is also true that humans have the capacity to forget gradually some previously learnt information at some point when they get older. Therefore, learning of new information rarely affect consolidated knowledge in human [1, 21]. Stability-plasticity dilemma is the degree whereby a system must be inclined to integrate and learn novel skills and, most of all, how these learning processes can be rewarded by internal mechanisms which stabilize and modulate neural activity just to avert catastrophic forgetting [1]. Artificial neural networks gain their principal structure by sensorimotor experiences, from the imitation of human brain which is mainly plastic during the crucial phase of early development [16]. Sensorimotor skill learning, like any other form of learning, happens through the general mechanism of experience-dependent synaptic plasticity. When new skill is learned via general training, synapses in the brain are revised to form a lasting motor memory of that skill learned [22].

Stability-Plasticity positioned at multiple brain areas are regulated by the mechanisms of neurosynaptic plasticity. Neurosynaptic plasticity mechanisms is such that it protects knowledge about previously learned tasks from forgetting, by decreasing the rates of synaptic plasticity. However, there are two types of plasticity needed for a stable continual process: Hebbian Plasticity [23] and Compensatory Homeostatic Plasticity [24]. When used together, both Hebbian learning and Compensatory Homeostatic Plasticity stabilize neural cells to shape the optimal patterns of experience-driven connectivity, integration, and functionality in a network [1, 16, 24].

Neurosynaptic plasticity is an important attribute in the brain because it produces physical changes in the neural structure and allows us to learn, remember, and adapt to any changing environments [16] as well as activity-dependent synaptic plasticity in learning and memory formation. Synaptic plasticity was first discovered in the hippocampus of the human brain in the early 1970s. It was concluded that an increase in the strength of the synaptic input of the stimulated connections only is produced by repeated, near-synchronous activation of both pre- and post-synaptic neurons [14] and this process is known as Long-Term Potentiation (LTP). These characteristics of synaptic plasticity suggests its role in learning new skills as well as being an information storage device [25].

However, memories may not be properly stabilized if synapses are easily bendable and in such state of perpetual flux, old learning can easily be overwritten by new learning. Hence, for any learning system, there is essentially constraint between the competing requirements of stability and plasticity [22].

B. The Hebbian Synaptic Plasticity

The brain can adapt to a changing environment and as well as providing important insights into the shape of cortex's connectivity and function. It has been shown that while fundamental designs of connectivity in the visual system are noticeable at early development, normal visual input is essential for the accurate development of the visual cortex [26]. Donald Hebb in 1949 was the first to propose the theory describing and explaining the mechanisms of synaptic plasticity in the adaptation of neurons to external stimuli. Hebb postulated that the connection between two neurons is strengthened, when one neuron pilots the activity of another neuron [27]. In the following years, Hebb's idea has been interpreted to the weight changes among nodes of a single layer perceptron in ANNs based on coincidence or the product of pre- and postsynaptic activity mimicked from the brain neurons, thereby altering the connection of neurons into changes relative to the coactivity of the input and output nodes in ANNs [14]. Thus, considering Hebb's theory from an ANN's standpoint, after a network has been trained using backpropagation successfully, the synapses between neurons that synchronous fires a given input are made stronger for as long as it takes, to maintain and improve its outputs [27, 28].

A simple formula for Hebbian plasticity considers a change in the synaptic weight w and it is updated as the product of the activities in pre-synaptic x and post-synaptic y with learning rate η is given as [1]:

$$\Delta w = x \cdot y \cdot \eta \tag{1}$$

Yet, Hebbian plasticity is unstable while alone, but depended on and requires compensatory mechanisms to stabilize its learning process. This is attainable by enhancing Hebbian plasticity with some constraints like upper limits on specific synaptic weights or regular neural activity, which can only be done by homeostatic plasticity [29, 27]. Homeostasis plasticity is also referred to as a compensatory process that stabilizes the neural firing rates in the brain [24]

III. Overcoming Catastrophic Forgetting With Continual Learning Algorithm

Catastrophic forgetting problem can occur in different ways. One way is between mini-batches when using stochastic gradient descent methods during the general training processes. Another way is the degradation of the generalization performance of a network [12, 30]. Similar to the continual learning methods, in Stochastic Gradient Descent (SGD) optimization, every mini-batch can be thought of as a mini-task offered sequentially to the network. In this context, the interest is describing the changes in the learning of the neural networks by analysing examples of forgetting events [14]. This happens when task that have been learned and correctly classified at some time t in the optimization process are afterward misclassified at a time $t' > t$ [31]. It should also be noted that catastrophic forgetting occurs to ANN models including SOMs as well as Deep Neural Networks, for example Transfer Learning in DNN [32].

Typically, the current approaches to overcome catastrophic forgetting in ANN have concurrently made data available from tasks during training. By passing in data from several tasks while training and learning, forgetting is prevented. This is attributed to the fact that the weights of the network can be mutually optimized for high performance on all training tasks. This case is frequently referred to as the multitask learning, and a good example can be seen in reinforcement learning method where a successfully trained single agent can be used to play many Atari games effectively. If data are introduced to the network sequentially, multitask learning can only be used if tasks are recorded by an episodic memory system and replayed during training to the network [19, 33]. However, this method can be impractical when dealing and learning a large number of tasks, as large number of memories would be required to stored and replayed, likewise been related to number of tasks [4, 14, 29].

A. Continual Learning Basics

Continual Learning is the basic step towards AI, because it permits an intelligent agent to continuously adapt to changes that occur in data and tasks. Nevertheless, there are some consequences during learning for both supervised and unsupervised learning. For example, when data are not properly represented or there is a mistake in the input distribution, a model can overfits the recently seen data, which is something continual learning systems aim to address [34, 35].

A series of desiderata are used to defined Continual Learning in practice which includes Firstly, online learning meaning that learning can occur at every moment, with no permanent tasks or datasets and with no clear boundaries/restrictions between tasks. Secondly,

forward/backward transfer of model from existing tasks to new tasks with the possibility of the new task improving the performance of older tasks. Furthermore, resistance to catastrophic forgetting, that is, new learning task does not degrade the performance on previous data, and lastly, there should be no direct access to previous tasks but be able to retain it [34, 35].

An infinite sequence of data is considered for a general continual learning setting, where at each timestep t the network accepts a new data $\{x_t, y_t\}$ to draw a non independent and identically distributed. from an existing distribution P that could by itself experience some rapid or gradual changes. The key goal is to learn a function F parameterized by θ that can minimize a predefined loss $\mathscr{L}$ on the new data without interfering on existing tasks and also with the possibility of improving on the tasks that were learned previously [34]:

$$\theta^t = \underset{\theta,\xi}{\operatorname{argmin}}\, \mathcal{L}(F(x_t;\theta), y_t) + \sum \xi_i \qquad (2)$$

Such that:

$$\mathcal{L}(F(x_t;\theta), y_t) \leq \mathcal{L}(F(x_t;\theta^{t-1}), y_t) + \xi_i, \qquad (3)$$

$$\xi_i \geq 0\,;\ \forall i \in [0..t-1] \qquad (4)$$

Where x_t is the input, y_t is the ouput and $\xi = \{\xi_i\}$ is the slack variable that allows some constraints to be violated like small increase in loss from previous tasks [34].

Some strategies have been designed for continual learning, which are: Firstly, the progressive/architectural strategy. Architectural strategy can be used to incrementally builds a network's structure for every single task being processed. In addition, it also tries to copy and re-use as much as possible the attributes of the previous model in the process. The second strategy is known as rehearsal methods, since it keeps a memory of data analyzed on previous tasks and continues to retrain the network on this memory to maintain its performance. And the third approach is regularization. Regularization strategy tries to re-use a single neural network, which is by including a few regularization penalties to alleviate the behaviour of the network with respect to previous tasks [1, 20]. Usually, rehearsal and progressive strategies, performs very well but always declines as the number of tasks increase, and might require a high computational power. With some differences from the first two approaches, the implementation of regularization strategy is quite simple, they require little memory, but it performance might not be up to that of rehearsal methods [7]. One main problem encountered when applying regularization strategy is determining what task best represents the behaviour of the network and, this can lead to the form of regularization penalty that would be taken [7].

Recently, a lot of attention has been shifted to the idea of using regularization function to fit the existing task for learning a new task in a network. This method can be understood as an approximation of sequential Bayesian. Some distinctive examples of this regularization approach include the elastic weight consolidation [4] and learning without forgetting [21].

B. A Review of Some Popular Continual Learning Strategies

Several algorithms have been proposed so far to mitigate catastrophic forgetting in neural networks and few are reviewed in this paper:

[4] proposed an algorithm that performs operation like synaptic consolidation used on the brain on ANNs by constraining some important parameters to stay close to their old values. This algorithm is known as Elastic Weight Consolidation (EWC).

In EWC, the performance in task A is protects by constraining its parameters to stay in a region of low error just for task A to be positioned mainly around θ_A^*. This constraint is implemented as a quadratic penalty and can exist as a spring anchoring the parameters to the previous solution, hence been called elastic. However, all parameters should not have the same stiffness of this spring, but it must be larger for parameters that are very much affected by the performance in task A.

To further explain the optimal choice of constraint and weights, the neural network training is considered from a probabilistic viewpoint using Bayes' rule and also noting that the log probability of the data $\mathcal{D}$ given the parameters log p $(|\theta)$ from the Bayes' rule equation is simply the negative of the loss function $-\mathscr{L}(\theta)$ [4]:

$$\log p(\theta\,|\,\mathcal{D}) = \log p(\mathcal{D}_B\,|\,\theta) + \log p(\theta\,|\,\mathcal{D}_A) - \log p(\mathcal{D}_B). \qquad (5)$$

The key to implement EWC is that all the information about task A, must have been accepted into the posterior distribution $p(\theta|\mathcal{D}_A)$. The true posterior probability is inflexible, so the posterior distribution was approximated as a gaussian distribution with average specified by parameters θ_A^* and a diagonal precision specified by the diagonal of the Fisher information matrix F. This matrix F is used because it has three key characteristics: Firstly, it is equivalent to the second derivative of the loss near a minimum. Secondly, it can be computed from first-order derivation alone and it is quite easy to compute even for big models. Thirdly, a positive semidefinite is guaranteed. Therefore, the loss $\mathscr{L}$ minimized in EWC is computed as:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \qquad (6)$$

where $\mathcal{L}_B(\theta)$ is the loss on task B only, λ determines how important the existing task is compared with the new task, and i gives labels to each parameter. However, when considering a third task C, the EWC algorithm might try to maintain the network parameters value close to the learned parameters of previous tasks A and B. This can be imposed either with two separate penalties or as one by observing that the sum of two quadratic penalties is itself a quadratic penalty [4]. Because computing over the diagonal of fisher requires summation of all possible outputs, thus EWC has complexity linear in the number of outputs, limiting its application to low-dimensional output spaces [10].

A simple structural regularizer that can be computed online was introduced by [10] and also implemented locally at each synapse/parameter (weights and biases). The authors developed an algorithm which can keep track of an *importance measure* ω_k^{μ}.

Considering the change in loss function $\mathscr{L}$ for an infinitesimal parameter update $\delta(t)$ at time t, where $\theta(t)$ is the trajectory in parameter space between task A and task B, and g is the gradient can be written as [10]:

$$\mathcal{L}(\theta(t) + \delta(t)) - \mathcal{L}(\theta(t) \approx \sum_k g_k(t)\delta_k(t) \quad (7)$$

However, to calculate the change in the loss over the whole trajectory, all the infinitesimal, and the changes are summed over, which amount to computing the path integral of the gradient vector from the start time t_0 to the end time t_1 and also the loss between the end and the start point $\mathscr{L}(\theta(t_1)) - \mathscr{L}(\theta(t_0)$ [10]):

$$\int_{t^{\mu-1}}^{t^{\mu}} g\big(\theta(t)\big) \cdot \theta'(t)dt = \sum_k \int_{t^{\mu-1}}^{t^{\mu}} g_k(\theta(t))\theta_k'(t)dt \equiv -\sum_k \omega_k^{\mu} \quad (8)$$

The authors tried to solve the problem of minimizing the total loss function summated on all tasks, $\mathscr{L} = \sum_{\mu} \mathcal{L}\mu$, with no contact to the loss function $\mathscr{L}\mu$ of the past training except the new task μ at any given time but with this minimization come catastrophic forgetting which led to a drastic weight changes between the old task and the new task ($\nu < \mu$) while training task μ. To avoid this problem, they introduced quadratic surrogate loss which approximates the summed loss function of old task $\mathscr{L}\nu$ ($\nu < \mu$). The implication of using the quadratic surrogate loss for training instead of the actual loss function, is that the final parameters will remain the same and change in loss during the training process [10]:

$$\tilde{\mathcal{L}}_{\mu} = \mathcal{L}\mu + c\sum_k \Omega_k^{\mu}(\tilde{\theta}_k - \theta_k)^2 \quad (9)$$

Where c is the dimensionless strength parameter, $\tilde{\theta}_k$ is the reference weight at the end of previous task, and Ω_k^{μ} is the per-parameter regularization strength. The equation 9 can only achieve two tasks.

Although [10] algorithm is similar to EWC in [4] in that more importance synapses are strongly directed towards the reference weight, however, the method computes the importance measure online including all the learning trajectory [10], considering that, EWC is about the point estimate of the diagonal of the Fisher information matrix at the final synapse values, that has to be calculated during a separate stage at the end of each task [4].

Inspired by Hebbian learning in neuroplasticity, [5] proposed Memory Aware Synapses (MASes). Unlike previous proposed research on synapses, their continual learning method can learn using unlabelled data and in online manner. The sensitivity of the output function was the main focus and not the loss while estimating importance weights for the network parameters. After the model has been trained on the approximation F of the true function $\bar{F}$, the function F output was preserved, and its sensitivity was measured for changes. A small perturbation δ in the parameters θ results in a change in the function output that can be approximated by

$$F(x_k;\theta + \delta) - F(x_k;\theta) \approx \sum_{i,j} g_{ij}(x_k)\delta_{ij} \quad (10)$$

Where $g_{ij}(x_k)$ is the gradient of the function learned and δ_{ij} is the change in parameter θ_{ij}. But the goal is to preserve the prediction of F. To do this, the gradients of all data point were accumulated to obtain importance weight Ω_{ij} [5]

$$\Omega_{ij} = \frac{1}{N}\sum_{k=1}^{N} \|g_{ij}(x_k)\| \quad (11)$$

Where N is the total number of data points. However, when function F is multi-dimensional, the gradients for each output can be computed by using the squared l_2 norm the learned function output. The importance is measured by the sensitivity of the squared l_2 norm over learned function output. To learn a new task, a new loss $\mathcal{L}_n(\theta)$, and a regularizer for penalty to change important parameters (high Ω_{ij})

$$\mathcal{L}(\theta) = \mathcal{L}_n(\theta) + \lambda\sum_{i,j} \Omega_{ij}\big(\theta_{ij} - \theta_{ij}^*\big)^2 \quad (12)$$

Where λ is the regularizer's hyperparameter and θ_{ij}^* is the previous network parameter [5].

[36] explicitly address the diagonal assumption made by EWC algorithm in [4]. They assumed that if the Fisher Information Matrix is not diagonal, EWC might fail to stop the network from drifting from "good parameter space". The proposed Rotated Elastic Weight Consolidation approach is based on rotating the parameter space of a network, that is, re-parameterization of the parameter space θ, in a way that the output of the forward pass is not changed, while the computed Fisher Information Matric from the gradients during the backpropagation is approximately diagonal [4]. To obtain reparameterization, the rotation matrix is computed using Singular Value Decomposition (SVD) even though computing SVD on a very large matrices is quite expensive. Applying chain rule on FIM and computing using SVD, the following equation was obtained where θ' is W' the new rotated weight matrix [36]:

$$E_{x\sim\pi}[XX^T] = U_1S_1V_1^T \quad (13)$$

$$E_{\substack{x\sim\pi \\ y\sim p}}\left[\left(\frac{\partial L}{\partial y}\right)\left(\frac{\partial L}{\partial y}\right)^T\right] = U_2S_2V_2^T \quad (14)$$

$$W' = U_2^T W U_2^T \quad (15)$$

The results obtained with rotated EWC is outstandingly more real at overcoming catastrophic forgetting in sequential task learning problems [36].

Variational continual learning [37] and Continual Learning with Adaptive Weights (CLAWs) [38] are another regularization strategy [37]. In [37], Bayesian inference provides a fundamental framework for continual learning with its algorithm where the posterior of the model parameters is learned and updated continually from a sequence of datasets. To achieve this algorithm, online Variational Inference (VI) was merged with Monte Carlo VI for neural networks to produce Variational Continual Learning (VCL). In addition, VCL was enhanced to contain a small episodic memory by the combination VI with the coreset data summarization process. The coreset can be compared to an episodic memory that holds important information from previous tasks, where the algorithm can go

back so as to refresh its memory of these important information.

Similar to [37], in [38], their approach is based on probabilistic modelling and variational inference [37]. But rather than strictly dividing the architecture into shared and task-specific parts, the approach adapts the contributions of each neuron using Gaussian distribution for the adaptation as the probabilistic model and afterwards the adaptation parameters are included within the variational parameter in Monte Carlo VI.

[39] is another form of regularization approach to continual learning. The method presents the option to control the stability and compactness of the learned task. This makes this method also agreeable for network compression applications and online learning. They proposed a task-based hard attention mechanism that can preserves learning from an existing task without affecting the learning of a new task. As well as learning tasks with the binary attention. A task can also be learned over gated task embeddings, using backpropagation and minibatch SGD. Some attributes of hard attention task are: It can store, as well as maintain a lightweight structure. Secondly, the task is learned instead of a heuristic approach or rule-driven. Thirdly, the mask is not necessarily binary, and this might be useful if the weights need to be re-used for learning other tasks, i.e., to overcome catastrophic forgetting.

[21] proposes Learning without Forgetting method which compose of Convolutional Neural Networks (CNN) and this approach can be perceived as combination of Distillation Networks (transfer of information from a large to a small model) [40] and fine-tuning. The main idea here is only used on new task data for training the network. The network learns from parameters that works fine on old task and uses this information to train the new tasks without the use of data from previous tasks.

To achieve this, the responses y_o on each new task object from the original network for outputs on the old tasks (defined by shared parameters θ_s and task-specific θ_o) were recorded, then the network was trained for the loss to be minimize for all tasks and regularization R by using SGD. To define the loss for a new task, the output $\hat{y}_n$ was merged with the one-hot ground truth y_n [21]:

$$L_{new}(y_n, \hat{y}_n) = -y_n \cdot log\ \hat{y}_n \tag{16}$$

To transfer the known, knowledge distillation loss must be introduced to the network [21]

$$L_{new}(y_n, \hat{y}_n) = -H(y_o', \hat{y}_o') \tag{17}$$
$$= -\sum_{i=1}^{l} y_o'^{(i)} \log \hat{y}_o'^{(i)}$$

[41] provided an architectural strategy algorithm called Reinforced Continual Learning (RCL). It comprises of three networks: *value network*, *controller*, and *task network*. The controller is executed as a Long Short-Term Memory network (LSTM) or as Recurrent Neural network to generate policies and determine how many filters/nodes will be added to each task. The value network was designed as a multilayer perceptrons/fully-connected network, that approximates the value of the state [41]:

$$\pi(a_{1:m}|s;\theta_c) = \prod_{i=1}^{m} p_{t,i,a_i'} \tag{18}$$

Where θ_c the controller network's parameter.

However, the task network, on the other hand, can be any network of interest for solving any task, for example object detection or image classification. Furthermore, RCL adaptively expands the network when a new task arrives, while using stochastic gradient descent with η as the learning rate [41]:

$$\min_{W_t/W_{t-1}^a} L_t(W_t; D_t) \tag{19}$$

$$W_t/W_{t-1}^a \leftarrow W_t/W_{t-1}^a - \eta \nabla W_t/W_{t-1}^a L_t \tag{20}$$

[42] propose Incremental Moment Matching (IMM) framework from Bayesian Neural networks, Here moments of posterior distribution which are trained on old and new task are matched together in an incremental way using Gaussian distribution. Considering that the objective is to determine the ideal parameter $\mu_{1:k}^*$ and $\Sigma_{1:k}^*$ of the gaussian approximation function $q_{1:k}$ from the posterior parameter of the kth task (μ_k, Σ_k), two different moment match method can be used: mean-IMM and mode-IMM. Mean-IMM finds the average of the parameters of two networks for both old and new task [42]:

$$\mu_{1:k}^*, \Sigma_{1:k}^* = \underset{\mu_{1:k}/\Sigma_{1:k}}{\text{argmin}} \sum_k^K \alpha_k KL(q_k \| q_{1:k}) \tag{21}$$

Mode-IMM is an alternative form of mean-IMM. It merges the parameter of old and new network using Laplacian approximation of the posterior of gaussian distribution [42]:

$$\log q_{1:k} \approx \sum_k^K \alpha_k \log q_k + C = -\frac{1}{2}\theta^T \left(\sum_k^K \alpha_k \Sigma_k^{-1}\right)\theta + \left(\sum_k^K \alpha_k \Sigma_k^{-1}\mu_k\right)\theta + C' \tag{22}$$

The result obtained from the experimental with both IMM on shows that Mode-IMM performed better than mean-IMM and other comparative models in the various dataset. The limitation is that IMM performance decreases with more complex dataset.

[43] introduced a model architecture called Progressive Neural Network (PNN) that support transfer of knowledge across sequence of tasks particularly in reinforcement learning. Progressive network makes use of transfer learning by retaining a pool of knowledge through training of an agent from a previous task, and also having the ability to transfer that knowledge to another agent to improve convergence speed. After PNN finishes training of a previous task, its parameter θ' is frozen when switching to the second task, after which another parameter θ is instantiated [43]:

$$h_i^k = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right) \tag{23}$$

Where $W_i^{(k)}$ is weight matrix, $U_i^{(k:j)}$ in the lateral connection, and f is an element-wise non-linearity.

PNN is robust to harmful features learned in incompatible tasks by the RL agent. A major downside of PNN is the growth in number of parameters with the number of tasks.

TABLE 1: SUMMARY OF SOME OTHER DIFFERENT APPROACHES TO ALLEVIATE CATASTROPHIC FORGETTING

Authors, Year, and Country:	Proposed Methods/Algorithms and Strategies/Approaches:	Important Note:	Limitation
[4] 2018, United State of America	Elastic Weight Consolidation (EWC), Regularization	The EWC uses only one network with static network capacity and nominal computational overhead which has a low computational cost.	EWC are very sensitive to the diagonal approximation of the FIM used in practice because of the large size of a full FIM and costly to compute weights for regularization penalty
[10] 2017, Australia	Synaptic Intelligence using Quadratic Surrogate Loss SI, Regularization	The method computes the per-synapse consolidation strength in an online manner and over an entire learning trajectory in parameter space and individual synapses act as higher dimensional dynamical systems.	SI can only learn importance weights during training, which leads to lack of adaptation to some particular subset.
[5] 2018, Germany	Memory Aware Synapses (MAS), Regularization	The important parameters (high Ω_{ij}) can be reused, through model sharing, which is only possible with a penalty when changing the parameters.	It is limited by brittleness caused by representation drift mostly common to regularization methods.
[36] 2018, Italy	Rotated Elastic Weight Consolidation, Regularization	The evaluation of the experiment on various learning tasks shows that the approach performed well compared to the standard EWC.	Rotated EWC can also suffer from brittleness caused by representation drift.
[39] 2018, Sweden	Hard Attention to Task (HAT), Regularization	HAT presents the option to monitor the used network capacity throughout different tasks and layers and it has only two hyperparameters, and are both referred to as the stability and compactness of the learned task.	HAT gradually declines in classification accuracy during training with no signs and hope of ever increasing
[21] 2016, Netherlands	Learning without Forgetting, Regularization	The method is only proposed for convolutional neural networks. It is a hybrid of knowledge distillation and fine-tuning.	Additional memory and computation are needed in LFL to compare activations
[44] 2020, Italy	Embedding Regularization for continual learning, Regularization	ER develops an efficient way to regularize the behaviour of the network by acting on its internal embeddings, i.e., the activations of one or more layers closer to the exit.	In ER, when the memory grows, the required training time also increases.
[45] 2020, Germany	Bayesian Neural Networks for Non-Stationary Data, Regularization	It makes use of Bayesian forgetting and a Gaussian diffusion process for adaptation to non-stationary data, leading to a better predictive performance	Bayesian neural networks with a uni-modal approximate posterior often find poor local minima if the dataset is small and models are complex, which is especially challenging in situation where data are streamed
[46] 2018, Canada	FearNet, Architectural	The basolateral amygdala is used to determine which memory system to use for recalling task and it is more memory efficient.	FearNet can suffer from recall when the number of classes to learn is high.
[41] 2019, Germany	Reinforced Continual Learning (RCL), Architectural	RCL explores the best neural network architecture for each upcoming task.	The training time of RCL is particularly important and high for large networks with more layers
[47] 2020, France	Move-to-Data: Incremental learning approach, Architectural	This approach does not require gradient based optimization	The Move-to-Data method is limited to only one fully connected layers
[33] 2017,	Gradient Episodic Memory (GEM),	The advanced memory management was not investigated, and the iteration	It may be less scalable and require many observations and complex generative model to

United State of America	Regularization, Rehearsal	requires one backward pass per task, which increases the computational time.	represent realistic tasks, and the Effective prioritized replay remains an unsolved problem
[48] 2019, United States of America	Average Gradient Episodic Memory (A-GEM), Regularization, Rehearsal	It is about 100 times faster and memory is 10 times less required; compared to regularization-based approaches, it achieved a significantly high average accuracy	The model is plausibly a little incremental over GEM
[49] 2017, United State of America	Incremental Classifier and Representation Learning (iCaRL), Regularization, Rehearsal	It comprises of 3 major components: a nearest-mean-of-exemplars, a herding to prioritize exemplars, and a representation learning step, and It learns strong classifiers and data representation at the same time	iCaRL's performance is still lower than what other systems achieve when trained in a batch setting,
[50] 2020, Virtual Conference	Functional Regularised Continual Learning (FRCL): Gaussian processes, Regularization, Rehearsal	When viewed from the regularisation perspective, it regularises the functional outputs of the neural network, while when viewed from a rehearsal method perspective, a principled way is provided for compressing data from previous task, by optimizing the selection of inducing points.	It suffers from a fixed memory buffer in which case the summaries of all the previous seen tasks need to be compressed one needs into a single summary.
[38] 2020, United State of America	Continual Learning with Adaptive Weights (CLAW), Regularization, Rehearsal and Architectural	It is based on variational inference from VCL.	This approach did not actually compare their result to other to VCL, every other.
[37] 2018, Canada	Variational Continual Learning, Regularization, Rehearsal and Architectural	VCL is most suitable for efficient model fine-tuning in sequential decision-making problems, and can be applied to generative model and discriminative model.	VCL also suffers from brittleness caused by representation drift

IV. OUR NOVEL RESEARCH PROPOSAL

According to [51], the human brain act as information filters. From the inward region of the brain (hippocampus), when new information is taken in, old irrelevant information is filtered out and the updated information are stored for long term retrieval and decision making. The unused pieces are however deleted to create space. It is called forgetting in neuroscience. Forgetting occurs when the synaptic connection between neurons weakens and are eliminated over time [51]. To effectively adapt, humans need to strategically forget, so also the need to forget in ANN for a successful continual learning.

The previously discussed works have approached the problem of catastrophic forgetting, specifically the continual learning with valuable strategies and algorithms. Most of works, tackled ways to achieve continual learning, but left out the aspect of forgetting. Forgetting some older knowledge is essential to accommodate information from new data. The novel idea here is to build a network to deploy learning and the same network will be re-purposed to learn a new task, forgetting some specific information that is irrelevant. Self-Organizing Map (SOM) will be used for this purpose of forgetting. The algorithm will be in such a way that the network learns and update in the opposite direction which will lead to forgetting in the SOM. In addition, we will measure 3 different performance metrics, which are: The Average Accuracy, The Backward Transfer, The Cumulative Backward Transfer Scores of Forgetting.

Forgetting can be beneficial in some cases: 1) it prevents overfitting to specific features and can improve generalization, 2) forgetting outdated data can enhance flexibility of decision made from learning with new data. [51]The main goal of this novel research is to control the forgetting process during learning to protect some vital information and in the process minimizing accuracy loss.

V. CONCLUSION AND FUTURE WORK

Continual Learning is the fundamental step towards AI, because it permits an intelligent agent to continuously adapt to a dynamic environment, a distinctive characteristic of natural intelligence. The goal for continual learning is to acquire knowledge across tasks particularly through model sharing and having a single model that can perform well on all the tasks, however, there is one challenge to achieving this, which is catastrophic forgetting of previous task learned, in the process of learning new task. In this paper, we presented continual learning in advanced biological animals and Artificial intelligent agents. We discussed plasticity-plasticity dilemma and taking it a little further, we talked about Hebbian plasticity and compensatory homeostatic plasticity process of learning and memory formation that occurs in the brain. Despite significant advancement, most of the currently proposed algorithms for continual learning are still far from providing a robust,

flexible, and scalable approach displayed in biological animals. However, we presented a state-of-the-art overview of several algorithms, from the most popular and recent literature on continual learning, where some significant progress has been made to tackle catastrophic forgetting in ANN. On top, we used a table to summarize these algorithms which included: the type of strategy/approach, the dataset used for the performance evaluation and some key notes about these algorithms. In addition, we introduced our novel research proposal on intentional forgetting, which is such that an intelligent system will chose to forget some irrelevant or old information when learning a new task. Evaluating with different dataset, we will measure different performance metrics with the new proposed algorithm.

REFERENCES

[1] German, P., Ronald, K., Jose, P., Christopher, K., & Stefan, W. (2019). Continual lifelong learning with neural networks: A review. *ScienceDirect- Neural Networks, 113*, 54-71. doi:10.1016/j.neunet.2019.01.012

[2] Z. Chen and B. Liu. (2018). *Continual Learning and Catastrophic Forgetting.* Morgan & Claypool Publishers. doi:10.2200/S00832ED1V01Y201802AIM037

[3] Nicolas, M., Gregory, G., & David, F. (2019). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences, 115*(44), E10467-E10475. doi:10.1073/pnas.1803839115

[4] Kirkpatrick, J., Pascanu, R., Rabinowitza, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2018). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*(13), 3521–3526. doi:10.1073/pnas.1611835114

[5] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory Aware Synapses: Learning what (not) to forget. *15th European Conference on Computer Vision ECCV'18*. doi:10.1007/978-3-030-01219-9_9

[6] Vincenzo, L., Davide, M., & Lorenzo, P. (2019). Fine-Grained Continual Learning. *Cornell University: Arxiv.org*, 1-12. Retrieved from https://arxiv.org/abs/1907.03799

[7] Pomponi, J., Scardapane, S., Lomonaco, V., & Uncini, A. (2020). Efficient Continual Learning in Neural Networks with Embedding Regularization. *ScienceDirect - NeuroComputing, 297*, 139-148. doi:10.1016/j.neucom.2020.01.093

[8] Michael, M., & Neal, C. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *ScienceDirect- The Psychology of Learning and Motivation, 24*, 109-165. doi:10.1016/S0079-7421(08)60536-8

[9] Andrew, P., Ryan, C., Patrick, M., Stephen, B., Renee, E., & MarioAguilar-Simon. (2019). Uncertainty-based modulation for lifelong learning. *ScienceDirect - Neural Networks, 120*, 129-142. doi:10.1016/j.neunet.2019.09.011

[10] Zenke, F., Poole, B., & Ganguli, S. (2017). Continual Learning Through Synaptic Intelligence. *Proceedings of the 34 th International Conference on Machine Learning, PMLR 70, 70*, pp. 3987–3995. Sydney, Australia. Doi:10.5555/3305890.3306093

[11] De, L. M., Rahaf, A., Marc, M., Sarah, P., Xu, J., Ales, L., . . . Tinne, T. (2019). Continual learning: A comparative study on how to defy forgetting in classification tasks. *Cornell University: arxiv.org*, 26. doi:10.1109/TPAMI.2021.3057446

[12] Heechul, J., Jeongwoo, J., Minju, J., & Junmo, K. (2016). Less-forgetting Learning in Deep Neural Networks. *IEEE*, 1-5. Retrieved from arXiv:1607.00122

[13] M.Stark, S., & E.L.Stark, C. (2016). *Chapter 67 - Introduction to Memory.* Academic Press. doi:10.1016/B978-0-12-407794-2.00067-5

[14] Magee, J. C., & Grienberger, C. (2020). Synaptic Plasticity Forms and Functions. *Annual Review of Neuroscience, 43*, 95-117. doi:10.1146/annurev-neuro-090919-022842

[15] Quentin, R., Awosika, O., & Leonardo, G. C. (2019). Plasticity and recovery of function. *ScienceDirect: Handbook of Clinical Neurology, 163*, 473-483. doi:10.1016/B978-0-12-804281-6.00025-2.

[16] Wickliffe, C. A., & Robins, A. (2005). Memory retention – the synaptic stability versus plasticity dilemma. *ScienceDirect*. doi:10.1016/j.tins.2004.12.003

[17] Junichiro, H., Junichiro, Y., & Shin, I. (2006). Balancing Plasticity and Stability of On-Line Learning Based on Hierarchical Bayesian Adaptation of Forgetting Factors. *ScienceDirect- NeuroComputing, 69*(16-18), 1954-1961. doi:10.1016/j.neucom.2005.11.020

[18] Sehgal, M., Song, C., L.Ehlers, V., & R.MoyerJr., J. (2013). Learning to learn – Intrinsic plasticity as a metaplasticity mechanism for memory formation. *Neurobiology of Learning and Memory, 105*, 186-199. doi:10.1016/j.nlm.2013.07.008

[19] Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., & Ranzato, M. (2019). On Tiny Episodic Memories in Continual Learning. *Cornell University*, 1-15. Retrieved from arXiv:1902.10486

[20] Lomonaco, V. (2019). *Continual Learning with Deep Architectures.* Bologna: Department of Computer Science and Engineering, University of Bologna.

[21] Li, Z., & Hoiem, D. (2016). Learning without Forgetting. *The 14th European Conference on Computer Vision ECCV2016.* doi:10.1109/TPAMI.2017.2773081

[22] Ajemiana, R., D'Ausilio, A., Moorman, H., & Bizzi, E. (2013). A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits. *Proceeding of the National Academy of Sciences of the United States of America*, 5078-5087. doi:10.1073/pnas.1320116110

[23] D.O., Hebbs. (1949). *The organization of behavior; a neuropsychological theory.* Psychology Press. doi:10.1007/978-3-642-70911-1_15

[24] Zenke, F., & Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of The Royal Society B Biological Sciences, 372*(1715). doi:10.1098/rstb.2016.0259

[25] Martin, S. J., Grimwood, P. D., & Morris, R. G. (2000). Synaptic Plasticity and Memory: An Evaluation of the Hypothesis. *Annual Review of Neuroscience, 23*, 649-711. doi:10.1146/annurev.neuro.23.1.649

[26] Nicolas Y. Masse, Gregory D. Grant, and David J. Freedman. (2019). Alleviating Catastrophic Forgetting using Context-Dependent Gating and Synaptic Stabilization. *Cornell University - arxiv.org*. doi:10.1073/pnas.1803839115

[27] Steven J.Cooper. (2005). Donald O. Hebb's synapse and learning rule: a history and commentary. *Neuroscience and Biobehavioral Reviews, 28*, 851-874. doi:10.1016/j.neubiorev.2004.09.009

[28] Abraham, W. C., Jones, O. D., & Glanzman, D. L. (2019). Is plasticity of synapses the mechanism of long-term memory storage? *Nature Partner Journal- Science of Learning, 4*, 9. doi:10.1038/s41539-019-0048-y

[29] German, P., Ronald, K., Jose, P., Christopher, K., & Stefan, W. (2019). Continual lifelong learning with neural networks: A review. *ScienceDirect- Neural Networks, 113*, 54-71. doi:10.1016/j.neunet.2019.01.012

[30] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., & Zhang, B.-T. (2017). Overcoming Catastrophic Forgetting by Incremental Moment Matching. *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, Califonia, USA. doi:10.5555/3294996.3295218

[31] Toneva, M., Sordoni, A., Tachet, d. C., Trischler, A., Bengio, Y., & Geoffrey, J. G. (2019). An Empirical Study of Example Forgetting During Deep Neural Network Learning. *The International Conference on Learning Representations (ICLR) 2019*. Retrieved from arXiv:1812.05159

[32] Fiona M Richardson, Michael S C Thomas(2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Developmental Science*, 371–389. doi:10.1111/j.1467-7687.2008.00682.x

[33] Lopez-Paz, D., & Ranzato, M. (2016). Gradient Episodic Memory for Continual Learning. *Facebook Artificial Intelligence Research*, 1-17. doi:10.5555/3295222.3295393

[34] Rahaf, A. (2019). *Continual Learning in Neural Networks*. Leuvan, Belgium: KU Leuven – Faculty of Engineering Science:. Retrieved from arXiv:1910.02718v2

[35] Pascanu, R., Teh, Y., Pickett, M., & Ring, M. (2018). Continual Learning. *Conference on Neural Information Processing Systems*. Montréal, Canada: NeurIPS.

[36] Liu, X., Masana, M., Herranz, L., Weijer, J. V., Lopez, A. M., & Bagdanov, A. D. (2018). Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting. *International Conference on Pattern Recognition'18*. doi:10.1109/ICPR.2018.8545895

[37] Nguyen, C. V., Li, Y., Bui, T. D., & Turner, R. E. (2018). Variational Continual Learning. *International Conference on Learning Representations (ICLR)*. doi:10.17863/CAM.35471

[38] Adel, T., Zhao, H., & Turner, R. E. (2020). Continual Learning with Adaptive Weights. *The International Conference on Learning Representations (ICLR)*. Retrieved from https://openreview.net/forum?id=Hklso24Kwr

[39] Serrà, J., Surís, D., Miron, M., & Karatzoglou, A. (2018). Overcoming Catastrophic Forgetting with Hard Attention to the Task. *International Conference on Machine Learning (ICML 2018)*. Retrieved from arXiv:1801.01423

[40] Hinton, G., Vinyals, O., & Dean, J. (2014). Distilling the Knowledge in a Neural Network. *NIPS 2014 Deep Learning Workshop: Neural and Evolutionary Computing*. Retrieved from https://arxiv.org/abs/1503.02531

[41] Ju, X., & Zhanxing, Z. (2019). Reinforced Continual Learning. *Cornell University*, 1-10. doi:10.5555/3326943.3327027

[42] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., & Zhang, B.-T. (2017). Overcoming Catastrophic Forgetting by Incremental Moment Matching. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, Califonia, USA. doi:10.5555/3294996.3295218

[43] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., . . . Hadsell, R. (2016). Progressive Neural Network. *Google DeepMind: arXiv:1606.04671*, 1-14. Retrieved from https://arxiv.org/abs/1606.04671

[44] Jary, P., Simone, S., Vincenzo, L., & Aurelio, U. (2020). Efficient Continual Learning in Neural Networks with Embedding Regularization. *ScienceDirect- Neurocomputing, 397*, 139-148. doi:10.1016/j.neucom.2020.01.093

[45] Richard, K., Botond, C., Alexej, K., der, S. P., & Stephan, G. (2020). Continual Learning with Bayesian Neural Networks for Non-Stationary Data. *International Conference on Learning Representations*. Virtual Conference. Retrieved from https://arxiv.org/abs/1910.04112

[46] Kemker, R., & Kanan, C. (2018). FearNet: Brain-Inspired Model for Incremental Learning. *The Sixth International Conference on Learning Representations*. Vancouver, Canada. Retrieved from https://arxiv.org/abs/1711.10563

[47] Miltiadis, P., Jenny, B.-P., Akka, Z., Boris, M., & de, R. A. (2020). Move to-Data: A new Continual Learning approach with Deep CNNs, Application for image-class recognition. *hal-02865878v1f*. Retrieved from https://arxiv.org/abs/2006.07152

[48] Arslan, C., Marc'Aurelio, R., Marcus, R., & Mohamed, E. (2019). Efficient Lifelong Learning with A-GEM. *International Conference on Learning Representations (ICLR)*. New Orleans. Retrieved from https://arxiv.org/abs/1812.00420

[49] Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). iCaRL: Incremental Classifier and Representation Learning. *Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii. doi:iCaRL: Incremental Classifier and Representation Learning

[50] Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, Yee Whye Teh(2020). Functional Regularisation for Continual Learning with Gaussian Processes. *International Conference on Learning Representations*. Virtual Conference. Retrieved from https://arxiv.org/abs/1901.11356

[51] Richards, B. A., & Frankland, P. W. (2017). The Persistence and Transience of Memory. *Cell Press journal, 94*(6), 1071-1084. doi:10.1016/j.neuron.2017.04.037

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 33–42
DOI: 10.15439/2021F72
ISSN 2300-5963 ACSIS, Vol. 26

Improvement of Story-telling Advertisement According to Screenwriting Techniques

Daiki Uehara
Graduation School of Information Science and Engineering
Ritsumeikan University
1-1-1 Nogihigashi, Kusatsu, Shiga, 525-8577, Japan
Email: d.uehara5151@de.is.ritsumei.ac.jp

Fumiko Harada, Hiromitsu Shimakawa
Connect Dot Ltd.
160-2 Saihojicho, Nakagyo-ku, Kyoto, Kyoto, 604-0866, Japan
Graduation School of Information Science and Engineering
Ritsumeikan University
1-1-1 Nogihigashi, Kusatsu, Shiga, 525-8577, Japan
Email: {harada, simakawa}@cs.ritsumei.ac.jp

***Abstract*—The study proposes a method for enterprises to enhance their scenarios for storytelling marketing. Scenarios to attain the sympathy of consumers are hard to write for enterprisers unfamiliar with scenario writing. In this research, we focus on the story development and scene development of the screenwriting technique. This method uses logistic regression to detect expressions lacking in each scene of the scenario. Repeated modification of the scenario according to the indication of the lacked expression leads to scenarios to an advertising document that resonates with the reader. In this study, we verified the effectiveness of the method for specific story development. As a result, it turns out the proposed method can support the creation of scenarios whose scene development is close to the model ones. The method enables enterprisers to refine the scenario into advertising documents that resonate with readers.**

I. Introduction

THE SPREAD of smartphones greatly increases our opportunities to view advertisements on the Internet. Smartphones promote privacy and immerse browsing on advertisements anytime and anywhere without worrying about what others are thinking.

Information on products and services enterprisers want to sell on the Web such as SNS that is called Social Network Services and reputation sites play an important role in their marketing [1]. Positive impressions from actual users of the products and services reduce skepticism of new customers against them, which makes it easier for the customers to get interested in them. It implies the advertisements should include the user's experiences of using the products and services. When enterprisers make advertisements, they should include experiences and impressions of customers actually using the products and services. Incorporation of the customer's view into the advertisement enables the reader of the advertisement to have a simulated experience of using the products and services.

On the other hand, in order for the readers to have simulated experiences with the advertisement, the advertisement scenario needs to attract the reader's interest. Advertisements on SNS and other Web media have been shown to have positive effects on consumers' motivation to buy. Especially, the storytelling marketing method has been away attracting attention to get consumers' empathy using narrative advertising [2] [3]. Its most prominent feature is that readers of the advertisement feel empathy with products and services through a scenario telling efforts of enterprisers who invented them. According to the study by Laer [4], commercial advertisements should use a narrative format to reduce customers' resistance to buying products and services. However, it is almost impossible for enterprisers who have no experience to write scenarios attracting the attention of consumers. On the other hand, a huge cost is necessary to commission professional scenario writers to make an advertisement. Fortunately, there is a book summarizing how to write scenarios. The study refers to a set of know-how explained there as screenwriting techniques. This study proposes a method to support writing narrative scenarios from the enterprisers' actual experiences based on the screenwriting technique. The purpose of screenwriting is to write a scenario through which the main character can get the reader's empathy. Advertisements based on a scenario attaining empathy from readers are posted on SNS so that it would show the enterprisers' efforts in their commercialization. It is expected to get the customers' empathy. If enterprisers follow screenwriting techniques to use appropriate expressions, they can write a scenario in which the reader can empathize with the business. Empathy leads to turning the reader into a customer.

The study focuses that scenarios should use words characteristic to each scene. A lack of words suitable for each scene would prevent readers from getting empathy. The development of scenes varies with a storyline. The proposed method creates a logistic regression model that classifies the scenes of a scenario for a specific storyline. The enterprisers select a storyline that fits their actual experiences to write the scenario for each scene. If there is a lack of the scene words in each scene scenario, the logistic regression model could not classify the scenes correctly. The proposed method detects the lack of expressions for the scenes which are not correctly classified. Through the detection of the lack of expressions, the method can support the enterprisers to modify scenarios that get empathy from readers.

In the experiment, a logistic regression model is created to classify scenes for a specific storyline. From the results of the experiment, the paper discusses how to point out the lack of the scene words. By evaluating the appropriateness of

words characteristic to each scene in the particular storyline, the method has turned out to have the potential to be extended to other storylines. As the enterprisers improve their scenarios, they can improve their scenarios to get the reader's empathy.

The scenario improvement by the enterprisers themselves will significantly reduce advertising costs.

If enterprisers follow the screenwriting techniques, they can write scenarios that can get the readers' empathy even without professional knowledge about scenario writing. This method enables even enterprisers of small businesses to create advertising documents which get the reader's empathy. These advertisements will contribute to business expansion.

II. Creating a Document to Get the Reader's Empathy

A. Importance of Empathy in Marketing

Painful experiences have the property of being easily transmitted to others. Understanding and sharing others' experiences full of effort, we get easier to empathize with them.

There is a marketing method called storytelling marketing, which uses narrative advertising to get the sympathy of consumers. Empathy in storytelling marketing is that the consumer gets emotionally involved with characters in the advertisement. The hard business experience of enterprisers is an experience full of efforts. The enterprisers can make advertisements that are easy to get the sympathy of readers by sending the stories of their experiences full of efforts as stories. Using narrative advertising makes it easier for the consumers to feel sympathy for the characters in the advertisement.

In marketing, it is important to get the empathy of customers. We would easily empathize with hard experiences where someone has commercialized products and services. For this reason, storytelling marketing often uses scenarios that tell the story of the enterprisers' own hard experiences as advertising documents. Empathy toward narrative advertising increases consumers' interest in the product and motivates them to purchase it.

B. Enterprisers with Poor Documentation Knowledge

There are many ways to create narrative advertising documents. Many of these documents are created from scratch. Enterprisers who write narrative advertising documents need to carefully consider the structure of the stories. However, most of the enterprisers who write narrative advertisements have little or no experience in writing stories. For enterprisers unfamiliar with writing scenarios, it is difficult to come up with a narrative structure specific to a storyline.

There are methods that can automatically generate new stories by learning several stories. The method takes the context into account to generate a story. However, the context of each sentence is incoherent because the words and storyline are taken from various kinds of stories. Incoherent sentences make it difficult for readers to understand the content. Such sentences fail to get empathy from the readers because the readers cannot understand the content of the user's document. Throughout the documentation, enterprisers themselves are required to write the contents so that almost all readers can understand it.

By writing the advertising documents for themselves, enterprisers can avoid losing the excitement and authenticity in their actual experience. The emotional changes of the characters are important. The enterprisers are characters in a narrative advertising document. In order to directly convey the enterprisers' hard experiences to readers, the enterprisers themselves should write the narrative advertisement documents. On the other hand, it is difficult for enterprisers themselves to evaluate the completeness of their documents. In many studies, the evaluation of the produced document depends on the reader's feelings. Since each reader's feelings are different, it is not appropriate to use feelings to evaluate a story. Enterprisers should be able to objectively evaluate whether the advertising document gets the readers' empathy. To write a scenario for storytelling marketing that appeals to many readers, we need a method satisfying the following two requirements. First, the method should help enterprisers to create documents with emotional changes so that the readers can understand their hard experiences. Second, the method should enable enterprises themselves to objectively evaluate the completeness of written documents.

C. Screenwriting Techniques

Let enterprisers write narrative scenarios for storytelling marketing to get readers' empathy from their own actual experiences.

Snyder has systematically organized the many famous Hollywood movies to summarize their scenario creation methods into the book, Save-the-cat [5]. This paper refers to the summary as screenwriting. The purpose of screenwriting techniques is to present secrets to writing a story attaining the reader's empathy for the main character. Screenwriting techniques help the enterprisers to write scenarios from their actual experiences to get the readers' empathy.

Snyder examined the narrative structure of many famous Hollywood movies to found the movies have commonalities. The commonalities are a combination of the small number of patterns relating to the overall flow of the storyline and a single pattern relating to the development of the scene in which the characters' experiences happen. The screenwriting techniques by Snyder present the combination of these two patterns. This paper refers to scenarios by experts according to the combination of these two patterns as pro-scenarios.

The pro-scenarios would be compliant with one of 10 storylines. The 10 storylines are shown in TABLE I. The left column specifies the name of the storyline. The content of each story is written in the right column.

The structure of many famous Hollywood movies has turned out that the characters' experiences are expressed by 15 scenes. TABLE II shows the development of the 15 scenes. The first column shows the names of the 15 scenes, while each of them is explained in the second. All scenarios exemplified in the screenwriting techniques use this scene development to represent scenes and their development.

TABLE I
STORYLINE;
COURTESY OF BLAKE SNYDER, SAVE THE CAT!: THE LAST BOOK ON SCREENWRITING YOU'LL EVER NEED, MICHAEL WIESE PRODUCTIONS, 2005

Storylines	Story
Monster in the House	The main character' escape from a monster that appers in a closed environment.
Golden Fleece	The main character who goes on a journey and obtains something important other than what he was initially seeking.
Out of the Bottle	The main character who obtains a mysterious or great power, goes through various experiences, and finally accomplishes something without relying on that power.
Dude with a Problem	The ordinary main character who gets caught up in an extraordinary event.
Rites of Passage	The main character stands at a crossroads in his life, and after suffering, he accepts his true self.
Buddy Love	The main characters are lacking in something, but they make up for their shortcomings and grow together.
Whydunit	When the essence of the story begins to emerge, the view thathad been visible changes completely.
The Fool Triumphant	The main character who is thought to be dumb and incompetent, but unexpectedly triumphs.
Institutionalised	The struggles of the main character who lives in a special family and organization.
Superhero	The main character has a special power and fate, but he suffers because he is special.

TABLE II
AN EXAMPLE OF SCENE DEVELOPMENT;
COURTESY OF BLAKE SNYDER, SAVE THE CAT!: THE LAST BOOK ON SCREENWRITING YOU'LL EVER NEED, MICHAEL WIESE PRODUCTIONS, 2005

Scine	Context
1. Opening Image	Sets the tone, mood and style. Often introduces the main character and their "before" state.
2. Theme State	Someone will pose a question or make a statement (usually to the main character) that is the theme of the movie, the thematic premise. An argument stated, and the rest of the film is the argument laid out, proving or disproving the statement.
3. Set-up	Introduce all the main characters. Introduce every character behaviour that will need to be addressed later on, that will need to change if the hero is to win. Introduce the things the hero needs, or need fixing, or are missing from their life (SHOW them). The world before the adventure starts. The thesis.
4. Catalyst	Something that arrives, a message, an event, that changes things. Bad news, but by the time the adventure is over, it's what leads the hero to happiness. The first moment when something happens.
5. Debate	Hero thinks this is crazy. Should they go? Is it possible? How can they do it? It should ask a question of some kind.
6. Break into two	We leave the old world behind, into the antithesis. No later than page 25 (of a 110 page script). The hero must make the decision themselves to step into Act Two.
7. B Story	A breather. A break from the A story. Carries the theme of the movie. Often "the love story". Often a bunch of entirely new characters, maybe opposites to those from Act One.
8. Fun and Games	Provides "the promise of the premise". Not as concerned with forward progress. Lighter than the rest of the movie. Where many of the trailer's moments are found. Set pieces. A break from the stakes of the story.
9. Midpoint	For the hero it seems like all is won or all is lost. Fun and games are over, back to the story.
10. Bad guys close in	Hardest to write. If all seemed won at the midpoint, now things start to go wrong. Dissent, doubt etc. disintegrate the hero's team and the defeated bad guys/thing regroup and return. There is nowhere for the hero to go.
11. All is lost	The opposite of the midpoint in terms of an "up" or a "down". A false defeat. It might seem like a total defeat. It can help to have something about death here. e.g. someone/something dies or thinks about death.
12. Dark night of the soul	Hero is hopeless, at their lowest point, no one to help them, no ideas.
13. Break into three	The solution is found. Thanks to characters in the B story, the conversations about the theme in the B story, the hero trying to find ways to beat the bad guys in the A story. The two stories meet and intertwine.
14. Finale	Act three. Lessons learned are applied, character tics are mastered, A and B stories end in triumph for the hero, the bad guys are dispatched (in ascending order), the old world is turned over, a new world is created.
15. final Image	The opposite of the opening image. Proof that change has occurred.

All storylines adopt the development of the 15 scenes. However, the contents to be stated in the scenes varies with each storyline. Therefore, the scene developments applicable to all storylines are too abstract for enterprisers to use as a reference for writing scenarios. When enterprisers write narrative scenarios for storytelling marketing, the scene development embodied for each storyline is preferable, because it gives enterprisers more specific information on each scene.

III. IMPROVEMENT OF ENTERPRISER SCENARIO

A. Improvement toward Pro-scenario

When a beginner tries to write a script, it is popular for the beginner to imitate the script of an expert in the way the expert writes the script. The beginner will progressively improve his script, increasing its similarity to the script of the expert. In a situation where the expert is unable to accompany the beginner for instruction, the following two methods are necessary. The first objectively evaluates the similarity between the beginner's script and the expert's script. The second makes it clear what parts of the beginner's script are not similar to the expert's script.

The proposed method evaluates the similarity between a scenario by an enterpriser and a pro-scenario using text analysis techniques. Snyder presents 10 kinds of storylines. The pro-scenario has detailed scene development for each storyline. The enterpriser selects one storyline that applies to their actual experience. After a choice of a storyline, the enterpriser is informed of the content of each storyline and its application example to the actual life. Since the 10 storylines are only rough explanations of scene development in famous fascinating stories, it is hard to find out which storyline is suitable to represent the actual experiences of the enterpriser. Therefore, the proposed method provides the enterpriser with examples of the actual events in the actual life in each storyline. An example of a "monster in the house" is the business management crisis caused by the COVID19 pandemic. The enterpriser can easily choose a storyline, comparing his actual experience with the examples. After the choice of the storyline, the enterpriser starts to write the scenario of the actual experience according to the scene development in the selected storyline.

The scenario for advertising is assumed to be for SNS and blogs. The total number of words in the scenario is much smaller than in the ones for movies. The division into 15 scenes in screenwriting is excessive for a scenario with few words. To accommodate advertisements in SNS and blogs, the number of scenes is reduced from 15 to 4, based on scene switching. The method takes it after the model of the four-part organization of Chinese poetry. The enterprisers write their actual experiences in the form of a scenario that follows the scene development corresponding to the chosen storyline.

There are five examples of movies for each of the 10 storylines, which are used as references to explain the screenwriting technique. The scene development of these movies is explained by presenting a scenario in each of the 15 scenes. In each of the movies, the scenario is presented for each scene of the movie. The scenario of the enterprisers according to the screenwriting technique should use expressions similar to these pro-scenarios. It makes the scenarios of the enterprisers more empathetic to readers.

Based on the similarity between the scenario written by the enterprisers and the pro-scenario, it is possible to improve the scenario of the enterprisers by giving specific examples of the pro-scenario to the enterprisers to improve their scenario.

According to this idea, the scenario of the enterprisers is improved by the following process.

1) The method finds which storyline the enterprisers' scenario is similar to.
2) The method shows the enterprisers some examples of the scenarios based on the similarity of storylines.
3) The method let the enterprisers improve their scenarios, comparing the given examples with their own scenarios and confirming the lack of expressions in the scenarios.

The proposed method identifies the deficient expressions in the enterprisers' scenario from the features of each scene. It also presents specific correction suggestions.

B. Improving the Scenario by Focusing on Words

In this study, the proposed method improves scenarios by detecting missing expressions from actual life scenarios described by employers according to screenwriting techniques. Fig. 1 shows the outline of this method.

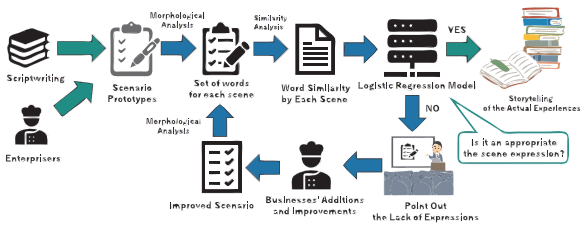


Fig. 1. Method

A screenwriting scenario has 15 scenes. Each scene has its content to be described. The enterprisers create the scenario according to the content. In addition, there are keywords for each scene, and these keywords represent the characteristics of the scene. This paper refers to this keyword as the scene word. By using appropriate scene words, the enterprisers can create a scenario that will gain the reader's empathy.

There are multiple words used depending on the context. Enterprisers do not necessarily need to use the scene words. The proposed method regards the words with meaning similar to that of the scene words as an appropriate expression. The word is defined as a synonym of the scene word. This method evaluates the scenario written by the enterprisers by the degree of similarity to the pro-scenario.

In this study, the scenario written by the business owner is defined as the target scenario. Morphological analysis is used to extract nouns, verbs, adjectives, and adverbs for each scene in the target scenario. From the morphologically analyzed target scenario, the number of synonyms of scene words can be found for each scene. In the same way, from the morphologically analyzed pro-scenario, the number of synonyms of scene words can be counted for each scene. The number of synonyms of scene words is used to detect the similarity of each scenario.

A logistic regression model is used to detect similarity. The model is trained to classify the scenes in the target scenario by the number of synonyms of the scene words in the pro-scenario.

When each scene in the target scenario is more correctly classified by the logistic regression model, the similarity with the pro-scenario is high. If there is a high degree of similarity between the scenarios, the target scenario can be concluded to be an appropriate representation of each scene. If the similarity between the scenarios is low, the lack of expressions in each scene is pointed out to the enterprisers. The lack of expressions is identified from the scene words in each scene. Based on the suggestions, the enterprisers modify the target scenario. The modified target scenario is analyzed again by morphological analysis, and the scene words are extracted for each scene. The modified scenarios are classified using a logistic regression model to detect similarities with the pro-scenario. In the repeated suggestion of the missing expressions which triggers modification of scenarios, the enterprisers improve the target scenario. By extracting the scene words in the pro-scenario, the enterprisers can find out what words should be placed in each scene.

C. Documentation of Actual Experiences

In marketing, it is important to create scenarios for advertisements to gain the empathy of readers. The book, Save-the-cat, presents screenwriting techniques to create scenarios to make readers feel empathy. This study values the 10 storylines and 15 scenes of the screenwriting techniques.

First, enterprisers are requested to select a storyline that matches their experience from the 10 storylines. Since the content the enterprisers write in each scene varies with each storyline, the enterprisers need to select a storyline at the beginning. It is difficult for enterprisers to have to decide for themselves which storyline best fits their experience. To help them, some explanations of the storyline development should be provided to the enterprisers in the study.

Second, the enterprisers create a scenario to fit the content of the 15 scene developments in the selected story. In this study, the fifteen scenes are classified into four following a storytelling style. When creating a scenario, the method provides enterprisers with explanations of the scene development. The explanations of each scene promote the enterprisers to create the scenario smoothly.

The scene words to be used for each scene are presented in the explanation of the scene development. For example, the first scene, which sets up the story, should state many places, times, and proper nouns relevant to the actual experiences. The scenario that follows the scene development should have a representation similar to the pro-scenario. The choice of storylines suitable for their actual experiences along with the description of scenarios compliant with the scene development would enable enterprisers to create scenarios that make readers empathize. Understandably, it is difficult for enterprises to create a scenario similar to professional scenarios with only a single trial. Therefore, the proposed method repeats the process where the enterprisers notified of the lack of expressions in each scene of the target scenario modifies the scenario by

themselves. Through the processes, the enterprisers gradually improve their target scenarios.

D. Finding Lack of Expressions in Each Scene

Enterprisers aiming at creating scenarios similar to the professional scenario need indications telling what parts to be modified. The proposed method points out the lack of expressions, depending on the number of synonyms of the scene words in the target scenario.

Logistic regression is used to find the lack of expressions from the similarity between the target scenarios and the pro-scenarios. The logistic regression model is trained so that it may predict which of the 4 scenes the current one belongs to, using the number of appearances of synonyms of the scene words. Let the logistic regression model classify the scenes in the target scenario. The model is expected to classify the scenes of the target scenario into ones intended by the enterprisers. The classification which is out of expectation implies the scene written by the enterprisers lacks proper expression.

The words to be learned by the model are taken from the pro-scenarios. For example, suppose the scene word for the first scene is "strange". The method counts up synonyms for "strange" in the first scene of the pro-scenario. It also counts up synonyms for "strange" in the remaining scenes. Since "strange" is involved in the scene words in the first scene, its synonyms should appear most in the first scene. In the same way, the number of synonyms for each scene word in the other scenes will be used to identify the scene. The model can classify scenes through the training of a logistic regression model with the number of occurrences of synonyms of scene words in each scene as explanatory variables. It also contributes to finding the lack of expressions.

The logistic regression model gives the classification results as probabilities. If the scene expression written in the enterprisers' scenario is not appropriate, the probability is high for an undesired scene.

The probabilities show the expressions that should be deleted and those that are lacked in the enterprisers' scenario. If the scenes in enterprisers' scenarios are not classified into appropriate ones, it is likely to contain the scene words from other scenes. Those words should be removed. At the same time, to make each scene of the enterprisers' scenarios classified appropriately, the enterprisers should add synonyms of the appropriate scene words to the scenarios. Since the number of synonyms is used as an explanatory variable, the model can calculate the occurrence ratio of the synonyms in each scene. Based on the ratio, the proposed method can suggest to the enterprisers what kinds of expressions lack in the scenes in their scenario. By specifically pointing out the lack of expressions in the scenes of the enterprisers' scenario, the method can support the enterprisers to modify the scenario by themselves.

IV. Experiment

A logistic regression model is created from the pro-scenario. Since the study aims at the improvement of target scenarios written in Japanese, it is assumed that both target scenarios and pro-scenarios are written in Japanese. From the scene words of the pro-scenario, a model is created to classify the four scenes of each storyline.

A. Purpose of the Experiment

In this study, the scene words are used to divide the story into four scenes. Since the scene words depend on the storylines, it will take a large amount of time and effort to check the effectiveness of this method for all storylines. To confirm the effectiveness of this method, it is necessary to create a model to classify the scenes for a specific storyline of screenwriting.

In this experiment, the storyline of "Monster in the House" is used to create the model. The "Monster in the House" scenario has five scenarios written by professional writers. The five scenarios are the stories of five movies: "Alien", "A Dangerous Affair", "Scream", "Saw", and "The Ring".

First, the text expressing the story is divided into some words by morphological analysis. For morphological analysis, mecab's -Ochasen is used. By reading the five scenarios for each of the four scenes divided according to the screenwriting technique, it is possible to find out the scene words for each of the four scenes. A logistic regression model is created by counting the number of synonyms of scene words in each of the five scenarios, using the number of scene words as the explanatory variable and the scene as the objective variable. The logistic regression model created classifies the four scenes of "Monster in the House". Analyzing the appropriateness of the scene words, it is necessary to check the coefficients of the scene words for each scene in the logistic regression model.

B. Word Extraction for a Specific Storyline

A logistic regression model is used to classify the scenes in the target scenario. The synonyms of the scene words are used as explanatory variables in the logistic regression model.

By reading the five pro-scenarios for each storyline, it is possible to identify the common expressions in each scene as scene words. Based on the similarity between the scene words and the words in the scenario, the number of times synonyms of the scene words appear is counted. That number of times synonyms is used as the feature value of each scene.

The fasttext was used to generate the word vectors. On the other hand, the scene words do not have to be words that appear in the scenario. This is because using fasttext does not necessarily mean that the scene words in the pro-scenarios accurately capture the meaning of the scene. To detect scene words in a scenario, computing the Euclidean distance between the multiple scene words and all words in the scenario is necessary. When the synonym of the scene words appears among words with close Euclidean distance, these scene words are appropriate.

The number of scene words and synonyms for each scene is shown in TABLE III to TABLE VI. The words in the top two columns of the table are scene words, and the column on the far left is the examples of the movies. The scene words used in

each scene are as follows. The scene words in the first scene are the number of nouns related to "place" and "time." The scene words in the second scene are a number of synonyms for "strange," "anxious," and "misbehave." The scene words in the third scene are the number of synonyms for " progressively " and "notice." The scene words in the fourth scene are the number of synonyms for "finally" and "courageously."

TABLE III
NUMBER OF SCENE WORDS AND THEIR SYNONYMS IN THE FIRST SCENE

movie	place	time	strange	anxious	misbehave	progressively	notice	finally	courageously
Alien	7	1	0	0	0	0	2	0	1
Danger	5	0	0	0	0	0	1	0	1
Ring	4	2	0	0	1	0	1	0	0
Scream	2	0	1	0	0	0	0	0	0
Saw	5	1	2	0	0	1	2	0	0

TABLE IV
NUMBER OF SCENE WORDS AND THEIR SYNONYMS IN THE SECOND SCENE

movie	place	time	strange	anxious	misbehave	progressively	notice	finally	courageously
Alien	3	0	6	1	2	2	0	0	0
Danger	5	1	0	4	2	1	0	0	0
Ring	1	1	5	0	1	3	6	0	0
Scream	5	1	1	4	2	1	4	4	2
Saw	3	1	1	0	0	1	9	0	1

TABLE V
NUMBER OF SCENE WORDS AND THEIR SYNONYMS IN THE THIRD SCENE

movie	place	time	strange	anxious	misbehave	progressively	notice	finally	courageously
Alien	0	0	0	3	0	3	4	1	0
Dangerous	4	1	0	2	0	3	3	1	0
Ring	3	2	4	0	2	4	4	0	1
Scream	3	0	0	2	0	2	0	0	0
Saw	2	0	1	0	0	0	8	1	0

TABLE VI
NUMBER OF SCENE WORDS AND THEIR SYNONYMS IN THE FOURTH SCENE

movie	place	time	strange	anxious	misbehave	progressively	notice	finally	courageously
Alien	1	0	0	0	1	0	1	2	4
Danger	4	0	0	0	0	0	3	0	1
Ring	3	0	0	0	2	0	4	0	1
Scream	2	0	0	0	0	0	2	0	4
Saw	5	0	1	1	0	0	1	1	4

In scenes 2 through 4, the number of synonyms of the scene words are taken as the explanatory variable. On the other hand, in the first scene, the number of words representing "place" and "time" is the feature value. The first scene is where the words related to the setting of the story scene or time should appear. Since each story has a different way of expressing "place" and "time", it is difficult to find the synonyms of scene words in the first scene. In the first scene, there should be many nouns that indicate "place" and "time". For this reason, only in the first scene, the noun types were referred to from the category information of Janome, and the number of these nouns was used as an explanatory variable. Morphological analysis of nouns in Janome can get the information of the category of the noun. The categorical information is given in the form of noun types, such as common and proper nouns, as well as personal names and places. The number of nouns that are "place, the organization" and "time" among the category information was used as the explanatory variable in the first scene.

The first scene is the scene that sets up the story, but the nouns related to people's names are not considered as features. The nouns about people also appear in other situations. Therefore, it is difficult to use them as features only for the first scene.

A logistic regression model is trained using the number of synonyms of the scene words as a feature. The appropriateness of the scene words for each scene is evaluated by the scale of the coefficients of the logistic regression. For the first scene, if the coefficients of "place" and "time" are large, the scene word is appropriate. The same method is used for scenes 2 through 4 to evaluate the appropriateness of the scene words.

C. Creating the Scene Classification Model

To find the synonyms of the scene words, the similarity between the scene words and all words in the pro-scenario is calculated. There are two main types of similarity calculations: Euclidean distance and cosine similarity. For the reasons given below, this study adopts the Euclidean distance. The cosine similarity between words with similar meanings should be close to 1.0. For example, the cosine similarity between "home" and "father" should be close to 1.0. However, the actual cosine similarity was 0.431. In fasttext, it was determined that "household" and "father" are not words that are close in meaning. On the other hand, the cosine similarity may be expected. As shown above, when using fasttext to vectorize words, it is difficult to uniquely define the threshold with cosine similarity.

In some cases, the word meanings that the fasttext for generating word vectors is trained with different word meanings expected. As a result, it is necessary to manually set an appropriate Euclidean distance threshold. When the threshold was 3.7, the most appropriate synonyms for the scene words were found. The number of words with Euclidean distance less than 3.7 was used as the features.

When the coefficient of the scene word set for each scene is large, the scene word is appropriate because the selected scene word is needed to classify each scene. When the scene word is appropriate, the threshold set to count synonyms of the scene word is also appropriate.

D. Word Vector

The delay in the response of a method to the user's input should be as small as possible. To reduce the delay in the response from the method, the processing in the method should be faster. In this method, fasttext is used to vectorize the words in the scenario. fasttext is fast computing large dimensional vectors such as natural language.

Although fasttext used in this study is trained on many Japanese words, it does not cover all of them. The words that have not been learned by fasttext are called undefined words. There are many undefined words in the scenario. In many situations in natural language processing, undefined words are ignored. Nevertheless, in our method, the undefined words may be in the scenario. Therefore, it is not possible to ignore the undefined words.

In this study, fasttet model is used by applying the python package "Magnitude" to fasttext, so that all words in a scenario may be vectorized. Magnitude gives a vector of unknown words in the fasttext that are similar to the undefined word in terms of letter order and the synonyms of the word. Magnitude considers the words with similar letter sequences to be close in meaning. By referring to vectors of the synonyms of the words with similar letter sequences to the undefined words, the meaning of the undefined words is more grounded.

E. Scene Word Selection for Each Scene

In the study, scene words are selected subjectively. Their effectiveness is confirmed with a logistic regression model.

In the five scenarios of "The Monster in the House," the common expressions and words appear in each scene. The story of "The Monster in the House" is about the main characters' survival from a monster that appears in a closed space. Monsters are not only monsters but can be people or disasters. Monsters can be seen as things that worsen the main character's situation. A closed space is not only a building, but also a region, a country, or even space if it is difficult for the main character and people around him to escape from the monster. Organizations, such as families, are also spaces.

The first scene of "The Monster in the House" introduced the characters and gave information about the place and time. The scene words in the first scene are the words related to "place, the organization" and "time. In the second scene, the protagonists face a monster. Not accidentally, the main characters have made a blunder that causes them to face the monster. However, the main character hasn't noticed his blunder yet. Therefore, in the second scene, the main character just feels unsafe. The characters feel strange about the situation and a little unsafe about their behavior. As a result, the three scene words for the second scene are "strange," "anxious," and " misbehave." In the third scene, the main character notices his "failure" and the situation around him progressively worsens. The main character is forced into an irreversible situation. In the third scene, there are two scene words, "gradually" and "notice". In the fourth scene, the main character forced recovers and overcomes the situation finally. There was a lot of content about the main character using his powers to escape from the monsters courageously. The two scene words in the fourth scene are "finally" and "courageously".

The scene words in each of the four scenes were selected subjectively. The appropriateness of these scene words is tested by the coefficients of a logistic regression model. When the coefficient of the scene word selected in each scene is large, the scene word is appropriate.

F. Experimental Results

From the scene words, a logistic regression model is created to classify the scenes in the target scenario. Table VII shows the results for the coefficients of the logistic regression model.

In the second scene, the coefficients of "strange" and " anxious" were 0.703 and 1.023, respectively, which were larger than those of the other scene words. The words "strange" and "anxious" are appropriate as scene words in the second scene. This result indicates that the second scene in which a disturbing atmosphere is generated can be represented.

TABLE VII
REGRESSION COEFFICIENT OF SCENE WORDS

	place	time	strange	anxious	misbehave	progressively	notice	finally	courageously
scene1	0.432	0.336	-0.012	-0.78	-0.335	-0.427	-0.605	-0.1	-0.413
scene2	0.078	0.075	0.703	1.024	0.209	-0.109	0.458	-0.18	-0.056
scene3	-0.41	-0.021	-0.301	0.044	-0.133	0.912	0.145	0.357	-0.359
scene4	-0.101	-0.39	-0.39	-0.288	0.258	-0.376	0.001	-0.078	0.828

On the other hand, the expression "misbehave" is a numerical value close to the coefficients in other situations. The third scene is the best scene word because the coefficient of "progressively " is 0.912 which is a large value. In other words, the expression " progressively increasing fear" is unique to the third scene.

The fourth scene is appropriate as a scene word because the coefficient of "courageously" is 0.827 which is a large value. However, since the coefficient of "finally" is smaller than that of the scene words in other scenes, the expression is not unique to the fourth scene.

The coefficients of the subjectively selected scene words are larger for the second through fourth scenes. In the first scene, the coefficients of "place" and "time" are smaller than those of the other scene words. The words related to "place" and "time" do not show the characteristics of the first scene. On the other hand, the coefficients of "anxiety" and " notice" are larger. Table VII shows that the number of synonyms for "anxious" and "notice" in the first scene is lower than in the other scenes. In other words, when classifying the first scene, there must be few synonyms of "anxious" and "notice". The first scene should be written with many words about the setting of the story. In addition, the words related to emotions and actions should be written less frequently.

It is possible to create a model for classifying scenes based on the words and expressions specific to each scene in the movie scenarios.

V. WORDS TO DESCRIBE SCENE DEVELOPMENT

A. The Importance of Words to Describe a Situation

The coefficients of the logistic regression model indicate that it is possible to classify scenes by scene words. In screenwriting, the story is developed into 15 scenes. In this study, the 15 scenes are divided into four categories.

The main character's emotions are considered to be a major factor in the development of the scene. Scenes in screenwriting are developed through the emotional ups and downs of the main character. The main character's emotions change in the following order: positive, negative, positive. The words that express emotion, such as "happy" and "sad," are called emotion words. Based on the assumption that the emotional ups and downs of the protagonist were responsible for the scene development, the choice of emotional words should have been important for the scene words. The scenes of the target scenario are classified according to the number of synonyms of the scene words appearing in the pro-scenario.

The scenes of the target scenario are classified according to the number of synonyms of the scene words appearing in the pro-scenario. Words that are not common to all pro-scenarios cannot be said to capture the characteristics of the scene. Therefore, the synonyms of scene words must occur in all pro-scenarios. Emotional words are not appropriate as common scene words because they do not appear often in pro-scenarios.

In movies and novels, emotions are often expressed through the facial expressions and actions of characters. In addition, the emotions of the characters may be expressed by the surrounding circumstances. For this reason, emotional words are rarely written in professional scenarios. For example, assume a situation where the main character is in a crisis. In novels and movies, the main character's sense of danger is conveyed to the reader through the main character's pained expression and the description of a situation from which there is no escape. Even if the main character is verbalizing his emotions, facial expressions and the situation around him should be described. The description of the main character acting hard gives the readers a detailed picture of the main character's situation. The adverbs that modify the main character's actions are also important in conveying the main character's emotions.

The enterprisers create a target scenario with a panoramic view of the actual experience. The expressions of the main character, the enterprisers themselves, and the situation around them must be described from a panoramic view. In the target scenario, it is important to describe the main character's facial expressions, surroundings, and adverbs. Scene words should be chosen to describe the surrounding situation or adverbs that describe the protagonist's actions rather than emotional words.

B. Meaning of the Words in the Story

The words selected as the scene words are based on fasttext. Fasttext can vectorize words and output synonyms of the vectorized words. In this study, synonyms of scene words were checked by fasttext, and appropriate scene words were selected subjectively.

Note that fasttext uses Facebook as its corpus. Due to it, discrepancies may happen in the meanings of words used in scenarios. For example, the word "place" is commonly used to mean "place" in English. However, the top 30 synonyms for "place" in fasttext were sumo-related words such as "yokozuna". This is probably because there were many posts about sumo on Facebook when fasttext model was learning the word "place". In "The Monster in the House," the first scene is characterized by "place," meaning "place". In our method, "place" was not selected as the scene word, but the number of nouns representing a place was used as a feature. This is because when "place" is used as the scene word, the similarity with the word meaning "place" is very low.

Due to the use of fasttext in our method, there are words such as "place" which have a different meaning in the scenario. Currently, the words selected as scene words must be checked in the fasttext to ensure that they have the appropriate meaning for the scenario. One of the ways to solve this problem is to actually create a corpus with many movie scripts and learn the model of fasttext. Learning from the movie script is also important for creating vectors with Magnitude. Magnitude generates vectors from words and their synonyms with a similar word-letter sequence to the undefined word. If the synonym of a word with a similar letter sequence to the unknown word is a word with an unsuitable meaning for the scenario, the meaning of the word will be different from the original meaning of the undefined word. It is important to learn fasttext in movie scripts in order for unidentified words to have proper meaning.

C. Expandability to Other Storylines

In the proposed method, when scenes in the target scenario are not properly classified, the lack expressions are identified based on the difference in the number of synonyms of scene words between the pro-scenario and the target scenario. A specific model was created for "monsters in the house," and the appropriateness of the scene words was tested by the size of the coefficient of determination.

The coefficients for the scene words "strange" and "anxiety" in the second scene were larger. This indicates that sentences containing "strange" or "anxiety" are classified as the second scene. In other words, by setting a scene word for each scene, the model is created to classify each scene.

The scene words from the pro-scenario were used to create the classification model. By using scene words, it is possible to identify the lack of expressions in the target scenario. The scene words enable the enterprisers to describe the scene appropriately. The presence or absence of the scene words in each scene can be used to identify the missing expressions in the target scenario. The scene words are considered to be present in every storyline. Therefore, for the other nine storylines, the model for the scene classification can be created by selecting the scene words for each scene and counting its synonyms.

VI. Approaches using Other Than Scene Words

A. Vectorization of Words and Documents

To calculate the similarity between texts, it is general to vectorize words or documents. For vectorization of natural language, word2vec is generally used. Word2vec vectors the documents from the number of appearances of words. However, the generated document vector depends on the number of appearances of the words. Therefore, word2vec ignores the context. Ignoring context makes it difficult to detect whether the scenario is appropriately expressed because the word order is not taken into account. To calculate the similarity by focusing on the scenario expressions, the vectorization of documents by word2vec is not appropriate. Then, doc2vec takes into account the context vectors of the scenario. The user's scenario and the pro-scenario are vectorized with doc2vec to calculate cosine similarity.

To research how effective doc2vec is in calculating the similarity of the enterprisers' scenario to the pro-scenario, an

experiment is conducted for 4 subjects. The subjects write scenarios from their own experiences, following the screenwriting technique.

The similarity of the subjects' scenarios to other storylines was calculated. For this purpose, the document vectors of the pro-scenario and the subject's scenario were generated for each storyline. The results of the similarity between the subject's scenario and the pro-scenario for each storyline development are shown in TABLE VIII. The columns on the far left are the scenarios of the four subjects and the top of the most column is the name of the storyline. For example, the similarity between Scenario 1 and "Monster in the House" is 0.138. When the similarity is close to 1.0, the two scenarios compared are similar. The similarity values were lower than 0.2 on the whole. In other words, the subjects' scenarios were not similar to any of the storylines. This is probably because the words used by the subjects and the pro-scenario are not similar. Although doc2vec takes context into account, it is based on word vectors, so the similarity will be low if no common words appear between the two scenarios.

TABLE VIII
COSINE SIMILARITY BETWEEN THE USER'S SCENARIO AND THE PROSCENARIO

	Monster	Fleese	Bottle	Problem	Rites	Buddy	Why	Fool	Institutional	Hero
Text1	0.138	0.143	0.161	0.140	0.132	0.139	0.140	0.138	0.137	0.145
Text2	0.029	0.034	0.034	0.025	0.035	0.033	0.028	0.042	0.041	0.028
Text3	0.051	0.042	0.043	0.047	0.032	0.061	0.042	0.010	0.046	0.053
Text4	-0.125	-0.118	-0.096	-0.124	-0.089	-0.115	-0.107	-0.094	-0.109	-0.134

The scenario is incomplete when the enterprisers begin to write it. Naturally, the enterprisers' unfinished scenario is dissimilar to the pro-scenario. Even if doc2vec succeeds in calculating the similarity, it is impossible to give the enterprisers specific advice on their scenarios. Because of the relationship between the words in the enterprisers' scenario and pro-scenario and the lack of specific advice to the expressions, doc2vec is not appropriate for detecting similarity. Since doc2vec focuses on the document as a whole, it is difficult to find the specific advice in the enterprisers' scenario. The following findings were found in this preliminary experiment. The enterprisers should be given something to rely on to write their scenarios.

To write a scenario, the enterprisers should follow the scene development of the pro-scenario. For improving the enterprisers' scenario to be similar to the pro-scenario, they should notice the words that describe the characteristics of the scene for each scene in the pro-scenario. By dividing the scene into detailed parts, it is expected that the lack of expressions can be found by the existence of words used in the pro-scenario within the enterprisers' scenario.

B. The Similarity Detection with Negative-Positive Analysis

The four scenes of the pro-scenario are written as follows. The first scene is the scene that the story is set up with characters, place, and time. In the second scene, the main character's emotions become temporarily positive due to the events to move the story forward. In the third scene, the main character's emotions change negatively as the crisis to him. In the fourth scene, the main character's emotions change to positive again because he has overcome the crisis. Although the first scene can be ignored because it is the setting of the story, the second through fourth scenes is written according to the emotional changes of the main character. The main character's emotions change in the following order: positive, negative, and positive. In other words, if the user's scenario is also such a change of emotion, the user can express the scene appropriately. Emotional changes in each scene are not enough to make specific advances about the user's scenario. Therefore, each scene is given positive or negative scores. Positive words are given a positive value. Negative words are given a negative value. This positive and negative score is called the negative-positive score. By comparing the negative-positive scores of each scene in the movie and the user's scenario, it is possible to notice how many expressions are lacked in the user's scenario. The sum of the negative-positive scores for each scene defines whether each scene is positive or negative. For example, a scene with many positive words is defined as a positive scene because the negative-positive score is positive. On the other hand, a scene with many negative words is a negative scene. An emotion polarity dictionary is used to give negative-positive scores to words in a scenario. In the emotion polarity dictionary, many words are given a negative-positive score before using. An emotion polarity dictionary gives negative-positive scores to words in the movie and the user's scenario. For each scene in each scenario, the negative-positive scores of the words are summed up and the average is divided by the number of words in the scene. The negative-positive score of a scene is calculated by the following formula.

$$\textit{Negative} - \textit{positive score of the scene} = \frac{\textit{Sum of the negative} - \textit{positive scores of the words in the scene}}{\textit{Number of words in the scene}}$$

The average makes it possible to reduce the effect of differences in the number of words in each scene. If the number of words is large, the negative-positive score may be biased either positively or negatively. The negative-positive scores for each of the ten storyline scenes are shown in Figure 2. The vertical axis is the negative-positive score, and the horizontal axis is the four scenes. The negative-positive score for the first scene of "Monster in the House" is -0.56. The shape of each graph is the change in the negative-positive score for each scene of the story development. The ideal transition of the negative-positive score for the four scene developments is as follows: the negative-positive score decreases from the second scene to the third scene and then increases again toward the fourth scene. The graphs of the expected form are "Golden Fleece" and "Buddy Love," but the graphs of the other storylines are not expected. Because the specific scenes differ in each storyline, the negative-positive scores for each scene are large or small. However, the shape of the graph should be a

V-shape like "Golden Fleece." Since the ten storylines follow the four scene developments of the screenwriting technique, the main character's emotional changes are common in all stories. However, many of the graphs are not expected. There are several reasons for this. The negative-positive score given in the emotion polarity dictionary may not match the meaning of the word used in the actual document. For example, the word "ocean" has a negative score in documents about natural disasters. On the other hand, the documents related to summer vacation and leisure are expected to have positive scores. The emotion polarity dictionary defines the negative-positive score of "sea" as -0.987. As shown above, it is difficult to use an emotion polarity dictionary in a pro-scenario using many words whose positive and negative scores can change with each context. In addition, many undefined words appear in the pro-scenario which are not in the emotion polarity dictionary. These undefined words may have a significant impact on the main character's emotions in the scenario. For this reason, the scenario similarity detection by negative-positive scores using the emotional polarity dictionary is inappropriate.

If the negative-positive score could be used to calculate the similarities between the pro-scenario and the enterprisers' scenario, the following issues would still exist. It is difficult to reevaluate the expressions of the scenario after the improvement by the enterprisers. By comparing the negative-positive scores for each scene in each scenario, the lack of expressions in the user's scenario is specified. For example, in the second scene, the negative-positive score of the pro-scenario is 0.600 and the negative-positive score of the enterprisers' scenario is 0.500. The user writes positive expressions so that his negative-positive score of the second scene becomes 0.600. However, the negative-positive score for each scene is averaged the number of words in the scene. There is little change in the negative-positive score even if the user writes a few more words. This means that it is difficult for users to specify the points of improvement in the scenario. Therefore, the detection of lack of expressions with negative-positive scores is inappropriate. The similarity detection using negative-positive scores had the risk of ignoring undefined words that may have important meanings in the context.

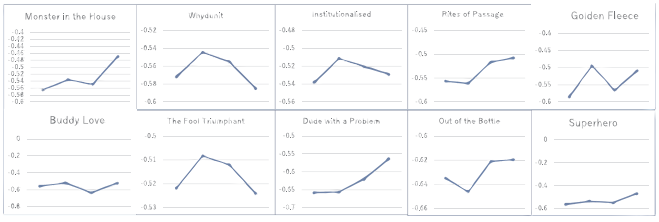


Fig. 2. Negative-positive score for each storyline

The following method can solve this problem. By giving the meaning from the words in the string close to the undefined word, the original meaning of the undefined word may be given in the context. It should be possible to notice the specific lack of expressions in the enterprisers' scenario by paying attention to the words that describe the characteristics of the scene for each scene.

VII. Conclusion

In this study, a logistic regression model was created to classify the scenes of the target scenario. The appropriateness of the scene words was tested by creating a concrete model with "Monsters in the House". Based on the coefficients of the scene words in each scene, the selection of appropriate scene words was important for classifying the scenes. In the target scenario, scene words should be used to describe the characteristics of each scene.

Since the content to be written in a scene is different in each story development, the scene words are considered to be different depending on the story development. Therefore, enterprisers should correctly select the storyline that corresponds to the experience. By selecting a storyline that corresponds to the experience and using appropriate scene words, the enterprisers can create a scenario that will gain the reader's empathy.

By testing the appropriateness of the scene words in "Monster in the house," the scene words can be extended to other storylines. However, it is necessary to verify that it is possible to create models with the scene words for all storylines. The next research is to create a model that can select scene words in other storylines and classifies the scenes in the target scenario.

References

[1] T. Y. CHUN, D. K. LEE, and N. H. PARK, "The effect of marketing activities on the brand recognition, brand familiarity, and purchase intention on the sns of franchise companies," *The Journal of Asian Finance, Economics, and Business*, vol. 7, no. 11, pp. 955–966, 2020.

[2] G. Freytag, *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs, 1895.

[3] A. G. Woodside, S. Sood, and K. E. Miller, "When consumers and brands talk: Storytelling theory and research in psychology and marketing," *Psychology & Marketing*, vol. 25, no. 2, pp. 97–145, 2008.

[4] T. Van Laer, S. Feiereisen, and L. M. Visconti, "Storytelling in the digital era: A meta-analysis of relevant moderators of the narrative transportation effect," *Journal of Business Research*, vol. 96, pp. 135–146, 2019.

[5] S. Blake, *Save the Cat: The Last Book on Screenwriting You'll Ever Need*. Michael Wiese Productions, 2005.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 43–52
DOI: 10.15439/2021F49
ISSN 2300-5963 ACSIS, Vol. 26

Having Avatar Nestle to User through Dialogues to Develop Exercise Habits with Intention Maintained

Tomoya Yuasa
Graduate School of Information Science and Engineering
Ritsumeikan University, Japan
Email: tomoya @ de.is. ritsumei.ac.jp

Fumiko Harada, Hiromitsu Shimakawa
Connect Dot Ltd., Japan
Graduate School of Information Science and Engineering
Ritsumeikan University, Japan
Email: {harada, simakawa}@ cs. ritsumei.ac.jp

***Abstract*—Continuation and habituation of exercises are in preventing lifestyle-related diseases. However, existing habituation applications fail to address mental factors of individual users. This study proposes a method to support habituation of exercise. It makes the best use of an accompanying avatars assigned to users according to their motivation. The avatar dialogues with a user every day. It proposes a goal based on the intention level and the current goal achievement of the user. It also shares the results of the analysis on collected data with the user at regular intervals. This method enables the user to continue the daily exercise easily. From the group of pre-survey subjects, we obtained 2 kinds of groups. We created avatars based on each group to verify their effectiveness. We found that the avatars have improved the goal achievements of the subjects.**

I. Introduction

LIFESTYLE-RELATED diseases are a major problem worldwide. Diabetes, hypertension and coronary artery disease are on the rise worldwide [1]. The risk of lifestyle-related diseases can be reduced by physical activities. For example, a study by Manson et al. shows that physical activity may be a promising approach to primary prevention of NIDDM [2]. A study by Pfaffenberger Jr. et al. shows that initiating moderately vigorous sports activity, quitting smoking, maintaining normal blood pressure, and avoiding obesity are individually associated with reduced mortality from all causes and coronary heart disease [3].

There are IT-based ways to help people exercise on a daily basis. One example is the use of habit-forming applications that run on smartphones. Regardless of the type of exercise, there are a number of apps that run on smartphones to support habit formation [4]. These apps support the implementation of continuous exercise in a variety of ways. For example, one app allows users to record their actions. It let them reflect on themselves in order to make exercise a habit. Others set an alarm time to remind the user to exercise at a certain time. Despite the fact that many studies have shown that exercise is effective against lifestyle-related diseases, one-fourth of the world's adults do very little physical activity [5]. The main reason is that conventional apps promote exercise habit uniformly to all users. Users have individual differences, such as those who can maintain their motivation by being praised by others and who can act with a strong will. The applications on devises should reflect these differences to promote exercise in a way appropriate for each user. In other words, there is a need for user-centered applications.

In this paper, we propose a method to support the habituation of exercise by using use of an accompanying avatars. In this paper, the avatar dialogues with the user to understand the user's state. It sets appropriate short-term goals to achieve long-term exercise goals. In addition, the avatar analyzes the collected data to share the analysis results with the user. The paper refers to avatars taking these behavior features of the users, as avatars accompanying them individually.

In the proposed method, users are classified into personas by motivation analysis based on (MSLQ), see[6]. Reflecting the analysis result, an avatar based on the persona dialogues with the user. The avatar calculates how positive the user's comments are, comparing them with a word-emotion polarity correspondence table [7]. The avatar calculates the user's motivation to exercise based on the difference of current states from the past ones. The avatar also asks the user to indicate the degree of achievement of the goal on a 5-point scale, indicating how much exercise the user take on that day. The avatar shows the optimal short-term goal to the user based on the user's motivation to exercise and the degree of goal achievement.

In this study, we first conducted a survey of what kind of avatars are effective for making exercise a habit. As a result, it is found that users need either avatars that sympathize with the user's utterances or ones that suggest specific things for the habituation of exercise. We developed two types of avatars based on the results of this survey. The study tries to address the following questions.

1) Does dialoguing with avatars improve users' motivation to exercise and achieve their exercise goals?
2) Is it effective for avatars to make concrete suggestions or to sympathize with the user's utterances, rather than to simply engage in dialogue?
3) Is there a relationship of the help seeking ability with long-term reflection which is derived from collected user utterances and exercise goal achievement?

We have conducted three experiments. The results of the experiments have revealed the followings. There is a significant difference in the increase in physical activity when the avatar is used. There is a significant difference in the increase in physical activity when a user is accompanied by an avatar that empathized with the user's utterances or one

that makes specific suggestions. To support habit formation, it is effective to provide long-term reflection with users of high help-seeking ability. The remaining parts of the paper is organized as follows. Chapter 2 discusses existing habit-forming applications and the need for avatars. In Chapter 3, a method is proposed to support the habituation of daily exercise using an avatar. Chapter 4 describes the survey method and the results of the survey on what kind of avatars users need in order to construct the avatar. In Chapter 5, the experiment and its results are explained to demonstrate the usefulness of the proposed avatar. Chapter 6 discusses the causes and future issues inferred from the results. In Chapter 7, the remarks of the paper is summarized as the importance of the research, suggests for applications, and future extensions.

II. Make daily exercise a habit

A. Current status of habit-forming apps

Currently, various applications on smartphones have been published for habituation. Katarzyna et al. the features of 115 habituation applications [4]. As a result, they found that many habituation applications have task tracking and graph display functions. On the other hand, we found that there are very few functions that motivate users through sending messages showing supports and comments from other users. This suggests that many habituation applications are designed for self-monitoring.

It is also important to set goals frequently. Locke et al. have suggested that goal setting should be an iterative process in which users evaluate their performance either to modify their goals or to set completely new goals [8]. However, the applications have no function that frequently presents the most appropriate goal to the users. The users must change their goals towards the most appropriate ones for themselves all the time. However, this places a large burden on the users.

Phillippa et al. found that it takes 60 days in average for specific behavior to get habitualized [9]. In addition to that, Navin et al. have reported that continuous participation of users in the gym activities is highly correlated with "their frequency of the participation" [10]. Therefore, it is necessary to work on specific behavior continuously to make it a habit. However, motivation can not always be kept high. Engagement in specific behavior with high frequency needs abilities to regulate oneself toward it. It is difficult to benefit from the existing habituation applications unless their users have strong wills.

B. The need for avatars

In order to maintain human motivation for a long time, dialogue is an important means because of persuading. For example, it has been found that persuaders who express positive emotions to persuade others are more likely to succeed in persuading others than those who do not express emotions [11][12]. On the other hand, a research has been conducted to increase the success rate of attempts to persuade users with dialogues considering their emotions [13].

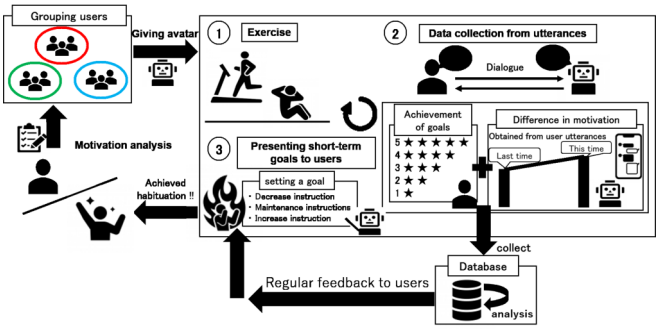


Fig. 1. a schematic diagram of the proposed method

If we focus only on dialogue for a small number of users, it is sufficient to prepare a human-to-human dialogue environment. However, in case of habituation applications to be used by tens of thousands of users or more, it is difficult to realize a function to respond immediately to the user's utterances. To address the problem, it is a promising way to implement avatars with chatbots. The chatbots make it easy to realize avatar-based applications respond immediately to many users 24 hours a day. According to Reeves et al., we would esteem exchanges of words. Humans have a tendency to treat any interaction as an interaction, between humans, even if the other party is a computer [14]. Chatbots that are available 24 hours a day are considered to be superior to human partners in that they can interact with users at any time.

In fact, chatbots that automatically interact are put into practical use in various fields. Fitzpatrick et al. developed Woebot, an automatic interactive chatbot. They found that it significantly reduced the severity of depression [15]. This suggests that the use of dialogue can be a great support for habituation.

III. Making behavioral habits with avatars that are close to users

A. Accompanying avatars

The paper proposes a method to support habituation of exercises. It makes the best use of an accompanying avatars assigned to users according to their motivation.

This study, realizes an "accompanying" avatar which understands the user's state through dialogues with the user. Its role is to set appropriate short-term exercise goals to achieve long-term goals toward a healthy life, promoting the user to continues exercises, while it analyzes the collected data from the user to sharing the analysis results with the user. Figure 1 shows a schematic diagram of the proposed method. This method allows users to continuously perform daily exercises that they determine themselves without difficulties.

In this method, we assume that users have a long-term exercise goal. First, users answer a questionnaire about their motivation. To the user, the method assigns both an appropriate persona based on the questionnaire result, and an avatar corresponding to the persona.

The persona is a classification of a user's personality in terms of motivation to take exercises. The avatar provides the user with motivation for exercise through an appropriate dialogue according to the persona. For example, an avatar can make specific suggestions or sympathize with the user's utterances.

Suppose users trying daily exercise. At the end of the day, a user dialogues with the avatar. The paper assumes the avatar presents the analysis results of user data. The dialogue is initiated by the avatar's utterance on the analysis results. It is followed by dialogues consisting of the user comments and succeeding responses from the avatar. The contents of responses depend on the avatar's persona. The user's comment and the avatar's response may be repeated several times.

From the content of the dialogue, the avatar obtains two values; the degree of achievement of the day's goal and the motivation of the user. The avatar sets a short-term goal based on the two values. It proposes the goal at the end of the dialogue.

B. Goal setting by avatar

The short-term goal values are calculated with a logistic regression, which takes the sum of the values of goal achievement and motivation as explanatory variables, while it. A logistic regression model is prepared for each persona. The short-term goal setting for exercise is one of the followings: to "reduce", "maintain" or "increase" the exercise load. It determines short-term goals that users can achieve the next day so that they may continue to exercise towards achievement of their long-term goals.

The degree of goal achievement shows how well the daily exercise is achieved. It is evaluated by the user in five levels at the end of the dialogue.

The degree of motivation of the user indicates the variation of the user feeling toward taking exercises. It is calculated from a numerical the user's utterances. It tells us whether the user's positive feeling toward exercises is increasing or not. If the avatar requests the user to increase the exercise load while the motivation is low, the user's motivation decreases, which makes it difficult for the user to continue exercising. On the other hand, an instruction to decrease the exercise load would make highly motivated users feel insufficient, which lessons their motivation to exercise. The degree of motivation of the user makes it possible to check whether exercise is painful or not.

Equation (1) is used to measure the degree of motivation of user P.

$$P = \frac{1}{k}\sum_{i=1}^{k} p_i - \frac{1}{k}\sum_{i=k+1}^{2k} p_i \tag{1}$$

Where is the positive degree of the user's utterances i times ago in the dialogue. To know the motivation of the user, it is better to examine the tendency rather than instantaneous values. The proposed method uses the moving average.

The avatar performs morphological analysis on each comment C obtained from the user. The word prototypes extracted through the morphological analysis are compared with the word sentiment polarity correspondence table [7] to obtain the positivity of each word. Summing up the positivity of each word presents, the degree of positive of each user's comment. The avatar takes the difference of the average of the positivity of the user's last k comments from the average of the positivity of their preceding k comments. If P is greater than 0, the user's motivation to exercise tends to increase, while P less than 0, means the user get demontivated. The user's comments are analyzed sequentially. The calculation of every utterance enables us to obtain the time series of the user's positivity.

C. Create personas based on user motivation

Avatar are different in terms of ways of interventions with their messages. Some avatars suggest numerical goals, while others present sympathy with the user's utterances. Since users also have different personalities, how they are motivated to exercise also varies with each use. It is necessary to provide users with avatars suitable for their personality in advance. This study has developed a questionnaire based on the [6] to measure the motivation of users. MSLQ is a questionnaire to examine the motivation of learning in education. The study adopt 5 kinds of motivational factors in MSLQ: Control of Learning Beliefs, Extrinsic, Intrinsic, Self-Efficacy, and Task Value. Modifying examples in MSLQ as they go well with engagement in healthy exercises, the study prepared 4 questions for each of the motivational factors; 20 questions in total are presented to users to know their motivation. Every user answers each question on a 7-points scale. This study uses the 20 answers of the questionnaire as explanatory variables of a logistic regression model. Separately prepared users who answer the same questionnaire will be interviewed in advance about which avatar is suitable. The avatars belong to either of one that specifically suggest to achieve numerical goals, or one that empathize with the user's utterances. The study trains the model using their questionnaire answers labeled with their suitable avatars. The model is used to determine the avatar suitable for each of new users.

D. Suggestions based on users' reply comments

Users do not always have time to exercise. When they are too busy to exercise, they are expected to engage in another exercise to get the same effect in a shorter time. Repeating the same thing every day is important, but doing different exercises with the same effect is one way to continue the daily exercise.

Avatars showing concrete numerical goals are suitable for the way. For example, avatars should avoid presenting stereotype messages, such as "Keep it up!", which encourage the same exercise over and over again. Instead of it avatars had better suggest another exercise, say, "Doing 10 sit-ups has the same effect as 1000 step walking exercise". In this way, users can engage in exercise in various ways, which enable them to continue their daily exercise according to their own schedule.

E. Sympathy based on user response comments

In order to encourage users, it is important to offer words of encouragement. However, depending on the user's personality, simply saying "Good work" may seem like a formulaic sentence, which may lessen the effect of encouragement. It is important to offer words of encouragement sympathy with the content of the user's utterances.

An avatar that sympathizes with the user's utterances will respond to a user statement such as "It was cold today, but I did my best to go outside to walk a lot", might say "Walking in the cold until your body gets warm burns a lot of calories. It is very effective for dieting. If you can lose weight, it will also have a preferable effect on your health". In this way, the user feels that the avatar understands him or her, which raises the user's will to continue daily exercise.

F. Long-Term Reflection

The avatar stores the user's utterances and the degree of achievement of the day's goal in a database. Using data collected for one week, avatars analyze the states of users and their exercise status, to share the analysis results with the users. The system provides the users with the exercise trends of how much exercise they have done in a week, goals for the next reviews based on the exercise trends, to concrete suggestions for achieving the long-term goals. This allows the users to understand their own conditions. It also contributes to their sustainable engagement in daily exercise.

IV. Implementing avatars from real-world examples

A. Searching for accompanying avatars

In order to create avatars that accompany to users, the study investigats what kind of avatars they need. The survey targets were 15 working adults (7 males and 8 females).

In this study, we investigated the avatars needed to improve the amount of walking. Since walking does not require any special skills or equipment, it is an exercise that is more accessible and can be incorporated into daily life compared to many sports[16][17].

The survey targets answered to a motivation questionnaire based on the MSLQ. We collected examples of dialogues which took place between the survey targets and the avatars. The survey provides as many kinds of messages as possible, to encourage engagement in walking exercise to the survey targets. After the dialogue, the survey targets accepted interviews. In the dialogues, the survey caused the survey targets to set a goal of walking 10000 steps every day. Mean while, the actual amount of walking of each survey target was classified in to the stages shown in TableI. The five stages in Table I are settled based on the results of the National Health and Nutrition Survey of the Ministry of Health, Labour and Welfare of Japan [18].

The survey used the "Wizard of Oz" testing, where the survey, instead of an avatar, sent expected various messages prepared in advance to the utterance of the survey targets. The dialogues are taken place, using the open chat function of LINE [19]. An example of the actual dialogue is shown in Figure 2.

TABLE I
5 LEVELS OF 10000 STEPS SETTING

	5 levels of 10000 steps setting
1	Less than 5000 steps
2	Less than 7000 steps∼More than 5000 steps
3	Less than 8000 steps∼More than 7000 steps
4	Less than 10000 steps∼More than 8000 steps
5	More than 10000 steps

Fig. 2. Example of a dialogue with a survey avatar

The interview was conducted on Zoom [20]. In the interview, the survey targets looked back at the actual dialogues in LINE. They answered whether the message from the avatar was appropriate, and if not, what would have been the ideal response. The survey collects appropriate messages in dialogues as well as answers or short-term goals. At the same time, they specify avatars needed to make habituation of exercises.

The K-means method is applied to the results of the questionnaire responses to classify the survey targets into personas. However, the classification of the questionnaire responses fails to find distinctive personas, such as a cluster with a high extrinsic motivation factor and a one with a low intrinsic motivation factor. The survey targets are also classified into personas based on the interview results. This time, the results showed that there were two types of personas: one who prefers to suggest specific things during dialogues, and one who prefers to empathize with the user's utterances. However, there are several survey targets those who could not be classified into either persona. The last group is referred to as the "miscellaneous persona". Due to the number of survey targets, common characteristics might not be found, because a result

TABLE II
PARTIAL REGRESSION COEFFICIENTS FOR PERSONAS WHO PREFER SPECIFIC SUGGESTIONS

Personas who prefer to suggest specific things	
Control of Learning Beliefs	-0.435
Extrinsic	-0.914
Intrinsic	0.261
Self-Efficacy	-0.660
Task Value	0.349

TABLE III
PARTIAL REGRESSION COEFFICIENTS FOR SYMPATHY-FAVORING PERSONAS

Personas who prefer to sympathize with the user	
Control of Learning Beliefs	0.0267
Extrinsic	0.0717
Intrinsic	-0.498
Self-Efficacy	0.1694
Task Value	-0.878

of the classification is only one person. The person might be assigned to miscellaneous persona. For miscellaneous persona, it is not possible to determine appropriate avatars because it is a mixture of different personas. Therefore, to check the effectiveness of avatars, the proposed method prepares avatars corresponding to two personas, one that prefers concrete suggestion of numeric goals, and the other that prefers sympathy of the avatar with the user's utterances.

To investigate the two personas, multiple regression analysis is applied to the results of the questionnaire. Let us check which of the five items is the questionnaire is important for each persona, looking at the partial regression coefficients of the multiple regression. The partial regression coefficients for both of the personas are shown in Table II and Table III.

Examination of the partial regression coefficients reveals, the persona preferring concrete suggestion consider the task value factor important. The task value favor corresponds to the evaluation on how assigned task is interesting, important, and useful. In other words, people who often think about how they feel the exercise to be performed or proposed tend to have preference on concrete suggestion. In the same way, the persona who prefers to avatar sympathy with the user's utterances values the self-efficacy factor. The factor is based on self-evaluation of one's ability to accomplish a task. It implies people who can make an effort without giving up easily even in difficult situations tend to prefer sympathy from avatars with their utterances.

B. Constract an accompanying avatar

For constracting a accompanying avatar, this study used Telegram [21], a tool for building messenger applications. In order to automatically respond to the user's dialogues by avatar, we created data based on the dialogues collected in the survey in Section 4.*A*, listing the pairs of the survey target's utterances and the survey avatar's responses or the ideal responses answered by the survey targets as examples. These data was stored in Elasticsearch [22]. Elasticsearch is a full-text search engine that can find appropriate examples from a large number of examples. The utterances-responses pairs of each avatar were mapped to each user persona, which was classified based on the interview results, to create a collection of examples for each avatar. The avatars dialogues each other using the example data stored in Elasticsearch. This is shown in Figure 3.

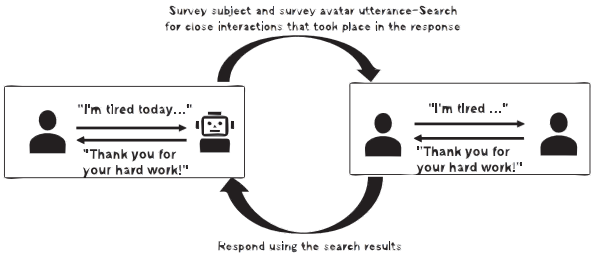


Fig. 3. How to dialogue with avatars

The avatar's response is generated as follows. To find sentences that are similar to the user's utterance, the similarity between the user's utterance and the survey target's utterance in each example is calculated using cos similarity. We calculate the cos similarity by assigning a vector with the frequency of words in the sentence as a component. By doing so, it finds the example query that is most similar to the user's utterance and responds to the user with a response to that example query.

C. Long-term reflection by an accompanying avatar

The user's utterances are stored in a database for long-term reflection. A user's long-term review is based on the subject's motor tendencies and the exercise tendencies based on the user's speech and motor status collected by Avatar over a period of time. This time, we couldn't implement the functionality shared by Avatar, so the overseer used her text message to share with the user the goals to reflect and specific suggestions for achieving them.

V. USEFULNESS OF EXERCISE WITH ACCOMPANYING AVATARS

A. Pilot field study Summary

In this study, we conducted a pilot field study to verify the following three points:

1) Does dialoguing with avatars improve users' motivation to exercise and achieve their exercise goals?
2) Is it effective for avatars to make concrete suggestions or to sympathize with the user's utterances, rather than to simply engage in dialogue?
3) Is there a relationship of the help seeking ability with long-term reflection which is derived from collected user utterances and exercise goal achievement?

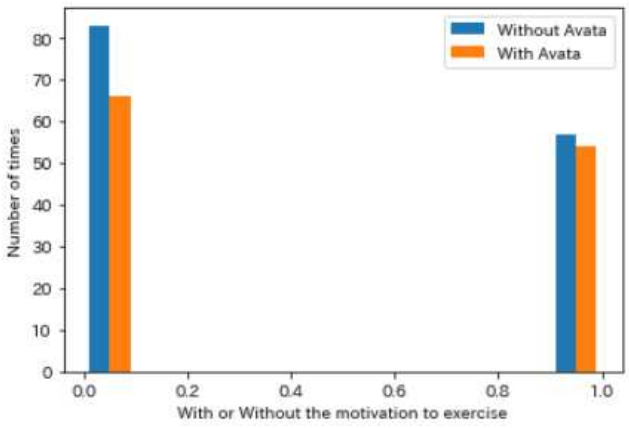


Fig. 4. Frequency distribution of subjects' motivation to exercise

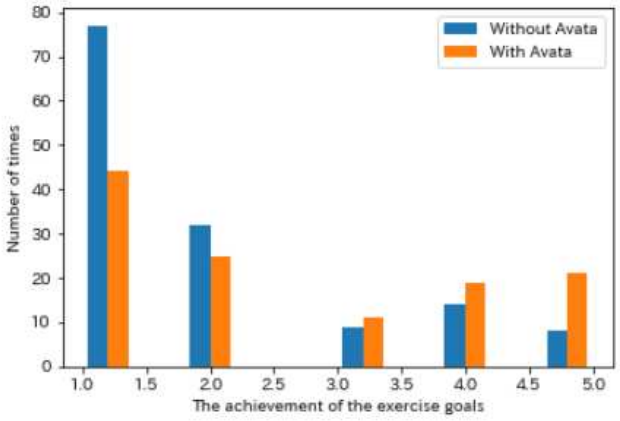


Fig. 5. Frequency distribution of subjects' goal achievement

The subjects were ten working adults (four males and six females). Each subject was answered a motivation questionnaire based on the MSLQ and a Help-Seeking questionnaire. Help-seeking is the act of asking for help from others [6]. Each subject submits daily walking and motivation to exercise for 14 days before the pre-pilot field study started. Motivation to exercise is represented with two options: motivation to exercise or not. The degree of goal achievement is calculated based on the submitted walking data, assuming a daily walk of 10000 steps. This period is referred to as the pre-pilot field study period. During the pre-pilot field study period, avatars were not used. For pilot field study period, the subjects were asked to take their daily activities for 12 days with the goal of walking 10000 steps every day. During the 12 days, each of the two kinds of avatars, as described in Section 4.B, is used for six days each. At the end of each day of the pilot field study, the subject reported the achievement of the exercise goal to the avatar and dialogued with it. The long-term reflection are realized using the Wizard of Oz test, that is, the supervisor of the pilot field study, instead of the avatar, reflected on the user's utterances and the degree of achievement of the exercise goals for 6 days periods. The results were shared with the subjects. We conducted questionnaires after the use of each avatar and after the completion of the entire pilot field study. The questionnaire included questions mainly about the avatars and whether or not the subject was motivated to exercise.

B. Usefulness of using leaning avatars

Figure 4 shows the frequency distributions of the data of "the degree of achievement of the goal of exercise during the 14 days during the pre-pilot field study". Figure 5 shows the frequency distributions of the data of "the degree of achievement of the goal during the 12 days during the pre-pilot field study" for all subjects.

In Figure 4, the horizontal axis indicates motivation. It means no motivation to exercise when the value is 0, while motivation to exercise when the value is 1. The vertical axis indicates the number of times the motivation to exercise. In Figure 5, the horizontal axis indicates the degree of achievement of the exercise goal. It indicates that the person is walking more steps when the value is higher. The vertical axis indicates the degree of achievement of the exercise goal.

The results of Figure 4 shows that the state of lack of motivation to exercise is lower than that of people who do not use avatars. The results of Figure 5 shows the number of times they achieved the highest value of the exercise goal increased compared to those who did not use the avatar. Therefore, we believe that users can improve their motivation to exercise and their achievement of exercise goals by dialoguing with avatars.

In order to verify whether the improvement in exercise motivation and goal achievement is significant, we conducted a statistical test. In the statistical test, we compared the motivation to exercise and goal achievement data of all subjects for each day of the 14-day pre-pilot field study period and the motivation to exercise and goal achievement data of all subjects for each day of the 12 days pilot field study period. As the distribution of the data was not normalized, the Mann-Whitney U test was used with a two-tailed significance level of 5%. As a result, a statistical value of 0.69 for motivation to exercise and 3.69 for goal achievement were calculated. We can say that there is a significant difference when the statistic value is 1.96 or higher. Therefore, it was found that the use of avatars did not significantly improve the motivation to exercise, but it significantly improve the achievement of exercise goals.

VI. Usefulness of using sympathizing avatars

As it is in 5.B, in order to test whether using an avatar that sympathizes with the user's utterances, rather than simply using an avatar, leads to an increase in the subject's motivation to exercise and achievement of exercise goals, we conducted a Mann-Whitney U test at a 5% two-tailed significance level. In the statistical test, we compared the motivation to exercise and goal achievement data of all subjects for 14 days pre-pilot field study period and the motivation to exercise and goal achievement data of all subjects for 6 days pilot field study periods using avatars that sympathize with the user's

TABLE IV
STATISTICAL TEST RESULTS OF EACH SUBJECT'S MOTIVATION TO EXERCISE IN THE SYMPATHIZING AVATARS

Subject	Motivation to exercise
A	1.89
B	2.21*
C	-
D	2.79*
E	0.20
F	2.09*
G	2.09*
H	1.24
I	4.35**
J	3.15**

TABLE V
STATISTICAL TEST RESULTS FOR EACH SUBJECT'S MOTIVATION TO EXERCISE IN THE CONCRETE PROPOSAL AVATAR

Subject	Motivation to exercise
A	1.89
B	-
C	-
D	1.52
E	1.55
F	2.09*
G	2.09*
H	2.52*
I	4.35**
J	1.19

statements. As a result, the statistical values of 0.09 for motivation to exercise and 3.68 for goal achievement were calculated. We can say that there is a significant difference when the statistical value is 1.96 or higher. Therefore, it was found that the use of avatars that sympathize with the user's utterances did not significantly improve the motivation to exercise, but it significantly improve the achievement of the exercise goals.

Next, we conducted a Mann-Whitney U test at the 5 two-tailed significance level for each subject in order to verify whether the use of avatars that sympathize with the user's utterances would not lead to an improvement in motivation to exercise. In the statistical test, we used the data of motivation to exercise of each subject for 14 days pre-pilot field study periods and the data of motivation to exercise each subject for 6 days pilot field study periods using avatars that sympathize with the user's utterances. The statistical values obtained from the statistical test results are shown in Table IV.

We can say that there is a significant difference if the value is 1.96 or higher. The values marked with * indicate that there is a significant difference this time. The values marked with ** indicate that there is a significant difference this time when the avatar was not used. The "-" indicates that there was no change before and after the use of avatar.

The results in Table IV shows that for some subjects, the use of avatars that sympathize with the user's utterances significantly improve the motivation to exercise.

A. Usefulness of using avatars as specifically proposed

As it is in the same way in 5.B, in order to test whether using an avatar that makes concrete suggestions, rather than simply using an avatar, leads to an increase in the subject's motivation to exercise and achievement of exercise goals, we conducted a Mann-Whitney U test at a 5% two-tailed significance level. In the statistical test, we compared the motivation to exercise and goal achievement data of all subjects for 14 days pre-pilot field study periods and the motivation to exercise and goal achievement data of all subjects for 6 days pilot field study periods using avatars that makes concrete suggestions. As a result, a statistical value of 1.21 for motivation to exercise and 2.17 for goal achievement were calculated. We can say that there is a significant difference when the statistical value is 1.96 or higher. Therefore, it was found that the use of avatars that suggest specific things did not significantly improve the motivation to exercise, but it significantly improve the achievement of the exercise goals. Next, we conducted a Mann-Whitney U test at the 5 two-tailed significance level for each subject in order to verify whether the use of avatars to make concrete suggestions would not lead to an improvement in motivation to exercise. In the statistical test, we used the data of motivation to exercise of each subject for 14 days pre-pilot field study periods and the data of motivation to exercise each subject for 6 days pilot field study periods using avatars to make concrete suggestions. The statistical values obtained from the statistical test results are shown in Table V.

We can say that there is a significant difference if the value is 1.96 or higher. The values marked with * indicate that there is a significant difference this time. The values marked with ** indicate that there is a significant difference this time when the avatar was not used. The "-" indicates that there was no change before and after the use of avatar. The results in Table V shows that for some subjects, the use of avatars that provided concrete suggestion significantly improved the motivation to exercise.

TABLE VI
EACH SUBJECT'S HELP-SEEKING VALUE AND EVALUATION OF LONG-TERM REFLECTION

Subject	Help-seeking	Long-term reflection
A	5.33	4
B	2.33	1
C	4.66	2
D	4.33	4
E	3.33	5
F	4.66	5
G	4.00	4
H	4.33	3
I	2.33	2
J	5.00	3

B. Sharing long-term reflections with users

As help-seeking is the act of asking for help from others, subjects with a strong help-seeking should feel that avatars are helping them with their own daily exercise habits. We researched the relationship between each subject's Help-Seeking value from the pre-experiment questionnaire and the 5-point evaluation of long-term reflection on the user obtained from the post pilot field study periods questionnaire.

In the post pilot field study periods questionnaire, the subjects answered whether they felt that the long-term reflection from avatars was effective and why. Subjects' opinions included "I feel that they understand me" and "I did not find the comments from avatars very appealing". It was found that not all the subjects had positive opinions. Table VI shows the values of Help-seeking and the 5 points evaluation of the reflections.

The higher the value of Help-seeking, the stronger the Help-seeking.

Subject A has a higher value of Help-seeking than the other subjects. Therefore, it is thought that Subject A is likely to feel that the long-term reflection is effective. In fact, subject A gave a high rating of 4 for the long-term reflection. On the other hand, because Subject B has a lower Help-seeking value than the other subjects, it is thought that Subject B is less likely to feel that the long-term reflection is effective. In fact, Subject B gave the lowest evaluation of long-term reflection among all subjects.

In order to research the relationship between Help-seeking values and reflection, the correlation coefficient between Help-seeking values and the 5 points rating for long-term reflection was researched. As a result, we obtained a positive correlation of 0.47. This indicates that sharing long-term reflections for subjects with high Help-seeking values is effective in supporting habit formation. The long-term reflection is based on the user's utterances and the achievement of the exercise goal. This indicates that dialogue is important for people with a strong help-seeking. On the other hand, people with low Help-Seeking values work to solve problems on their own, so they do not need to receive advice from collected data.

VII. DISCUSSION

A. Non-significant increase in motivation to exercise

In the present study, the use of avatars did not significantly improve the motivation to exercise. The subjects answered "yes" or "no" whether they wanted to exercise their motivation to exercise. However, each person's consciousness of movement is different. Even in the same situation, one subject may be "yes" while another may be "no". Therefore, it was necessary to clarify the situation setting.

In addition to clarifying the situation, it was also necessary to set the evaluation in more detail, such as five levels, instead of two levels. For example, a questionnaire has "Do you want to exercise today?". It is possible that a person who answers "yes" if there is enough time, but "no" if he does not want to go outside because it is raining, cannot answer clearly if there is enough time and it is raining. It would be nice if the situation setting could be subdivided, but it is difficult to subdivide everything appropriately. Therefore, we think it is necessary to use a rank out of 5 or a rank out of 7 in order to have a certain range of responses.

B. Appropriateness of avatars to be attached

Based on the preliminary questionnaire, we investigated which avatars matched the persona of each subject: avatars that sympathize with the user's utterances, avatars that suggest concrete things, or other avatars. The five items of Control of Learning Beliefs, Extrinsic, Intrinsic, Self-Efficacy, and Task Value in the questionnaire results obtained in section 4.A were used as explanatory variables, and the persona classification of preference concrete, empathy, other as the objective variable.

A logistic regression model was created from the explanatory and objective variables. The regression equation was used to predict the optimal avatar by applying it to the subjects' questionnaire results. As training data, the subjects answered a questionnaire after the pilot field study : "What kind of avatars do you feel is most appropriate for you in supporting your exercise?". They answer from three types of avatars.

Figure 6 shows the prediction results of the logistic regression and the results of the questionnaire.

In Figure 6, the horizontal axis is the number of the most suitable avatar for each subject predicted by the logistic regression, and the vertical axis is the number of avatars that each subject answered as suitable for him/herself obtained from the results of the subject's questionnaire. 0 represents avatars that sympathize with the user's utterances, 1 represents other avatars, and 2 represents avatars that suggest specific things. Figure 6 shows that the prediction by logistic regression is not very adequate, 3/10.

In the questionnaire based on the MSLQ described in section 3.B, five motivational factors were employed: Control of Learning Beliefs, Extrinsic, Intrinsic, Self-Efficacy, and Task Value. However, there are other motivational factors that were not used in this study. Therefore, in order to improve the

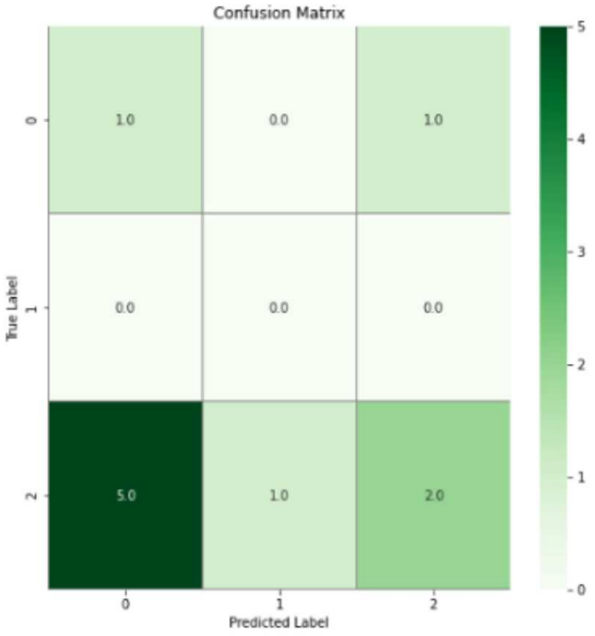


Fig. 6. Confusion matrix of persona prediction results for all subjects

accuracy, it is necessary to prepare explanatory variables that correspond to more motivational factors.

C. Improving avatars with specific suggestions

Looking at the True Label in Figure 6, most of the subjects preferred the number 2. This indicates that the subject needs an avatar to suggest specific things. However, looking at the results in Table V, few subjects showed a significant improvement in motivation and achievement in exercise by avatars who made specific suggestions. We think the reason is that the proposal was not feasible for them.

In fact, in the free descriptions of the subjects after the experiment, some of the subjects for whom no significant difference wrote that "the proposed concrete plan was unrealistic and difficult". This means that the concrete suggestions made by avatars to users need to be appropriate for the users, not just suggestions. For example, a user who likes to go outside can be suggested to take a walk, and a user who likes housework can be suggested to do housework. By incorporating information about the fields that are not difficult for each user in advance, significant differences in motivation to exercise and achievement by avatars who suggest specific things are expected to appear in more people.

Next, the means of the values for each item of the Control of Learning Beliefs, Extrinsic, Intrinsic, Self Efficacy, and Task Value of the questionnaire using the MSLQ of all subjects were calculated. The mean values were used to classify the participants into "high" and "low" groups. Based on the these groups, the relationship between the groups of subjects with significant differences in exercise motivation and achievement was examined. The results showed that there was no commonality between the groups of subjects with significant differences in any of the items. This may also be due to the fact that there were only five explanatory variables, so it is necessary to prepare more explanatory variables.

D. Improving the interaction between the user and avatar

When we checked the content of the dialogue between the subject and avatar described in Chapter 5 from the perspective of linguistics, we found that the avatar responded to the user in a way that was out of line and that the same phrases were responded to over and over again.

In this case, the interview was conducted by one person based on the interview in 4.A. Therefore, since there were cases where different survey targets had similar dialogues, the phrases were similar and the answers were commonplace. Therefore, when adding a new response content, rather than creating a response to the content of the user's utterances by one person, multiple people come up with the response content individually. By doing this, it is possible to give responses from various people to one of her remarks by the user, which not only increases the types of responses, but also eliminates mundane responses.

One of the shortcomings of our method is that the dialogue is not conducted in a time-series. For example, if you were sick the day before, in human-to-human situations, you would worry about the other person. On the next day, taking into account the previous day's condition, we will ask, "Are you feeling better?". However, our method does not have such a function. Consider the other party's condition in time series, it is the embodiment of a more "accompanying avatars". We believe that it is necessary to consider the time series in dialogue processing. In this paper, we will discuss how to use the time series in the dialogue.

E. Study Limitation

In this experiment, we focused on the number of steps and examined the effectiveness of the goal achievement of exercise. However, in reality, exercise varies. For example, some of the subjects did exercise other than walking, such as riding a road bike on a regular basis. Such people get enough exercise because they ride road bikes, even if the number of steps is small. By taking into account items other than the number of steps, such as the amount of exercise and the time of exercise, rather than just the number of steps, we were able to verify the usefulness of the avatars. For this purpose, it is more appropriate to measure the number of steps using a smartwatch instead of using the acceleration sensor of a smartphone. Since few people wear smartwatches on a daily basis, this study mainly used a pedometer on a smartphone. But then, it is highly likely that people do not wear their smartphones during the time they are doing housework or exercising at the gym, for example. Those exercises are a very effective way to acquire the number of steps. Therefore, in order to measure the number of steps more accurately, the smartwatch should have been adapted in the experiment.

Since there were 10 subjects, the sample size was small. Therefore, there are limitations in generalization. We received

volunteers to help us with the experiment. The subjects were also asked to participate in the study as volunteers. Therefore, they varied from those who exercise on a daily babecause we are more interested in increasing the awareness of exercise and the amount of exercise for those who do not have exercise habits than in increasing the awareness of exercise and the amount of exercise for those who already have exercise habits.

VIII. Conclusion

In this paper, we proposed a method of using an avatar to accompany users to help them make daily exercise a habit, by having users interact with the avatar. The strength of this research is not to adapt the application in a uniform manner, but to assign an appropriate avatar to each individual and to present short-term goals on a daily basis.

We investigated what kind of avatar is needed through interaction with the research avatar. As a result of the survey, we found that some personas in the research group preferred to receive specific suggestions to help them make daily exercise a habit, while others preferred to be sympathetic to the speech.

We also conducted an experiment to test the effectiveness of avatars. The results showed that the use of the avatar did not significantly improve the subjects' motivation to exercise, but it significantly improve their achievement of their daily exercise goals.

The use of avatars that empathize with the user's speech and suggest specific things to do was found to significantly improve the achievement of daily exercise goals. On the other hand, motivation to exercise was not significantly improved. However, some of the subjects significantly increased their motivation to exercise.

In the future, the avatars need to be improved by increasing the number of explanatory variables in the regression equation for predicting appropriate avatars, and by having multiple people think about and add the response content of what the user is expected to say, rather than having one person create it. In addition, it is necessary to collect information on feasible motions from the user in advance and consider specific proposals based on this information. Furthermore, since the dialogue was not conducted in time-series data, it is necessary to process the avatar's responses in time-series.

References

[1] Arora, Charu, et al. Development and validation of health education tools and evaluation questionnaires for improving patient care in lifestyle related diseases. Journal of clinical and diagnostic research: JCDR, 2017, 11.5: JE06.

[2] Manson, Joann E., et al. Physical activity and incidence of non-insulin-dependent diabetes mellitus in women. The Lancet, 1991, 338.8770: 774-778.

[3] Paffenbarger Jr, Ralph S., et al. The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. New England journal of medicine, 1993, 328.8: 538-545.

[4] Stawarz, Katarzyna; COX, Anna L.; BLANDFORD, Ann. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. 2015. p. 2653-2662.

[5] World Health Organization. Global action plan on physical activity 2018-2030: more active people for a healthier world. World Health Organization, 2019.

[6] Pintrich, Paul R., et al. A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). 1991.

[7] Takamura, Hiroya; Inui, Takashi; Okumura, Manabu. Extracting semantic orientations of words using spin model. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). 2005. p. 133-140.

[8] Locke, Edwin A.; Latham, Gary P. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American psychologist, 2002, 57.9: 705.

[9] Lally, Phillippa, et al. How are habits formed: Modelling habit formation in the real world. European journal of social psychology, 2010, 40.6: 998-1009.

[10] Kaushal, Navin; Rhodes, Ryan E. Exercise habit formation in new gym members: a longitudinal study. Journal of Behavioral Medicine, 2015, 38.4: 652-663.

[11] Mazzotta, Irene; De Rosis, Fiorella; Carofiglio, Valeria. Portia: A user-adapted persuasion system in the healthy-eating domain. IEEE Intelligent systems, 2007, 22.6: 42-51.

[12] Carnevale, Peter JD; Isen, Alice M. The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiation. Organizational behavior and human decision Processes, 1986, 37.1: 1-13.

[13] Forgas, Joseph P. On feeling good and getting your way: Mood effects on negotiator cognition and bargaining strategies. Journal of personality and social psychology, 1998, 74.3: 565.

[14] Reeves, Byron; Nass, Clifford. The media equation: How people treat computers, television, and new media like real people. Cambridge, UK: Cambridge university press, 1996.

[15] Fitzpatrick, Kathleen Kara; Darcy, Alison; Vierhile, Molly. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 2017, 4.2: e19.

[16] Morris, Jeremy N.; Hardman, Adrianne E. Walking to health. Sports medicine, 1997, 23.5: 306-332.

[17] Ogilvie, David, et al. Interventions to promote walking: systematic review. bmj, 2007, 334.7605: 1204.

[18] https://www.mhlw.go.jp/content/10900000/000687163.pdf

[19] LINE, LINE | always at your side.https://line.me/en/.

[20] Zoom, Zoom Meetings - Zoom. https://zoom.us/meetings.

[21] Telegram, Telegram Messenger. https://telegram.org/.

[22] Elasticsearch, Elasticsearch: The Official Distributed Search & Analytics Engine | Elastic. https://www.elastic.co/elasticsearch/.

14th International Workshop on Computational Optimization

MANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

TOPICS

The list of topics includes, but is not limited to:

- combinatorial and continuous global optimization
- unconstrained and constrained optimization
- multiobjective and robust optimization
- optimization in dynamic and/or noisy environments
- optimization on graphs
- large-scale optimization, in parallel and distributed computational environments
- meta-heuristics for optimization, nature-inspired approaches and any other derivative-free methods
- exact/heuristic hybrid methods, involving natural computing techniques and other global and local optimization methods
- numerical and heuristic methods for modeling

The applications of interest are included in the list below, but are not limited to:

- classical operational research problems (knapsack, traveling salesman, etc)
- computational biology and distance geometry
- data mining and knowledge discovery
- human motion simulations; crowd simulations
- industrial applications
- optimization in statistics, econometrics, finance, physics, chemistry, biology, medicine, and engineering
- environment modeling and optimization

BEST PAPER AWARD

The best WCO'21 paper will be awarded during the social dinner of FedCSIS 2021.

The best paper will be selected by WCO'21 co-Chairs by taking into consideration the scores suggested by the reviewers, as well as the quality of the given oral presentation.

TECHNICAL SESSION CHAIRS

- **Fidanova, Stefka,** Bulgarian Academy of Sciences, Bulgaria
- **Mucherino, Antonio,** INRIA, France
- **Zaharie, Daniela,** West University of Timisoara, Romania

PROGRAM COMMITTEE

- **Abud, Germano,** Universidade Federal de Uberlândia, Brazil
- **Bonates, Tibérius,** Universidade Federal do Ceará;, Brazil
- **Breaban, Mihaela,** West University of Timisora, Romania
- **Gruber, Aritanan,** Federal University of ABC, Santo André, Brazil
- **Hosobe, Hiroshi,** Hosei University, Japan
- **Kouichi, Hirata,** Kyushu Institute of Technology, Kawazu, Japan
- **Lavor, Carlile,** IMECC-UNICAMP, Brazil
- **Micota, Flavia,** West University of Timisora, Romania
- **Muscalagiu, Ionel,** Politehnica University Timisoara, Romania
- **Stoean, Catalin,** University of Craiova, Romania
- **Tami, Tamir,** School of Computer Science, The Interdisciplinary Center, Herzliya, Israel
- **Wang, Yifei,** Georgia Institute of Technology, USA
- **Zilinskas, Antanas,** Vilnius University, Lithuania

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 55–60
DOI: 10.15439/2021F111
ISSN 2300-5963 ACSIS, Vol. 26

Towards Evolutionary Emergence

Jörg Bremer
Department of Computing Science
Carl von Ossietzky University
Oldenburg, Germany
joerg.bremer@uni-oldenburg.de

Sebastian Lehnhoff
Department of Computing Science
Carl von Ossietzky University
Oldenburg, Germany
sebastian.lehnhoff@uni-oldenburg.de

***Abstract*—With the upcoming era of large-scale, complex cyber-physical systems, also the demand for decentralized and self-organizing algorithms for coordination rises. Often such algorithms rely on emergent behavior; local observations and decisions aggregate to some global behavior without any apparent, explicitly programmed rule. Systematically designing these algorithms targeted for a new orchestration or optimization task is, at best, tedious and error prone. Suitable and widely applicable design patterns are scarce so far. We opt for a machine learning based approach that learns the necessary mechanisms for targeted emergent behavior automatically. To achieve this, we use Cartesian genetic programming. As an example that demonstrates the general applicability of this idea, we trained a swarm-based optimization heuristics and present first results showing that the learned swarm behavior is significantly better than just random search. We also discuss the encountered pitfalls and remaining challenges on the research agenda.**

I. Introduction

CYBER-PHYSICAL systems (CPS) are equipped with a steadily increasing degree of autonomy (cf. [1], [2]). The technical viability of such systems has already achieved broad attention (see for example [3]). Often, the autonomy in CPS emerges from self-organization principles that are used for coordination as well as from integrating artificial intelligence (AI) -enabled algorithms – as also stipulated in [4] for the example of the European Union. Today's cyber-physical systems already comprise a huge number of physical operation and sensing equipment that has to be orchestrated for secure and reliable operation – prominent examples are the electric energy grid, modern transportation systems, or environmental management systems [5].

As yet, often human operators monitor and control a hierarchically organized CPS and aggregate information from lower level subsystems. Supervisory control and data acquisition (SCADA) systems – as an example from the energy sector – provide a view on and allow for control of a decentralized process and are thus a state-of-the-art means [6].

As complexity grows, more autonomy is desirable in future CPS [7]; desiring for algorithms with self-*-properties [8]. A targeted design of algorithms with specific emergent behavior is difficult to achieve, especially with standard programming languages [9]. Design patterns like [10], [11] may ease the design process, but are often limited in applicability. There are meta-heuristics like the combinatorial optimization heuristic for decentralized agents [12] that are best on self-organization principles, but they need to be manually adapted to each new use case. Having a systematic methodology describing at the design time the construction of self-organizing algorithms or systems step-by-step, would be highly desirable but is hard to achieve for general applicability [13]. Few paradigms and pattern on a rather high abstraction level exist, but the individual adaption to a specific algorithmic or functional goals is left to the designer [13]. Nevertheless, as soon as changes in the system occur, manual adaption might be necessary. The concept of controlled self-organization addresses this issue by introducing a observer-controller architecture for automated correction of self-organized behavior of the system. The concept works well, if correction can be achieved by (re-)parametrizing the self-organizing entities/ agents according to changes in the environment. Sometimes, it might be necessary to change the algorithmic behavior on a level that needs a redesign.

For the initial design of an algorithm addressing a given task by self-organizing mechanism as well as for automated redesign at runtime for proper situational tracking, we propose an automated design of emergent behavior by machine learning approaches. As this goal is a huge field with many aspect to be addressed, we here start by discussing a first example: the applicability of Cartesian genetic programming [14] to the automated design of a swarm-based optimization algorithm for solving global optimization problems.

Thus, we propose to choose a different approach for designing purposeful emergence in self-organizing systems by using machine learning. Machine learning in multi-agent systems is already used for problems that are difficult to solve with preprogrammed agent behavior. The agents must instead discover a solution to the problem on their own, using machine learning [15]; often by reinforcement learning. We go for automatically discovering mechanisms for emergent behavior and self-organization by genetic programming [16]. In this way, we learn control programs for individually acting entities in a decentralized system with the goal to jointly solve a specific problem. As test scenario, we started with swarm-based optimization.

The rest of the paper is organized as follows. We start with a brief review on related work with a focus on multi-agent reinforcement learning and Cartesian genetic programming that we use for learning control programs for particles in our optimization swarm. After describing our share of pitfalls on the way to the first successfully trained swarms, we present

preliminary results comparing our swarm with random search and real particle swarm optimization.

II. Related Work

In general, machine learning algorithms automatically build a mathematical model using sample data. These models are then used to make decisions without a need for specifically programming rules to make these decisions. Starting from the first works of [17] many different algorithms and approaches have been developed. Among them are reinforcement learning [18], [19], classifiers like support vector machines [20], or artificial neural networks [21], to name just a few. A good overview can for example be found in [22].

Reinforcement learning is often applied in intelligent agents for learning to take appropriate actions based on observations from the environment that the agent interacts with [18]. An extension is multi-agent reinforcement learning (MARL) [23]. In MARL, many agents independently learn how to decide on the most rewarding action in a dynamic environment that is disturbed by the other agents. Many MARL algorithms are designed for static and thus stateless games [15]. But, also use cases for cooperative games are scrutinized and may generate emergent behavior [24], [25]. Nevertheless, the application is limited as agents still just learn to choose from a predefined set of (singular) actions [23].

A subset of machine learning algorithms is made up by a special type of evolutionary algorithms. Genetic programming (GP) is used to discover solutions to problems automatically by using evolutionary mechanisms like random mutation, crossover, a fitness function, and multiple generations of evolution. Alan Turing was probably the first to raise the question, whether programs might be evolved by something like evolution [26]. After a first implementation by [27] for logical functions represented as tree programs, many improvements were made [28]–[31].

One of this improvements is the use of a special phenotype representation that allows leaving computational nodes unused. In general, Cartesian genetic programming (CGP) is a more efficient version of genetic programming and encodes computer programs as graph representation [32]. CGP is an enhancement of a method originally developed for evolving digital circuits [33], [34]. CGP already demonstrated its capabilities in synthesizing complex functions in several different use cases for example for image processing [35], neural network training [36], or for the synthesis of Bent functions for cryptography [37].

III. Learning emergence with Cartesian genetic programming

Our goal was to automatically generate a swarm-based heuristics for optimization similar to the particle swarm algorithm, i. e. to derive a swarm of individually acting particles that may include observations from neighboring particles into their own move decisions. To achieve this, we implemented particles that can be equipped with a control program learned by CGP. Figure 1 shows the general architecture. A swarm

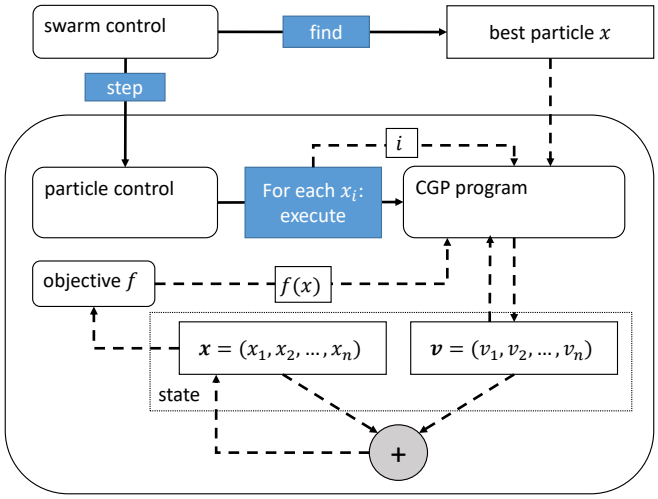


Fig. 1. General architecture of the swarm and the incorporated particles with the embedded CGP program.

consists of an arbitrary number of particles. In each iteration during optimization, each particle is stepped by a global swarm control; just like PSO. The global control is also responsible for ranking the particles and detecting the best one (in terms of fitness). When a particle is stepped, the CGP program that determines the new position of the particle is executed and the new fitness value is calculated. We experimented with different input to the CGP program and different internal particle states as described later

Currently we are only considering the observation of other particles in the swarm. Two succeeding stages of extension will be the integration of inter-entity coordination (1) by using stigmergy and (2) by communication by exchanging messages. Finally, we are opting for problem solving with multi-agent systems.

When learning the control program by CGP, the same swarm setting is used. For each CGP solution candidate, several swarms were set up. Each particle was equipped with the solution candidate program. Each swarm was run for several iterations. Finally, the mean achieved optimization result evaluated the solution candidate.

Cartesian genetic programming is an advanced form of genetic programming (GP) designed to evolve acyclic graphs [38]. The nodes are indexed by their Cartesian coordinates and represent functions of a computational structure (the graph) [39]. Many traditional GP approaches suffered from the so called bloat effect [40] – programs steadily growing in complexity without any significant objective improvement [41]. CGP does not suffer from this problem [40].

A chromosome comprising function as well as connection genes and output genes encodes the computational graph that represents the executable program. Figure 2 shows an example with six computational nodes, two inputs and two outputs. The gene of a function represents the index in an associated lookup-table (0 to 3 in the example). Each computation node is encoded by a gene sequence consisting of the function

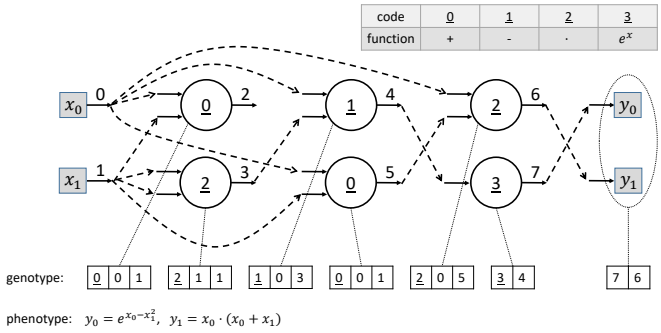


Fig. 2. Computational graph and its genotype representation in Cartesian genetic programming.

look-up index and the connected input (or output of another computation node) that is fed into the function. Thus, the length of each function node gene sequence is $n + 1$ with n being the arity of the function. The graph in traditional CGP is acyclic. Parameters that are fed into a computation node may only be collected from previous nodes or from the inputs into the system. Outputs are connected to any computation node output or directly to any input. Not all outputs of computational nodes are used as input for other functions. In fact, usually many of such unused computational nodes occur in evolved CGP [33]. These nodes are inactive, do not contribute to the encoded program's output, and are not executed during interpretation of the program. In this way, phenotypes are of variable length whereas the size of the chromosome is static.

A computational graph in CGP is typically evolved using a $(1 + \lambda)$-evolution strategy, i. e. with probabilistic mutation but no crossover [42]. CGP allows for unused nodes. Thus, the maximum number of nodes is a priori specified. It has been shown to be advantageous to evolution to overestimate the number of nodes due to an induced higher genetic drift [39].

For our experiments, we used an extension to the ECJ-toolkit [43], [44]. In addition to the traditional integer representation as in Fig. 2, the ECJ version also supports a real-valued representation. For each gene, alleles are allowed to range from $[0, 1]$. Prior to executing the program, all real values are rounded back to integer for interpretation as described above. With real-valued encoding, it becomes possible to apply a real-valued crossover operator. In this way, the performance of convergence is significantly improved at least for regression and [45]. In the integer-encoded case, crossover is usually left out. Nevertheless, more operators are possible with continuous encoding and discovering improved genetic operators for other problems remains an open area of research. We chose to use real-valued encoding.

For learning the internal particle control, we started by setting up a standard CGP scenario. For a start, as function set we chose the four basic arithmetic operations, a generator for normal distributed random numbers, the classical if-then-else-statement, and the set of standard order relations. As input, we gave the current position (in search space), the current objective value, and the position of the best particle. The output of the program was set to be the new particle position. Initially, we introduced an additional parameter $\boldsymbol{v}$ meant to be comparable to the velocity in particle swarm optimization [46], that was output and input to the next iteration as well. In this way, it was meant to enable the particle have a more complex inner state apart from the mere position. But, we were not able to make train CPG to make any targeted use of it. Thus, we changed it to be an automatic increment of the current position.

Instead, we extended the functions set by a function that is able to determine the current rank of the particle (compared with all other). Moreover, the numbers 0-9 were given as constant functions. In many training process we observed that CGP learned to construct needed constants by itself. This was for example achieved by using the if-statement to construct a 1 and then adding it up several times. Usually, this seemed to be a waste of necessary evolutions as well as of needed computation nodes. With introducing the constants, CGP could use the numbers directly. A further improvement was to reuse the same learned program for all dimensions of a multi-variate problem. Figure 1 shows the final architecture of a particle and its embedding in the swarm.

The next challenge was the decision for the objective function. First, we tried evaluating the fitness of a swarm by a single optimization problem. The swarm solves each problem several times and the mean achieved residual problem error is taken to evaluate the performance of the swarm in solving the problem. This approach failed, because CGP learned to solve the given optimization problem directly and made the swarm output the problem solution hard-coded. Actually, this was to be expected. With the next try, we handed a bunch of different optimization problems with optimal solutions at different positions – otherwise it would have resulted in a directly learned result again. With a given set of objectives that are all to be solved independently by the swarm, a sort of swarm behavior could already be generated – but not as expected. The swarm learned to move along a trace that passes through all the optima of the different problems. Again, this was not optimization. In order to tackle this problem, we introduced a random offset. For each problem instance f_i, a random offset $\boldsymbol{r}$ uniformly sampled from the problem domain was generated and added to $\boldsymbol{x}$. The offset is fixed for one training episode. So the swarm solves $f_i(\boldsymbol{x} + \boldsymbol{r})$ resulting in a randomly translated optimum $\boldsymbol{x}^*$. Now we were able to observe an optimization behavior within the trained swarms.

When just using the goodness of the optimization results as criterion for training, the achieved swarm behavior resembles more or less a random search. As our goal was to generate a swarm behavior that exhibits some emergent characteristics and shows self-organization, further criteria evaluating these characteristics need to be added.

Criteria for quantifying emergence are for example known from biology [47] or neuroscience [48]. Applications in computing science are scarce. An example for detecting emergence

in technical systems can be found in [49]. Many fractal analysis tools from chaos theory have a rather high computational complexity. At least, for the application within an objective function for training that has to be called millions of times. For our experiments we tested the so called correlation length as known from fitness landscape analysis [50]. When analyzing fitness landscapes, the correlation length is a criterion that measures the number of iterations after which the majority of succeeding solutions is statistically no longer correlated. It is calculated by $\lambda = -\frac{1}{\ln(\rho(1))}$ from the autocorrelation

$$\rho(\sigma) = \frac{E[f_k f_{k+\sigma}] - E[f_k]E[f_{k+\sigma}]}{V[f_k]} \tag{1}$$

of a series of consecutively sampled objective values f_{km}. When using the inverse version, we can maximize this distance. As additional indicators for desired swarm behavior we used the improvement relation

$$r_{\text{imp}} = \frac{1 + \frac{n_{\text{dec}} - n_{\text{inc}}}{n_{\text{dec}} + n_{\text{inc}}}}{2} \tag{2}$$

to maximize the number of improvements n_{inc} over decreasing optimization steps n_{dec}. Finally, we integrated the eventually reached swarm diameter to measure contraction. All criteria were combined in a scalarization approach.

IV. Results

For our experiments we used an islanding model for CGP training with two $(\mu + \lambda)$-ES. One was set to $\mu = 20$ and $\lambda = 100$ with a mutation probability of 0.04. The other was set to $\mu = 8$ and $\lambda = 16$ with a mutation probability of 0.4. Thus, we had a rather steadily evolving ES sending individuals every 1000 iteration and a small rather fast-paced, fluctuating one sending every 100 iterations; thus ensuring liveliness in exploration. The number of nodes was set to 20. As training optimization problems we used Rosenbrock, Bohachevsky, Alpine and Booth [51]. Because each candidate has to be evaluated several times for each of these functions, we limited the number of swarm iterations during the learning phase to 200. The number of particles was set to 5 during training due to performance issues.

Table I and II show the best result. We compared the learned optimization algorithm with a random search and with a real PSO. Random search was our bottom line that needs to be beaten. Table I compares the performance of the swarm, achieved with the number of particles set to 10 and with a budget of 10000 objective evaluations. The performance was tested on six different objective functions; three of which had not been used for learning. Compared with the pure random search, the learned optimization algorithm already behaves rater good, except for the Booth function. For the Rosenbrock function (4-dimensional) and the Six Hump Camel Back functions (2-dimensional), it is already competitive to the PSO. Table II shows the results when using a budget of 200000 objective evaluations; demonstrating that the learned algorithm is significantly better than random search.

In order to detect emergent behavior or at least to distinguish from pure random behavior in the system, we did a quick analysis using two criteria: The correlation dimension [52] and the Hurst exponent [53], [54]. The correlation dimension is a characteristic measure describing the geometry of chaotic attractors. One of the main applications of the Grassberger-Procaccia-algorithm is to distinguish between stochastic and deterministically chaotic time sequences [55]. We use it to analyze the fitness sequence generated along the path of particles. Table III shows example results for some test runs revealing that the behavior of the particles in the learned algorithm behave similar to the ones from PSO when attracted from good solutions. Each run reflects a different objective function. Although, when attacking function 4 from De Jong's test suite [56] which incorporates noise, the learned particle behavior seems to be attracted from more local optima at the same time (larger correlation dimension). The random approach shows no attraction behavior at all.

The Hurst exponent is a measure for the long-term memory of a time series. In this way the long-term statistical dependencies (excluding dependencies from cycles) seen in the series are evaluated [57]. A Hurst exponent of 0.5 denotes white noise. Larger values denote positive dependency, smaller negative dependency. The results in Table IV suggest that the PSO as well as the learned algorithm show a behavior of systematically improving solutions whereas the random search (as expected) exhibits mostly white noise.

V. Conclusion and further work

With the upcoming era of large scale cyber physical systems, the need for controlling numerous entities will in future most likely be accompanied by a growing demand of self-organizing algorithms. We presented a first approach to learn emergent swarm behavior. In a first step, individuals of a swarm were trained to jointly solve global optimization on arbitrary problem instances. So far, mere observation of other swarm members was incorporated. Nevertheless, CGP already was able to come up with solutions that are probable better than a mere random search.

Recently, recurrent CGP has been developed to foster the evolution of recurrent artificial neural networks [58]. For some other use cases, the recurrent version also showed superior performance [42]. On the other hand, necessary control and reduction of the number of recurrent connections introduces new challenges into learning [42]. Nevertheless, one of the next tasks will be to test this version for our use case. Other variants also provide promising extensions or modification [59].

Looking at the mid-term agenda, several challenges still have to be addressed.

- The question for detecting the desired emergent behavior is still open to future research.
- Moreover, if the desired emergent behavior is present, it needs to be quantified to generate appropriate guidance for sampling new solutions.
- What is the best objective function? Obviously, it is a mix of different criteria that would best be addressed

TABLE I
COMPARISON OF THE BEST LEARNED ALGORITHM WITH RANDOM SEARCH AND PSO WHEN USING A BUDGET OF 10.000 OBJECTIVE EVALUATIONS.

function	learned algorithm	random	PSO
Sphere	$5.555 \times 10^{-3} \pm 2.865 \times 10^{-3}$	$1.638 \times 10^{-2} \pm 1.667 \times 10^{-2}$	$4.692 \times 10^{-5} \pm 1.251 \times 10^{-4}$
Rosenbrock	$2.928 \times 10^{-1} \pm 1.824 \times 10^{-1}$	$1.06 \times 10^{-1} \pm 1.187 \times 10^{-1}$	$2.351 \times 10^{-1} \pm 1.045 \times 10^{0}$
Alpine	$1.822 \times 10^{-1} \pm 3.984 \times 10^{-1}$	$7.27 \times 10^{-3} \pm 6.257 \times 10^{-3}$	$2.335 \times 10^{-4} \pm 4.163 \times 10^{-4}$
Six Hump Camel Back	$-1.013 \times 10^{0} \pm 1.242 \times 10^{-2}$	$-9.927 \times 10^{-1} \pm 4.269 \times 10^{-2}$	$-1.031 \times 10^{0} \pm 9.789 \times 10^{-4}$
Booth	$3.837 \times 10^{0} \pm 5.941 \times 10^{0}$	$2.714 \times 10^{-2} \pm 2.603 \times 10^{-2}$	$3.529 \times 10^{-4} \pm 1.4 \times 10^{-3}$
DeJong f4	$3.478 \times 10^{-5} \pm 5.81 \times 10^{-5}$	$4.114 \times 10^{-4} \pm 8.626 \times 10^{-4}$	$1.365 \times 10^{-9} \pm 3.471 \times 10^{-9}$

TABLE II
COMPARISON OF THE BEST LEARNED ALGORITHM WITH RANDOM SEARCH AND PSO WHEN USING A BUDGET OF 200.000 OBJECTIVE EVALUATIONS.

function	learned algorithm	random	PSO
Sphere	$4.971 \times 10^{-7} \pm 3.86 \times 10^{-7}$	$7.158 \times 10^{-4} \pm 7.738 \times 10^{-4}$	$1.2 \times 10^{-12} \pm 4.33 \times 10^{-12}$
Rosenbrock	$1.668 \times 10^{-5} \pm 1.902 \times 10^{-5}$	$6.514 \times 10^{-3} \pm 6.835 \times 10^{-3}$	$5.476 \times 10^{-10} \pm 1.347 \times 10^{-9}$
Alpine	$7.914 \times 10^{-4} \pm 1.199 \times 10^{-3}$	$1.329 \times 10^{-3} \pm 6.917 \times 10^{-4}$	$6.947 \times 10^{-8} \pm 1.532 \times 10^{-7}$
Six Hump Camel Back	$-1.032 \times 10^{0} \pm 1.798 \times 10^{-6}$	$-1.03 \times 10^{0} \pm 1.453 \times 10^{-3}$	$-1.032 \times 10^{0} \pm 2.323 \times 10^{-12}$
Booth	$1.553 \times 10^{-6} \pm 1.293 \times 10^{-6}$	$1.042 \times 10^{-3} \pm 1.079 \times 10^{-3}$	$1.127 \times 10^{-11} \pm 3.635 \times 10^{-11}$
DeJong f4	$3.85 \times 10^{-13} \pm 6.305 \times 10^{-13}$	$4.378 \times 10^{-7} \pm 8.11 \times 10^{-7}$	$5.304 \times 10^{-20} \pm 2.652 \times 10^{-19}$

TABLE III
FRACTAL CORRELATION DIMENSION AS CRITERION TO DISTINGUISH STOCHASTIC AND DETERMINISTIC BEHAVIOR.

function	learned algorithm	random	PSO
Sphere	2.805	1.135×10^{-15}	2.659
Alpine	0.081	-6.107×10^{-16}	1.447
DeJong f4	2.295	2.928×10^{-16}	0.276

TABLE IV
HURST EXPONENT AS INDICATOR FOR LONG TERM MEMORY OF THE SWARM'S DYNAMIC SYSTEM.

function	learned algorithm	random	PSO
Sphere	0.936	0.556	0.872
Alpine	0.933	0.544	0.901
DeJong f4	0.949	0.497	0.758

in a multi-objective approach. In general, CGP could be solved as multi-objective optimization problem, but this would most likely generate severe performance problems.

- One major performance issue is the objective function. For each evaluation of a CGP solution candidate, an optimization procedure has to be run several times and different evaluation criteria have to be calculated.
- Finally, the question for the best set of functions is still open. Presumably, this set can be divided into an always necessary base set and problem specific extensions.

Then, further steps will be the inclusion of first stigmergy and second message-based information exchange. First simple tests of evolving agent-based negotiations via message are already promising for two agents.

REFERENCES

[1] M. Jipp and P. L. Ackerman, "The impact of higher levels of automation on performance and situation awareness," *Journal of Cognitive Engineering and Decision Making*, vol. 10, no. 2, pp. 138–166, 2016.

[2] T. B. Sheridan and R. Parasuraman, "Human-automation interaction," *Reviews of Human Factors and Ergonomics*, vol. 1, no. 1, pp. 89–129, 2016.

[3] R. Parasuraman and C. D. Wickens, "Humans: still vital after all these years of automation," *Human factors*, vol. 50, no. 3, pp. 511–520, 2008.

[4] European Commission, " Draft Ethics Guidelines for Trustworthy AI ," European Commission, Tech. Rep., 12 2018.

[5] B. Rapp, A. Solsbach, T. Mahmoud, A. Memari, and J. Bremer, "It-for-green: Next generation cemis for environmental, energy and resource management," in *EnviroInfo 2011 - Innovations in Sharing Environmental Observation and Information, Proceedings of the 25th EnviroInfo Conference 'Environmental Informatics'*, W. Pillmann, S. Schade, and P. Smits, Eds. Shaker Verlag, 2011, pp. 573 – 581.

[6] L. Cardwell and A. Shebanow, "The efficacy and challenges of scada and smart grid integration," *Journal of Cyber Security and Information Systems*, vol. 1, no. 3, pp. 1–7, 2016.

[7] D. W. McKee, S. J. Clement, J. Almutairi, and J. Xu, "Survey of advances and challenges in intelligent autonomy for distributed cyber-physical systems," *CAAI Transactions on Intelligence Technology*, vol. 3, no. 2, pp. 75–82, 2018.

[8] J. Collier, "Fundamental properties of self-organization," *Causality, emergence, self-organisation*, pp. 287–302, 2003.

[9] H. Parzyjegla, A. Schröter, E. Seib, S. Holzapfel, M. Wander, J. Richling, A. Wacker, H.-U. Heiß, G. Mühl, and T. Weis, "Model-driven development of self-organising control applications," in *Organic Computing – A Paradigm Shift for Complex Systems*. Springer, 2011, pp. 131–144.

[10] J. Dormans *et al.*, *Engineering emergence: applied theory for game design*. Universiteit van Amsterdam [Host], 2012.

[11] M. Parhizkar, G. D. M. Serugendo, and S. Hassas, "Leaders and followers: a design pattern for second-order emergence," in *2019 IEEE 4th International Workshops on Foundations and Applications of Self* Systems (FAS* W)*. IEEE, 2019, pp. 269–270.

[12] C. Hinrichs and M. Sonnenschein, "A distributed combinatorial optimisation heuristic for the scheduling of energy resources represented by self-interested agents," *International Journal of Bio-Inspired Computation*, vol. 10, no. 2, pp. 69–78, 2017.

[13] C. Prehofer and C. Bettstetter, "Self-organization in communication networks: principles and design paradigms," *IEEE Communications Magazine*, vol. 43, no. 7, pp. 78–85, 2005.

[14] J. F. Miller and S. L. Harding, "Cartesian genetic programming," in *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation*, 2008, pp. 2701–2726.

[15] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[16] J. R. Koza and R. Poli, "Genetic programming," in *Search methodologies*. Springer, 2005, pp. 127–164.

[17] D. O. Hebb, "The organization of behavior; a neuropsycholocigal theory," *A Wiley Book in Clinical Psychology*, vol. 62, p. 78, 1949.

[18] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[19] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[20] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*. Elsevier, 2020, pp. 101–121.

[21] M. Van Gerven and S. Bohte, "Artificial neural networks as models of neural information processing," *Frontiers in Computational Neuroscience*, vol. 11, p. 114, 2017.

[22] A. Burkov, *The hundred-page machine learning book*. Andriy Burkov Canada, 2019, vol. 1.

[23] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[24] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.

[25] F. Martinez-Gil, M. Lozano, and F. Fernandez, "Emergent behaviors and scalability for multi-agent reinforcement learning-based pedestrian models," *Simulation Modelling Practice and Theory*, vol. 74, pp. 117–133, 2017.

[26] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 49, no. 236, pp. 433–460, Oct. 1950.

[27] R. Forsyth, "BEAGLE a Darwinian approach to pattern recognition," *Kybernetes*, vol. 10, no. 3, pp. 159–166, 1981.

[28] N. L. Cramer, "A representation for the adaptive generation of simple sequential programs," in *Proceedings of an International Conference on Genetic Algorithms and the Applications*, J. J. Grefenstette, Ed., Carnegie-Mellon University, Pittsburgh, PA, USA, 24-26 Jul. 1985, pp. 183–187.

[29] J. R. Koza, "Non-linear genetic algorithms for solving problems," United States Patent 4935877, 19 Jun. 1990, filed may 20, 1988, issued june 19, 1990, 4,935,877. Australian patent 611,350 issued september 21, 1991. Canadian patent 1,311,561 issued december 15, 1992.

[30] ——, "Hierarchical genetic algorithms operating on populations of computer programs," in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCAI-89*, N. S. Sridharan, Ed., vol. 1. Detroit, MI, USA: Morgan Kaufmann, 1989, pp. 768–774.

[31] ——, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[32] L. F. D. P. Sotto, P. Kaufmann, T. Atkinson, R. Kalkreuth, and M. P. Basgalupp, "A study on graph representations for genetic programming," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, ser. GECCO '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 931–939. [Online]. Available: https://doi.org/10.1145/3377930.3390234

[33] J. Miller, *Cartesian Genetic Programming*, 06 2003, vol. 43.

[34] J. F. Miller, P. Thomson, and T. Fogarty, "Designing electronic circuits using evolutionary algorithms. arithmetic circuits: A case study," *Genetic algorithms and evolution strategies in engineering and computer science*, pp. 105–131, 1997.

[35] S. Harding, J. Leitner, and J. Schmidhuber, "Cartesian genetic programming for image processing," in *Genetic programming theory and practice X*. Springer, 2013, pp. 31–44.

[36] M. M. Khan, A. M. Ahmad, G. M. Khan, and J. F. Miller, "Fast learning neural networks using cartesian genetic programming," *Neurocomputing*, vol. 121, pp. 274–289, 2013.

[37] R. Hrbacek and V. Dvorak, "Bent function synthesis by means of cartesian genetic programming," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2014, pp. 414–423.

[38] J. R. Koza and J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992, vol. 1.

[39] J. F. Miller and S. L. Smith, "Redundancy and computational efficiency in cartesian genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 2, pp. 167–174, 2006.

[40] J. Miller, "What bloat? cartesian genetic programming on boolean problems," in *2001 Genetic and Evolutionary Computation Conference Late Breaking Papers*. San Francisco, California, USA, 2001, pp. 295–302.

[41] A. J. Turner and J. F. Miller, "Cartesian genetic programming: Why no bloat?" in *European Conference on Genetic Programming*. Springer, 2014, pp. 222–233.

[42] ——, "Recurrent cartesian genetic programming applied to famous mathematical sequences," in *Proceedings of the Seventh York Doctoral Symposium on Computer Science & Electronics*, 2014, pp. 37–46.

[43] E. O. Scott and S. Luke, "Ecj at 20: toward a general metaheuristics toolkit," in *Proceedings of the genetic and evolutionary computation conference companion*, 2019, pp. 1391–1398.

[44] J. Miller and N. C. Series, "Resources for cartesian genetic programming," *Cartesian Genetic Programming*, p. 337.

[45] J. Clegg, J. A. Walker, and J. F. Miller, "A new crossover technique for cartesian genetic programming," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 2007, pp. 1580–1587.

[46] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[47] V. Balaban, S. Lim, G. Gupta, J. Boedicker, and P. Bogdan, "Quantifying emergence and self-organisation of enterobacter cloacae microbial communities," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[48] E. P. Hoel, L. Albantakis, and G. Tononi, "Quantifying causal emergence shows that macro can beat micro," *Proceedings of the National Academy of Sciences*, vol. 110, no. 49, pp. 19 790–19 795, 2013. [Online]. Available: https://www.pnas.org/content/110/49/19790

[49] A. Shahbakhsh and A. Nieße, "Modeling multimodal energy systems," *Automatisierungstechnik : AT*, vol. 67, no. 11, pp. 893–903, 2019.

[50] J.-P. Watson, "An introduction to fitness landscape analysis and cost models for local search," in *Handbook of metaheuristics*. Springer, 2010, pp. 599–623.

[51] M. Jamil, X.-S. Yang, and H.-J. Zepernick, "8 - test functions for global optimization: A comprehensive survey," in *Swarm Intelligence and Bio-Inspired Computation*, X.-S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, and M. Karamanoglu, Eds. Oxford: Elsevier, 2013, pp. 193–222.

[52] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical review letters*, vol. 50, no. 5, p. 346, 1983.

[53] H. E. Hurst, "The problem of long-term storage in reservoirs," *Hydrological Sciences Journal*, vol. 1, no. 3, pp. 13–27, 1956.

[54] ——, "A suggested statistical model of some time series which occur in nature," *Nature*, vol. 180, no. 4584, pp. 494–494, 1957.

[55] P. Grassberger, T. Schreiber, and C. Schaffrath, "Nonlinear time sequence analysis," *International Journal of Bifurcation and Chaos*, vol. 1, no. 03, pp. 521–547, 1991.

[56] K. A. De Jong, "Analysis of the behavior of a class of genetic adaptive systems," Ph.D. dissertation, University of Michigan, Ann Arbor, 1975.

[57] R. Weron, "Estimating long-range dependence: finite sample properties and confidence intervals," *Physica A: Statistical Mechanics and its Applications*, vol. 312, no. 1-2, pp. 285–299, 2002.

[58] A. J. Turner and J. F. Miller, "Recurrent cartesian genetic programming of artificial neural networks," *Genetic Programming and Evolvable Machines*, vol. 18, no. 2, pp. 185–212, 2017.

[59] A. Manazir and K. Raza, "Recent developments in cartesian genetic programming and its variants," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–29, 2019.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 61–63
DOI: 10.15439/2021F55
ISSN 2300-5963 ACSIS, Vol. 26

Multicriterial evaluation and optimization of an algorithm for charging energy storage elements

Krasimir Kishkin *, Dimitar Arnaudov *, Venelin Todorov†‡, Sefka Fidanova ‡
*Technical University of Sofia 8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria
†Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
8 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
‡Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Email: k.kishkin@abv.bg , dda@tu-sofia.bg, vtodorov@math.bas.bg, venelin@parallel.bas.bg, stefka@parallel.bas.bg

***Abstract*—This study compares optimized active voltage balancing algorithms, applicable for energy storage systems made of supercapacitor cells connected in series. The results presented herein are obtained from a simulation model and confirmed on an experimental stand.**

I. Introduction

NOWADAYS, the most widely used elements for energy storage systems (ESS) are either Li-Ion cells or Supercapacitor cells. The control system or the Battery Management System (BMS) [1] has the task to charge and discharge them and manufactured a battery pack without any damage due to over voltages or over currents. Nowadays all BMSs use active voltage balancing techniques based on different methods [2]–[7].

II. A short description of the studied energy storage system

A simplified block diagram of the studied energy storage system (ESS) is shown in Fig. 1.

The DC/DC converters linked in parallel to each cell are additional charging sources (ACS). The DC/DC converter linked in parallel to the whole string is a main charging source (MCS). I_{main} is the main charging current. It charges the whole battery module. I_{add} is an additional charging current. I_{cell} is the cell charging current. For more details see [8], [9]. We stress on the fact that the simulations use technical information for 58F/16V module supercapacitor by Maxwell Technologies Inc. [10]. t_{fch} is the time necessary for the cell

Venelin Todorov is supported by KP-06-M32/2-17.12.2019 "Advanced Stochastic and Deterministic Approaches for Large-Scale Problems of Computational Mathematics" and by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" (ICTinSES), contract No. D01-205/23.11.2018, financed by the Ministry of Education and Science. The work is supported by the Bulgarian National Science Fund under Project DN 12/5-2017 "Efficient Stochastic Methods and Algorithms for Large-Scale Problems" and by the Project KP-06-Russia/17 "New Highly Efficient Stochastic Simulation Methods and Applications" funded by the National Science Fund - Bulgaria.

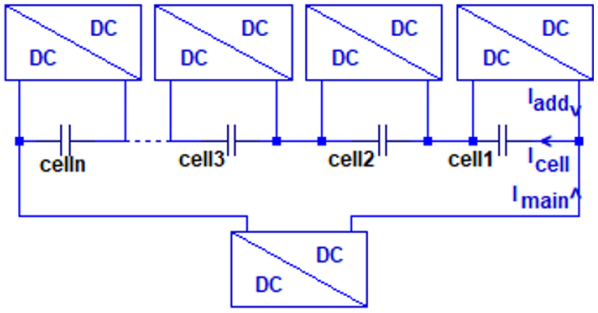


Fig. 1. ESS block diagram

with the highest capacitance C_{max} to charge from its initial voltage up to its rated voltage. It is described as

$$t_{fch} = f(C_{max}, I_{max}, \triangle U_{(C_{max})}). \quad (1)$$

III. The basic algorithm

In the basic algorithm (BASIC) the charging current is less than the maximum charging current I_{max}.

$$I_{cell} = I_{main} + I_{add} < I_{max} \quad (2)$$

I_{main} and I_{add} are strictly fixed. For each cell C_n and I_{add} is different and depends on the capacitance of the cell.I_{main} is given by:

$$I_{main} = I_{max}.C_{min}/C_{max}, [A] \quad (3)$$

I_{add} is given by:

$$I_{ACC} = I_{max}.((C_n - C_{min})/C_{max}), [A] \quad (4)$$

Here I_{max} is the maximum charging current. C_n is the capacitance of the n-th cell. C_{min} is the lowest capacitance and C_{max} is the highest capacitance. A detailed description of this algorithm can be found in [11]. Fig. 2 shows a typical charging process by using this algorithm.

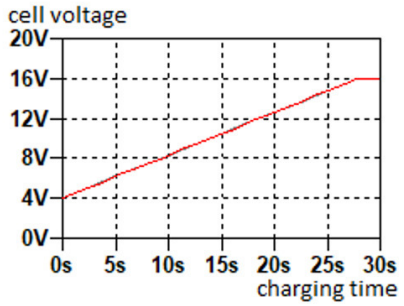


Fig. 2. Voltages across cells during charging

IV. The optimized algorithm

The second algorithm (OPTIMIZED) is optimized because it has a value $\triangle U$ of the voltage. It cannot be changed indefinitely. There are some optimal values, which, if skipped, result in other phenomena. Also some limit values for which the algorithm works optimally can be mentioned. Thats why for the future work we will make a 3D visualization, as a dependence on several quantities and to show their optimal value. The description of the optimized algorithm is described below. We begin by loading all the cells with a maximum charging current, being achieved through the synchronized work of the MCS and all ACSs which currents are equal. The charging current for each cell is:

$$I_{cell} = I_{main} + I_{add} = I_{max} \tag{5}$$

A detailed description of this algorithm can be found in [12]. Fig. 3 shows a typical charging process by using this algorithm.

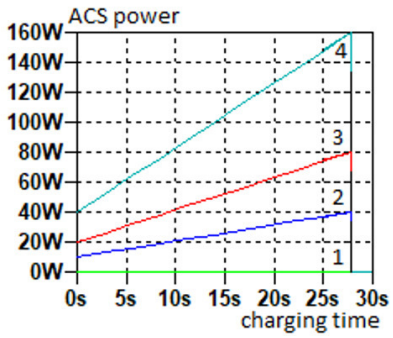


Fig. 3. Cell voltages during charging

The key point of the OPTIMIZED method is turning the ACSs off one after another and loaded the cells only by the MCS. For each cell C_n there is a specific moment $t_{x_{(C_n)}}$ when its ACS is turned off. For the time from $t_{x_{(C_n)}}$ to t_f, the charging current is provided only from the MCS. The specific for each cell moment t_x is given by:

$$t_x = t_{fch}.[I_{main}/I_{add}.(C_n/C_{max} - 1) + C_n/C_{max}] \tag{6}$$

On Fig. 4 it is shown what happens when the capacitance difference changes.

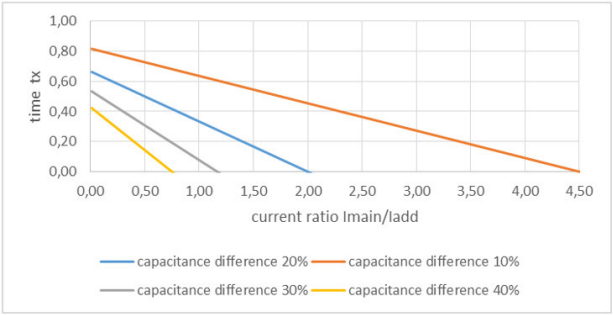


Fig. 4. t_x as a function of β

V. Simulation results and comparison of the studied algorithms

On Fig. 5 it is shown how the power of the additional charging sources changes.

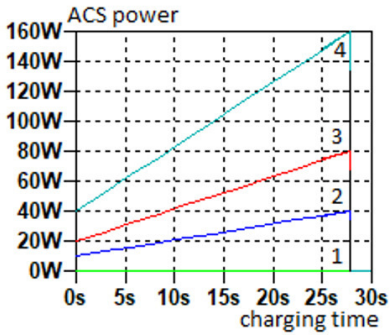


Fig. 5. ACS power

On Fig. 6 it is shown the power of the MCS.

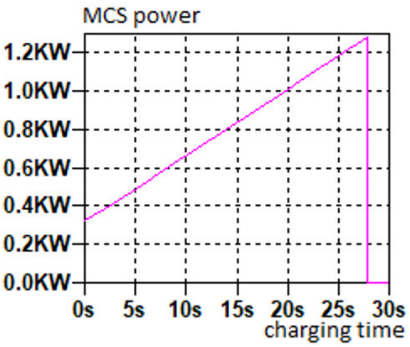


Fig. 6. MCS power

On Fig. 7 it is shown how the power of the ACSs changes during the process of charging.

Fig. 8 shows the power of the main charging source.

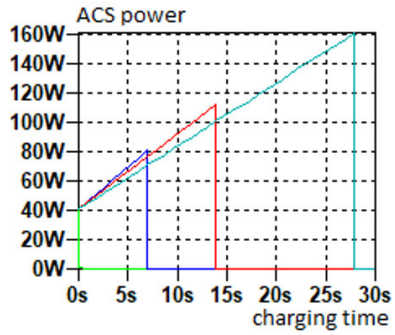


Fig. 7. ACS power

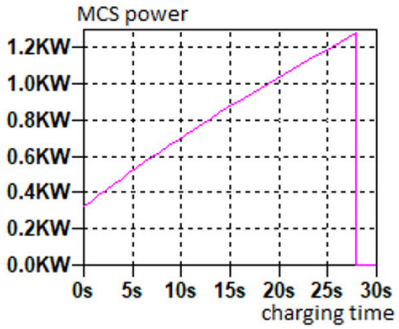


Fig. 8. MCS power

The summarized data are shown below. The notation b is for the BASIC algorithm and o is for the OPTIMIZED algorithm. The OPTIMIZED method gives better results for the sources. Table 1 shows calculated values for daily mean value of consumed energy and its statistical characteristics -– standard deviation, coefficient of variation, and confidence intervals at 0.95 confidence probability, where: P_a -– an average power; P_p — a peak power; E -– the energy;

TABLE I

source	Pa[W],b	Pa[W],o	Pp[W],b	Pp[W],o	E[J],b	E[J],o
ACS1	0	0	0	0	0	0
ACS2	25	60	40	80	696	422
ACS3	50	77	80	113	1392	1066
ACS4	100	101	160	160	2784	2800
MCS	800	827	1280	1282	22272	23019

VI. Conclusion

In the BASIC algorithm, the main charging current is fixed as a function of the cell with the lowest capacitance. A key element in the OPTIMIZED algorithm has the capacitance difference between the capacitor with the highest capacitance and the capacitor with the lowest capacitance. For the future work we will make a 3D visualization, as a dependence on several quantities and to show their optimal value.

References

[1] D. Andrea, "Battery Management Systems for Large Lithium-Ion Battery Packs", ARTECH HOUSE 2010
[2] P. Barrade "Series connection of supercapacitors: Comparative study of solution for the active equalization of the voltages"
[3] Yasser Diab, Pascal Venet, Gerard Rojat, "Comparison of the Different Circuits Used for Balancing the Voltage of Supercapacitors: Studying Performance and Lifetime of Supercapacitors", Author manuscript, published in "ESSCAP, Lausanne: Switzerland (2006)"
[4] Chunhe CHANG, Jiangping YANG, Yu LI, Zhongni ZHU, "Research of Supercapacitor Voltage Equalization Strategy on Rubber-Tyred Gantry Crane Energy Saving System", Energy and Power Engineering, 2010, 25-30
[5] S. Moore, P.Schneider, "A Review of Cell Equalization Methods for Lithium Ion and Lithium Polymer Battery Systems", Society of Automotive Engineers, Inc, Delphi Automotive Systems, 2001
[6] M. Daowd , N. Omar , P. Van Den Bossche, J. Van Mierlo, „Passive and Active Battery Balancing comparison based on MATLAB Simulation", Vehicle Power and Propulsion Conference (VPPC), 2011 IEEE
[7] C. Ionescu, A. Drumea, A. Vasile and N. Codreanu, "Investigations on Active Balancing Circuits for Supercapacitor Banks," 2018 41st International Spring Seminar on Electronics Technology (ISSE), Zlatibor, 2018, pp. 1-5, doi: 10.1109/ISSE.2018.8443679.
[8] D. Arnaudov, K. Kishkin and V. Dimitrov, "An Algorithm and Circuits for Active Balancing Systems," 2020 21st International Symposium on Electrical Apparatus & Technologies (SIELA), Bourgas, Bulgaria, 2020, pp. 1-4, doi: 10.1109/SIELA49118.2020.9167066.
[9] D. Arnaudov and K. Kishkin, "Modelling and Research of Synchronous Converter for Active Balancing System," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), Varna, Bulgaria, 2019, pp. 1-4, doi: 10.1109/ELMA.2019.8771689.
[10] https://files.ev-power.eu/inc/_doc/attach/StoItem/1123/ThunderSky-Winston-LIFEPO4-40Ah-Datasheet.pdf
[11] K. Kishkin, D. Arnaudov and D. Penev, "Algorithm for Charging a Supercapacitor Energy Storage System," 2020 43rd International Spring Seminar on Electronics Technology (ISSE), Demanovska Valley, Slovakia, 2020, pp. 1-6, doi: 10.1109/ISSE49702.2020.9120958.
[12] D. Arnaudov, D. Penev and K. Kishkin, "Management of Supercapacitor Battery Charging," 2020 43rd International Spring Seminar on Electronics Technology (ISSE), Demanovska Valley, Slovakia, 2020, pp. 1-7, doi: 10.1109/ISSE49702.2020.9121001.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 65–74
DOI: 10.15439/2021F35
ISSN 2300-5963 ACSIS, Vol. 26

The Extended Shift Minimization Personnel Task Scheduling Problem

Nico Kyngäs
Satakunta University of Applied Sciences, Satakunnankatu 23, 28130 Pori, Finland
Email: nico.kyngas@samk.fi

Kimmo Nurmi
Satakunta University of Applied Sciences, Satakunnankatu 23, 28130 Pori, Finland
Email: nico.kyngas@samk.fi

***Abstract*—In workforce scheduling, shift generation is the process of determining the shift structure, along with the tasks to be carried out in particular shifts. Application areas of shift generation include hospitals, retail stores, contact centers, cleaning, home care, guarding, manufacturing and delivery of goods. We present an extension to the Shift Minimization Personnel Task Scheduling Problem that is a problem in which a set of tasks with fixed start and finish times have to be allocated to a heterogeneous workforce. The objective in the SMPTSP is to minimize the number of employees required to carry out the given set of tasks. In the ESMPTSP, another objective is to maximize the number of feasible (shift, employee) pairs. We provide a mathematical formulation of the extended problem. We present an efficient ruin and recreate heuristic along with computational results for existing SMPTSP data sets and to a new data set. The presented heuristic is suitable for application in large real-world scenarios. The new instances, along with our best solutions, have been made available online.**

I. Introduction

SHIFT generation is the process of transforming the determined workload into shifts as accurately as possible. For labor-intensive industries, such as hospitals, retail stores, contact centers, cleaning, home care, guarding, manufacturing and delivery of goods, it is crucial to find a good match between the predicted and scheduled workload. The generated shifts form an input for the staff rostering, where employees are assigned to the shifts (see e.g. [1], [2] and [3]).

The generation of shifts is based on either the varying number of required employees working during the planning horizon or the tasks that the shifts must cover. We call these employee-based and task-based shift generation problems. The first major contribution for the employee-based shift generation problem was the study by Musliu et al. [4]. They introduced a problem, in which the workforce requirements for a certain period of time were given, along with constraints about the possible start times and the length of shifts, and an upper limit for the average number of duties per week per employee. Di Gaspero et al. [5] proposed a problem in which the most important issue was to minimize the number of different kinds of shifts used.

Kyngäs et al. [6] introduced the unlimited shift generation problem in which the most important goal is to minimize understaffing and overstaffing. They define a strict version of the problem, in the sense that each timeslot should be exactly covered by the correct number of employees. In the person-based multitask shift generation problem with breaks presented in [7], employees can have their personal shift length constraints and competences. The goal is to ensure that the employees can execute the shifts later in the staff rostering phase.

In the task-based shift generation problem the goal is to create shifts and assign tasks to these shifts so that the employees can be assigned to the shifts. The first major contribution of the task-based problem was the study by Dowling et al. [8]. They developed a day-to-day planning tool and to estimate a minimal staff set capable of operating as the ground staff of an international airport. Valls et al. [9] introduced a model where they minimized the number of workers required to perform a machine load plan. They presented a coloring approach to identify possible allocations along with bounds on the branch-and-bound search tree.

Krishnamoorthy and Ernst [10] introduced a similar group of problems, which they called Personnel Task Scheduling Problems (PTSP). Given the staff that are rostered on a particular day, the PTSP is to allocate each individual task, with specified start and end times, to available staff who have skills to perform the task. Later, Krishnamoorthy et al. [11] introduced a special case referred as Shift Minimization Personnel Task Scheduling Problem (SMPTSP) in which the goal is to minimize the number of employees used to perform the shifts. The SMPTSP has been studied under a few other names. Jansen [12] called SMPTSP the license and shift class design problem. Kroon et al. [13] called SMPTSP tactical fixed interval scheduling problem, and showed that solving it to optimality is NP-hard. The SMPTSP is also similar to the basic interval scheduling problem presented in [14] where the goal is to decide which jobs to process on which machines.

The General Task-based Shift Generation Problem (GTSGP) was defined in [15]. Given the tasks that should be rostered on a particular day, the GTSGP is to create anonymous shifts and assign tasks to these shifts so that employees can be assigned to the shifts. The targeted tasks must be completed within a given time window. For example, shelving in retail stores is often carried out in the forenoon. Some tasks are so-called back-office tasks. For example, in a contact center answering emails might require a given number of working hours per day dedicated to the activity but these tasks can be carried out any time of the day.

The main contributions of this paper are the following:

- a mathematical formulation of the Extended Shift Minimization Personnel Task Scheduling Problem (ESMPTSP)
- a ruin and recreate heuristic, which can successfully solve ESMPTSP instances
- a new benchmark set for the SMPTSP.

The paper is organized as follows. Section 2 first describes the Shift Minimization Personnel Task Scheduling Problem and the General Task-based Shift Generation Problem. Then we define the Extended Shift Minimization Personnel Task Scheduling Problem as an extension to the SMPTSP and as a highly simplified version of the GTSGP. In Section 3, we give the mathematical formulation of the ESMPTSP. We also present a simplified instance of the problem. Section 4 describes the most challenging SMPTSP benchmark instances, which we solve as ESMPTSP instances. Furthermore, we introduce a new benchmark instance set for the SMPTSP. This data set is generated especially for the ESMPTSP. In Section 5, we describe a ruin and recreate heuristic, which can successfully solve instances of SMPTSP, ESMPTSP and GTSGP. Finally, Section 6 presents the first computational results for solving the ESMPTSP. We also compare the results to the best-known SMPTSP results.

II. Problem Description

The Shift Minimization Personnel Task Scheduling Problem [11] can be defined formally as follows. A set of tasks $J = t_1,...,t_n$ needs to be allocated to a set of heterogeneous employees $E = e_1,...,e_m$ over a specified planning horizon. The processing time interval at which a task t has to be performed is determined by a timetable with fixed start time s_t and finish time f_t. Each employee e has a set of tasks $J_e \subseteq J$ that e can carry out. Each task t has a set of employees $E_t \subseteq E$ that can carry out t. All sets J_e and E_t are defined based on skills of employees/skill requirements of tasks and availability of employees/time windows of tasks. The objective is to minimize the number of employees required to perform the given set of tasks. The following basic assumptions hold:

A1. Preemption of tasks is not allowed.
A2. There are no precedence constraints among the tasks.
A3. Each task is processed only once without interruption.
A4. Each employee can execute only one task at a time.

The General Task-based Shift Generation Problem (GTSGP) [15] has the same assumptions besides (A2). However, the problem differs from the SMPTSP in several important ways:

B1. Tasks are not explicitly assigned to employees.
B2. Tasks are not fixed in time.
B3. Tasks may have shift-local precedence constraints.
B4. Transition times between tasks are considered.
B5. Employees have total working time restrictions.
B6. Employees have availability restrictions.

The GTSGP is to create anonymous shifts and assign tasks to these shifts so that employees can be assigned to the shifts. Instead of minimizing the number of employees required to carry out the given set of tasks, the objective is to maximize the number of feasible (shift, employee) pairs. The mathematical formulation of the problem was first given in [16]. The idea is to ensure that the resulting set of shifts can be carried out by the employees, i.e. each shift can be assigned to an employee s.t. all shifts are assigned to someone and no employee is assigned to multiple shifts.

In practical applications of the GTSGP, the full-time permanent and temporary employees are expected to cover 100% of the total workload in the shift generation, and later in the staff rostering phase. This is opposite to the idea behind the SMPTSP, where a large pool of casual staff is expected to be available and management would like to minimize the pool usage.

By including only requirements (B1) and (B2), we obtain the following simplified version of the GTSGP. The set of shifts S is to be generated. A set of tasks T is to be assigned to the shifts. Each task t has a duration d_t (in timeslots) and a time window $[lb_t, ub_t]$. A task t must not start before lb_t and must not end after ub_t. Each task is related to the collection of skills required by the tasks, which is a subset of the skill set C. Respectively, each employee e from the set of employees E has a collection K_e of skills. The number of shifts is usually the same as the number of available employees. In case of understaffing, additional pseudo employees can be used.

A solution to the simplified GTSGP is feasible if the following three hard constraints have no violations:

H1. The tasks in the shift do not overlap in time (overlap).
H2. Each shift can be executed by one or more employees, i.e. the skill set required by the tasks in each shift is possessed by one or more employees (shift).
H3. Each shift can be assigned to an employee s.t. all shifts are assigned to someone and no employee is assigned to multiple shifts (combination).

Modifying (B2) to allow only fixed timetables for the tasks, we have a further simplified version of the GTSGP. We call this problem the *Extended Shift Minimization Personnel Task Scheduling Problem (ESMPTSP)*. The objective is to first minimize the number of employees required to carry out the given set of tasks and then to maximize the number of feasible (shift, employee) pairs.

III. ESMPTSP Formulation

In this section, we give the mathematical formulation of the ESMPTSP. We also present a small instance of the ESMPTSP along with an example solution. We start with introducing some additional definitions and the decision variables:

P_j The set of employees that may carry out task j.
K_t The set of tasks active at time t.
K_t^w The set of tasks active at time t that worker w can perform.
C The family of sets containing all such sets K_t that $K_t \not\subseteq K_{t'}\ \forall t \neq t'$.
C^w The family of sets containing all such sets K_t^w that $K_t^w \not\subseteq K_{t'}^w\ \forall t \neq t'$.

$$x_{jwv} = \begin{cases} 1 & \text{if task } j \in J \text{ is assigned to employee } w \in W \\ & \text{and } v \in P_j \\ 0 & \text{otherwise.} \end{cases}$$

$$y_{wv} = \begin{cases} 1 & \text{if employee } w \in W \text{ is active and employee} \\ & v \in W \text{ can carry out the shift of } w \\ 0 & \text{otherwise.} \end{cases}$$

For $w \neq v$, we call x_{jwv} pseudoassignments of v to w with respect to j, as they represent whether j could be assigned to v assuming j is assigned to w. Similarly, we call y_{wv} pseudoassignments of v to w, as they represent whether all the tasks (and thus the entire shift) assigned to w could be assigned to v.

$$Z_{ESMPTSP} = min\left(\alpha * \sum_{w \in W} y_{ww} - \beta * \sum_{w,v \in W} y_{wv}\right) \quad (1)$$

$$\text{s.t.} \sum_{w \in P_j} x_{jww} = 1 \quad \forall j \in J \quad (2)$$

$$\sum_{j \in K_t^w} x_{jww} \leq y_{ww} \quad \forall w \in W, K_t^w \in C_w \quad (3)$$

$$y_{wv} \leq x_{jwv} - x_{jww} + 1 \quad \forall (j,w,v) \in J \times W \times W : w \neq v \quad (4)$$

$$y_{wv} \leq \sum_{w \in W} x_{jww} \quad \forall (j,w,v) \in J \times W \times W \quad (5)$$

$$x_{jwv} = 0 \quad \forall (j,w,v) \in J \times W \times W : w \notin P_j \text{ or } v \notin P_j \quad (6)$$

$$x_{jwv} = x_{jww} \quad \forall (j,w,v) \in J \times W \times W : w,v \in P_j \quad (7)$$

$$0 \leq y_{wv} \leq 1 \quad \forall w \in W, v \in W \quad (8)$$

$$x_{jwv} \in \{0,1\} \quad \forall (j,w,v) \in J \times W \times W \quad (9)$$

The objective function (Equation 1) consists of the weighted sum of the number of used employees and the number of able (employee, shift) pairs. The rest of the equations ensure the following:

(2) Each task will be carried out by exactly one able employee.
(3) No overlapping tasks are assigned to a single employee, and the indicator for using an employee indicates employee usage, i.e. that at least one task is assigned to the employee.
(4) A shift cannot be pseudoassigned to an employee if it has shifts the employee is unable to do.
(5) Empty shifts are not counted as pseudoassignments.
(6,7) Tasks are pseudoassigned according to both actual assignments and the abilities of the employees.
(8,9) Variables must be binary.

Fig. 1 shows a small instance of the ESMPTSP along with an example solution. The instance and the presented example of feasible (shift, employee) pairs have the following characteristics:

- The tasks in the shift do not overlap in time (overlap).
- The planning period is divided into 18 timeslots.
- The number of tasks is 14 and the number of employees is 7 (indicated by letters from A to G).
- The duration of the tasks is given by the length of the corresponding rectangles.
- The employees able to carry out a task are indicated by the letters in the rectangles.
- The parentheses indicate a (non-unique) feasible assignment between tasks/shifts and employees.
- The colors indicate which tasks are assigned to the same shift.
- The solution is clearly optimal in the number of shifts, as at least 6 concurrent shifts are needed during slot 12.
- Furthermore, employee E can carry out the blue shift, A and C the brown shift, D and E the green shift, E the yellow shift, B, D and F the violet shift, and A, D and E the red shift, totalling 18 feasible (shift, employee) pairs.

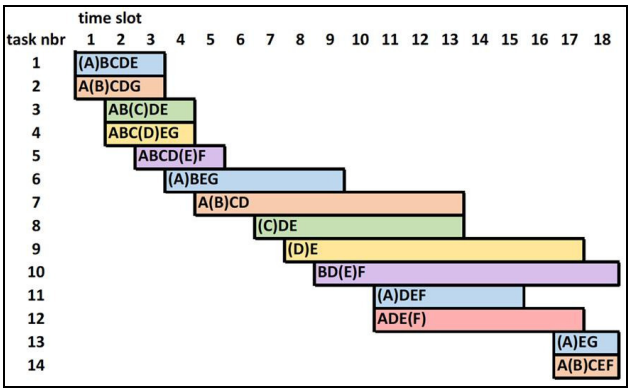


Fig. 1 A small instance of the ESMPTSP and a feasible solution. The letters indicate employees able to carry out a task. The colors indicate generated shifts.

IV. Benchmark Instances

We have basically two possibilities to create benchmark instances for the ESMPTSP: either to extend SMPTSP instances or to simplify GTSGP instances. We use both approaches. First, we use instances selected from the three current SMPTSP data sets. This approach has two benefits. The instances need no modifications whatsoever and a significant number of studies and algorithms have been designed to solve the problem. This enables the authors of the current SMPTSP algorithms to try their ideas to the ESMPTSP more easily. Second, we generate a new fourth data set for the SMPTSP, which is derived from the simplified GTSGP instances.

Krishnamoorthy et al. [11] presented the first data set of 137 instances for the SMPTSP. The data set is referred to as KEB instances. They used a Lagrangian approach to solve large instances of the SMPSTP. Smet et al. [17] generated the second data set of ten instances, because they were able to solve all KEB instances to optimality. The data set is referred to as SWMB instances. Fages and Lapegue [18] generated the third data set of 100 instances, because KEB and SWMB instances were trivial with respect to finding good quality lower bounds. This data set is referred to as FL instances.

As the first benchmark set for the ESMPTSP, we decided to select 56 very challenging instances from the three SMPTSP data sets. From the KEB data set, we selected 25 instances based on the performance of the three heuristics on the data set. An FL instance was selected if at least one of the following criteria holds:

C1. The solution (number of employees) obtained by the constructive heuristic of Lin and Ying [19] was at least 7% inferior to the optimum solution.

C2. The solution obtained by the iterative heuristic of Lin and Ying [19] was at least 2% inferior to the optimum solution.

C3. The greedy heuristic of Hojati [20] was unable to find the optimum solution.

Some of these instances are quite easy to solve as SMPTSP instances. This enables the authors of the SMPTSP algorithms to try their ideas to the ESMPTSP more easily. Note however, that we are not aware of how easy or difficult the instances are to solve as ESMPTSP instances.

From the SWMB data set, we selected all the ten instances.

From FL instances, we selected the 21 instances. We excluded instances #5 and #89, for which the minimum number of shifts is not known. Note that the instances were also unsolved by the Smet et al. method in [17]. An instance was selected if either one of the following criteria holds:

D1. The solution obtained by Fages and Lapegue [18] was at least 3% inferior to the known optimum solution.

D2. The greedy heuristic of Hojati [20] was not able find the optimum solution.

Tables I, II and III show the characteristics of the selected instances. The number of shifts denotes the known optimum value for the SMPTSP, i.e. the minimum number of shifts derived from the recent paper by Chandrasekharan et al. [21]. The @AVG measure indicates the estimated average number of tasks per non-empty shift, i.e. the number of tasks is divided by the minimum number of shifts.

The tightness level is defined as the total length of all tasks as a percentage of the total availability of all employees. The task skill level is defined as the average percentage of the total number of tasks each employee is qualified for. In addition to the task skill level, the shift skill level describes, how qualified an average employee is to carry out all the tasks of an average shift. We define shift skill level = t^a, where t = task skill level and a = @AVG. The overlap level is the probability of two random tasks to overlap. To calculate the probability, we need to iterate all task pairs once to check whether they overlap or not.

In order to make an instance challenging, the following guidelines can be drawn from the discussions and results in [11, 17, 18, 19 and 20]:

- The number of tasks and the number of employees influence the hardness of the instance, since they enlarge the search space.
- The combinatorial search space increases when the average number of tasks per shift increases.
- The tightness level should be closer to 90%.
- The task skill level should be at most 33%.
- The shift skill level should be way below 1%.
- The overlap level should be at most 40%.

However, these are general observations and no statistical evidence can be drawn from the guidelines. We generated a new fourth data set for the SMPTSP and the ESMPTSP based on the above guidelines. We had two goals for the new data set. First, the instances should be more challenging than the existing ones, and second, the characteristics of the instances should be distinctly different compared to the current data sets.

We have created an elaborate random test instance generator for the GTSGP (see [16]). The generator is guided by five parameter sets having a total number of seventeen parameters. We disabled and omitted most of the parameters to generate SMPTSP and ESMPTSP instances. The instances were generated using the following guidelines:

- The instances should be versatile.
- The minimum number of shifts should be the same as the number of employees.
- The number of employees varies from 20 to 500.
- The average number of tasks per shift is close to 5. This is about the same as the average in the other data sets.
- The tightness level should be at least 93% and in most of the cases close to 100%. This is clearly higher than in the other data sets. The level should decrease when the number of employees increases.

- The task skill level should vary from 5% to 65%. In some cases, the level should be clearly lower than in the other data sets. A higher value should increase the ESMPTSP solution value.
- The shift skill level should be at most 1%, and in some cases less than 0.001%. The variation should be about the same as in the other data sets altogether. Note, that seven instances have values close to 10%, which should increase the ESMPTSP solution value.
- The overlap level should be between 30% and 40%. This is on average higher than in the other data sets.

We decided to create two distinct instance sets. The first set of instances should have a unique optimum SMPTSP solution. When the objective is treated hierarchically, i.e. $\alpha = m^2 + 1$ and $\beta = 1$, this results in a unique optimum

TABLE I.
THE CHARACTERISTICS OF THE SELECTED KEB INSTANCES

#	#Emps	#Shifts	#Tasks	@AVG	Tightness level	Task skill level	Shift skill level	Overlap level
4	23	20	59	3.0	88.3	34.4	4.3	46.7
5	25	20	60	3.0	90.2	36.2	4.7	46.9
9	49	40	104	2.6	89.9	35.0	6.5	57.3
11	24	20	119	6.0	90.0	36.2	0.2	26.0
13	25	20	120	6.0	90.2	35.8	0.2	25.9
15	72	60	126	2.1	74.5	34.2	10.5	59.6
17	23	20	139	7.0	90.0	67.7	6.7	23.6
22	47	40	180	4.5	89.9	67.4	16.9	36.1
28	75	60	208	3.5	90.1	66.8	24.7	45.0
29	22	20	219	11.0	89.8	67.2	1.3	15.0
30	25	20	219	11.0	90.1	68.8	1.7	15.0
35	171	140	280	2.0	70.3	33.2	11.0	59.3
45	67	60	420	7.0	90.0	33.9	0.05	24.1
59	70	59	525	8.9	91.4	34.4	0.01	19.4
68	359	300	613	2.0	72.7	66.1	42.9	60.2
75	72	60	665	11.1	90.0	34.2	0.001	15.4
77	180	160	688	4.3	90.0	33.4	0.9	36.8
79	94	80	689	8.6	90.1	33.6	0.01	19.8
80	112	99	691	7.0	90.9	33.8	0.05	24.4
89	88	70	788	11.3	90.1	34.0	0.001	15.3
94	93	80	881	11.0	90.0	33.8	0.001	15.6
98	91	80	896	11.2	90.0	34.2	0.001	15.3
106	121	100	1096	11.0	90.0	33.3	0.001	15.6
107	114	100	1112	11.1	90.0	33.7	0.001	15.4
108	162	128	1115	8.7	91.4	33.6	0.008	19.9

TABLE II.
THE CHARACTERISTICS OF THE SWMB INSTANCES

#	#Emps	#Shifts	#Tasks	@AVG	Tightness level	Task skill level	Shift skill level	Overlap level
1	50	40	258	6.5	89.6	19.5	0.003	25.6
2	44	40	510	12.4	87.6	19.6	0.0000002	13.3
3	102	77	525	6.8	93.5	30.0	0.03	25.4
4	113	98	647	6.6	91.7	20.0	0.002	25.7
5	77	59	777	13.2	91.5	29.7	0,00001	13.2
6	135	116	777	6.7	92.9	19.9	0.002	25.8
7	70	59	781	13.2	88.5	19.9	0.00000006	13.1
8	88	79	1022	12.8	90.0	19.9	0,0000001	13.5
9	125	98	1308	13.2	90.9	19.8	0,00000005	13.2
10	153	116	1577	13.6	93.1	19.9	0,00000003	13.1

TABLE IV.
THE CHARACTERISTICS OF THE SELECTED FL INSTANCES

#	#Emps	#Shifts	#Tasks	@AVG	Tightness level	Task skill level	Shift skill level	Overlap level
28	262	105	402	3.8	19.1	26.0	0.6	11.0
29	248	95	355	3.7	18.5	28.3	0.9	11.6
31	290	116	488	4.2	21.0	25.7	0.3	10.4
33	338	132	534	4.0	20.3	25.7	0.4	10.8
35	308	118	469	4.0	19.8	26.6	0.5	11.1
39	284	108	446	4.1	19.8	25.7	0.4	10.6
45	376	144	586	4.1	20.5	27.0	0.5	11.4
46	409	157	635	4.0	20.2	26.6	0.5	11.2
54	498	190	850	4.5	22.5	25.8	0.2	10.7
60	443	173	783	4.5	21.9	27.0	0.3	10.7
61	551	222	891	4.0	20.0	26.3	0.5	11.0
62	610	262	1096	4.2	20.8	25.5	0.3	10.2
63	524	203	905	4.5	21.9	26.5	0.3	11.1
64	366	140	570	4.1	20.2	26.2	0.3	11.0
68	561	219	958	4.4	21.4	27.3	0.4	11.0
69	550	211	891	4.2	21.1	26.1	0.3	10.8
77	648	248	1123	4.5	22.1	26.6	0.3	10.7
79	638	246	1052	4.3	21.0	26.3	0.3	10.8
80	578	222	885	4.0	19.6	27.0	0.5	11.2
84	644	247	1121	4.5	21.9	26.1	0.2	10.4
94	812	313	1394	4.5	21.9	25.7	0.2	10.6

TABLE III.
THE CHARACTERISTICS OF THE NEW KN INSTANCES. THE INSTANCES DENOTED BY * SHOULD HAVE A UNIQUE OPTIMAL SMPTSP SOLUTION. LB = LOWER BOUND OBTAINED USING SOLYALI PROCEDURE.

#	#Emps	#Shifts LB	#Tasks	@AVG	Tightness level	Task skill level	Shift skill level	Overlap level
1*	20	20	105	5.3	100	24.5	0.06	33.0
2	25	25	125	5.0	93.9	19.8	0.03	32.8
3*	30	30	146	4.9	100	60.3	8.5	35.9
4	35	35	174	5.0	94.5	62.0	9.3	33.8
5*	40	40	200	5.0	100	18.8	0.02	35.2
6	45	45	216	4.8	95.5	17.7	0.02	35.2
7*	50	50	266	5.3	100	32.1	0.2	33.5
8	55	55	265	4.8	95.0	63.3	11.1	35.0
9*	60	60	297	5.0	99.5	20.3	0.04	35.6
10	65	65	335	5.2	95.5	54.9	4.6	33.2
11*	70	70	352	5.0	99.5	19.8	0.03	35.2
12	75	75	363	4.8	94.0	39.9	1.2	34.6
13*	80	80	417	5.2	99.0	15.0	0.005	34.0
14	85	85	434	5.1	95.0	36.5	0.6	33.3
15*	100	100	506	5.1	98.5	10.8	0.001	34.7
16	110	110	583	5.3	95.5	42.5	1.1	32.4
17*	120	120	613	5.1	98.0	7.7	0.0002	34.4
18	130	130	641	4.9	94.5	31.1	0.3	34.5
19*	140	140	670	4.8	97.6	5.2	0.00007	36.4
20	150	149	782	5.2	95.0	29.4	0.2	32.9
21	160	160	802	5.0	97.3	62.7	9.6	34.8
22	180	180	843	4.7	97.0	37.8	1.1	36.9
23	200	199	967	4.8	95.0	19.9	0.04	35.3
24	240	239	1157	4.8	96.6	57.9	7.1	35.9
25	280	279	1419	5.1	96.3	32.5	0.3	34.3
26	320	309	1611	5.0	95.0	5.4	0.00004	32.7
27	360	356	1763	4.9	96.0	63.7	11.0	35.2
28	400	399	2012	5.0	96.0	40.0	1.0	34.4
29	450	443	2231	5.0	94.5	12.0	0.003	34.4
30	500	488	2473	4.9	93.5	8.6	0.0005	34.1

ESMPTSP solution as well. The second set of instances can have many different optimum SMPTSP solutions per instance, from which we should find the best ESMPTSP solution. These instances should be easier to solve as tightness level decreases and shift skill level increases.

The generator creates instances where the minimum number of shifts should be the same as the number of employees. The basic idea is to construct an instance along with a feasible solution. Applying the lower bounding procedure of Solyali [22] confirmed this for 21 instances. However, when the number of employees increases and the tightness level decreases, the generator may create instances where the minimum number of shifts is less than the number of employees.

Table IV shows the characteristics of the thirty instances generated. We refer to this data set as KN instances. The characteristics clearly show that the instances are distinctly different from the existing data sets. The instances denoted by * belong to the first set of instances, which should have a unique optimum SMPTSP solution.

The existing three data sets can be found in [23] and the new KN data set in [24]. We provide the new KN instances in the traditional SMPTSP format as well as in the GTGSP format.

V. Ruin and Recreate Heuristic

We use a ruin and recreate heuristic similar to that described in [25] to solve the ESMPTSP. Pseudo code for the algorithm is given in Fig. 2. Our version of ruin and recreate heuristic (2RH) has been created to solve practical GTSGP instances. In practical applications of the GTSGP, we should

- generate as versatile shifts as possible to
- ensure that the rostering of the staff can be completed, so that
- the computation time is still acceptable considering the release time of the rosters.

Therefore, we do not seek the fastest possible solution method. Instead, it is advantageous to use more computation time in order to achieve shifts that are more versatile.

The ruin operator first chooses a random task t_0 assigned to some shift s_0 and removes a random sequence of adjacent tasks from s_0 containing t_0. For each task sorted by closeness to t_0 w.r.t. time windows and skills, similar removal is done if the corresponding shift has not been removed from yet. The underlying idea is that by removing whole strings of tasks from shifts at a time, room is created for new tasks to be inserted, and due to the way removed tasks are selected, at least some of them are more or less interchangeable between shifts. When the total number of removed tasks exceeds the given parameter, the ruin operator quits.

The recreate operator adds free tasks one by one to their respective best positions in the incumbent solution. First, all free tasks are sorted in order of how many times they have been unassigned after recreation during the solution process. This is done in order to emphasize tasks that seem more difficult to place in the solution. For each task *t*, all feasible addition positions in the incumbent solution are evaluated.

The concept of a position depends on the exact problem. For the GTSGP, a position is determined by an immediate predecessor, e.g. a task or an employee. Note that the tasks can have wide time windows in the GTSGP, hence the order of tasks within a shift is not predetermined. In the SMPTSP there are no time windows, which fixes the order of tasks within a shift. Task *t* is then added to the position that leads to the best objective function value, with a small chance to skip over to the next best position. Consecutive skipping is not constrained in any way, so *t* might not get assigned even if it has feasible addition positions. When all free tasks have been processed, the recreate operator quits.

Essential parameters of 2RH and the values used in our computational experiments include

- average number of tasks removed per ruin operator (10),
- maximum length of task string to remove from a single shift (8),
- recreate skipping chance (1%), and
- move skipping chance (1%).

```
round ← 0, bestSol ← null
while round < f do
    storedSol ← currentSol
    seed ← random currently assigned task from T
    tm ← maximum number of tasks to ruin per shift
    tasks ← list of all tasks ordered by distance from seed
    S = ∅
    for t ∈ tasks
        if S(t) ∉ S
            l ← U (1, min(|S(t)|, tm))
            Remove a random string of l tasks from S(t)
            Update ruin quota
            S ← S ∪ S(t)
        end if
        if ruin quota is full
            break
        end if
    end for
    tasks ← list of all unassigned tasks in unassignment
            count order
    for t ∈ tasks
        P ← all spots in all current shifts where adding t is
            feasible in worsening order of objective
        for p ∈ P
            if U (1, 100) ≥ lowLevelSkipChance
                Add task t to spot p
                break
            end if
        end for
    end for
    Update bestSol if necessary
    if U (1, 100) ≤ highLevelSkipChance
        currentSol ← storedSol
    end if
    round ← round + 1
end while
```

Fig. 2 The pseudo-code of the ruin and recreate heuristic.

VI. Computational Results

This section presents our computational results for the ESMPTSP benchmark instances introduced in Section 4. As was stated in previous section, our implementation of 2RH requires sufficient running time to tackle the largest practical instances. For each ESMPTSP benchmark instance, we execute eight parallel 2RH runs exactly four hours. We also register the running time elapsed to reach the first solution to the SMPTSP as well as the value for the ESMPTSP at that time.

Note that our implementation of 2RH is such that the best solution is generally reached at the later stages of the optimization run. We could find the first solution to the SMPTSP faster if we used less maximum running time for 2RH. It should also be noted here, that running parallel 2RHs one hour instead of four hours seems to weaken ESMPTSP solutions only at most 5%. This could be acceptable for practical applications, since after the shift generation and staff rostering have completed, new tasks will most certainly arise and some of the tasks need to be changed or removed.

Tables V, VI, VII and VIII show our results for the benchmark instances. We ran the instances using hierarchical objective, i.e. $\alpha = m^2 + 1$ and $\beta = 1$, i.e. only solutions optimal w.r.t. to the underlying SMPTSP need to be considered. The optimum values for the underlying SMPTSPs were derived from the recent paper by Chandrasekharan et al. [21]. The test runs were carried out on a workstation with AMD Ryzen 9 3950X 16-Core Processor at 3.49GHz and with 64GB RAM running Windows 10 using default settings. 2RH is implemented in C++.

We used no domain specific knowledge in order to generate better solutions, nor did we fine-tune any parameter. However, we also experimented with different numbers of parallel runs, different parameter values and different running times. We publish the best solutions we have found during these experiments, but we emphasize that these solutions are shown only for comparison purposes. For comparison purposes, we also conducted 30-minute test runs with Gurobi 8.1. Gurobi was able to verify, that we have found the optimum ESMPTSP solution for 16 instances. For the other 70 instances, we do not know the optimum values. Note that Gurobi was not able to find optimum solutions to any of the SWMB and FL instances.

First, the results show that 2RH can solve the underlying SMPTSP instances extremely well and sufficiently fast. With respect to the SMPTSP, the heuristic can successfully solve the most challenging existing benchmark instances as well as the new KN instances. The SWMB instances are very challenging. These instances have a high average number of tasks per shift and low task skill levels, which implies very low shift skill levels.

With respect to the running time required to reach the SMPTSP optimum, SWMB4 and SWMB8 are easier, and SWMB3 and SWMB6 harder. We could not solve SWMB7 and SWMB10 instances within the given time limit using the given parameter set. However, we did find the SMPTSP optimum by increasing the time limit.

The FL instances are quite easy to solve. The instances have very low tightness levels and shift skill levels are not too low. The KEB instances are easy to solve. This is true even for those instances that have low shift skill levels. We have no other reason for this than the high task skill levels. With respect to the running time required to reach the SMPTSP optimum, KEB080 and FL77 are the hardest ones.

Our goal in creating the KN data set was to generate instances that are more challenging. The results indicate this to be true at least for our ruin and recreate heuristic. There are mainly two reasons for this: higher tightness levels and lower task skill levels.

With respect to the running time required to reach the SMPTSP optimum, the easiest instances are KN5 and KN19. Among the instances with equal number of shifts and employees, the hardest instances are KN7 and KN22. We could not solve KN26, KN29 and KN30 instances to the lower bound value. We suspect that our solutions to these instances are not optimal.

We define the ratio $r = e/s$, where s = the best-known SMPTSP value and e = the number of feasible pairs in the best-known ESMPTSP solution. In general, we could argue that instances with high r values should be easier to solve as SMPTSP and as ESMPTSP instances, because several employees may carry out several shifts. This seems to be true, because the FL instances have a very high r value, and the SWMB instances have r values very close to one. However, for the KEB and KN instances, there is no correlation between r values and hardness of instances.

We only know the optimum ESMPTSP values for 16 of the 86 benchmark instances. It should be clear, that the combinatorial search space for an ESMPTSP instance increases when its r value increases. Therefore, we speculate that our solution to the ESMPTSP should be closer to the optimum value for those instances that have low r values.

The KN instances intended to have a unique optimum SMPTSP solution turned out to be almost trivial for Gurobi. The corresponding SMPTSP instances were solved in a few seconds each. It seems likely that the instances are so heavily constrained that methods focused on reducing the search space are far superior to any pure heuristics.

VII. Conclusions

We presented a mathematical formulation of the Extended Shift Minimization Personnel Task Scheduling Problem (ESMPTSP), which in turn is a highly simplified version of the GTSGP. We showed that the presented 2RH heuristic can successfully solve ESMPTSP benchmark instances. Furthermore, we showed that the heuristic was able to find optimal solutions to the SMPTSP instances.

We published a new data set for the SMPTSP and the ESMPTSP. We provide the new instances in the traditional SMPTSP format as well as in the GTGSP format. The instances, along with our best solutions, have been made available online [24].

This was the first encounter of solving the SMPTSP instances as ESMPTSP instances. Even though the computational results were encouraging, we suspect that better solutions for most of the instances exist. Furthermore, there should be room for both more efficient solution methods and efficient lower bounding methods. These would also bring more insight to the hardness of the problem as well as to the hardness of the current benchmark instances.

No matter the point during the planning process at which the problem is solved, there will always be changes, be it to the tasks themselves due to e.g. changed customer expectations or the employees at our disposal due to e.g. sick leaves. In practice, solutions with equal number of shifts and larger number of feasible pairs are better because the flexibility of the assignment of the shifts is increased, making it easier to assign existing shifts to different employees. This justifies the ESMPTSP in the big picture of the workforce optimization and the real-world workforce scheduling process. We believe that the presented method is suitable for application in large real-world scenarios.

TABLE V.
THE RESULTS FOR KEB INSTANCES. TIMES GIVEN IN MINUTES. THE FINAL ESMPTSP SOLUTION IS THE BEST OF EIGHT FOUR-HOUR PARALLEL 2RH RUNS. THE BEST SOLUTIONS WE HAVE FOUND IN OUR OTHER TEST RUNS ARE ALSO REPORTED. THE SOLUTIONS DENOTED BY ^{O} ARE OPTIMUM.

#	SMPTSP optimum	Time to reach the SMPTSP optimum	# of feasible pairs at that time	Final # of feasible pairs	Best # we have found
4	20	0.001	30	53 ^{o}	
5	20	0.001	41	62 ^{o}	
9	40	0.06	161	233	
11	20	0.03	20	29	30 ^{o}
13	20	0.03	20	31	32 ^{o}
15	60	0.02	506	684	694
17	20	0.1	58	97	
22	40	0.1	351	518	524
28	60	0.3	1152	1489	1497
29	20	0.2	33	68	69
30	20	0.3	31	78	
35	140	0.06	2701	3534	3558
45	60	2	63	110	114
59	59	1.7	60	79	85
68	300	0.2	46422	49448	49539
75	60	3	60	76	84
77	160	1.8	485	1021	1053
79	80	3	84	127	131
80	99	9	121	214	219
89	70	0.9	70	92	95
94	80	4	80	110	115
98	80	6	81	109	113
106	100	3	100	140	146
107	100	4	100	140	146
108	128	7	136	220	232

TABLE VI.
THE RESULTS FOR SWMB INSTANCES. TIMES GIVEN IN MINUTES. THE FINAL ESMPTSP SOLUTION IS THE BEST OF EIGHT FOUR-HOUR PARALLEL 2RH RUNS. THE BEST SOLUTIONS WE HAVE FOUND IN OUR OTHER TEST RUNS ARE ALSO REPORTED. SWMB7 AND SWMB10 COULD NOT BE SOLVED WITHIN THE GIVEN TIME LIMIT.

#	SMPTSP optimum	Time to reach the SMPTSP optimum	# of feasible pairs at that time	Final # of feasible pairs	Best # we have found
1	40	67	40	41	45
2	40	30	40	41	42
3	77	139	83	101	117
4	98	15	98	111	114
5	59	86	59	60	
6	116	207	116	129	
7	59	*	*	*	59
8	79	30	79	80	
9	98	102	98	99	
10	116	*	*	116	

TABLE VII.
THE RESULTS FOR FL INSTANCES. TIMES GIVEN IN MINUTES. THE FINAL ESMPTSP SOLUTION IS THE BEST OF EIGHT FOUR-HOUR PARALLEL 2RH RUNS. THE BEST SOLUTIONS WE HAVE FOUND IN OUR OTHER TEST RUNS ARE ALSO REPORTED.

#	SMPTSP optimum	Time to reach the SMPTSP optimum	# of feasible pairs at that time	Final # of feasible pairs	Best # we have found
28	105	0.1	3181	4034	4051
29	95	0.1	3389	4285	4296
31	116	0.1	3614	4505	4535
33	132	0.2	4728	5915	5933
35	118	0.1	4418	5374	5393
39	108	0.2	3318	4215	4224
45	144	0.3	6526	8456	8461
46	157	0.7	7306	9699	9702
54	190	1	9560	11321	
60	173	0.6	8102	9648	9698
61	222	2	14084	17590	17646
62	262	3	16776	19419	19497
63	203	1	11140	13088	13191
64	140	0.3	5981	7499	7524
68	219	2	14331	18576	18682
69	211	1	12506	15945	16045
77	248	14	17619	20824	20986
79	246	2	16974	21917	22056
80	222	1	16096	20034	20154
84	247	4	17515	21953	22110
94	313	6	25574	30006	30342

TABLE VIII.

THE RESULTS FOR KN INSTANCES. THE INSTANCES DENOTED BY * SHOULD HAVE A UNIQUE OPTIMAL SMPTSP SOLUTION. TIMES GIVEN IN MINUTES (X = NOT FOUND WITHIN THE GIVEN TIME LIMIT). THE FINAL ESMPTSP SOLUTION IS THE BEST OF EIGHT FOUR-HOUR PARALLEL 2RH RUNS. THE BEST SOLUTIONS WE HAVE FOUND IN OUR OTHER TEST RUNS ARE ALSO REPORTED. KN26, KN29 AND KN30 COULD NOT BE SOLVED WITHIN THE GIVEN TIME LIMIT. THE SOLUTIONS DENOTED BY ° ARE OPTIMUM.

#	Best found SMPTSP solution	Time to reach the best found SMPTSP solution	# of feasible pairs at that time	Final # of feasible pairs	Best # we have found
1*	20	0.002	43	43 °	
2	25	0.03	55	57 °	
3*	30	26	235	235 °	
4	35	1	365	456	
5*	40	0.08	89	89 °	
6	45	0.3	96	101°	
7*	50	192	147	147 °	
8	55	0.6	1063	1369	1375
9*	60	21	76	76 °	
10	65	3.8	753	1259	1271
11*	70	36	118	118 °	
12	75	0.9	945	1361	1378
13*	80	43	220	220 °	
14	85	29	791	1106	1160
15*	100	37	128	128 °	
16	110	34	1446	2562	2621
17*	120	11	131	131 °	
18	130	10	1471	3012	3150
19*	140	1	152	152 °	
20	149	128	1620	2513	2933
21	160	42	8473	11858	12101
22	180	179	4411	5952	6888
23	199	148	1852	2163	2654
24	239	92	18041	24858	25835
25	279	110	6452	9674	11778
26	316	*	*	*	795
27	356	156	47213	63392	66171
28	399	92	24160	44071	51277
29	445	*	*	*	5190
30	495	*	*	*	2731

REFERENCES

[1] A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: A review of applications, methods and models", European Journal of Operational Research vol. 153, no. 1, pp. 3-27, 2004.

[2] J. Van den Bergh, J. Belin, P. De Bruecker, E. Demeulemeester and L. De Boeck, "Personnel scheduling: A literature review", European Journal of Operational Research vol. 226, no. 3, pp 367-385, 2013.

[3] L. Kletzander and N. Musliu, "Solving the General Employee Scheduling Problem", Computers and Operations Research, vol. 13: 104794, 2020.

[4] N. Musliu, A. Schaerf and W. Slany, "Local search for shift design", European Journal of Operational Research, vol. 153, no. 1, pp. 51-64, 2004.

[5] L. Di Gaspero, J. Gärtner, G. Kortsarz, N. Musliu, A. Schaerf and W. Slany, "The minimum shift design problem", Annals of Operations Research, vol. 155, no. pp. 79-105, 2007.

[6] N. Kyngäs, D. Goossens, K. Nurmi and J. Kyngäs, "Optimizing the unlimited shift generation problem", In: Di Chio C. et al. (eds) Applications of Evolutionary Computation. EvoApplications, Lecture Notes in Computer Science, vol. 7248, pp. 508-518, 2012.

[7] N. Kyngäs, K. Nurmi and J. Kyngäs, "Solving the person-based multitask shift generation problem with breaks", Proceedings of the 5th International Conference On Modeling, Simulation And Applied Optimization, pp. 1-8, 2013.

[8] D. Dowling, M. Krishnamoorthy, H. Mackenzie and H. Sier, "Staff rostering at a large international airport", Annals of Operations Research vol. 72, pp. 125-147, 1997.

[9] V. Valls, A. Perez and S. Quintanilla, "A graph colouring model for assigning a heterogenous workforce to a given schedule", European Journal of Operations Research vol. 90, pp. 285-302, 1996.

[10] M. Krishnamoorthy and A.T. Ernst, "The personnel task scheduling problem", Optimization Methods and Applications, pp. 343–367, 2001.

[11] M. Krishnamoorthy, A.T. Ernst and D. Baatar, "Algorithms for large scale Shift Minimisation Personnel Task Scheduling Problems", European Journal of Operational Research, vol. 219, no. 1, pp. 34-48, 2012.

[12] K. Jansen, "An approximation algorithm for the license and shift class design problem", European Journal of Operational Research vol. 73, pp. 127-131, 1994.

[13] L.G. Kroon, M. Salomon and L.N. Van Wassenhove, "Exact and approximation algorithms for the tactical fixed interval scheduling problem", Operations Research vol. 45, no. 4, pp. 624-638, 1997.

[14] A.W.J. Kolen, J.K. Lenstra, C.H. Papadimitriou and F.C.R. Spieksma, "Interval Scheduling: A Survey", Naval Research Logistics vol. 54, no. 5, pp. 530-543, 2007.

[15] K. Nurmi, N. Kyngäs and J. Kyngäs, "Workforce Optimization: the General Task-based Shift Generation Problem", IAENG International Journal of Applied Mathematics, vol. 49, no. 4, pp. 393-400, 2019.

[16] N. Kyngäs, K. Nurmi and D. Goossens, "The General Task-based Shift Generation Problem: Formulation and Benchmarks", Proceedings of the 9th Multidisciplinary International Scheduling Conference: Theory and Applications (MISTA), pp. 301-319, 2019.

[17] P. Smet, T. Wauters, M. Mihaylov and G. Vanden Berghe, "The shift minimization personnel task scheduling problem: A new hybrid approach and computational insights", Omega vol. 46, pp. 64-73, 2014.

[18] J.G. Fages and T. Lapegue, "Filtering Atmostnvalue with Difference Constraints: Application to the Shift Minimisation Personnel Task Scheduling Problem", Lecture Notes in Computer Science, vol. 8124, pp. 63-79, 2013.

[19] S.-W. Lin and K.-C. Ying, "Minimizing Shifts for Personnel task Scheduling Problems: A three-Phase Algorithm", European Journal of Operational Research vol. 237, pp. 323-334, 2014.

[20] M. Hojati, "A greedy heuristic for shift minimization personnel task scheduling problem", Computers and Operations Research vol. 100, pp. 66-76, 2018.

[21] R. Chirayil Chandrasekharan, P. Smet, and T. Wauters, "An automatic constructive matheuristic for the shift minimization personnel task scheduling problem", J Heuristics, Feb. 2020, doi: 10.1007/s10732-020-09439-9.

[22] O. Solyali, "The Shift Minimization Personnel Task Scheduling Problem: An Effective Lower Bounding Procedure", Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, vol. 34, no. 2, Jun. 2016, doi: 10.17065/huniibf.259136.

[23] T. Lapègue: "Personnel Task Scheduling Problem Library", [Online]. Available: https://sites.google.com/site/ptsplib/smptsp/instances, (Last access 15-January-2021).

[24] K. Nurmi: "The General Task-based Shift Generation Problem - Benchmark Instances", [Online]. Available: http://web.samk.fi/public/tkiy/GTSGP/, (Last update 26-July-2021).

[25] K. Sörensen, M. Sevaux and F. Glover, "A History of Metaheuristics", In: R. Martí, P. Pardalos and M. Resende (eds) Handbook of Heuristics, pp. 791-808, 2018.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 75–80
DOI: 10.15439/2021F94
ISSN 2300-5963 ACSIS, Vol. 26

Optimized lattice rule and adaptive approach for multidimensional integrals with applications

Venelin Todorov *†, Ivan Dimov†, Stefka Fidanova †, Stoyan Poryazov *
*Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
8 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
†Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Email: vtodorov@math.bas.bg, venelin@parallel.bas.bg, ivdimov@bas.bg, stefka@parallel.bas.bg, stoyan@math.bas.bg

***Abstract*—In this work we make a comparison between optimized lattice and adaptive stochastic approaches for multidimensional integrals with different dimensions. Some of the integrals has applications in environmental safety and control theory.**

I. Introduction

MONTE Carlo methods are one of the most commonly used numerical methods. Their advantages are enhanced by increasing the dimensionality. For this reason, they are a major tool for numerically solving classes of problems in such important areas as particle physics, engineering chemistry, molecular dynamics, and financial mathematics. A major scientific challenge in the development of modern Monte Carlo methods is their relatively slow rate of convergence, which in many cases has the asymptotic $O(N^{-1/2})$, where N is the sample size. There are two approaches to improve convergence - reducing the variance of the estimated value and reducing the discrepancy of the sequence used. Adaptive strategy and lattice rules are two different ways to improve the convergence and has never been compared on this type of multidimensional integrals before.

II. The stochastic approaches

A. Adaptive approach

Adaptive strategy [1], [3], [4], [7] is well known method for evaluation of multidimensional integrals, especially when the integrand function has peculiarities and peaks. Let p_j and I_{Ω_j} be the following expressions: $p_j = \int_{\Omega_j} p(\mathrm{x})\,\mathrm{dx}$ and $I_{\Omega_j} = \int_{\Omega_j} f(\mathrm{x})p(\mathrm{x})\,\mathrm{dx}$. Consider now a random point $\xi^{(j)} \in \Omega_j$ with a density function $p(\mathrm{x})/p_j$. In this case $I_{\Omega_j} = \mathrm{E}\left[\frac{p_j}{N}\sum_{i=1}^{N} f(\xi_i^{(j)})\right] = \mathrm{E}\theta_N$. This adaptive algorithm gives an approximation with an error $\varepsilon \leq c\ N^{-1/2}$, where $c \leq 0.6745\sigma(\theta)$ ($\sigma(\theta)$ is the standard deviation).

Venelin Todorov is supported by the Bulgarian National Science Fund under Project KP-06-M32/2 - 17.12.2019 "Advanced Stochastic and Deterministic Approaches for Large-Scale Problems of Computational Mathematics" and by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICT in SES)", contract No DO1-205/23.11.2018, financed by the Ministry of Education and Science in Bulgaria. The work is also supported by Bulgarian National Science Fund under Project DN 12/5-2017 "Efficient Stochastic Methods and Algorithms for Large-Scale Problems" and by the Project KP-06-Russia/17 "New Highly Efficient Stochastic Simulation Methods and Applications" funded by the National Science Fund - Bulgaria.

Algorithm

1. **Input data**: *total number of points $N1$, constant M(the initial number of subregions taken), constant ε (max value of the variance in each subregion), constant δ (maximal admissible number of subregions), d-dimensionality of the initial region/domain, f - the function of interest.*
   - 1.1. **Calculate** *the number of points to be taken in each subregion $N = N1/\delta$.*
2. **For** $j = 1,\ M^d$:
   - 2.1. **Calculate** *the approximation of I_{Ω_j} and the variance D_{Ω_j} in subdomain Ω_j based on N independent realizations of random variable θ_N;*
   - 2.2. **If** $(\mathrm{D}_{\Omega_j} \geq \varepsilon)$ **then**
     - 2.2.1. **Choose** *the axis direction on which the partition will perform,*
     - 2.2.2. **Divide** *the current domain into two (G_{j_1}, G_{j_2}) along the chosen direction,*
     - 2.2.3. **If** *the length of obtained subinterval is less than δ* **then go to** *step 2.2.1* **else** *$j = j_1$ G_{j_1} is the current domain right and* **go to** *step 2.1;*
   - 2.3. **Else if** $(\mathrm{D}_{\Omega_j} < \varepsilon)$ *but an approximation of $I_{G_{j_2}}$ has not been calculated yet,* **then** *$j = j_2$ G_{j_2} is the current domain along the corresponding direction right and* **go to** *step 2.1;*
   - 2.4. **Else if** $(\mathrm{D}_{\Omega_j} < \varepsilon)$ *but there are subdomains along the other axis directions,* **then go to** *step 2.1;*
   - 2.5. **Else** *Accumulation in the approximation I_N of I.*

B. Lattice rules

We will use the following lattice point sets

$$x_n = \left\{ \frac{n}{N} z \right\}, \ n = 0, \ldots, N-1$$

where $z = (z_1, \ldots, z_n)$ is the generating vector with dimensionality s. By a lattice rule we mean a rule of the form

$$I_N(f) = \frac{1}{N} \sum_{j=0}^{N-1} f(x_j), \tag{1}$$

where $x_0, \ldots, x_{N-1}$ are all points of a lattice L in the unit hypercube. Lets deal with the problem for approximate evaluation of the integral:

$$I(f) = \int_{[0,1)^s} f(x)dx,$$

where f is a real function in $[0,1)^s$. We consider the case when f has a periodic extension $\widetilde{f}$ in R^s,

$$\widetilde{f} = f(x), x \in [0,1)^s,$$

$$\widetilde{f}(x+z) = \widetilde{f}, x \in \mathbb{R}^s, z \in \mathbb{Z}^s.$$

Let

$$I_N(f) = \frac{1}{N} \sum_{j=0}^{N-1} f(x_j), \tag{2}$$

where $x_0, x_1, \ldots, x_{N-1}$ are points from the lattice $L \in [0,1)^s$. We define the dual lattice L as

$$L^{\perp} = m \in \mathbb{R}^s : m.x \in \mathbb{Z}, x \in L.$$

In the case when the lattice has rank 1

$$L^{\perp} = m \in \mathbb{Z}^s : m.x \equiv 0 \ (\mod N).$$

Let f can be presented in Fourie series as:

$$f(x) = \sum_{m \in \mathbb{Z}^s} a(m) e^{2\pi i m.x}, x \in [0,1)^s,$$

where

$$a(m) = \int\limits_{[0,1)^s} e^{-2\pi i m.x} f(x)dx,$$

and the scalar product $m.u = m_1x_1 + m_2x_2 + \ldots m_sx_s$. Let E_s^{α} for $\alpha > 1$ and $c > 0$ is a class of functions f, for which the coefficients of Fourier [8] satisfies:

$$|a(m)| \leqq \frac{c}{(\overline{m}_1 \ldots \overline{m}_s)^{\alpha}}, \tag{3}$$

where

$$\overline{m} = |m|, \ |m|, \ m \geqq 1, \ \overline{m} = 1, \ m = 0.$$

We define the Zaremba index [12] as

$$\rho = min_{m \in L^{\perp}, m \neq 0} (\overline{m}_1 \ldots \overline{m}_s).$$

The following theorems are key points in analysis the error of integration in the lattice rule:

Theorem 1: [10] Let L is a lattice with points $x_0, x_1, \ldots, x_{N-1}$ in $[0,1)^s$ and $m \in \mathbb{Z}^s$. Then

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{2\pi i m.x_j} = 1, m \in L^{\perp},$$

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{2\pi i m.x_j} = 0, m \notin L^{\perp}.$$

Theorem 2: [12] Let L is a lattice with points $x_0, x_1, \ldots, x_{N-1}$ in $[0,1)^s$. Then for the error of integration is fulfilled

$$I_N(f) - I(f) = \sum_{m \in L^{\perp}, m \neq 0} a(m).$$

When we replace $f(x) = \sum\limits_{m \in \mathbb{Z}^s} a(m) e^{2\pi i m.x}$ in $I_N(f) = \frac{1}{N} \sum\limits_{j=0}^{N-1} f(x_j)$:

$$I_N(f) = \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) = \frac{1}{N} \sum_{j=0}^{N-1} \sum_{m \in \mathbb{Z}^s} a(m) e^{2\pi i m.x_j} =$$

$$\sum_{m \in \mathbb{Z}^s} a(m) \frac{1}{N} \sum_{j=0}^{N-1} e^{2\pi i m.x_j} = \sum_{m \in L^{\perp}} a(m).$$

and using the definition $I(f) = a(0)$ we will obtain

$$I_N(f) - I(f) = \sum_{m \in L^{\perp}, m \neq 0} a(m).$$

Theorem 3: [12] Let L is a lattice with points $x_0, x_1, \ldots, x_{N-1}$ in $[0,1)^s$ and let $f \in E_s^{\alpha}(c), \alpha > 1$. Then

$$|I_N(f) - I(f)| \leqq c \sum_{m \in L^{\perp}, m \neq 0} (\overline{m}_1 \ldots \overline{m}_s)^{-\alpha}.$$

The proof of this fact directly leads from the previous theorem

$$|I_N(f) - I(f)| = \sum_{m \in L^{\perp}, m \neq 0} |$$

$$a(m)| \leqq c \sum_{m \in L^{\perp}, m \neq 0} \frac{1}{(\overline{m}_1 \ldots \overline{m}_s)^{\alpha}}$$

.

Theorem 4: [12] Let L is a lattice with $x_0, x_1, \ldots, x_{N-1}$ in $[0,1)^s$ and let $f \in E_s^{\alpha}(c), \alpha > 1$ and $\rho \geqq 2$ is the Zaremba index. Then

$$|I_N(f) - I(f)| \leqq cd(s,\alpha)\rho^{-\alpha}(\log \rho)^{s-1}.$$

Here we can define the number $R_l, l = 1, 2, \ldots$, which will show the number of points $m \in L^{\perp}$ for which

$$\overline{m}_1 \ldots \overline{m}_s < l\rho$$

Here we will use the lemma proven by Hua and Wang (1981) [6]:

$$R_l \leq e(s) l (\log 3l\rho)^{s-1}, l = 1, 2, \ldots,$$

where $e(s)$ depend only on s.

From Theorem 3

$$|I_N(f) - I(f)| \leq c \sum_{m \in L^{\perp}, m \neq 0} \frac{1}{(\overline{m}_1 \dots \overline{m}_s)^{\alpha}}$$

The sum on m can be broken down into a sum of collectibles by $E_1, E_2, \dots$, where E_l is defined by the inequalities

$$l\rho \leq \overline{m}_1 \dots \overline{m}_s < (l+1)\rho, l = 1, 2, \dots$$

By the definition of R_l we have the following inequalities:

$$\sum_{m \in L^{\perp}, m \neq 0} \frac{1}{(\overline{m}_1 \dots \overline{m}_s)^{\alpha}} \leq$$

$$\sum_{l=1}^{\infty} \frac{R_{l+1} - R_l}{(l\rho)^{\alpha}} \leq$$

$$\frac{1}{\rho^{\alpha}} \sum_{l=1}^{\infty} R_{l+1} \left(\frac{1}{l^{\alpha}} - \frac{1}{(l+1)^{\alpha}} \right).$$

We have that

$$\frac{1}{l^{\alpha}} - \frac{1}{(l+1)^{\alpha}} = \alpha \int_l^{l+1} x^{-\alpha-1} dx \leq \frac{\alpha}{l^{\alpha+1}},$$

and using the Lemma of Hua and Wang

$$|I_N(f) - I(f)| \leq \frac{c\alpha}{\rho^{\alpha}} \sum_{l=1}^{\infty} \frac{R_{l+1}}{l^{\alpha+1}} \leq$$

$$\leq \frac{c\alpha e(s)}{\rho^{\alpha}} \sum_{l=1}^{\infty} \frac{(l+1)(\log 3(l+1)\rho)^{s-1}}{l^{\alpha+1}}$$

$$\leq cd(s,\alpha)\rho^{-\alpha}(\log \rho)^{s-1},$$

where $d_1(s,\alpha)$ depend only on s and α. This proves theorem 4. In the theory of integration lattice rule a key point play the functions $f_\alpha, \alpha = 2, 4, \dots$. Every function f_α is the worst function [12] for appropriate class $E_s^{\alpha}(1)$. This functions are defined by

$$f_\alpha(x) = \sum_{m \in L^{\perp}} \frac{1}{(\overline{m}_1 \dots \overline{m}_s)^{\alpha}} e^{2\pi i m.x}.$$

Furthermore $f_\alpha \in E_s^{\alpha}(1), I(f_\alpha) = 1$. Let $P_\alpha(z,N) = P_\alpha$ means the error in $I(f_\alpha)$. From Theorem 2

$$P_\alpha(z,N) = I_N(f_\alpha) - I(f_\alpha) = \sum_{m \in L^{\perp}, m \neq 0} \frac{1}{(\overline{m}_1 \dots \overline{m}_s)^{\alpha}}.$$

Now for $f \in E_s^{\alpha}(c)$ according Theorem 3 the error is defined by

$$|I_N(f) - I(f)| \leq cP_\alpha(N,z), \tag{4}$$

where $\alpha = 2, 4, \dots$ and the error is fulfilled when $f = f_\alpha$. The values of $P_\alpha(z,N)$ for fixed α are using as indication of the relative quality of the particular lattice. In the case of rank-1 lattice

$$P_\alpha(z,N) = \sum_{z.a \equiv 0 \pmod N, a \neq 0} \frac{1}{(\overline{m}_1 \dots \overline{m}_s)^{\alpha}}.$$

Bakhvalov proves that [12]:

Theorem 5: If P is a lattice point set, with an optimal generating vector z, for the error of integration we have

$$\left| \frac{1}{N} \sum_{k=0}^{N-1} f\left(\frac{k}{N} z\right) - \int_{[0,1)^s} \right| \leq Cd(s,\alpha) \frac{(\log N)^{\beta(s,\alpha)}}{N^{\alpha}} \tag{5}$$

for $f \in E_s^{\alpha}(c), \alpha > 1$ and $d(s,\alpha), \beta(s,\alpha)$ does not depend on N.

Bakhvalov (1959) [12] prove that:

Theorem 6: If N is a prime number, there exists generating vector z, such that

$$D(N) = O(N^{-1} \log^s N),$$

$$P_\alpha(z,N) = O(N^{-\alpha} \log^{\alpha s} N).$$

Niederreiter shows [11], if N is not a prime number,there exist lattice point sets for which:

$$P_\alpha(z,N) = O(N^{-\alpha} (\log N)^{\alpha(s-1)+1} (\frac{N}{\phi(N)})), s \geq 2,$$

$$P_\alpha(z,N) = O(N^{-\alpha} (\log N)^{\alpha} (\frac{N}{\phi(N)} + \frac{\tau(N)}{log(N)})), s = 2,$$

$$P_\alpha(z,N) = O(N^{-\alpha} (\log N)^{\alpha} (s-1)(1 + \frac{\tau(N)}{log^{s-1}(N)})), s \geq 3,$$

where $\phi(N)$ is the Euler's totient function and $\tau(N)$ is the number of divisors of N. For prime number from this formulas leads that there exist z, for which

$$P_\alpha(z,N) = O(N^{-\alpha} \log^{\alpha(s-1)} N).$$

It is fulfilled the following theorem of Sharygin (1963) [9]:

Theorem 7: For a given lattice rule it is fulfilled that

$$P_\alpha(z,N) \geq O(N^{-\alpha} \log^{s-1} N). \tag{6}$$

When $s = 2$ there is an optimal construction. Bakhavalov (1959), Hua and Wang (1960) introduced construction, based on Fibonacci numbers, which are defined recursively by

$$F_0 = 0, F_1 = 1, F_l = F_{l-1} + F_{l-2}, l \geq 2.$$

Let $N = F_l$ and $z = (1, F_{l-1})$. For the obtained lattice Bakhavalov and Hua and Wang show that

$$P_\alpha((1, F_{l-1}), F_l) = O(F_l^{-\alpha} \log F_l),$$

which is optimal according to Sharygin. In 1966 Zaremba [12] shows that

$$D(F_l) = O(F_l^{-1} \log F_l),$$

which is optimal according to Schmidt (1972) [10]. It is important that for finding F_l are necessary only $O(\log F_l)$ elementary operations. There are different techniques for optimal constructions when $s \geq 2$. Let $s = \frac{p-1}{2}$, where $p \geq 5$ is a prime number. If we have the set $Q(2\cos\frac{2\pi}{p})$, which is an algebraic field of degree s with basis functions

$2\cos(2\pi j/p) \mid j = 1, \ldots, s$, we construct the sequence $\eta_l, l = 1, 2, \ldots,$ which satisfies:

$$c_s^{-1}e^l < \eta_l < c_s e^l, c_s^{-1}e^{-l/(s-1)} \leq |\eta_l^{(j)}| \leq c_s^{-1}e^{-l/(s-1)}, \quad j = 2, \ldots, s,$$

where c_s is a constant and $\eta^{(j)}$ is the conjugate of η. Define the generating vector by:

$$\eta_l = \sum_{j=1}^{s} \eta_l^{(j)}, h_j^{(l)} = [\eta_l 2\cos(2\pi j/p)], j = 2, \ldots, s,$$

where η_l is the number of points and [.] is a function whole part. With such a choice of z Hua and Wang show that

$$D(\eta_l) = O(\eta_l^{-\frac{1}{2}-\frac{1}{2(s-1)+\varepsilon}}), P_\alpha(z, N) = O(\eta_l^{-\frac{\alpha}{2}-\frac{\alpha}{2(s-1)+\varepsilon}}),$$

where ε is a preliminary given positive number.

We will construct a lattice L with the following optimized generating vector. for positive number n:

$$z = (1, F_n(2), \ldots, F_n(s)) \tag{7}$$

It is fulfilled that $F_n(j) = F_{n+j-1} - F_{n+j-2} - \ldots - F_n$, where F_i are the generalized Fibonacci numbers with dimensionality s:

$$F_{l+s} = F_l + F_{l+1} + ... + F_{l+s-1}, l = 0, 1, \ldots \tag{8}$$

with initial condition:

$$F_0 = F_1 = \ldots = F_{s-2} = 0, F_{s-1} = 1, \tag{9}$$

for $l = 0, 1, \ldots$.

After simplifying:

$$z = (1, F_{n-1} + F_{n-2} + \ldots + F_{n-s+1}, \ldots, F_{n-1} + F_{n-2}, F_{n-1}) \tag{10}$$

III. Numerical examples

We will test the optimized lattice rule into the following examples:

Example 1. s=3.

$$\int_{[0,1]^3} \exp(x_1 x_2 x_3) \approx 1.14649907. \tag{11}$$

Example 2. s= 4.

$$\int_{[0,1]^4} x_1 x_2^2 e^{x_1 x_2} \sin(x_3)\cos(x_4) \approx 0.1089748630. \tag{12}$$

Example 3.

$$\int_{[0,1]^5} \exp(-100 x_1 x_2 x_3)(\sin(x_4) + \cos(x_5)) \approx 0.1854297367. \tag{13}$$

Example 4. s= 7.

$$\int_{[0,1]^7} e^{1-\sum_{i=1}^{3} sin(\frac{\pi}{2}.x_i)}.arcsin(sin(1) + \frac{\sum_{j=1}^{7} x_j}{200}) \approx 0.75151101. \tag{14}$$

Table I
Relative error for 3 dimensional integral

N	crude	t,s	adapt	t,s	lattice	t,s
10^3	3.62e-2	0.007	4.82e-3	0.17	1.21e-3	0.006
10^4	1.67e-3	0.07	1.07e-3	1.44	5.04e-4	0.07
10^5	8.60e-4	0.74	1.52e-4	10.9	5.34e-6	0.66
10^6	5.12e-4	6.12	5.11e-5	131	7.85e-7	7.02
10^7	3.15e-4	60.1	2.34e-5	1094	8.89e-8	79.7

Table II
Relative error for 3 dimensional integral for preliminary given time

time in sec.	crude	adapt	lattice
1	1.05e-3	7.96e-3	2.34e-6
5	6.84e-4	8.14e-4	8.47e-7
10	4.79e-4	1.82e-4	4.89e-7
100	1.57e-4	7.04e-5	6.53e-9

Example 5. s= 15.

$$\int_{[0,1]^{15}} (\sum_{i=1}^{10} x_i^2)(x_{11} - x_{12}^2 - x_{13}^3 - x_{14}^4 - x_{15}^5)^2 \approx 1.96440666. \tag{15}$$

Example 6. s= 25.

$$\int_{[0,1]^25} \frac{4x_1 x_3^2 e^{2x_1 x_3}}{(1 + x_2 + x_4)^2} e^{x_5 + \cdots + x_{20}} x_{21} \ldots x_{25} \approx 108.808. \tag{16}$$

Example 7. s= 30.

$$\int_{[0,1]^30} \frac{4x_1 x_3^2 e^{2x_1 x_3}}{(1 + x_2 + x_4)^2} e^{x_5 + \cdots + x_{20}} x_{21} \ldots x_{30} \approx 3.244540. \tag{17}$$

The results are given in the tables below. We have used laptop CPU Core i7 4710HQ at 2.5GHz. The first group of tables contains information about the method used, the relative error obtained, the number of conversions required, and the CPU time required to compute the integral. The second group table contains information about the computational complexity. A comparison is made, which shows what relative error each of the used algorithms gives at a predetermined time. From these results it can be concluded that the lattice method is the most efficient for computing multidimensional integrals of smooth subintegral functions due to the low computational complexity and high accuracy in comparison with the simple Monte Carlo algorithm (crude) and the adaptive approach (adapt). The crude Monte Carlo is the basic and simplest possible Monte Carlo approach. Such kind of applications appear also in some important problems in control theory.

It can be seen that by increasing the dimension, the optimized lattice rule gives the best results (Table I-VII), and the advantage is more clearly pronounced for a preliminary given computational time (Table II, Table IV).

Table III
RELATIVE ERROR FOR 4 DIMENSIONAL INTEGRAL

N	crude	t,s	adapt	t,s	opt. lattice	t,s
10^4	9.31e-3	0.08	1.11e-3	1.97	8.61e-5	0.07
10^5	4.37e-3	0.78	1.44e-4	20.1	3.69e-5	0.99
10^6	7.87e-4	5.86	5.63e-5	210	2.86e-6	5.22
10^7	4.31e-5	50.1	9.11e-6	2035	3.38e-7	58

Table IV
RELATIVE ERROR FOR 4 DIMENSIONAL INTEGRAL FOR PRELIMINARY GIVEN TIME

time in sec.	crude	adapt	opt. lattice
5	8.61e-4	5.24e-3	8.47e-7
20	2.31e-4	1.44e-4	4.89e-7
100	2.21e-5	8.21e-5	4.53e-8

The lattice method is not applicable to functions with singularities as we will see from the numerical experiments in this section. Let the following model function be given:

$$f(x) = (1 + \sum_{i=1}^{d} a_i \ x_i)^{-(s+1)}. \tag{18}$$

The class of test functions in question belongs to a package proposed by Genz [5]. Each individual class of the package is characterized by a peculiarity in computational terms. The selected set of functions has a single local maximum near one of the vertices of the multidimensional single cube, similar to some model functions describing the change in the concentrations of pollutants in the air. The parameters a_i are evaluated, using variables a_i', uniformly distributed in $[\frac{1}{20}; 1-\frac{1}{20}]$, and the relation $a = c\ a'$. The constant c is *parameter of computational complexity* [1], selected so that the "sharpness" of the local maximum is controlled by the following norm $||a||_1 = \frac{600}{s^2}$. The adaptive approach is effective for such a class of functions - functions with computational features in a local subdomain of the field of integration.

Table V
RELATIVE ERROR FOR 5 DIMENSIONAL INTEGRAL

N	crude	t,s	adapt	t,s	opt. lattice	t,s
10^3	2.10e-2	0.007	2.15e-3	0.27	1.75e-4	0.007
10^4	4.52e-3	0.07	2.01e-3	2.43	1.28e-5	0.06
10^5	1.19e-3	0.64	8.91e-4	25.2	9.50e-6	0.61
10^6	9.47e-4	6.06	2.92e-4	219.5	5.47e-7	5.98
10^7	6.38e-4	59.9	8.21e-5	2043	7.71e-8	58.4

Table VI
RELATIVE ERROR FOR 7 DIMENSIONAL INTEGRAL

N	crude	t,s	adapt	t,s	opt. lattice	t,s
10^4	1.47e-2	0.11	1.07e-3	2.07	2.19e-4	0.11
10^5	8.26e-3	1.02	7.51e-4	19.3	6.87e-5	0.99
10^6	1.76e-3	10.1	6.30e-5	194	7.39e-6	9.81
10^7	9.85e-4	96.3	2.34e-5	1861	8.89e-7	94.2

Table VII
RELATIVE ERROR FOR 15-DIMENSIONAL INTEGRAL

N	crude	t,s	adapt	t,s	opt. lattice	t,s
10^3	6.31e-2	0.09	3.16e-3	9.24	5.34e-2	0.08
10^4	4.30e-2	0.95	1.49e-3	88	1.22e-3	0.93
10^5	2.77e-2	9.70	5.76e-4	847	3.08e-4	9.65
10^6	2.13e-3	95.8	1.29e-4	8235	1.37e-5	96.9

The results obtained after applying the simple(crude) and adaptive Monte Carlo algorithm for integrals of 5 and 18 are given in Table VIII and Table IX, respectively. The efficiency of the adaptive and lattice algorithms is studied.

In both tables, the value N denotes the total number of conversions in the entire domain for the ordinary algorithm, for the adaptive algorithm, and for the algorithm using a plurality of lattice types. The total number of conversions and approximately the same time to calculate the integrals is actually the basis for comparing the presented results. A number of realizations of the random variable have been chosen so that the times for obtaining an approximate value of the integral are close. The obtained results confirm the reduction of the variance - the adaptive algorithm needs much fewer implementations and gives more accurate results than the ordinary Monte Carlo and the lattice type algorithm, but it is significantly slower (see Table IX).

IV. CONCLUSION

A comprehensive experimental study is done for multidimensional integrals with applications in ecology. The numerical experiments show that the optimized lattice rule is more efficient for multidimensional integrals of smooth functions. The adaptive approach is more efficient for multidimensional integrals with peculiarities and peaks which have applications in air pollution modelling.

REFERENCES

[1] Berntsen J., Espelid T.O., Genz A. (1991) An adaptive algorithm for the approximate calculation of multiple integrals, ACM Trans. Math. Softw. 17: 437–451.

[2] Dimov I. (2008) Monte Carlo Methods for Applied Scientists, New Jersey, London, Singapore, World Scientific, 291 p., ISBN-10 981-02-2329-3.

[3] Dimov I., Karaivanova A., Georgieva R., Ivanovska S. (2003) Parallel Importance Separation and Adaptive Monte Carlo Algorithms for Multiple Integrals, Springer Lecture Notes in Computer Science, 2542, 99–107.

[4] Dimov I., Georgieva R. (2010) Monte Carlo Algorithms for Evaluating Sobol' Sensitivity Indices. Math. Comput. Simul. 81(3): 506–514.

[5] A. Genz, Testing multidimensional integration routines. Tools, *Methods and Languages for Scientific and Engineering Computation* (1984) 81–94.

[6] Hua L.K. and Wang Y. (1981) Applications of Number Theory to Numerical analysis.

[7] Pencheva, V., I. Georgiev, and A. Asenov. "Evaluation of passenger waiting time in public transport by using the Monte Carlo method." AIP Conference Proceedings. Vol. 2321. No. 1. AIP Publishing LLC, 2021.

[8] Raeva, E., & Georgiev, I. R. (2018, October). Fourier approximation for modeling limit of insurance liability. In AIP Conference Proceedings (Vol. 2025, No. 1, p. 030006). AIP Publishing LLC.

Table VIII
RELATIVE ERROR FOR $s = 5$, $I[f]$ = 2.12**e**-06, $a = (5, 5, 5, 5, 4)$.

adapt			crude			opt. lattice		
N	$I_N[f]$	(s)	N	$I_N[f]$	(s)	N	$I_N[f]$	(s)
10^2	3.7735e-03	0.33	10^5	5.4858e-02	0.27	1346269	9.7135e-02	0.38
10^3	1.2877e-03	1.44	10^6	3.8207e-02	1.22	3524578	6.7594e-02	1.32
10^4	4.2452e-04	10.75	10^7	3.3962e-03	12.3	14930352	1.5377e-02	15.07
10^5	4.7169e-05	142.18	10^8	9.4339e-04	124.2	102334155	2.9245e-03	134.58

Table IX
RELATIVE ERROR FOR $s = 18$, $I[f]$ = 9.919**e**-06, $a = \left(\frac{1}{9}, \frac{2}{27}, \frac{2}{27}, \frac{1}{9}, \frac{2}{27}, \frac{1}{9}, \frac{1}{9}, \frac{4}{27}, \frac{2}{27}, \frac{1}{9}, \frac{1}{9}, \frac{2}{27}, \frac{2}{27}, \frac{1}{9}, \frac{1}{9}, \frac{4}{27}, \frac{1}{9}, \frac{1}{9}\right)$.

adapt			crude			opt. lattice		
N	$I_N[f]$	(s)	N	$I_N[f]$	(s)	N	$I_N[f]$	(s)
10	9.2341e-04	15.7	10^7	4.5367e-03	13.6	14930352	7.1579e-02	14.7
10^2	8.0653e-05	142	10^8	2.0163e-03	140	102334155	1.3096e-02	144.1
10^3	1.0081e-05	1408	10^9	5.0480e-04	1353.5	1134903170	7.8883e-03	1344.3

[9] I.F. Sharygin (1963) A lower estimate for the error of quadrature formulas for certain classes of functions, *Zh. Vychisl. Mat. i Mat. Fiz.* **3**, 370–376.

[10] I.H. Sloan and P.J. Kachoyan (1987) Lattice methods for multiple integration: Theory, error analysis and examples, *SIAM J. Numer. Anal.* **24**, 116–128.

[11] I.H. Sloan and S. Joe, Lattice Methods for Multiple Integration, *Lattice methods for multiple Integration*, (Oxford University Press 1994).

[12] Y. Wang and F. J. Hickernell (2000) *An historical overview of lattice point sets*, in Monte Carlo and Quasi-Monte Carlo Methods 2000, Proceedings of a Conference held at Hong Kong Baptist University, China.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 81–84
DOI: 10.15439/2021F56
ISSN 2300-5963 ACSIS, Vol. 26

Optimized Nano Grid Approach for Small Critical Loads

Daniel Todorov *, Venelin Todorov†‡, Sefka Fidanova ‡
*University of Ruse Angel Kanchev, 8 Studentska Str, 7017 Ruse, Bulgaria
†Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
8 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
‡Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Email: dltodorov@uni-ruse.bg, vtodorov@math.bas.bg, venelin@parallel.bas.bg, stefka@parallel.bas.bg

***Abstract*—This study is focused on the possibility of utilizing solar energy nano grids for feeding small scale critical loads. The reasons conditioning this necessity are reviewed and the strengths of small scale micro grids over the centralized type of powering are stated. On the basis of studies on renewable energy sources, photovoltaic converters are pointed to be most appropriate for small scale generation in urban conditions at specific geographical region. The loads of typical traffic lights show that they are relatively constant, which makes such consumers very suitable for micro or nano grids.**

I. Introduction

RENEWABLE energy sources (RES) have been widely discussed and applied last decades. Recently on focus have become full or semi autonomous smart systems providing electricity for number of applications. Depending on region specific nature resources, different RES my be utilized. It is important to study the best economic option, and also to study the load profiles in order to match production with demand. Because of erratic nature of some RES, it is necessary to design a storage system providing in periods of lacking RES generation or to relay on other energy at that time. Best option is to use multiple energy sources, however at a cost. For each application there is best option, but regarding complexity, price, space, convenience, etc., there is an optimal option.

II. The need of Photovoltaics

Last years there is great interest in solar energy. Solar cells may not be most efficient among renewable energy converters, but place a lot of convenient aspects, that led to their wide use. A study concerning the registers and issued licenses for RES energy trade of Sustainable Energy Development Agency (SEDA), Bulgaria [1], shows the distribution of renewable facilities in Bulgaria.

The work is supported by KP-06-M32/2-17.12.2019 'Advanced Stochastic and Deterministic Approaches for Large-Scale Problems of Computational Mathematics" and by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" (ICTinSES), contract No. D01-205/23.11.2018, financed by the Ministry of Education and Science. The work is supported by the Bulgarian National Science Fund under Project DN 12/5-2017 "Efficient Stochastic Methods and Algorithms for Large-Scale Problems".

In this section we will discuss the reasons for using photovoltaics. The results show that one of the main utilizable RES is solar power – Fig 1. On Fig. 2 this distributions are displayed graphically in percents for Bulgaria and Central North region. For central north region, solar energy is most used, because there are not much winds or possibilities of hydro generation. In urban conditions there are lots of limitations for building large RES structures, however the existing roof constructions or poles are appropriate for installation of photovoltaic (PV) modules. Furthermore for small scale generation PVs are one of the most suitable RES. However, in cases where an abundance of another energy source is present – like winds, small scale wind turbines may be more appropriate than solar cells. With the development of energy harvesting technologies, new energy sources may emerge and as soon as they become cost comparative to solar cells the current situation may change. We believe, that the road itself and the weight of the vehicles may be used for better generation, when piezo, vibration or pneumatic harvesters are used, however the price of such implementation is prohibitive in the current situation. The yearly global irradiation for optimally inclined PV cells in central north region is estimated to be in the range (1350...1500) kWh/m2. For the other regions in Bulgaria and for some European countries it is respectively (1200...1500) kWh/m2 and (800...2200) kWh/m2. From [2] it can be noted, that significant amounts of solar irradiation are available during months April – September. During this period, solar cells are expected to be most efficient.

III. The need of microgrids

The existing electrical grid has mainly centralized structure, providing decent quality of power and reliability, but because of its complexity and largely spread transmission lines, there exist possible failure points. Some publications blame mainly the grid itself instead of the central power plants [3], [4]. The modern consumers have lots of digital equipment that require high level of reliability and quality of the supplied energy. Economy and industry also relay on digital equipment. For certain types of consumers, outages can impose unwanted risks,

	Total for Bulgaria		North-west		Central-North		North-east		South-west		Central-south		South-east	
Type of RES	Number of RES	Total Installed power, MW	Number of RES	Installed power, MW	Number of RES	Installed power, MW	Number of RES	Installed power, MW	Number of RES	Installed power, MW	Number of RES	Installed power, MW	Number of RES	Installed power, MW
Wind energy	173	656.9	8	17.17	1	0.95	108	528.4	2	0.95			54	109.3
Hydro Energy	222	2307.5	51	68.41	16	17.04	3	3.66	85	302.6	56	1861	11	54.69
Solar Energy	1027	897.59	104	111.33	62	80.37	49	58.45	115	53.36	300	243	397	351.24
Landfill Gas	1	0.83							1	0.83				
Biomass	4	16.8							1	5	2	12	1	0.33
Waste water	2	34.7			1	0.29			1	3.19				
Total	1429	3883.2	163	196.9	80	98.65	160	590.6	205	366	358	2116	463	515.6

Fig. 1. Distribution of RES facilities in Bulgaria and in each of its 6 regions (According to SEDA)

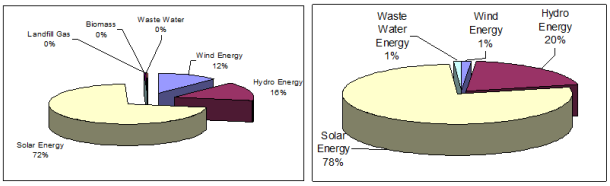


Fig. 2. The distribution of RES energy facilities by types in Bulgaria and central north region

or can threaten safety. Such consumers are facilities part of national defense and security, water treatment, communication, hospitals, traffic signals, data storage facilities, etc. Providing high reliability to such consumers would be too expensive if only the main grid was used with its current structure [5]. If possible, additional bypass wiring is done together with auto transfer switches, or diesel backup generators or large scale UPS systems. One of the main types of failure in the electrical grids are outages. They are affecting negatively various processes. For e.g. in 2012, for Bulgaria the number of reported grid outages is 1375 with total duration of 5005 h. The Electricity System Operator, has reported only the outages of 110kV substations – which is small number for each city, but there also exist more outages that occur locally, due to accidents or power line interruption, that are not reported. Outages are also present due to local or hierarchically bigger node failures or scheduled maintenance of electricity grids. Practically there is no available information about the real number of outages for end consumer in the reports. Most of outages are with short duration, but they can still interrupt or stop various processes in the consumer side. So there must be used another solution for the consumers that require high reliability and quality of power. Microgrids may be capable of fulfilling this demand, because of their ability to use multiple locally distributed power sources (generators and energy storage devices) and to isolate from the grid whenever there is a problem in it. Further more modern controllers make it possible to automatically control loads by priority, leaving only crucial loads online. A smaller structure, recently defined as nanogrid [6], [7], resembles the microgrid approach, however it is much smaller in terms of capacity and structure complexity. Nanogrids are also capable of maintaining good power quality during sags, surges and outages in the outside grid. On Fig. 3 it is presented a block schematic of a small scale microgrid. The controller makes decision for switching between sources of energy – utility grid(s), local generators and storage. Depending on the control algorithm critical load will always stay powered. If no utility power is present and not enough RES generation is available it may switch to alternative grid if available (like auto transfer switches) or internal energy storage devices. In the worst case when storage begins depletion, controller may disconnect non critical loads. Furthermore additional communication is provided with maintenance services, server or another microgrid's controller.

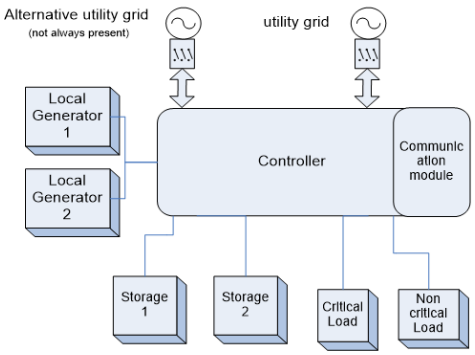


Fig. 3. Small scale microgrid. The controller is responsible for switching between power sources – utility grid, local generators and storage and communication within and outside the grid

The microgrid controller also compensates the erratic nature of RES, making the whole system operate constantly and reliably. PV systems part of small local micro grid are becoming increasingly attractive, because of their advantages concerning sustainability, efficiency and price. PV systems are very suitable for supplying small loads in cities and remote areas, because solar irradiation is nearly ubiquitous, unlike wind or hydro energy. The efficiency of solar energy converters is subject to constant improvements and the prices going down, because of demand and high manufacturing rates. However a microgrid may operate with any other RES or battery energy storage devices if there is abundance of some other local source of energy. It can also make use of several sources at the same time.

IV. Determine the Loads

In this paper, the possibility of using microgrids for supplying power to publicly significant consumers – traffic lights (TL) is considered. An research of the typical loads of TL is required for this purpose. A study concerning the electrical consumption of sample of TL is made. For comparison purpose there is included old consumption data obtained in the past from non modernized TLs using non energy efficient loads, because unfortunately they are still present in some parts of the world due to lack of funds. Some TLs are studied theoretically, by using data about operation cycle and loads switched during each phase (Fig. 4). As seen the load of typical LED-only TL is very small (0,28kW), while the load of inefficient TL with inefficient lighting modules is several times higher (1,2kW). In the theoretical load models, the transients during loading capacitors of LED power supplies are not taken into account, because they are different for the various LED modules, depending mainly on wattage rating of the power supply module and are relatively short. Furthermore most of the SMPS found in these LED modules have inrush current limiter and power factor correction. In some LED heads, the manufacturers has included active loads to obtain compatibility with controllers monitoring current, also to minimize cable and terminals voltage leaks and transfers between signals, but this increases consumption while dissipates heat and worsens efficiency for the sake of safety. Statistical data for the real electrical energy consumption is obtained from archives of periodic meter readings. The statistical characteristics of the loads of two LED and two old data from inefficient TLs – mean value, standard deviation, coefficient of variation and confidence interval at confidence probability 0.95 are shown in Table I. We use the following notations: Mean Value (MV), Standard Deviation (SV), Coefficient of Variation (CV), Confidence + limit (C+l), Confidence – limit (C-l)

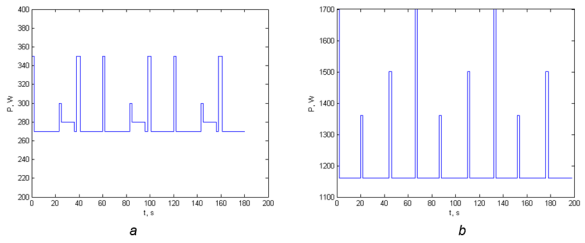


Fig. 4. Theoretical load profile of: a) LED TL; b) old inefficient TL

The small deviations are caused mainly by inaccuracy in data collection and others factors like luminaire type or group of luminaires replacement during maintenance. This means, that probabilistic and unpredictable nature of the load is not observed, which is very convenient in designing micro grids, that fulfills the demand of each consumer. As seen for the case with old TLs, it was practically impossible to implement such PV system, while on LED efficient TLs it is easily achievable. Most of the TL are operating in normal mode in the time

TABLE I
Calculated values for daily mean value of consumed energy and its statistical characteristics – standard deviation, coefficient of variation, and confidence intervals at 0.95 confidence probability

Object	MV, kWh	SD , kWh	CV	C+l, kWh	C–l, kWh
TL LED1	4,62	1,12	0,24	5,34	3,91
TL LED2	6,83	0,39	0,06	7,08	6,58
TL old1	24,43	2,16	0,09	25,81	23,06
TL old2	39,67	2,84	0,07	41,47	37,87

period (6:00…22:30h), during rest of the time, they are in flashing mode, with lowered consumption, however there are also cases where they operate constantly as well, this means that the electrical energy produced from the PV modules would feed, for considerable part of time, the TLs that operate only daily, which will Lead to significantly minimized grid consumption, and extended accumulating media lifecycles.

V. Numerical Example

Having determined the Load Profiles and with the case of deterministic nature of loads is an easy way to implement a microgrid. For preliminary generated values of solar irradiation we use database from [8] for a point in North Bulgaria with tilt angle -30 deg. In our example we use 1kW PV installation, with battery backup, but instead of using battery energy to increase income we would prefer to reserve it for safety in case of emergency outage. Thats why we do not include charge/discharge cycles in our graphs on Fig. 5 and Fig. 6. We will take two cases for solar irradiation at summer maximum about whole June, and minimum at 20 December.

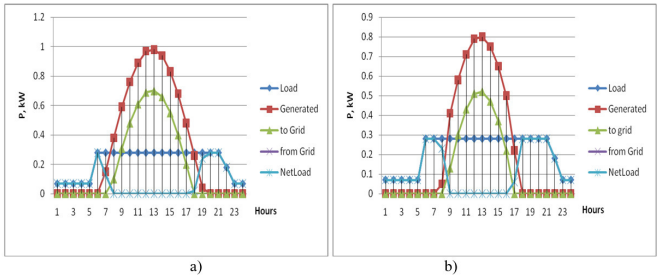


Fig. 5. Hourly diagram Load Profiles of traffic light, Generated solar energy, Energy from and to utility grid: a) June; b) December.

As seen in summer time the expected solar generation starts after 6 till 20 o'clock, and in winter time it is smaller period from 7 till 17 o'clock. Also the maximum generation energy is reduced 20%. Without using the storage the daily income is from 0.31 to 2.8kW, respectively for winter and summer. If bigger PV generators are used this will increase income but price, and space required also increase.

The flowchart of proposed optimized control algorithm for the specific Nanogrid is given on Fig. 7, where Ppv is the generated power from photovoltaics, Pl - measured power of

		Summer June 1st			Winter Dec 20th		
H, h	Load, kW	Generated, kW	to Grid, kW	from Grid, kW	Generated, kW	to grid, kW	from Grid, kW
1	0.07	0	0	0.07	0	0	0.07
2	0.07	0	0	0.07	0	0	0.07
3	0.07	0	0	0.07	0	0	0.07
4	0.07	0	0	0.07	0	0	0.07
5	0.07	0	0	0.07	0	0	0.07
6	0.28	0	0	0.28	0	0	0.28
7	0.28	0.15	0	0.13	0	0	0.28
8	0.28	0.38	0.1	0	0.05	0	0.23
9	0.28	0.59	0.31	0	0.41	0.13	0
10	0.28	0.76	0.48	0	0.58	0.3	0
11	0.28	0.89	0.61	0	0.71	0.43	0
12	0.28	0.97	0.69	0	0.79	0.51	0
13	0.28	0.98	0.7	0	0.8	0.52	0
14	0.28	0.94	0.66	0	0.75	0.47	0
15	0.28	0.83	0.55	0	0.65	0.37	0
16	0.28	0.68	0.4	0	0.5	0.22	0
17	0.28	0.48	0.2	0	0.22	0	0.06
18	0.28	0.26	0	0.02	0	0	0.28
19	0.28	0.04	0	0.24	0	0	0.28
20	0.28	0	0	0.28	0	0	0.28
21	0.28	0	0	0.28	0	0	0.28
22	0.18	0	0	0.18	0	0	0.18
23	0.07	0	0	0.07	0	0	0.07
24	0.07	0	0	0.07	0	0	0.07
Total	5.15	7.94	4.7	1.9	5.46	1.99	1.68

Fig. 6. Hourly diagram Load Profiles of traffic light, Generated solar energy, Energy from and to utility grid: a)June; b)December.

the load, Pgrid - power from to grid, Pbatt - power from to storage media. Fig 7 presents the optimized control algorithm under consideration.

VI. Conclusion

Knowing Load Profiles, together with the information about solar irradiation in the area of application, can be used for optimal design and implementation of PV nano grids. That can help increase reliability and efficiency locally. Such micro grids are necessary even only as protection, not counting the efficiency aspect, because there are cases when power quality worsens for short times within or outside the standard requirements. Modern telecommunications and embedded devices make it possible to easily implement not only power protection for short times like traditional UPS, but such controllers can also reconfigure the whole local grid so that it can maintain stability for prolonged time. Controllers can always send signals to the maintenance service when they have limited amount of reserved resources left, or can send signals to the connected devices to securely switch to emergency mode, stop or hibernate if possible. Unexpected power outage is extremely dangerous for digital equipment, many industrial processes and

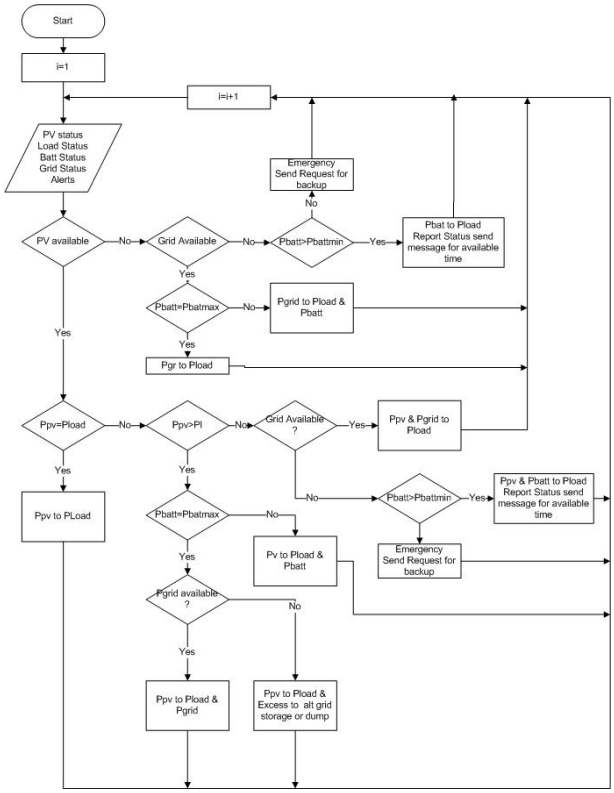


Fig. 7. The flowchart of proposed control algorithm for the specific Nanogrid

traffic safety. Putting all crucial loads inside protected and self-available (for specific time) network, ensures long term efficient operation during unfavorable conditions.

References

[1] Sustainable Energy Development Agency, http://www.seea.government.bg/;
[2] Mihailov N 2014 Renewable Energy Sources and Intelligent Electrical Grids, Presentation, University of Ruse and Regional Academic Council of BAS, Ruse;
[3] McLinn J 2009 Major Power Outages in the US and around the World, Annual technology report of the IEEE Reliability Society http://paris.utdallas.edu/IEEE-RS-ATR;
[4] Electricity System Operator Annual report http://tso.bg/Reports;
[5] Fereidoon P 2011 Smart Grid Integrating Renewable, Distributed & Efficient Energy, Academic Press, 2011;
[6] Nordman B, Christensen K 2013 Local Power Distribution with Nanogrids, International Green Computing Conference, Arlington, VA;
[7] Schonberger J, Round S, Duke R 2005 Decentralized source scheduling in a model nanogrid using DC bus signaling, , AJEEE.
[8] http://www.debarel.com/

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 85–88
DOI: 10.15439/2021F51
ISSN 2300-5963 ACSIS, Vol. 26

Optimized stochastic methods for sensitivity analysis for large-scale air pollution model

Venelin Todorov
Bulgarian Academy of Sciensec
Institute of Mathematics and Informatics
ul. G. Bonchev 8, 1113 Sofia, Bulgaria
Bulgarian Academy of Sciences
Institute of Information and Communication Technologies
ul. G. Bonchev 25A, 1113 Sofia, Bulgaria
Email: vtodorov@math.bas.bg,venelin@parallel.bas.bg

Tzvetan Ostromsky
Bulgarian Academy of Sciences
Institute of Information and Communication Technologies
ul. G. Bonchev 25A, 1113 Sofia, Bulgria
Email: ceco@parallel.bas.bg

Ivan Dimov
Bulgarian Academy of Sciences
Institute of Information and Communication Technologies
ul. G. Bonchev 25A, 1113 Sofia, Bulgria
Email: ivdimov@bas.bg

Rayna Georgieva
Bulgarian Academy of Sciences
Institute of Information and Communication Technologies
ul. G. Bonchev 25A, 1113 Sofia, Bulgria
Email: rayna@parallel.bas.bg

***Abstract*—Environmental security is rapidly becoming a significant topic of present interest all over the world, and environmental modelling has a very high priority in various scientific fields, respectively. Different optimizations of the Latin Hypercube Sampling algorithm have been used in our sensitivity studies of the model output results for some air pollutants with respect to the emission levels and some chemical reactions rates.**

I. Introduction

HIGH levels of pollution can disrupt ecosystems and cause harm to plants, animals and humans. Therefore, it is extremely important to investigate accurately the levels of contamination [9], [11]. It is necessary to know whether the pollution levels are below some critical values and if so - to develop a reliable control system to keep them within these limits. Mathematical models are used to study and predict the behavior of a variety of complex systems - engineering, physical, economic, social, environmental. They determine the most important quantities that control the state and behavior of a system, as well as the quantitative regularities, that is, the mathematical laws that underlie the change of these quantities. On the other hand, it is of significant importance to develop techniques to determine the reliability, robustness and efficiency of a mathematical model [10]. One such approach is sensitivity analysis (SA). Sensitivity studies are nowadays applied to some of the most complicated mathematical models from various intensively developing areas of application [1], [2], [3], [4]. In the present study we use UNI-DEM [12], [13], [14] which is a powerful large-scale air pollution model for calculation of the concentrations of a large number of pollutants and other chemical species in the air along a certain time period. Its results can be used in various application areas (environmental protection, agriculture, health care, etc.). The large computational domain covers completely the European region and the Mediterranean.

UNI-DEM simulates the long-range transport of the air pollutants, their transition in time as a result of the chemical and photochemical reactions between them and the interaction with the other components of the environment. It takes into account the basic physical processes - advection, diffusion, deposition, harmful emissions, as well as the chemical reactions. It provides an opportunity to study over time the concentrations of the main types of pollutants (sulfur, nitrogen, ammonia, ammonium ions, nitrogen, free radicals, hydrocarbons), which is important for environmental safety, agriculture, healthcare. The model domain includes the whole Europe and the Mediteranean plus parts of Asia and Africa. The approximate area covered is $4800 \times 4800\,km$.

II. Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) is a type of stratified sampling and in the case of integral approximation we must simply divide the domain $[0,1]^d$ into m^d disjoint subdomains, each of volume $\frac{1}{m^d}$ and to sample one point from each of

Venelin Todorov is supported by the Bulgarian National Science Fund under Project KP-06-M32/2 - 17.12.2019 "Advanced Stochastic and Deterministic Approaches for Large-Scale Problems of Computational Mathematics" and by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICT in SES)", contract No DO1-205/23.11.2018, financed by the Ministry of Education and Science in Bulgaria. The work is also supported by Bulgarian National Science Fund under Project DN 12/5-2017 "Efficient Stochastic Methods and Algorithms for Large-Scale Problems", Project DN 12/4-2017 "Advanced Analytical and Numerical Methods for Nonlinear Differential Equations with Applications in Finance and Environmental Pollution", and by the Project KP-06-Russia/17 "New Highly Efficient Stochastic Simulation Methods and Applications" funded by the National Science Fund - Bulgaria.

them. Let this sample be $\mathbf{x}_{k,j}$, for dimensions $k = 1,\ldots,m^d$, $j = 1,\ldots,d$. LHS does not require more samples for more dimensions (variables) - it is one of the main advantages of this scheme. Examples of random, stratified and Latin Hypercube Samplings with 16 points are presented on Figure 1 [5].

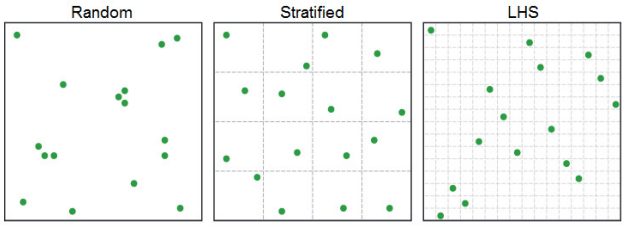


Fig. 1. Comparison of random, stratified and Latin Hypercube Samplings with 16 points (d = 2, m = 4).

Two different optimization versions of Latin Hypercube Sampling are compared with different seeds (seed is an integer used to initialize the corresponding pseudorandom number generator). The algorithms Latin Hypercube Sampling (random) [7], [8] LHSR1 (seed=1) and LHSR-1 (seed=-1) and Latin Hypercube Sampling (edge) [5], [6] LHSE1 (seed=1) and LHSE-1 (seed=-1) have been used in our comprehensive experimental study for the first time for the particular model. The difference is that the simple Latin edge algorithm returns edge points in a Latin square, while Latin random algorithm returns random points in a Latin square.

III. Sensitivity Studies with Respect to Emission Levels

In this section the results for the sensitivity of UNI-DEM output (in particular, the ammonia mean monthly concentrations) with respect to the anthropogenic emissions input data variation are shown and discussed. The anthropogenic emissions input consists of 4 different components

$$\mathbf{E^A} - \text{ammonia } (NH_3);$$
$$\mathbf{E^N} - \text{nitrogen oxides } (NO + NO_2);$$

$$\mathbf{E^S} - \text{sulphur dioxide } (SO_2);$$
$$\mathbf{E^C} - \text{anthropogenic hydrocarbons.}$$

Results of the relative error estimation for the quantities f_0, the total variance $\mathbf{D}$, first-order (S_i) and total (S_i^{tot}) sensitivity indices are given in Tables I, II, III, IV, V, respectively. The quantity f_0 is presented by a 4-dimensional integral, while the rest of the above quantities are presented by 8-dimensional integrals, following the ideas of *correlated sampling* technique to compute sensitivity measures in a reliable way. The results show that the computational efficiency of the algorithms depends on integrand dimension and magnitude of estimated quantity. The order of relative error is different for different quantities of interest (see column *Reference value*) for the same sample size.

TABLE I
Relative error for the evaluation of $f_0 \approx 0.048$.

# of samples n	**LHSE1** Relative error	**LHSE-1** Relative error	**LHSR1** Relative error	**LHSR-1** Relative error
2^{10}	6.37e-05	2.45e-04	3.22e-04	2.43e-04
2^{12}	6.18e-05	1.55e-04	2.67e-04	7.11e-05
2^{14}	5.21e-06	8.44e-06	3.62e-05	8.30e-05
2^{16}	1.34e-05	1.58e-05	1.62e-05	1.01e-05
2^{18}	1.29e-05	5.31e-07	5.24e-05	1.52e-05
2^{20}	1.53e-05	8.48e-07	8.78e-06	2.24e-05

Table V is similar to Table IV and Table III, with the only difference – the increased number of samples $n = 2^{20}$ (instead of $n = 2^{16}$ in Table IV and $n = 2^{10}$ in Table III). In general, this increases the accuracy of the estimated quantities. The exceptions are S_4, S_4^{tot} and S_{15}, which have extremely small reference values. All LHS methods produce similar results, but for in small in value sensitivity indices LHSE-1 has the edge - see for example the value of S_2^{tot} in Table V.

TABLE II
Relative error for the evaluation of the total variance $\mathbf{D} \approx 0.0002$.

# of samples n	**LHSE1** Relative error	**LHSE-1** Relative error	**LHSR1** Relative error	**LHSR-1** Relative error
2^{10}	4.91e-02	2.78e-03	4.08e-02	4.68e-02
2^{12}	5.10e-03	9.60e-03	1.68e-02	1.40e-02
2^{14}	8.82e-03	7.55e-03	1.26e-02	1.12e-02
2^{16}	8.36e-03	9.72e-03	4.43e-03	4.27e-03
2^{18}	1.32e-04	7.90e-04	7.55e-04	1.14e-03
2^{20}	1.03e-03	8.64e-04	3.43e-05	7.19e-04

IV. Sensitivity Studies with Respect to Chemical Reactions Rates

In this section we will study the sensitivity of the ozone concentration values in the air over Genova with respect to the rate variation of some chemical reactions of the condensed CBM-IV scheme [12], namely: # $1, 3, 7, 22$ (time-dependent)

TABLE III
Relative error for estimation of sensitivity indices of input parameters ($n \approx 2^{10}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	**9e**-01	1.13e-02	5.33e-03	4.39e-02	1.85e-01
S_2	**2e**-04	8.53e+00	7.23e+00	1.22e+00	5.36e+00
S_3	**1e**-01	1.48e-01	4.78e-02	4.53e-01	4.62e-02
S_4	**4e**-05	1.85e+01	1.22e+01	1.55e+01	1.06e+01
S_1^{tot}	**9e**-01	1.54e-02	4.63e-03	5.32e-02	6.91e-03
S_2^{tot}	**2e**-04	1.21e+01	4.87e+00	1.15e+01	6.36e+00
S_3^{tot}	**1e**-01	1.27e-01	4.53e-02	3.74e-01	3.26e-02
S_4^{tot}	**5e**-05	2.25e+01	1.71e+01	1.35e+01	5.66e+00

TABLE IV
RELATIVE ERROR FOR ESTIMATION OF SENSITIVITY INDICES OF INPUT PARAMETERS ($n \approx 2^{16}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	9**e**-01	1.19e-03	2.72e-03	2.75e-03	5.98e-03
S_2	2**e**-04	6.16e-01	4.53e-01	1.99e-01	5.40e-02
S_3	1**e**-01	7.19e-03	1.49e-02	1.38e-02	3.66e-03
S_4	4**e**-05	9.65e-01	2.22e+00	4.11e+00	5.30e-01
S_1^{tot}	9**e**-01	7.04e-04	1.79e-03	1.81e-03	5.15e-04
S_2^{tot}	2**e**-04	9.78e-01	3.45e-01	2.36e-01	1.06e+00
S_3^{tot}	1**e**-01	7.67e-03	1.76e-03	2.02e-02	6.24e-04
S_4^{tot}	5**e**-05	8.66e-01	2.72e+00	1.08e+00	7.54e-01

TABLE V
RELATIVE ERROR FOR ESTIMATION OF SENSITIVITY INDICES OF INPUT PARAMETERS ($n \approx 2^{20}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	9**e**-01	4.66e-04	1.46e-04	1.40e-04	2.99e-04
S_2	2**e**-04	6.38e-03	5.63e-02	1.98e-01	2.55e-01
S_3	1**e**-01	4.01e-03	1.27e-03	4.06e-04	4.05e-04
S_4	4**e**-05	7.09e-01	4.81e-01	5.24e-01	1.77e-01
S_1^{tot}	9**e**-01	4.97e-04	1.34e-04	2.26e-05	1.31e-04
S_2^{tot}	2**e**-04	5.40e-02	1.59e-03	3.09e-01	3.49e-01
S_3^{tot}	1**e**-01	4.42e-03	3.09e-02	6.25e-04	1.71e-03
S_4^{tot}	5**e**-05	5.85e-01	1.08e+00	3.11e-01	1.02e-01

and # 27,28 (time independent). The simplified chemical equations of those reactions are:

$$[\#\mathbf{1}] \quad NO_2 + h\nu \Longrightarrow NO + O;$$
$$[\#\mathbf{3}] \quad O_3 + NO \Longrightarrow NO_2;$$
$$[\#\mathbf{7}] \quad NO_2 + O_3 \Longrightarrow NO_3;$$

$$[\#\mathbf{22}] \quad HO_2 + NO \Longrightarrow OH + NO_2;$$
$$[\#\mathbf{27}] \quad HO_2 + HO_2 \Longrightarrow H_2O_2;$$
$$[\#\mathbf{28}] \quad OH + CO \Longrightarrow HO_2.$$

The relative error estimation for the quantities f_0, the total variance $\mathbf{D}$ and some sensitivity indices are given in Tables VI, VII, VIII, IX, X respectively. The quantity f_0 is presented by 6-dimensional integral, while the rest are presented by 12-dimensional integrals. Table X is similar to Table IX and Table VIII, with the only difference – the increased number of samples $n = 2^{20}$ (instead of $n = 2^{16}$ in Table IX and $n = 2^{10}$ in Table VIII). In general, this increases the accuracy of the estimated quantities. Exceptions are S_5, S_5^{tot} and S_{15}, which have extremely small reference values. All LHS methods produce similar results, but for in small in value sensitivity indices LHSE-1 sometimes has the edge - see for example the value of S_{45} in Table X and sometimes LHSE1 gives the best results - see for example the value of S_5 and S_4^{tot} in Table X.

V. CONCLUSION

The present study focuses on the so-called environmental safety. The computational efficiency (in terms of relative error and computational time) of the optimization versions of Latin Hypercube Sampling Random and Edge algorithms with different seeds for multidimensional numerical integration have been studied to analyze the sensitivity of UNI-DEM model output to variation of input emissions of the anthropogenic pollutants and of rates of several chemical reactions. The various optimization versions of Latin Hypercube Sampling techniques have been successfully applied to compute global Sobol sensitivity measures corresponding to the influence of several input parameters on the concentrations of important air pollutants. The novelty of the proposed approaches is that Latin Hypercube Sampling Edge algorithm with different seeds have been applied for the first time to sensitivity studies of the particular air pollution model. The numerical tests show that the presented stochastic approaches is efficient for the multidimensional integrals under consideration and especially for computing small by value sensitivity indices.

TABLE VI
RELATIVE ERROR FOR THE EVALUATION OF $f_0 \approx 0.27$.

# of samples n	**LHSE1** Relative error	**LHSE-1** Relative error	**LHSR1** Relative error	**LHSR-1** Relative error
2^{10}	4.90e-05	5.86e-04	9.90e-05	1.19e-03
2^{12}	7.47e-05	1.05e-04	7.21e-04	1.04e-04
2^{14}	2.43e-04	1.89e-04	3.38e-04	2.26e-04
2^{16}	2.99e-05	1.73e-05	2.92e-05	6.15e-05
2^{18}	1.29e-04	2.09e-05	3.04e-05	3.68e-05
2^{20}	2.07e-05	2.99e-06	1.31e-05	1.21e-05

TABLE VII
RELATIVE ERROR FOR THE EVALUATION OF THE TOTAL VARIANCE $\mathbf{D} \approx 0.0025$.

# of samples n	**LHSE1** Relative error	**LHSE-1** Relative error	**LHSR1** Relative error	**LHSR-1** Relative error
2^{10}	5.30e-02	5.01e-02	3.86e-03	1.04e-01
2^{12}	1.70e-02	2.90e-03	6.63e-02	4.79e-02
2^{14}	1.47e-02	1.41e-02	2.48e-02	3.25e-02
2^{16}	5.81e-03	2.91e-03	7.13e-03	1.51e-02
2^{18}	6.91e-03	1.73e-03	7.22e-04	4.72e-03
2^{20}	2.30e-03	2.84e-04	9.33e-05	1.07e-03

TABLE VIII
RELATIVE ERROR FOR ESTIMATION OF SENSITIVITY INDICES OF INPUT PARAMETERS ($n \approx 2^{10}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	4**e**-01	8.83e-03	2.40e-03	1.35e-01	1.41e-01
S_2	3**e**-01	1.68e-01	3.04e-02	6.03e-02	1.17e-01
S_3	5**e**-02	1.30e-01	4.19e-01	2.52e-01	2.20e-02
S_4	3**e**-01	1.65e-01	1.76e-02	7.06e-02	8.60e-02
S_5	4**e**-07	3.81e+03	5.17e+03	3.10e+03	6.53e+03
S_6	2**e**-02	5.97e-01	1.13e+00	4.29e-01	1.82e-01
S_1^{tot}	4**e**-01	6.77e-02	5.56e-02	1.04e-02	1.82e-02
S_2^{tot}	3**e**-01	2.39e-01	2.50e+00	1.47e-01	1.80e-01
S_3^{tot}	5**e**-02	3.74e-02	1.36e-01	3.18e-01	4.92e-01
S_4^{tot}	3**e**-01	2.40e-01	1.17e-02	4.71e-02	1.10e-01
S_5^{tot}	2**e**-04	4.79e+00	6.65e+00	8.59e+00	2.65e+01
S_6^{tot}	2**e**-02	3.72e-01	5.90e-01	7.67e-01	3.64e-02
S_{12}	6**e**-03	2.97e+00	1.48e+00	9.64e+00	5.18e+00
S_{14}	5**e**-03	4.97e+00	6.74e-01	2.61e+00	5.77e-01
S_{15}	8**e**-06	8.60e+02	9.12e+02	1.00e+03	8.17e+02
S_{24}	3**e**-03	8.40e-02	3.58e+00	3.49e+00	2.72e+00
S_{45}	1**e**-05	1.33e+01	2.64e+01	1.12e+02	9.94e+01

TABLE IX
RELATIVE ERROR FOR ESTIMATION OF SENSITIVITY INDICES OF INPUT PARAMETERS ($n \approx 2^{16}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	4e-01	3.19e-03	5.38e-03	9.94e-03	1.49e-02
S_2	3e-01	2.23e-02	2.13e-02	4.19e-03	2.34e-03
S_3	5e-02	8.71e-02	6.84e-02	1.24e-02	1.58e-02
S_4	3e-01	5.16e-03	5.60e-03	2.14e-02	1.62e-02
S_5	4e-07	1.02e+03	1.01e+03	2.37e+01	2.29e+02
S_6	2e-02	7.27e-02	4.22e-02	1.71e-02	1.03e-01
S_1^{tot}	4e-01	7.30e-03	3.85e-03	1.88e-02	9.89e-03
S_2^{tot}	3e-01	1.05e-02	1.73e-02	1.10e-02	4.60e-04
S_3^{tot}	5e-02	1.00e-01	1.07e-01	2.71e-03	6.85e-03
S_4^{tot}	3e-01	6.82e-03	1.57e-02	5.89e-03	8.41e-04
S_5^{tot}	2e-04	5.13e-01	1.21e+00	9.24e-01	1.61e+00
S_6^{tot}	2e-02	2.59e-02	1.01e-01	4.23e-02	1.82e-01
S_{12}	6e-03	2.75e-02	1.42e-01	5.70e-01	2.68e-01
S_{14}	5e-03	2.35e-02	1.10e-01	8.29e-01	1.29e+00
S_{15}	8e-06	9.25e+02	9.33e+02	9.06e+02	9.25e+02
S_{24}	3e-03	3.52e-01	6.26e-02	1.53e-01	5.18e-01
S_{45}	1e-05	2.55e+00	3.88e+00	2.29e+00	4.13e+00

TABLE X
RELATIVE ERROR FOR ESTIMATION OF SENSITIVITY INDICES OF INPUT PARAMETERS ($n \approx 2^{20}$).

Est. qnt.	Ref. val.	**LHSE1**	**LHSE-1**	**LHSR1**	**LHSR-1**
S_1	4e-01	7.83e-04	7.26e-05	6.12e-03	2.73e-03
S_2	3e-01	4.62e-03	3.65e-03	1.70e-03	2.91e-03
S_3	5e-02	6.86e-03	5.05e-03	7.73e-04	7.33e-03
S_4	3e-01	2.98e-04	2.65e-03	2.46e-03	2.21e-03
S_5	4e-07	9.28e+00	1.28e+02	1.53e+02	3.06e+02
S_6	2e-02	5.68e-03	6.60e-03	2.54e-02	1.09e-02
S_1^{tot}	4e-01	1.23e-03	1.23e-03	2.39e-03	2.37e-03
S_2^{tot}	3e-01	2.25e-03	3.68e-03	7.43e-03	4.91e-03
S_3^{tot}	5e-02	1.39e-02	8.33e-03	1.21e-02	1.16e-02
S_4^{tot}	3e-01	2.91e-04	3.47e-03	2.88e-03	3.23e-03
S_5^{tot}	2e-04	4.68e-01	1.12e+00	2.68e-01	3.52e-01
S_6^{tot}	2e-02	2.33e-02	6.06e-03	4.22e-02	9.54e-03
S_{12}	6e-03	3.06e-01	1.22e-01	3.05e-01	1.39e-03
S_{14}	5e-03	9.18e-02	5.75e-02	9.69e-02	6.95e-02
S_{15}	8e-06	9.31e+02	9.32e+02	9.29e+02	9.27e+02
S_{24}	3e-03	7.99e-02	3.96e-01	3.16e-02	2.87e-01
S_{45}	1e-05	2.01e+00	4.83e-01	1.96e+00	2.41e+00

REFERENCES

[1] G. Dimitriu: Global Sensitivity Analysis for a Chronic Myelogenous Leukemia Model: Proc. 9th International Conference NMA'2018, Borovets, Bulgaria, August 20-24, 2018, LNCS 11189, Springer, Jan 2019. DOI: 10.1007/978-3-030-10692-8_42

[2] Gocheva-Ilieva, Snezhana G., Atanas V. Ivanov, and Ioannis E. Livieris. "High Performance Machine Learning Models of Large Scale Air Pollution Data in Urban Area." Cybernetics and Information Technologies 20.6 (2020): 49-60.

[3] Gocheva-Ilieva, S. G., Voynikova, D. S., Stoimenova, M. P., Ivanov, A. V., & Iliev, I. P. (2019). Regression trees modeling of time series for air pollution analysis and forecasting. Neural Computing and Applications, 31(12), 9023-9039.

[4] H. Hamdad, Ch. Pézerat, B. Gauvreau, Ch. Locqueteau, Y. Denoual, Sensitivity analysis and propagation of uncertainty for the simulation of vehicle pass-by noise, Applied Acoustics Vol. 149, Elsevier, pp. 85-98 (June 2019). DOI: 10.1016/j.apacoust.2019.01.026

[5] Kroese, D.P., Taimre, T., Botev, Z.: Handbook of Monte Carlo Methods, Wiley Series in Probability and Statistics, (2011)

[6] McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2), 239–45 (1979)

[7] Minasny B., McBratney B.: A conditioned Latin hypercube method for sampling in the presence of ancillary information Journal Computers and Geosciences archive, Volume 32 Issue 9, November, 2006, Pages 1378-1388.

[8] Minasny B., McBratney B.: Conditioned Latin Hypercube Sampling for Calibrating Soil Sensor Data to Soil Properties, Chapter: Proximal Soil Sensing, Progress in Soil Science, pp. 111-119, 2010.

[9] Pencheva, Velizara, Ivan Georgiev, and Asen Asenov. "Evaluation of passenger waiting time in public transport by using the Monte Carlo method." AIP Conference Proceedings. Vol. 2321. No. 1. AIP Publishing LLC, 2021.

[10] I. M. Sobol', Sensitivity estimates for nonlinear mathematical models, *Matem. Modelirovanie* **2** (1) (1990), 112–118.

[11] S. L. Zaharieva, I. Radoslavov Georgiev, A. N. Borodzhieva and V. Angelov Mutkov, "Classical Approach For Forecasting Temperature In Residential Premises Part 1," 2021 20th International Symposium Infoteh-Jahorina (infoteh), 2021, pp. 1-6.

[12] Z. Zlatev, Computer treatment of large air pollution models, KLUWER Academic Publishers, Dorsrecht-Boston-London, 1995.

[13] Z. Zlatev, I. T. Dimov, Computational and Numerical Challenges in Environmental Modelling, Elsevier, Amsterdam, 2006.

[14] Z. Zlatev, I. Dimov, K. Georgiev, Three-dimensional version of the Danish Eulerian Model, *Zeitschrift für Angewandte Mathematik und Mechanik*, **76** (1996) S4, 473-476.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 89–96
DOI: 10.15439/2021F76
ISSN 2300-5963 ACSIS, Vol. 26

Two-Stage Intuitionistic Fuzzy Transportation Problem through the Prism of Index Matrices

Velichka Traneva
"Prof. Asen Zlatarov" University
"Prof. Yakimov" Blvd, Burgas 8000, Bulgaria
Email: veleka13@gmail.com

Stoyan Tranev
"Prof. Asen Zlatarov" University
"Prof. Yakimov" Blvd, Burgas 8000, Bulgaria
Email: tranev@abv.bg

***Abstract*—In today's market environment not all the parameters of the transportation problems may not be known precisely. Uncertain data can be represented by fuzzy sets (FSs). Intuitionistic FSs (IFSs) are an extension of FSs with a degree of hesitancy. The paper presents a new approach for solution of a two-stage intuitionistic fuzzy transportation problem (2-S IFTP) through the prism of index matrices (IMs). Its main objective is to find the quantities of delivery from manufacturers and resselers to buyers to maintain the supply and demand requirements at the cheapest transportation costs. The solution procedure is demonstrated by a numerical example.**

I. Introduction

THE TP originally was proposed by Hitchcock in 1941 [14].

In conventional TP, values of the transportation cost, the demanded and offered quantities of the product are precisely defined. In real-life TPs, some of their parameters are uncertain due to climatic, road conditions or other market conditions. In some TPs, destinations cannot get all the required quantity of product due to limited storage capacity. In this case, the necessary quantities of products are sent to the destinations in two stages. Initially, the minimum destination requirements are sent from the sources to the destinations. Once part of the entire initial shipment has been used up, they are ready to receive the remaining quantity in the second stage. This type of transportation problem is known as two-stage TPs (2-S TPs). The main purpose of the 2-S TP is to transport the items from the origins to the destinations in two stages such way that the total transportation costs in the two stages are minimum [43]. In real life 2-S TPs, information about the parameters of the problem is uncertain due to weather and road conditions, lack of good communications, traffic jams, etc. For description of imprecise information, Zadeh has developed the theory of fuzzy sets (FSs) [29]. An extension of FSs is intuitionistic fuzzy sets (IFSs), which was proposed by Atanassov in 1983 [18]. The main difference between FSs and IFSs is that the IFSs have a degree of hesitancy.

Let us give a brief literature overview of the works on the topic fuzzy (FTPs) and intuitionistic fuzzy transportation problems (IFTPs). Chanas et al., in 1984, has proposed a fuzzy linear programming model for solving TPs with clear transportation costs, fuzzy supply and demand values [39]. Jimenez and Verdegay, in 1999, researched fuzzy Solid TP with trapezoidal FNs and presented a genetic approach for solving FTP [13]. Liu and Kao [41] have demonstrated a method, based on Zadeh's extension principle, to find the optimal solution of the trapezoidal FTPs. Dinagar and Palanivel [11] have described a fuzzy modified distribution method to find the fuzzy optimal solution of FTPs in which all the parameters are represented by trapezoidal fuzzy numbers. Pandian and Natarajan, in 2010, developed zero-point method for solution for FTP with trapezoidal fuzzy parameters [34]. In [1] was proposed a new method based on fuzzy zero-point method for finding a more-or-less fuzzy optimal solution for such FTPs in which all the parameters are represented by trapezoidal fuzzy numbers.

Kaur and Kumar, in 2012, introduced fuzzy least cost method, fuzzy north west corner rule and fuzzy Vogel approximation method for determining of an optimal solution of FTP [33]. Basirzadeh [17] has found a fuzzy optimal solution of fully FTPs by transforming the fuzzy parameters into the crisp parameters using classical algorithms. Gani et al. [2] used Arsham and Khan's simplex algorithm [16] to find a fuzzy optimal solution of FTPs with trapezoidal fuzzy parameters. Patil and Chandgude, in 2012, performed "Fuzzy Hungarian approach" for TP with trapezoidal FNs [7]. A modified Vogel's approximation method for finding an optimal solution of FTPs was proposed in [8]. Aggarwal and Gupta, in 2013, described a procedure for solving intuitionistic fuzzy TP (IFTP) with trapezoidal IFNs via ranking method [15]. Jahihussain and Jayaraman, in 2013, presented a zero-suffix method for obtaining an optimal solution for FTPs with triangular and trapezoidal FNs (see [36], [37]). Zero suffix method to solve FTP after its converting into the crisp problem was applied in [32] and [44]. A fuzzified version of zero suffix method was performed and applied in [30], in 2018, to FTPs. Shanmugasundari and Ganesan, in 2013, proposed a fuzzy modified distribution algorithm and a fuzzy approximation method of Vogel to solve FTP with FNs [31]. Gani and Abbas, in 2014 [4], and Kathirvel, and Balamurugun, in 2012 (see [27], [28]), proposed a method for solving TP in which the quantities demanded and offered are represented in the form

Work on Sect. I and Sect. II is supported by the Asen Zlatarov University through project Ref. No. NIX-440/2020 "Index matrices as a tool for knowledge extraction". Work on Sect. III and Sect. IV is supported by the Ministry of Education and Science under the Programme "Young scientists and postdoctoral students", approved by DCM # 577/17.08.2018.

of the trapezoidal intuitionistic FNs (IFNs). In well-known and commonly used methods, proposed by Basirzadeh [17], Gani et al. [2], Pandian and Natarajan [34] and Dinagar and Palanivel [11], there is a problem that, in a general case, neither the cost values, nor the obtained fuzzy optimal solution need necessarily to be non-negative fuzzy numbers. These are shortcomings of these methods, as in real life problems there is no physical meaning of a negative value of the cost and a negative quantity of the product transported. In [41] was developed a method for solution fully FTPs with both the inequality and equality constraints in which all the parameters are represented by non-negative trapezoidal fuzzy numbers. Fully FTPs was resolved in [40], in 2017, using a new method, based on the Hungarian and MODI algorithm. The methods for finding a fuzzy optimal solution of TPs with the LR flat fuzzy numbers were proposed in [6], based on the tabular representation and on the fuzzy linear programming formulation to overcome these shortcomings. Antony et al. used Vogel's approximation method for solving triangular IFTP in 2014 [35]. Fuzzy methods of 2-S time minimizing TPs are presented in [3], [33]. The work [42] was focused on presenting an innovative study of a multi-stage multi-objective solid trapezoidal IFTP with a green supply chain network system. 2-S time minimizing TP have considered in [38] over triangular intuitionistic fuzzy (IF) numbers. Trapezoidal and triangular IFSs are special cases of IFSs.

In our previous works [46], [47], [48], [52], we have proposed for the first time an intuitionistic fuzzy modified distribution algorithm, a zero-suffix and a zero-point method to determine an optimal solution of the IFTP, interpreted by the IFSs and IMs [18], [19] concepts. The concept of index matrices was introduced to enable two matrices with different dimensions to be summed. Later, IMs concept was extended and were defined operations, relations and operators over IMs. The IMs theoretical apparatus was described in [21], [51]. Here, we propose a novel approach to the formulation and solution 2-S IFTP, in which the transportation costs, supply and demand quantities are IFPs, depending on the climatic, road conditions and economic factors. The proposed algorithm uses IMs toolkit for modeling the 2-S IFTP and for finding of its optimal solution. The advantages of the algorithm are indicated.

The remainder of this paper is as follows: Section 2 describes some initial definitions of the theories of the IMs and the IFPs. In Section 3, we formulate 2-S IFTP and propose an algorithm for its solution by the concepts of IMs and IFSs. The effectiveness of the approach is demonstrated by an example in Section 4. Section 5 drawns conclusions and outlines directions for future research.

II. Preparatory Definitions on Intuitionistic Fuzzy Pairs and IMs

In this section we recall some basic definitions on intuitionistic fuzzy pairs (IFPs) from [12], [20], [22], [26], [49] and on index matrices tool from [21], [51].

2.1. Basic Remarks on IFPs

An **IFP** is under the form of an ordered pair $\langle a,b\rangle = \langle \mu(p), \nu(p)\rangle$, where $a,b \in [0,1]$ and $a+b \leq 1$, that is used as an evaluation of a proposition p [22], [26]. $\mu(p)$ and $\nu(p)$ respectively determine the "truth degree" (degree of membership) and "falsity degree" (degree of non-membership).

In the works [10], [12], [20], [26], [24] were proposed some basic operations over two IFPs $x = \langle a,b\rangle$ and $y = \langle c,d\rangle$:

$$\begin{gathered} \neg x = \langle b,a\rangle; \\ x \wedge_1 y = \langle \min(a,c), \max(b,d)\rangle; \\ x \vee_1 y = \langle \max(a,c), \min(b,d)\rangle; \\ x \wedge_2 y = x+y = \langle a+c-a.c, b.d\rangle; \\ x \vee_2 y = x.y = \langle a.c, b+d-b.d\rangle; \\ \alpha.x = \langle 1-(1-a)^{\alpha}, b^{\alpha}\rangle (\text{for } \alpha = n \text{ or } 1/n \ (n \in N)); \\ x-y = \langle \max(0, a-c), \min(1, b+d, 1-a+c)\rangle \end{gathered} \tag{1}$$

$$x : y = \begin{cases} \langle \min(1, a/c), \min(\max(0, 1-a/c), \\ \max(0, (b-d)/(1-d))) \text{if } c \neq 0 \ \& d \ \neq 1 \\ \langle 0,1\rangle \text{ otherwise} \end{cases} \tag{2}$$

The forms of the relations with IFPs are the following

$$\begin{array}{ll} x \geq y \text{ iff } a \geq c \text{ and } b \leq d; & x \leq y \text{ iff } a \leq c \text{ and } b \geq d; \\ x \geq_{\square} y \text{ iff } a \geq c; & x \leq_{\square} y \text{ iff } a \leq c; \\ x \geq_{\diamond} y \text{ iff } b \leq d; & x \leq_{\diamond} y \text{ iff } b \geq d; \\ x = y & \text{iff } a = c \text{ and } b = d \\ x \geq_R y & \text{iff } R_{\langle a,b\rangle} \leq R_{\langle c,d\rangle}, \end{array} \tag{3}$$

where $R_{\langle a,b\rangle} = 0.5(2-a-b)(1-a)$ [12].

The IFP x is an **"intuitionistic fuzzy false pair" (IFFP)** if and only if $a \leq b$.

2.2. Definition, Operations and Relations over Intuitionistic Fuzzy Index Matrices

One of the basic IM-types are intuitionistic fuzzy IMs (IFIMs) whose elements are IFPs. Let $\mathscr{S}$ be a fixed set. The definition of two-dimensional IFIM (2-D IFIM) $[K,L,\{\langle \mu_{k_i,l_j}, \nu_{k_i,l_j}\rangle\}]$ with index sets K and L $(K,L \subset \mathscr{S})$ is the following:

	l_1	$\ldots$	l_j	$\ldots$	l_n
k_1	$\langle \mu_{k_1,l_1}, \nu_{k_1,l_1}\rangle$	$\ldots$	$\langle \mu_{k_1,l_j}, \nu_{k_1,l_j}\rangle$	$\ldots$	$\langle \mu_{k_1,l_n}, \nu_{k_1,l_n}\rangle$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
k_m	$\langle \mu_{k_m,l_1}, \nu_{k_m,l_1}\rangle$	$\ldots$	$\langle \mu_{k_m,l_j}, \nu_{k_m,l_j}\rangle$	$\ldots$	$\langle \mu_{k_m,l_n}, \nu_{k_m,l_n}\rangle$

,

where for $i = 1,...,m; j = 1,...,n$:

$$0 \leq \mu_{k_i,l_j}, \nu_{k_i,l_j}, \mu_{k_i,l_j} + \nu_{k_i,l_j} \leq 1.$$

The basic operations over two IMs

$$A = [K,L,\{\langle \mu_{k_i,l_j}, \nu_{k_i,l_j}\rangle\}] \text{ and } B = [P,Q,\{\langle \rho_{p_r,q_s}, \sigma_{p_r,q_s}\rangle\}]$$

are as follows [21]:

Negation: $\neg A = [K,L,\{\langle \nu_{k_i,l_j}, \mu_{k_i,l_j}\rangle\}]$.

Addition-$(\circ,*)$: $A \oplus_{(\circ,*)} B = [K \cup P, L \cup Q, \{\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle\}]$, where $\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle$

$$= \begin{cases} \langle \mu_{k_i,l_j}, \nu_{k_i,l_j} \rangle, & \text{if } t_u = k_i \in K \text{ and } v_w = l_j \in L - Q \\ & \text{or } t_u = k_i \in K - P \text{ and } v_w = l_j \in L; \\ \langle \rho_{p_r,q_s}, \sigma_{p_r,q_s} \rangle, & \text{if } t_u = p_r \in P \text{ and } v_w = q_s \in Q - L \\ & \text{or } t_u = p_r \in P - K \\ & \text{and } v_w = q_s \in Q; \\ \langle \circ(\mu_{k_i,l_j}, \rho_{p_r,q_s}), & \text{if } t_u = k_i = p_r \in K \cap P \\ *(\nu_{k_i,l_j}, \sigma_{p_r,q_s}) \rangle, & \text{and } v_w = l_j = q_s \in L \cap Q; \\ \langle 0, 1 \rangle, & \text{otherwise.} \end{cases}$$

where $\langle \circ, * \rangle \in \{\langle \max, \min \rangle, \langle \min, \max \rangle, \langle \text{ average,average} \rangle\}$.

Termwise subtraction-(max,min):

$$A -_{(\max,\min)} B = A \oplus_{(\max,\min)} \neg B.$$

Termwise multiplication-$(\min, \max)$:

$$A \otimes_{(\min,\max)} B = [K \cap P, L \cap Q, \{\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle\}],$$

where

$$\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle = \langle \min(\mu_{k_i,l_j}, \rho_{p_r,q_s}), \max(\nu_{k_i,l_j}, \sigma_{p_r,q_s}) \rangle.$$

Multiplication:

$$A \odot_{(\circ,*)} B = [K \cup (P - L), Q \cup (L - P)\{\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle\}], \quad (4)$$

where $\langle \phi_{t_u,v_w}, \psi_{t_u,v_w} \rangle$ is defined in [21] and $\langle \circ, * \rangle \in \{\langle max, min \rangle, \langle min, max \rangle, \langle \wedge_2, \vee_2 \rangle\}$.

Transposition: A^T is the transposed IM of A.

Reduction: The symbol "$\perp$" denotes the lack of some component in the definitions. The operation $(k, \perp)$-reduction of the IM A is defined by: $A_{(k,\perp)} = [K - \{k\}, L, \{c_{t_u,v_w}\}]$,
where $c_{t_u,v_w} = a_{k_i,l_j}$ for $t_u = k_i \in K - \{k\}$ and $v_w = l_j \in L$.

Projection: Let $M \subseteq K$ and $N \subseteq L$. Then,

$$pr_{M,N} A = [M, N, \{b_{k_i,l_j}\}],$$

where for each $k_i \in M$ and each $l_j \in N$, $b_{k_i,l_j} = a_{k_i,l_j}$.

Substitution: Let IM $A = [K, L, \{a_{k,l}\}]$ be given. Some forms of the substitution over A are defined for the couples of indices (p,k) by

$$\left[\frac{p}{k}; \perp\right] A = \left[(K - \{k\}) \cup \{p\}, L, \{a_{k,l}\}\right].$$

Index type operations [45]:

$$AGIndex_{\{(\min/\max)/(\min_\square/\max_\square)/(\min_\diamond/\max_\diamond)(\min_R/\max_R)\}(\not<)}(A) = \langle k_i, l_j \rangle$$

finds the index of the minimum/ maximum element of A with no empty value in accordance with the relations (3).

$$AGIndex_{\{(\min/\max)/(\min_\square/\max_\square)/(\min_\diamond/\max_\diamond)(\min_R/\max_R)\}(\not<)(\notin F)}(A) = \langle k_i, l_j \rangle$$

presents the index of the minimum/ maximum element between the elements of A, whose indexes $\notin F$, with no empty value in accordance with the relations (3).

$$Index_{\{(\min/\max)/(\min_\square/\max_\square)/(\min_\diamond/\max_\diamond)(\min_R/\max_R)\}(\not<),k_i}(A) = \{\langle k_i, l_{v_1} \rangle, \ldots, \langle k_i, l_{v_x} \rangle, \ldots, \langle k_i, l_{v_V} \rangle\},$$

where $\langle k_i, l_{v_x} \rangle$ (for $i = 1, ..., m; j = 1, ..., n; x = 1, ..., V$) are the indices of the minimum/ maximum IFFP of k_i-th row of A with no empty value in accordance with the relations (3).

$$Index_{(\not<)}(A) = \{\langle k_1, l_{v_1} \rangle, \ldots, \langle k_i, l_{v_i} \rangle, \ldots, \langle k_m, l_{v_m} \rangle\},$$

where $\langle k_i, l_{v_i} \rangle$ (for $1 \le i \le m$) are the indices of the element of A, whose cell is full.

Aggregation operations

Let us use the operations $\#_q, (q \le i \le 3)$ from [50] for scaling aggregation operations over two IFPs $x = \langle a, b \rangle$ and $y = \langle c, d \rangle$:
$x \#_1 y = \langle min(a,c), max(b,d) \rangle$;
$x \#_2 y = \langle average(a,c), average(b,d) \rangle$;
$x \#_3 y = \langle max(a,c), min(b,d) \rangle$.

Let $k_0 \notin K$ be a fixed index. The definition of the aggregation operation by the dimension K is [21], [50]: is:

$$\alpha_{K,\#_q}(A, k_0) = \begin{array}{c|ccc} & l_1 & \ldots & l_n \\ \hline k_0 & \overset{m}{\underset{i=1}{\#_q}} \langle \mu_{k_i,l_1}, \nu_{k_i,l_1} \rangle & \ldots & \overset{m}{\underset{i=1}{\#_q}} \langle \mu_{k_i,l_n}, \nu_{k_i,l_n} \rangle \end{array},$$

where $1 \le q \le 3$.

Aggregate global internal operation [45]:

$$AGIO_{\oplus_{(\circ,*)}}(A), \quad (5)$$

where $\langle \circ, * \rangle \in \{\langle max, min \rangle, \langle min, max \rangle, \langle \wedge_2, \vee_2 \rangle\}$.

Non-strict relation "inclusion about value" The form of this type of relations between two IMs A and B is as follows:

$$A \subseteq_v B \text{ iff } (K = P) \ \& \ (L = Q) \ \& \ (\forall k \in K)(\forall l \in L)(a_{k,l} \le b_{k,l}).$$

III. Intuitionistic Fuzzy Index Matrix Approach to Two-Stage IFTP

Let us extend the IFTP from [48] into a two-stage one as follows:

A. Generalized 2-S IFTP

A trader supplies a product to different companies after delivery of that product from different producers in an uncertain environment. Destinations cannot get all the required quantity of product due to limited storage capacity. In this case, the necessary quantities of products are sent to the destinations in two stages. Initially, the minimum destination requirements are sent from the sources to the destinations. Once part of the entire initial shipment has been used up, they are ready to receive the remaining quantity in the second stage. The trader wants to find optimal solutions for the 2-S IFTP.

First stage A trader supplies a product to n different companies (consumers) $\{l_1, \ldots, l_j, \ldots, l_n\}$ after delivery of that product from different m manufacturers (producers) $\{k_1, \ldots, k_i, \ldots, k_m\}$ in quantities $c_{k_i,R}$ (for $1 \le i \le m$). Let the consumers (destinations) need this product in quantities of c_{Q,l_j} (for $1 \le j \le n$).

Let c_{k_i,l_j} be the intuitionistic fuzzy cost for transporting a unit quantity of the product from the k_i-th producer to the l_j-th consumer; x_{k_i,l_j} - the number of units of the product, transported from k_i-th source to l_j-th destination and c_{pl,l_j} (for $1 \le j \le n$) are limits to the transportation costs of the delivery a product from the k_i-th manufacturer to the l_j-th destination under form of IFPs.

Second stage Let some of the buyers $RS = \{l_1^*,\ldots,l_{j^*}^*,\ldots,l_{n^*}^*\}$ ($RS \subset L$) become resellers. The resellers $\{l_1^*,\ldots,l_{j^*}^*,\ldots,l_{n^*}^*\}$ want to sell quantities of the product not only purchased, but also from own production or stocks at a surplus charge $c^*_{l_{j^*}^*,q^*}$ for a product unit to other consumers $\{u_1,\ldots,u_g,\ldots,u_f\}$, in quantities $c^*_{l_{j^*}^*,R^*}$ (for $1 \le j^* \le n^*$).

Consumers need this product in an amount of $c^*_{Q^*,u_g}$ (for $1 \le g \le f$). Let $c^*_{l_{j^*}^*,u_g}$ (for $1 \le j^* \le n^*, 1 \le g \le f$) be the total cost for the purchase of one unit quantity of the product from the $l_{j^*}^*$-th reseller to u_g-th destination; $x^*_{l_{j^*}^*,u_g}$ – the number of units of the product, transported from the $l_{j^*}^*$-th reseller to u_g-th destination; $c^*_{l_{j^*}^*,pu^*}$, (for $1 \le j^* \le n^*$) – is the price of a product unit of the $l_{j^*}^*$-th reseller; $c^*_{pl^*,u_g}$, (for $1 \le g \le f$) – upper limit of the price at which the u_g-th consumer wish to purchase the product.

For estimating the parameters of 2-S IFTP, we can use the expert approach described in detail in [20]. The experts are not sure about the transportation costs, the quantities of offered and demanded goods due to uncontrollable factors. The transportation costs are evaluated as intuitionistic fuzzy numbers after a thorough discussion, interpreted by the intuitionistic fuzzy concept. The purpose of the 2-S IFTP is to meet the requests of all users $\{l_1,\ldots,l_j,\ldots,l_m\}$ and $\{u_1,\ldots,u_g,\ldots,u_f\}$ from the two stages so that the intuitionistic fuzzy transportation cost is minimum.

B. Solution of the 2-S IFTP

The proposed algorithm for modeling of 2-S IFTP and finding of its optimal solution is based on IMs concept [21].

1) Solution of the First Stage of the 2-S IFTP: **Step 1.** At starting of the algorithm for solution of the 2-S IFTP, the cost IM $C[K,L]$ is created:

$$= \begin{array}{c|cccccc} & l_1 & \ldots & l_n & R & pu \\ k_1 & \langle \mu_{k_1,l_1}, \nu_{k_1,l_1}\rangle & \cdots & \langle \mu_{k_1,l_n}, \nu_{k_1,l_n}\rangle & \langle \mu_{k_1,R}, \nu_{k_1,R}\rangle & \langle \mu_{k_1,pu}, \nu_{k_1,pu}\rangle \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ k_m & \langle \mu_{k_m,l_1}, \nu_{k_m,l_1}\rangle & \cdots & \langle \mu_{k_m,l_n}, \nu_{k_m,l_n}\rangle & \langle \mu_{k_m,R}, \nu_{k_m,R}\rangle & \langle \mu_{k_m,pu}, \nu_{k_m,pu}\rangle \\ Q & \langle \mu_{Q,l_1}, \nu_{Q,l_1}\rangle & \cdots & \langle \mu_{Q,l_n}, \nu_{Q,l_n}\rangle & \langle \mu_{Q,R}, \nu_{Q,R}\rangle & \langle \mu_{Q,pu}, \nu_{Q,pu}\rangle \\ pl & \langle \mu_{pl,l_1}, \nu_{pl,l_1}\rangle & \cdots & \langle \mu_{pl,l_n}, \nu_{pl,l_n}\rangle & \langle \mu_{pl,R}, \nu_{pl,R}\rangle & \langle \mu_{pl,pu}, \nu_{pl,pu}\rangle \\ pu_1 & \langle \mu_{pu_1,l_1}, \nu_{pu_1,l_1}\rangle & \cdots & \langle \mu_{pu_1,l_n}, \nu_{pu_1,l_n}\rangle & \langle \mu_{pu_1,R}, \nu_{pu_1,R}\rangle & \langle \mu_{pu_1,pu}, \nu_{pu_1,pu}\rangle \end{array},$$

where $K = \{k_1,k_2,\ldots,k_m,Q,pl,pu_1\}$, $L = \{l_1,l_2,\ldots,l_n,R,pu\}$ and for $1 \le i \le m$, $1 \le j \le n$, $\{c_{k_i,l_j}, c_{k_i,R}, c_{k_i,pu}, c_{pl,l_j}, c_{pl,R}, c_{pl,pu}, c_{Q,l_j}, c_{Q,R}, c_{Q,pu}, c_{pu_1,l_j}, c_{pu_1,R}, c_{pu_1,pu}\}$ are IFPs.

Let we denote by $|K| = m+3$ the number of elements of the set K; then $|L| = n+2$.

We also define the IM

$$X[K^I,L^I] = \begin{array}{c|ccccc} & l_1 & \ldots & l_j & \ldots & l_n \\ k_1 & x_{k_1,l_1} & \cdots & x_{k_1,l_j} & \cdots & x_{k_1,l_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ k_m & x_{k_m,l_1} & \ldots & x_{k_m,l_j} & \ldots & x_{k_m,l_n} \end{array},$$

$K^I = \{k_1,k_2,\ldots,k_m\}$, $K^I = \{l_1,l_2,\ldots,l_n\}$, and for $1 \le i \le m$, $1 \le j \le n$: $x_{k_i,l_j} = \langle \rho_{k_i,l_j}, \sigma_{k_i,l_j}\rangle$. Go to *Step 2.*

Step 2. For solving the first stage on the 2-S IFTP we can apply one of the algorithms, outlined in our papers [47], [48], [52]. In the program code of the developed algorithms was used a part of Microsoft Visual Studio.NET 2010 C project's.

After an application of the algorithm for finding an optimal solution of IFTP, the following conditions are checked: $D = Index_{\not\perp} X = \{\langle k_{i*_1}, l_{j*_1}\rangle,\ldots,\langle k_{i*_f}, l_{j*_f}\rangle,\ldots,\langle k_{i*_\varphi}, l_{j*_\varphi}\rangle\}$.

If the intuitionistic fuzzy feasible solution is degenerated (it contains less than $m+n-1$ (the total number of producers and consumers decreased by 1) occupied cells in the X i.e. $|D| < m+n-1$) [9] then increase the basic cells x_{k_i,l_j} with one to which the minimum transportation cost corresponds. Let us the recorded delivery of this cell is $\langle 0,1\rangle$. The IMs operations are:

If

$$|D| < m+n-1, \text{ then}$$

$$\{AGIndex_{\{(\min/\max)/(\min_\square/\max_\square)/(\min_\circ/\max_\circ)(\min_R/\max_R)\}(\not\perp)(\notin D)}(C) = \langle k_\alpha, l_\beta\rangle; x_{k_{al},l_\beta} = \langle 0,1\rangle\}.$$

for $i = 1$ to m
for $j = 1$ to n
If $x_{k_i,l_j} = \langle \perp,\perp\rangle$ then $x_{k_i,l_j} = \langle 0,1\rangle$.
Go to *Step 3.*

Step 3. The optimal intuitionistic fuzzy transportation cost at the first stage is calculated by:

$$AGIO^1_{\oplus_{(\max,\min))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})} \otimes_{(\min,\max)} X_{opt}\right)$$

or $AGIO^2_{\oplus_{(\wedge_2))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})} \otimes_{(\vee_2)} X_{opt}\right)$,
where $\vee_2$ and $\wedge_2$ are the operations from (1).

2) Solution of the Second Stage of the 2-S IFTP: To find the optimal solution for the second stage of the problem, we propose the following algorithm, described by a program code, which is a part of Microsoft Visual Studio.NET 2010 C project.

Step 4. Let us create the following cost IFIM $C^*[L^*,U]$

$$= \begin{array}{c|cccccc} & u_1 & \cdots & u_f & R^* & q^* & pu^* \\ l_1^* & c^*_{l_1^*,u_1} & \cdots & c^*_{l_1^*,u_f} & c^*_{l_1^*,R^*} & c^*_{l_1^*,q^*} & c^*_{l_1^*,pu^*} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\ l_{j^*}^* & c^*_{l_{j^*}^*,u_1} & \cdots & c^*_{l_{j^*}^*,u_f} & c_{l_{j^*}^*,R^*} & c^*_{l_{j^*}^*,q^*} & c^*_{l_{j^*}^*,pu^*} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\ l_{n^*}^* & c^*_{l_{n^*}^*,u_1} & \cdots & c^*_{l_{n^*}^*,u_f} & c^*_{l_{n^*}^*,R^*} & c^*_{l_{n^*}^*,q^*} & c^*_{l_{n^*}^*,pu^*} \\ Q^* & c^*_{Q^*,u_1} & \cdots & c^*_{Q^*,u_f} & c^*_{Q^*,R^*} & c^*_{Q^*,q^*} & c^*_{Q^*,pu^*} \\ pl^* & c^*_{pl^*,u_1} & \cdots & c^*_{pl^*,u_f} & c^*_{pl^*,R^*} & c^*_{pl^*,q^*} & c^*_{pl^*,pu^*} \\ pu_1^* & c^*_{pu_1^*,u_1} & \cdots & c^*_{pu_1^*,u_f} & c^*_{pu_1^*,R^*} & c^*_{pu_1^*,q^*} & c^*_{pu_1^*,pu^*} \end{array}$$

where $L^* = \left\{l_1^*, \ldots, l_{j^*}^*, \ldots, l_{n^*}^*, Q^*, pl^*, pu_1^*\right\}$, $U = \{u_1, \ldots, u_g, \ldots, u_f, R^*, q^*, pu^*\}$ and $L^* \subset L$ and for $1 \leq j^* \leq n^*$, $1 \leq g \leq f$, $\{c^*_{l^*_{j^*},u_g}, c_{l^*_{j^*},R^*}, c^*_{l^*_{j^*},q^*}, c^*_{l^*_{j^*},pu^*}, c^*_{pu_1^*,q^*}, c^*_{pu_1^*,u_g}, c^*_{pu_1^*,R}, c^*_{Q^*,u_f}, c^*_{pl^*,u_f}\}$ and $c^*_{pu_1^*,pu^*}$ are IFPs, having meaning as defined in the generalized 2-S IFTP.

We also define the IFIM

$$X[L^J,U] = \begin{array}{c|ccccc} & u_1 & \ldots & u_g & \ldots & u_f \\ \hline l_{1^*}^* & x_{l_{1^*}^*,u_1} & \cdots & x_{l_{1^*}^*,u_g} & \cdots & x_{l_{1^*}^*,u_f} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ l_{n^*}^* & x_{l_{n^*}^*,u_1} & \ldots & x_{l_{n^*}^*,u_g} & \ldots & x_{l_{n^*}^*,u_f} \end{array},$$

where $L^J = \{l_{1^*}^*, l_{2^*}^*, \ldots, l_{n^*}^*\}$, $U = \{u_1, u_2, \ldots, u_f\}$, and for $1 \leq j^* \leq n^*$, $1 \leq g \leq f$: $x_{l_{j^*}^*,u_g} = \langle \rho_{l_{j^*}^*,u_g}, \sigma_{l_{j^*}^*,u_g} \rangle$ are the number of units of the product, transported from the $l_{j^*}^*$-th reseller to u_g-th destination.

Go to *Step 5*.

Step 5. Let construct IFIM matrix:

$$C_1 = pr_{RS,R^*}(\alpha_{K^I,\#_q}(X_{opt}, R^*)^T)).$$

Then $C^* := C^* \oplus_{(\circ,*)} C_1$.

So, the quantities of product purchased in this way by the resellers from the set RS are set in the column R^* of the matrix C^*. Also, the elements $\{c^*_{l^*_{j^*},q^*}, c^*_{Q^*,u_g}, c^*_{pl^*,u_g}\}$ (for $1 \leq j^* \leq n^*$, $1 \leq g \leq f$) are introduced in C^*.

Construct the matrix $E[K/\{Q, pl, pu_1\}, L/\{R, pu\}]$

$$= C_{(\{Q,pl,pu_1\},\{R,pu\})} \otimes_{(\min,\max)} X_{opt}.$$

Go to *Step 6*.

Step 6. Through the following operations we will find the average price of the $l_{j^*}^*$-th reseller $\in RS$ to purchase a single quantity of product.

Then construct the IM $C_{2a} = \alpha_{K^I,\#_2}(E, pu^*)^T$;

For $1 \leq j^* \leq n^*$ do:

{Construct the matrices:

$$C_{2b}[l_{j^*}^*, R_{j^*}^*, \{C_{2a_{l_{j^*}^*,pu^*}}/C^*_{l_{j^*}^*,R_{j^*}^*}\}],$$

in which we use the operation division of IFPs (2):

$$C^* := C^* \oplus_{(\circ,*)} \left[\perp; \frac{pu^*}{R^*}\right] C_{2b}.\}$$

Go to *Step 7*.

Step 7. The following operations will reflect in the column pu^* of the matrix C^* the final selling prices of a unit quantity of the product together with its surplus charge above the purchase price.

Let us construct the matrices

$$C_3 = \left[\perp; \frac{pu^*}{q^*}\right] \{pr_{RS,q^*} C^*\}$$

and $C_4 = pr_{RS,pu^*} C^*$.

Let us perform operation $C^* := C^* \oplus_{(\circ,*)} C_3 \otimes C_4$.

Go to *Step 8*.

Step 8. Through the following operations, the elements $c^*_{l^*j^*,u_g}$ (for $1 \leq g \leq f$) of the matrix C^* will contain the final selling price per unit of product, including the unit price and its transportation price from the l^*j^*-th reseller to u_g-th destination.

For $1 \leq j^* \leq n^*$, $1 \leq g \leq f$, do following:

$$\{C^*_{l^*j^*,u_g} = \{pr_{l^*_{j^*},u_g} C^*\} \oplus_{(\circ,*)} \left[\perp; \frac{u_g}{pu^*}\right] \{pr_{l^*_{j^*},pu^*} C^*\};$$

$$C^* := C^* \oplus_{(\circ,*)} C^*_{l^*_{j^*},u_g}.\}$$

Go to *Step 9*.

Step 9. Determining the optimal plan at second stage of the 2-S IFTP - $X^*[L^J, U, \{x^*_{l^*_{j^*},u_g}\}]$ after execution of one of the algorithms, presented in [47], [48], [52] with the obtained cost IFIM C^*. The optimal intuitionistic fuzzy transportation cost at the second stage is calculated by:

$$AGIO^1_{\oplus_{(\max,\min))}}\left(C^*_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})} \otimes_{(\min,\max)} X^*{}_{opt}\right)$$

or $AGIO^2_{\oplus_{(\wedge_2))}}\left(C^*_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})} \otimes_{(\vee_2)} X^*{}_{opt}\right)$, where $\vee_2$ and $\wedge_2$ are the operations from (1).

Step 10. The optimal intuitionistic fuzzy transportation cost for the problem is calculated by:

$AGIO^1_{\oplus_{(\max,\min))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})} \otimes_{(\min,\max)} X_{opt}\right) \quad \oplus_{(\max,\min)}$
$AGIO^1_{\oplus_{(\max,\min))}}\left(C^*_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})} \otimes_{(\min,\max)} X^*{}_{opt}\right)$
or $AGIO^2_{\oplus_{(\wedge_2))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})} \otimes_{(\vee_2)} X_{opt}\right) \quad \oplus_{(\max,\min)}$
$AGIO^2_{\oplus_{(\wedge_2))}}\left(C^*_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})} \otimes_{(\vee_2)} X^*{}_{opt}\right)$
where $\vee_2$ and $\wedge_2$ are the operations from (1).

IV. An Application of the Algorithm for Solution of 2-S IFTP

In this section we will define 2-S IFTP extending the IFTP from [48]: A trader supplies a product to 4 different companies $\{l_1, l_2, l_3, l_4\}$. Let a product be produced at the manufacturers $\{k_1, k_2, k_3\}$ in quantities $c_{k_i,R}$ (for $1 \leq i \leq 3$). Let the companies ($\{l_1, l_2, l_3, l_4\}$) demand this product in an quantity of c_{Q,l_j} (for $1 \leq j \leq 4$) and c_{pl,l_j} (for $1 \leq j \leq 4$) are intuitionistic fuzzy limits to the transportation costs of delivery a particular product from the k_i-th source to the l_j-th destination. Let some of the buyers $RS = \{l_1, l_2, l_3\}$ ($RS \subset L$) become resellers. The resellers $\{l_1, l_2, l_3\}$ want to sell quantities of the product not only purchased, but also from own production or stocks at a surplus charge $c^*_{l^*_{j^*},q^*}$ (for $1 \leq j^* \leq 3$) for an product unit to other consumers $\{u_1, u_2, u_3, u_4\}$, in quantities $c^*_{l^*_{j^*},R^*}$ (for $1 \leq j^* \leq 3$). Consumers need this product in an amount of $c^*_{Q^*,u_g}$ (for $1 \leq g \leq 4$). Let $c^*_{l^*_{j^*},u_g}$ (for $1 \leq j^* \leq n^*, 1 \leq g \leq f$) be the total cost for the purchase of one unit quantity of the product from the $l^*_{j^*}$-th reseller to u_g-th destination; $x^*_{l^*_{j^*},u_g}$ – the number of units of the product, transported from the $l^*_{j^*}$-th reseller to u_g-th destination; $c^*_{l^*_{j^*},pu^*}$, (for $1 \leq j^* \leq 3$) – is the price of a product unit of the $l^*_{j^*}$-th reseller; $c^*_{pl^*,u_g}$, (for $1 \leq g \leq 4$) –

upper limit of the price at which the u_g-th consumer wish to purchase the product.

The purpose of the 2-S IFTP is to meet the requests of all users $\{l_1,\ldots,l_4\}$ and $\{u_1,u_2,u_3\}$ so that the intuitionistic fuzzy transportation cost is minimum.

All elements of the transportation problem are intuitionistic fuzzy due to several uncertainties.

Let us apply the proposed approach in the Sect. III.

1) Solution of the First Stage of the 2-S IFTP: **Step 1.** At starting of the algorithm for solution of the problem, the cost IM C is created. c_{k_i,l_j} (for $1\le i\le 3, 1\le j\le 4$) is the IF cost for transporting a unit quantity of the product from the k_i-th producer to the l_j-th user.

$$C[K,L]=\left\{\begin{array}{c|cccc} & l_1 & l_2 & l_3 & \ldots \\ k_1 & \langle 0.6,0.2\rangle & \langle 0.7,0.1\rangle & \langle 0.3,0.1\rangle & \ldots \\ k_2 & \langle 0.5,0.3\rangle & \langle 0.4,0.1\rangle & \langle 0.5,0.1\rangle & \ldots \\ k_3 & \langle 0.4,0.2\rangle & \langle 0.3,0.2\rangle & \langle 0.6,0.1\rangle & \ldots \\ Q & \langle 0.4,0.2\rangle & \langle 0.5,0.3\rangle & \langle 0.6,0.2\rangle & \ldots \\ pl & \langle 0.65,0.3\rangle & \langle 0.6,0.4\rangle & \langle 0.75,0.1\rangle & \ldots \\ pu_1 & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \ldots \end{array}\right.$$

…	l_4	R	pu
…	$\langle 0.8,0.1\rangle$	$\langle 0.5,0.2\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.3,0.2\rangle$	$\langle 0.7,0.1\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.7,0.2\rangle$	$\langle 0.4,0.5\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.06,0.02\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.75,0.1\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$

Let x_{k_i,l_j} is the number of units of the product, transported from the k_i-th producer to l_j-th destination (for $1\le i\le 3$ and $1\le j\le 4$) and is an element of IFIM X with initial elements $\langle \bot,\bot\rangle$. The trader wants to satisfy the required quantities of the users so that the intuitionistic fuzzy transportation cost is minimum.

Step 2. The conditions for limiting the transportation costs are checked according to proposed approach in [48]. The problem is also balanced.

The IM C is transformed in this form following the IF algorithm in [48]:

$$C[K,L]=\left\{\begin{array}{c|cccc} & l_1 & l_2 & l_3 & \ldots \\ k_1 & \langle 0.6,0.2\rangle & \langle 1,0\rangle & \langle 0.3,0.1\rangle & \ldots \\ k_2 & \langle 0.5,0.3\rangle & \langle 0.4,0.1\rangle & \langle 0.5,0.1\rangle & \ldots \\ k_3 & \langle 0.4,0.2\rangle & \langle 0.3,0.2\rangle & \langle 0.6,0.1\rangle & \ldots \\ Q & \langle 0.4,0.2\rangle & \langle 0.5,0.3\rangle & \langle 0.6,0.2\rangle & \ldots \\ pl & \langle 0.65,0.3\rangle & \langle 0.6,0.4\rangle & \langle 0.75,0.1\rangle & \ldots \\ pu_1 & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \ldots \end{array}\right.$$

…	l_4	R	pu
…	$\langle 1,0\rangle$	$\langle 0.5,0.2\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.3,0.2\rangle$	$\langle 0.7,0.1\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.7,0.2\rangle$	$\langle 0.4,0.5\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.06,0.02\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.65,0.3\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$

For solving the first stage we can apply the zero-point algorithm for IFTP with IFIMs C and X, outlined in [48].

Step 3. The intuitionistic fuzzy optimal solution, presented by the IM X_{opt} is non-degenerated, it includes 6 occupied cells. The IM X_{opt} has the following form:

$$X_{opt}=\left\{\begin{array}{c|cccc} & l_1 & l_2 & l_3 & l_4 \\ k_1 & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0.5,0.2\rangle & \langle 0,1\rangle \\ k_2 & \langle 0.4,0.2\rangle & \langle 0.1,0.8\rangle & \langle 0.1,0.4\rangle & \langle 0.06,0.02\rangle \\ k_3 & \langle 0,1\rangle & \langle 0.4,0.5\rangle & \langle 0,1\rangle & \langle 0,1\rangle \end{array}\right. . \tag{6}$$

The optimal intuitionistic fuzzy optimal solution $X_{opt}[K*,L*,\{x_{k_i,l_j}\}]$ is obtained. The optimal intuitionistic fuzzy transportation cost is:

$$AGIO^1_{\oplus_{(\max,\min))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})}\otimes_{(\min,\max)}X_{opt}\right)=\langle 0.4,0.2\rangle \tag{7}$$

or

$$AGIO^2_{\oplus_{(\wedge_2))}}\left(C_{(\{Q,pl,pu_1\},\{R,pu\})}\otimes_{(\vee_2)}X_{opt}\right)=\langle 0.464,0.006\rangle. \tag{8}$$

The degree of membership (acceptance) of this optimal solution is equal to 0.4 (or 0.464) and its degree of non-membership (non-acceptance) is equal to 0.2 (or 0.006).

The ranking function R, defined in (3), we can use to rank alternatives of decision-making process. For the obtained optimal solution by IFZPM, the distance between the optimal solution to the pair $\langle 1,0\rangle$ is equal to $R_{\langle 0.4;0.2\rangle}=0.42$ ($R_{\langle 0.464;0.006\rangle}=0.41$).

2) Solution of the Second Stage of the 2-S IFTP: To find the optimal solution for the second stage, we propose the following algorithm, described by program code, which is a part of Microsoft Visual Studio.NET 2010 C project.

Step 4. The following cost IFIM $C^*[L^*,U]$ is created:

$$C^*=\left\{\begin{array}{c|cccc} & u_1 & u_2 & u_3 & \ldots \\ l_1 & \langle 0.27,0.73\rangle & \langle 0.23,0.77\rangle & \langle 0.19,0.81\rangle & \ldots \\ l_2 & \langle 0.17,0.83\rangle & \langle 0.29,0.71\rangle & \langle 0.29,0.71\rangle & \ldots \\ l_3 & \langle 0.24,0.65\rangle & \langle 0.24,0.6\rangle & \langle 0.2,0.65\rangle & \ldots \\ Q^* & \langle 0.45,0.3\rangle & \langle 0.4,0.2\rangle & \langle 0.15,0.013\rangle & \ldots \\ pl^* & \langle 0.82,0.1\rangle & \langle 0.8,0.1\rangle & \langle 0.85,0.1\rangle & \ldots \\ pu_1^* & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \langle \bot,\bot\rangle & \ldots \end{array}\right.$$

…	u_4	R^*	q^*	pu^*
…	$\langle 0.65,0.35\rangle$	$\langle 0.4,0.2\rangle$	$\langle 0.1,0\,\bot\rangle$	$\langle 0.5,0.3,\bot\rangle$
…	$\langle 0.67,0.33\rangle$	$\langle 0.5,0.3\rangle$	$\langle 0.1,0\rangle$	$\langle 0.31,0.27\rangle$
…	$\langle 0.56,0.1\rangle$	$\langle 0.6,0.2\rangle$	$\langle 0.15,0\rangle$	$\langle 0.25,0\rangle$
…	$\langle 0.2,0.013\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle 0.85,0.1\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$
…	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$	$\langle \bot,\bot\rangle$

where $L^*=\{l_1,\ldots,l_3,Q^*,pl^*,pu_1^*\}$, $U=\{u_1,\ldots,u_4,R^*,q^*,pu^*\}$ and all elements are IFPs. The quantities of product purchased on the first stage by the resellers from the set RS are set in the column R^* of the matrix C^*.

We also define

$$X[L^J,U]=\left\{\begin{array}{c|cccc} & u_1 & u_2 & u_3 & u_4 \\ \hline l_1 & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle \\ l_2 & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle \\ l_3 & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle & \langle 0,1\rangle \end{array}\right\},$$

$L^J=\{l_1,l_2,l_3\}$, $U=\{u_1,u_2,u_3,u_4\}$ and for $1\le j^*\le 3$, $1\le g\le 4$: $x_{l^*_{j^*},u_g}=\langle \rho_{l^*_{j^*},u_g},\sigma_{l^*_{j^*},u_g}\rangle$ are the number of units of the product, transported from the $l^*_{j^*}$-th reseller to u_g-th destination.

Step 5. The average prices of the resellers l_1,l_2,l_3 to purchase a single quantity of product are calculated. The IFIM $C^*[L^*,U]$ is changed as follows:

$$\left\{\begin{array}{c|ccccc} & \ldots & R^* & q^* & pu^* \\ \hline l_1 & \ldots & \langle 0.4,0.2\rangle & \langle \perp,\perp\rangle & \langle 0.5,0.3\rangle \\ l_2 & \ldots & \langle 0.5,0.3\rangle & \langle \perp,\perp\rangle & \langle 0.31,0.27\rangle \\ l_3 & \ldots & \langle 0.6,0.2\rangle & \langle \perp,\perp\rangle & \langle 0.25,0\rangle \\ Q^* & \ldots & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle \\ pl^* & \ldots & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle \\ pu_1^* & \ldots & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle \end{array}\right\}$$

Step 6. The column pu^* of the matrix C^* contains the final selling prices of a unit quantity of the product together with its mark-up above the purchase price. The elements $c^*_{l^*j^*,u_g}$ (for $1\le j^*\le 3$, $1\le g\le f$) of the matrix C^* contain the final selling price per unit of product, including the unit price and its transportation price from the l^*j^*-th reseller to u_g-th destination. C^* obtains the following form:

$$C^*=\left\{\begin{array}{c|cccc} & u_1 & u_2 & u_3 & \ldots \\ \hline l_1 & \langle 0.32,0.55\rangle & \langle 0.28,0.55\rangle & \langle 0.24,0.7\rangle & \ldots \\ l_2 & \langle 0.2,0.7\rangle & \langle 0.32,0.55\rangle & \langle 0.32,0.6\rangle & \ldots \\ l_3 & \langle 0.28,0.65\rangle & \langle 0.28,0.6\rangle & \langle 0.2,0.65\rangle & \ldots \\ Q^* & \langle 0.45,0.3\rangle & \langle 0.4,0.2\rangle & \langle 0.15,0.013\rangle & \ldots \\ pl^* & \langle 0.82,0.1\rangle & \langle 0.8,0.1\rangle & \langle 0.85,0.1\rangle & \ldots \\ pu_1^* & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle & \langle \perp,\perp\rangle & \ldots \end{array}\right.$$

…	u_4	R^*	q^*	pu^*
…	$\langle 0.7,0.1\rangle$	$\langle 0.4,0.2\rangle$	$\langle 0.1,0\perp\rangle$	$\langle 0.05,0.3,\perp\rangle$
…	$\langle 0.7,0.1\rangle$	$\langle 0.5,0.3\rangle$	$\langle 0.1,0\rangle$	$\langle 0.03,0.27\rangle$
…	$\langle 0.6,0.1\rangle$	$\langle 0.6,0.2\rangle$	$\langle 0.15,0\rangle$	$\langle 0.038,0\rangle$
…	$\langle 0.2,0.013\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$
…	$\langle 0.85,0.1\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$
…	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$	$\langle \perp,\perp\rangle$

Step 7. The problem is balanced. Then the requirements for an upper limit on the price at which consumers have the opportunity to purchase the necessary quantities of the product are checked. After execution of the algorithm, presented in [48], with the obtained cost IFIMs C^* and X^*, we obtain the following optimal plan $X^*[L^J,U,\{x^*_{l^*_{j^*},u_g}\}]$ for the second stage of the problem:

$$\left\{\begin{array}{c|cccc} & u_1 & u_2 & u_3 & u_4 \\ \hline l_1 & \langle 0,1\rangle & \langle 0.35,0.65\rangle & \langle 0,1\rangle & \langle 0,1\rangle \\ l_2 & \langle 0.45,0.3\rangle & \langle 0.05,0.6\rangle & \langle 0,1\rangle & \langle 0,1\rangle \\ l_3 & \langle 0,1\rangle & \langle 0.15,0.23\rangle & \langle 0.15,0.013\rangle & \langle 0.2,0.013\rangle \end{array}\right\},$$

The intuitionistic fuzzy optimal solution, presented by the IM X^*_{opt} is non-degenerated, it includes 6 occupied cells. The optimal intuitionistic fuzzy transportation cost at the second stage is calculated by:

$$AGIO^1_{\oplus_{(\max,\min)})}\left(C^*{}_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})}\otimes_{(\min,\max)}X^*{}_{opt}\right)=\langle 0.28,0.1\rangle$$

$$\text{or } AGIO^2_{\oplus_{(\wedge_2)})}\left(C^*{}_{(\{Q^*,pl^*,pu_1^*\},\{R^*,q^*,pu^*\})}\otimes_{(\vee_2)}X^*{}_{opt}\right)=\langle 0.31,0.04\rangle,$$

where $\vee_2$ and $\wedge_2$ are the operations from (1). The degree of membership (acceptance) of this optimal solution is equal to 0.28 (or 0.31) and the its degree of non-membership (non-acceptance) is equal to 0.1 (or 0.04). For the obtained optimal solution by IFZPM, the distance between the optimal solution to the pair $\langle 1,0\rangle$ is equal to $R_{\langle 0.28;0.1\rangle}=0.58$ ($R_{\langle 0.31;0.04\rangle}=0.57$).

Step 8. The optimal intuitionistic fuzzy transportation cost for the problem is calculated by: $\langle 0.4,0.2\rangle\oplus_{(\max,\min)}\langle 0.28,0.1\rangle$(or $\langle 0.31,0.4\rangle$) $=\langle 0.4,0.1\rangle$(or $\langle 0.4,0.4\rangle$).

The degree of membership (acceptance) of this optimal solution is equal to 0.4.

The example illustrates the reliability of the proposed algorithm in Section III to the studied 2-S IFTP.

V. Conclusion

The apparatus of IMs provides the ability to expand the existing transportation problems to formulate non-existent ones to find strategic decisions for logistics management in uncertain environment. It is proposed for the first time, extending the approach in [48], to model and find the optimal solution of a 2-S IFTP using the concepts of the IMs and IFSs. The formulated IFTP has additional constraints: upper limits to the transportation costs and a surplus charge on reseller sales prices. The proposed algorithm for solution of the 2-S IFTP is illustrated with a numerical example. The advantages of the proposed algorithm are that it can be easy generalized to the multidimensional intuitionistic fuzzy TPs [23] and also can be applied to both the TP with crisp parameters and with intuitionistic fuzzy ones. In the future, we will extend the proposed approach to the interval-valued intuitionistic fuzzy TPs [25] and will apply it over real life TPs.

References

[1] A. Edwuard Samuel, "Improved zero point method," *Applied mathematical sciences,* vol. 6 (109), 2012, pp. 5421–5426.

[2] A. Gani, A. Samuel, D. Anuradha, "Simplex type algorithm for solving fuzzy transportation problem," *Tamsui Oxf. J. Inf. Math. Sci.,* vol. 27, 2011, pp. 89–98.

[3] A. Gani, K. Razak, "Two stage fuzzy transportation problem," *J Phys Sci,* vol. 10, 2006, pp. 63–69.

[4] A. Gani, S. Abbas, "A new average method for solving intuitionistic fuzzy transportation problem," *International Journal of Pure and Applied Mathematics,* vol. 93 (4), 2014, pp. 491-499.

,[5] A. Kaur, A. Kumar, "A new approach for solving fuzzy transportation problems using generalized trapezoidal fuzzy numbers," *Applied Soft Computing,* vol. 12 (3), 2012, pp. 1201-1213.

[6] A. Kaur, J. Kacprzyk and A. Kumar, *Fuzzy transportation and transshipment problems,* Studies in fuziness and soft computing, vol. 385, 2020.

[7] A. Patil, S. Chandgude, "Fuzzy Hungarian Approach for Transportation Model," *International Journal of Mechanical and Industrial Engineering,* vol. 2 (1), pp. 77-80, 2012.

[8] A.E. Samuel, M. Venkatachalapathy, "Modified vogel's approximation method for fuzzy transportation problems," *Appl. Math. Sci.,* vol. 5, 2011, pp. 1367–1372.

[9] B. Atanassov, *Quantitative methods in business management,* Publ. houseTedIna, Varna; 1994. (in Bulgarian)

[10] B. Riecan, A. Atanassov, "Operation division by *n* over intuitionistic fuzzy sets," *NIFS,* vol. 16, No. 4, 2010, pp. 1-4.

[11] S. Dinagar, K. Palanivel, "The transportation problem in fuzzy environment," *Int. J. Algorithms Comput. Math.,* vol. 2, 2009, pp. 65–71.

[12] E. Szmidt, J. Kacprzyk, "Amount of information and its reliability in the ranking of Atanassov's intuitionistic fuzzy alternatives," *in: Rakus-Andersson, E., Yager, R., Ichalkaranje, N., Jain, L.C. (eds.),* Recent Advances in Decision Making, SCI, Springer, Heidelberg, vol. 222, 2009, pp. 7–19.

[13] F. Jimenez, J. Verdegay, "Solving fuzzy solid transportation problems by an evolutionary algorithm based parametric approach," *European Journal of Operational Research,* vol. 117 (3),1999, pp. 485-510.

[14] F. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematical Physics,* vol. 20, 1941, pp. 224-230.

[15] G. Gupta, A. Kumar, M. Sharma, "A Note on A New Method for Solving Fuzzy Linear Programming Problems Based on the Fuzzy Linear Complementary Problem (FLCP)," *International Journal of Fuzzy Systems,* 2016, pp. 1-5.

[16] H. Arsham, A. Khan, "A simplex type algorithm for general transportation problems-An alternative to stepping stone," *Journal of Operational Research Society,* vol. 40 (6), 2017, pp. 581-590.

[17] H. Basirzadeh, "An approach for solving fuzzy transportation problem, " *Appl. Math. Sci.,* vol. 5, 2011, pp. 1549–1566.

[18] K. Atanassov, "Intuitionistic Fuzzy Sets," VII ITKR Session, Sofia, 20-23 June 1983 (Deposed in Centr. Sci.-Techn. Library of the Bulg. Acad. of Sci., 1697/84) (in Bulgarian). Reprinted: *Int. J. Bioautomation,* vol. 20(S1), 2016, pp. S1-S6.

[19] K. Atanassov, "Generalized index matrices," *Comptes rendus de l'Academie Bulgare des Sciences,* vol. 40(11), 1987, pp. 15-18.

[20] K. Atanassov, *On Intuitionistic Fuzzy Sets Theory,* STUDFUZZ. Springer, Heidelberg, vol. 283; DOI:10.1007/978-3-642-29127-2, 2012.

[21] K. Atanassov, *Index Matrices: Towards an Augmented Matrix Calculus. Studies in Computational Intelligence*, Springer, Cham, vol. 573; DOI: 10.1007/978-3-319-10945-9, 2014.

[22] K. Atanassov, "Intuitionistic Fuzzy Logics," *Studies in Fuzziness and Soft Computing,* Springer, vol. 351, DOI:10.1007/978-3-319-48953-7, 2017.

[23] K. Atanassov, "n-Dimensional extended index matrices Part 1," *Advanced Studies in Contemporary Mathematics*, vol. 28 (2), 2018, pp. 245-259.

[24] K. Atanassov, "Remark on an intuitionistic fuzzy operation division," *Annual of Informatics Section, Union of Scientists in Bulgaria,* vol. 10, 2019 (in press)

[25] K. Atanassov, G. Gargov, "Interval valued intuitionistic fuzzy sets," *Fuzzy sets and systems,* vol. 31 (3), 1989, pp. 343-349.

[26] K. Atanassov, E. Szmidt, J. Kacprzyk, "On intuitionistic fuzzy pairs," *Notes on Intuitionistic Fuzzy Sets,* vol. 19 (3), 2013, pp. 1-13.

[27] K. Kathirvel, K. Balamurugan, "Method for solving fuzzy transportation problem using trapezoidal fuzzy numbers," *International Journal of Engineering Research and Applications,* vol. 2 (5), 2012, pp. 2154-2158.

[28] K. Kathirvel, K. Balamurugan, "Method for solving unbalanced transportation problems using trapezoidal fuzzy numbers," *International Journal of Engineering Research and Applications,* vol. 3 (4), 2013, pp. 2591-2596.

[29] L. Zadeh, *Fuzzy Sets,* Information and Control, vol. 8 (3), 338-353; 1965.

[30] M. Purushothkumar, M. Ananthanarayanan, S. Dhanasekar, "Fuzzy zero suffix Algorithm to solve Fully Fuzzy Transportation Problems," *International Journal of Pure and Applied Mathematics,* vol. 119 (9), 2018, pp. 79-88.

[31] M. Shanmugasundari, K. Ganesan, "A novel approach for the fuzzy optimal solution of fuzzy transportation problem," *International journal of Engineering research and applications,* vol. 3 (1), 2013, pp. 1416-1424.

[32] P. Jayaraman, R. Jahirhussain, "Fuzzy optimal transportation problem by improved zero suffix method via Robust Ranking technique," *International Journal of Fuzzy Mathematics and systems,* vol. 3 (4), 2013, pp. 303-311.

[33] P. Kaur, K. Dahiya, "Two-stage interval time minimization transportation problem with capacity constraints," *Innov Syst Des Eng,* vol. 6, 2015, pp.79–85.

[34] P. Pandian, G. Natarajan, "A new algorithm for finding a fuzzy optimal solution for fuzzy transportation problems," *Applied Mathematical Sciences,* vol. 4, 2010, pp. 79- 90.

[35] R. Antony, S. Savarimuthu, T. Pathinathan, "Method for solving the transportation problem using triangular intuitionistic fuzzy number," *International Journal of Computing Algorithm,* vol. 03, 2014, pp. 590-605.

[36] R. Jahirhussain, P. Jayaraman, "Fuzzy optimal transportation problem by improved zero suffix method via robust rank techniques," *International Journal of Fuzzy Mathematics and Systems (IJFMS),* vol. 3, 2013, pp. 303-311.

[37] R. Jahihussain , P. Jayaraman, "A new method for obtaining an optinal solution for fuzzy transportation problems," *International Journal of Mathematical Archive,* vol. 4 (11), 2013, pp. 256-263.

[38] S. K. Bharati, R. Malhotra, "Two stage intuitionistic fuzzy time minimizing TP based on generalized Zadeh's extension principle," *Int J Syst Assur Eng Manag*, vol. 8, 2017, pp. 1142-1449. DOI: 10.1007/s13198-017-0613-9

[39] S. Chanas, W. Kolodziejckzy, A. Machaj, "A fuzzy approach to the transportation problem," *Fuzzy Sets and Systems,* vol. 13, 1984, pp. 211-221.

[40] S. Dhanasekar, S. Hariharan, P. Sekar, "Fuzzy Hungarian MODI Algorithm to solve fully fuzzy transportation problems," *Int. J. Fuzzy Syst.,* vol. 19 (5), 2017, pp. 1479-1491.

[41] S. Liu, C. Kao, "Solving fuzzy transportation problems based on extension principle," *Eur. J. Oper. Res.,* vol. 153, 2004, pp. 661–674.

[42] S. Midya, SS. K. Roy, V. F. Yu, "Intuitionistic fuzzy multi-stage multi-objective fixed-charge solid transportation problem in a green supply chain," *Int. J. Mach. Learn. & Cyber.,* vol. 12, 2021, pp. 699–717.

[43] S. Malhotra, R. Malhotra, "A polynomial Algorithm for a Two – Stage Time Minimizing Transportation Problem, *OPSEARCH,* vol. 39, 2002, pp. 251–266.

[44] V. Sudhagar, V. Navaneethakumar, "Solving the Multiobjective two stage fuzzy transportation problem by zero suffix method," *Journal of Mathematics Research,*vol. 2 (4), 2010, pp. 135-140.

[45] V. Traneva, "Internal operations over 3-dimensional extended index matrices," *Proceedings of the Jangjeon Mathematical Society,* vol. 18 (4), 2015, pp. 547-569.

[46] V. Traneva, "One application of the index matrices for a solution of a transportation problem," *Advanced Studies in Contemporary Mathematics,* vol. 26 (4), 2016, pp. 703-715.

[47] V. Traneva, P. Marinov, K. Atanassov, "Index matrix interpretations of a new transportation-type problem," *Comptes rendus de l'Academie Bulgare des Sciences,* vol. 69 (10), 2016, pp. 1275-1283.

[48] V. Traneva, S. Tranev, "Intuitionistic Fuzzy Transportation Problem by Zero Point Method," *Proceedings of the 15th Conference on Computer Science and Information Systems (FedCSIS),* Sofia, Bulgaria, 2020, pp. 345–348. DOI: 10.15439/2020F61

[49] V. Traneva, S. Tranev, V. Atanassova, "An Intuitionistic Fuzzy Approach to the Hungarian Algorithm," *in: G. Nikolov et al. (Eds.): NMA 2018,* LNCS 11189, Springer Nature Switzerland, AG, 2019, pp. 1–9. DOI: 10.1007/978-3-030-10692-8_19

[50] V. Traneva, S. Tranev, M. Stoenchev, K. Atanassov, " Scaled aggregation operations over two- and three-dimensional index matrices," *Soft computing,* vol. 22, 2019, pp. 5115-5120. DOI: 10.1007/s00500-018-3315-6

[51] V. Traneva, S. Tranev, *Index Matrices as a Tool for Managerial Decision Making,* Publ. House of the Union of Scientists, Bulgaria; 2017 (in Bulgarian).

[52] V. Traneva, S. Tranev, "An Intuitionistic fuzzy zero suffix method for solving the transportation problem," *in: Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019,* Studies in computational intelligence, Springer, Cham, vol. 902. DOI: 10.1007/978-3-030-55347-0_7, 2020

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 97–102
DOI: 10.15439/2021F14
ISSN 2300-5963 ACSIS, Vol. 26

Flexible job shop scheduling problem with sequence-dependent transportation constraints and setup times

Sacha Varone, David Schindl, Corentin Beffa
University of Applied Sciences Western Switzerland (HES-SO),
HEG Genève, Switzerland
Rue de la Tambourine 17, 1227 Carouge, Switzerland
Email: {sacha.varone, david.schindl}@hesge.ch

***Abstract*—We study a production scheduling problem, which adresses on the one hand the usual operational constraints such as the precedence of operations, time windows, delays, uniqueness of treatment, availability of resources, and waiting times. On the other hand, the problem takes into account possible restricted movements according to production orders. This problem is a variant of a flexible job shop scheduling problem with several types of sequence-dependent constraints. We consider additional sequence-dependent setup times, as well as sequence-dependent transportation and assignment restrictions. We propose a mixed integer programming model (MIP). It is based on the MIP model of a flexible job shop scheduling problem, in which we add those sequence-dependent constraints. We solve it with a general purpose MIP solver.**

I. Introduction

THE metallic pieces production needs several steps in different machines, as warming the metal, soak the pieces and drain them. For a company needing to produce batches of different pieces, the optimal production scheduling is difficult to create, as different pieces need different machine's temperatures. The production line may have different equivalent machines. The choice of one of them may have an impact on the production, as the use of a machine may hinder the displacement between several machines. This problem may be seen as a variation of the well studied job shop scheduling problem (JSP) with additional constraints.

The usual job shop scheduling problem consists in scheduling a set of jobs on machines in order to minimize the makespan, defined as the completion time of the last job. These jobs are composed of different operations to be realized in a determined order, with some known processing times. Only one operation may be realized on a machine at the same time and no preemption is allowed. In the classical Job Shop Scheduling problem (JSP), the operations have to be scheduled on predefined machines, whereas in the Flexible Job Shop Scheduling problem (FJSP), each operation can be carried out on one machine from a set of compatible machines.

A review of the problem with different methods to obtain exact or approximated solutions is presented by Zhang et al. [1], with an analysis of the problem for the requirements of current industries. The result of a comparison of four different models for the JSP problem, realized by Ku and Beck [2], highlights the performances of the disjunctive model. Contrary to the majority of the research, Karimi et al. [3] chose to take into account the transportation time between the machines in the purpose to be closer to the real case. Benttaleb et al. [4] propose a model taking into account the non-availability of the machines that can be caused by maintenance for example. In addition to take into account the periodic maintenance of the machines, Krim et al. [5] model the setup time for a one machine problem. The electrical consumption reduction can save significant costs to companies and reduce the impact on the environment. Therefore, Mansouri et al. [6] propose a multi objective optimization problem, while Assia et al.[7] propose a bi-objective function, to find the trade-off between minimizing the makespan and the total energy consumption.

This article presents an usual mixed integer linear programming model to solve the job shop scheduling problem, with different extensions to handle multi machines re-entrant jobs, time-windows, setup times, hindering movement and dedicated waiting machines. The problem is described in Section II, and its mathematical model in Section III. The model's evaluation is illustrated in Section IV, followed by a conclusion in Section V.

II. Problem description

Our study considers the flexible job shop scheduling problem with sequence-dependent constraints. It is stated as follows, in which we use a similar notation and vocabulary as in [8]. A set J of n jobs has to be scheduled on a set M of m machines. A job j is composed of a sequence of n_j consecutive operations, noted $O_{j1}, \ldots, O_{n_j}$. The l^{th} operation of job j, noted O_{jl}, can be processed on any of the compatible machines from the set $M_{jl} \subset M$. We denote $p_{O_{jl}}$ the processing time of operation O_{jl}, which is supposed to be independent of the machine on which the operation is carried out and no preemption is allowed. A sequence-dependent setup time is incurred between any two consecutive operations carried out on a same machine..

Each machine can perform at most one operation at a time. We suppose that each machine and each job are available

at time 0. The problem is to assign each operation O_{jl} to an eligible machine $k \in M_{jl}$ starting at a time t_{jl}. The objective is to minimize the sum of the starting times of the last operation of each job. This particular objective function limits the waiting time of jobs. This objective function is stricter than the minimization of the makespan $C_{\max}$, which is the time necessary to complete all jobs.

Our real case study is based on the FJSP, to which additional constraints have to be fulfilled. The machines are only available during determined periods of time and between two operations on a same machine, a setup-time is allowed to recondition the machine. Sometimes, the operations may not be processed right after the end of the previous one, so waiting machines are used with waiting operations without determined duration. A waiting time is not systematically allowed. Finally, regarding the topology of the machines, certain displacements are not allowed according to the use of certain machines. An example of topology for metal piece production is presented on Figure 1 with two ovens (O_1 and O_2) and where a drain operation (H'_1 or H'_2) is executed over the soak machine (H_1 or H_2). These drain operations may hinder moving a job between two machines as illustrated on Figure 2. Such operations are called blocking operations.

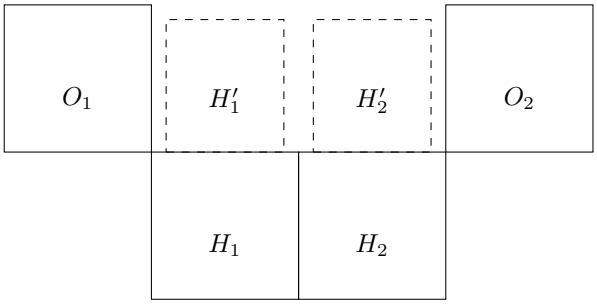


Fig. 1. Example of a machine line topology with two ovens (O_1 and O_2), two soak machines (H_1 and H_2) and two drain machines (H'_1 and H'_2).

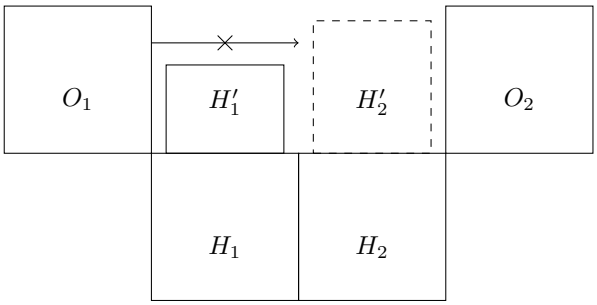


Fig. 2. Example of forbidden movement between the machines O_1 and H_2, induced by the draining operation on the machine H'_1

The notation is presented on Table I. The constant $bigM$ is defined as follows:

$$bigM = \sum_{j \in J} \sum_{l=1}^{|O_j|} p_{O_{j,l}}$$

This is representing the time needed if the recipe of all the jobs are composed of all operations and only one job can be done at a time.

III. Model

A. Variables

a) Usual variables: The proposed model works with a continuous representation of the time. Three principal sets of variables are needed. They express the beginning time of each operation of the recipe of a job on a machine, the assignment of a machine to a job's operation and the precedence between two jobs' operations.

$x_{j,O_{j,l}} \in \mathbb{R}$: start time of the l^{th} operation of job j

$$y_{j,O_{j,l},m} = \begin{cases} 1 & \text{if the } l^{th} \text{ operation of job } j \text{ is carried out on} \\ & \text{machine } m \\ 0 & \textit{otherwise} \end{cases}$$

$$z_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}} = \begin{cases} 1 & \text{if the couple } (O_{j_1,l_1},\, j_1) \text{ occurs before} \\ & \text{the couple } (O_{j_2,l_2},\, j_2) \\ 0 & \textit{otherwise} \end{cases}$$

As a machine may be visited several times during the recipe execution, the predecessor variables take into account the operation in addition to the job.

b) Waiting machines: The proposed model uses waiting shelves, permitting to liberate the machines in the case of waiting time. The idea is to recreate the x, y and z variables dedicated to the waiting operations. To each operation, we associate a waiting operation "prime" with no determined duration. This duration is the difference between the end of the corresponding job's previous operation and the beginning of the next one. Currently, all waiting machines are equivalent and can therefore be used independently of the last job's operation.

$x'_{j,O_{j,l}} \geq 0, \forall j \in J, l \in \{1, ..., |\ O_j\ |\}$

$y'_{j,O_{j,l},m} \in \{0,1\}, \forall j \in J, l \in \{1, ..., |\ O_j\ |\}, m \in StationWaiting$

$z'_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}} \in \{0,1\}, \forall j_1 \in J, l_1 \in \{1, ..., |\ O_{j_1}\ |\}, j_2 \in J, l_2 \in \{1, ..., |\ O_{j_2}\ |\}$

c) Time windows: Machines may be unavailable for different reasons. To take into account these unavailabilities, boolean variables are needed to express which time window is used for each operation's execution.

$tw_{j,O_{j,l},k} \in \{0,1\}$

$\forall j \in J, l \in \{1, ..., |\ O_j\ |\}, k \in TW$ such that $Comp_{m_k,O_{j,l}==1}$

B. Constraints

1) Precedence: to ensure the coherence of the precedence variable, only one of two jobs may be executed before the other:

$z_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}} + z_{j_2,O_{j_2,l_2},j_1,O_{j_1,l_1}} \leq 1$

$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, \forall l_1 \in \{1, ..., |$

TABLE I
TABLE OF NOTATIONS

Notations:	Signification:
$j \in J$:	indices of jobs
$m \in M$:	indices of machines
$w \in TW$:	indices of the time windows
d_j:	deadline of job j
O_{jl}:	l^{th} operation for job j
O'_{jl}:	l^{th} waiting operation for job j
$Tb_{O_{jl}}$:	temperature at the beginning of operation O_{jl}
$Te_{O_{jl}}$:	temperature at the end of operation O_{jl}
$p_{O_{jl}}$:	processing time of operation O_{jl}
$Comp_{m,O_{jl}}$:	set of compatibility with value 1 if operation O_{jl} may be executed on machine m and 0 otherwise.
$prep_{t1,t2,m}$:	preparation time (or setup time) to go from temperature t1 to temperature t2 on machine m.
$bigM$:	constant representing a big number
$hinder_{m_iO_{jl}m_jm_k}$	inconvenience caused by the use of machine m_i for operation O_{jl} on the motion between machines m_j and m_k. A value of 1 is given if such a case occurs and 0 otherwise.
$nodelay_{O_{jl}}$	set of boolean values indicating if a delay is permitted between l^{th} operation and the next one
StationWaiting	set of index of the waiting machines
CycleWaiting	set of indexes of the waiting operations. To each operation O_{jl} correspond a waiting operation O'_{jl}
mc_w	machine corresponding to time window w
$earliest_w$	beginning of time window w
$latest_w$	end of time window w

$O_{j_1} |\}, l_2 \in \{1, ..., | O_{j_2} |\}$
The same idea has to be applied between two operations of a same job:
$z_{j,O_{j,l_1},j,O_{j,l_2}} + z_{j,O_{j,l_2},j,O_{j,l_1}} \leq 1$
$\forall j \in J, \forall (l_1, l_2) \in \{(l_1, l_2) \in \{1, ..., | O_j |\}^2 \mid l_1 \neq l_2\}$

2) Precedence uniqueness on the waiting machines:
$z'_{j_1,O_{j_1,l_1},j_2,O'_{j_2,l_2}} + z'_{j_2,O_{j_2,l_2},j_1,O'_{j_1,l_1}} \leq 1$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, l_1 \in \{1, ..., | O_{j_1} | \}, l_2 \in \{1, ..., | O_{j_2} |\}$

3) Operation order: the operations order in the recipe has to be respected. So the beginning time of an operation has to be larger than the ending time of the previous operation of the same job.
$x_{j,O_{j,l+1}} \geq x_{j,O_{j,l}} + p_{O_{j,l}}$
$\forall j \in J, l \in \{1, ..., | O_j \mid -1\}$
If two operations are realized on the same machine, the preparation time of the second one has to be taken into account. So the starting time of that operation has to be larger than the ending time of the previous one plus the preparation time. If they are not realized on the same machine, the constraint is relaxed due to the "M" term:
$x_{j,O_{j,l+1}} \geq x_{j,O_{j,l}} + p_{O_{j,l}} + prep_{Te_{O_{j,l}},Tb_{O_{j,l+1}},m} - bigM(2 - y_{j,O_{j,l},m} - y_{j,O_{j,l+1},m})$
$\forall j \in J, l \in \{1, ..., | O_j \mid -1\}, m \in \{m \in M \mid Comp_{m,O_{j,l}} = 1, comp_{m,O_{j,l+1}} = 1\})$

4) No conflict on the same machine: two operations O_{j_1,l_1} and O_{j_2,l_2} may not overlap on the same machine, so O_{j_1,l_1} has either to finish before the beginning of O_{j_2,l_2}, or to begin after the end of O_{j_2,l_2}.
$x_{j_1,O_{j_1,l_1}} \geq x_{j_2,O_{j_2,l_2}} + p_{O_{j_2,l_2}} + prep_{Te_{O_{j_2,l_2}}Tb_{O_{j_1,l_1}}m} - bigM * (pred_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}} + \sum_{m' \neq m}(y_{j_1,O_{j_1,l_1},m'} + y_{j_2,O_{j_2,l_2},m'}))$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, \forall l_1 \in \{1, ..., | O_{j_1} |\}, l_2 \in \{1, ..., | O_{j_2} |\}, \forall m \in M$
and:
$x_{j_2,O_{j_2,l_2}} \geq x_{j_1,O_{j_1,l_1}} + p_{O_{j_1,l_1}} + prep_{Te_{O_{j_1,l_1}}Tb_{O_{j_2,l_2}}m} - bigM * ((1 - z_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}}) + \sum_{m' \neq m}(z_{j_1,O_{j_1,l_1},m'} + y_{j_2,O_{j_2,l_2},m'}))$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, \forall l_1 \in \{1, ..., | O_{j_1} |\}, l_2 \in \{1, ..., | O_{j_2} |\}, \forall m \in M$

5) No conflict on a waiting machine: the use of a waiting machine may happen only after or before the use of the same machine by another job.
$x'_{j_1,O_{j_1,l_1}} \geq x_{j_2,O_{j_2,l_2+1}} - bigM * (z'_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}} + (2 - y'_{j_1,O_{j_1,l_1},m} - y'_{j_2,O_{j_2,l_2},m}))$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, l_1 \in \{1, ..., | O_{j_1} | \}, l_2 \in \{1, ..., | O_{j_2} \mid -1\}$
and:
$x'_{j_1,O_{j_1,l_1}} \leq x' j_2, O_{j_2,l_2} + bigM((1 - z'_{j_1,O_{j_1,l_1},p_2,O_{j_2,l_2}} + (2 - y'_{j_1,O_{j_1,l_1},m} - y' j_2, O_{j_2,l_2}, m))$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, l_1 \in \{1, ..., | O_{j_1} | \}, l_2 \in \{1, ..., | O_{j_2} |\}$

6) Respect of compatibilities: an operation may be executed on a machine only if they are compatible.
$y_{j,O_{j,l},m} \leq Comp_{m,O_{j,l}}$
$\forall j \in J, l \in \{1, ..., | O_{j,l} |\}, m \in M$

7) Each operation of a job has to be realised on exactly one compatible machine:
$\sum_{m \in \{m \in M | Comp_{m,O_{j,l}} = 1\}} y_{j,O_{j,l},m} = 1$
$\forall l \in \{1, ..., | O_j |\}, j \in J, m \in M$

None of the other machines are used for the operation, so all other $y_{j,O_{j,l},m}$ variables are set to 0:
$\sum_{m \notin \{m \in M | Comp_{m,O_{j,l}}=1\}} y_{j,O_{j,l},m} = 0$
$\forall l \in \{1, ..., \mid O_j \mid\}, j \in J, m \in M$

8) At most one waiting machine per operation and job:
$\sum_{m \in machineWaiting} y'_{j,O_{j,l},m} \leq 1$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}$

9) No delay: if the nodelay value is set to true, the beginning of the following operation has to be the same as the ending time of the previous operation.
$\sum_{(j,O_{j,l}) \in NDL} (x_{j,O_{j,l+1}} - (x_{j,O_{j,l}} + p_{O_{j,l}})) = 0$
Where NDL represents the set of all the couples $(J \times O)$ preceding the operation where the nodelay value is set to true in the associated recipe.

10) Required succession of machines: sometimes, the use of a certain machine for an operation implies the use of the same machine for the next operation.
$y_{j,O_{j,l},m} = y_{j,O_{j,l+1},m}$
$\forall j \in J, l \in \{l \in \{1, ..., \mid O_j \mid\} \mid nodelay_{O_{jl}=1}\}, m \in M$

11) Impossible machine successions: regarding the disposition of the machines, some of them may not be reached from another machine.
$y_{j,O_{jl},m_1} + y_{j,O_{jl+1},m_2} = 0$
$\forall j \in J, \forall l \in \{1, ..., \mid O_j \mid -1\}, \forall (m_1, m_2) \in incompatibleSuccession$
Where incompatibleSuccession is the set of all ordered pairs of incompatible machine successions.

12) Blocking operation: if a blocking operation is running, the concerned movement has to be realized either before the beginning of the blocking operation, or after the end of the blocking operation.
$x_{j_1,O_{j_1,l_1}} \geq (x_{j_2,O_{j_2,l_2}} + p_{O_{j_2,l_2}}) * hinder_{m_i,O_{j_2,l_2},m_j,m_k} - bigM * (3 - y_{j_2,O_{j_2,l_2},m_i} - y_{j_1,O_{j_1,l_1+1},m_j} - y_{j_1,O_{j_1,l_1},m_k}) - bigM * z_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}}$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, \forall l_1 \in \{1, ..., \mid O_{j_1} \mid -1\}, l_2 \in \{1, ..., \mid O_{j_2} \mid\}, \forall m_i, m_j, m_k \in M$
and:
$x_{j_2,O_{j_2,l_2}} \geq x_{j_1,O_{j_1,l_1}} * hinder_{m_i,O_{j_2,l_2},m_j,m_k} - bigM(3 - y_{j_2,O_{j_2,l_2},m_i} - y_{j_1,O_{j_1,l_1+1},m_j} - y_{j_1,O_{j_1,l_1},m_k}) - bigM * (1 - z_{j_1,O_{j_1,l_1},j_2,O_{j_2,l_2}})$
$\forall (j_1, j_2) \in \{(j_1, j_2) \in J^2 \mid j_1 \neq j_2\}, \forall l_1 \in \{1, ..., \mid O_{j_1} \mid -1\}, l_2 \in \{1, ..., \mid O_{j_2} \mid\}, \forall m_i, m_j, m_k \in M$

13) Deadline: each job has to be completed before the deadline.
$x_{j,O_{j,end}} + p_{O_{j,end}} \leq d_j$
$\forall j \in J$

14) Relation between the x and x' variables:
$x'_{j,O_{j,l}} = x_{j,O_{j,l}} + p_{O_{j,l}}$
$\forall x \in J, l \in \{1, ..., \mid O_{j,l} \mid\}$

15) If there is a waiting time between two operations, a waiting machine has to be used:
$bigM * \sum_{m \in machineWaiting} y'_{j,O_{j,l},m} \geq x_{j,O_{j,l+1}} - (x_{j,O_{j,l}} + p_{O_{j,l}})$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}$
and:
$\sum_{m \in machineWaiting} y'_{j,O_{j,l},m} \leq bigM * (x_{j,O_{j,l+1}} - (x_{j,O_{j,l}} + p_{O_{j,l}}))$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}$

16) Only one time window per couple operation job is allowed:
$\sum_{k \in \{w \in TW \mid \exists m' \in mc_w, m' == m\}} tw_{j,O_{jl},k} = y_{j,O_{jl},m}$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}, m \in M$

17) Each operation cannot start earlier than the beginning of the time window:
$\sum_{k \in TW} tw_{j,O_{jl},k} * earliest_k \leq x_{j,O_{jl}}$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}$

18) Each operation has to be completed before the end of the time window:
$\sum_{k \in TW} tw_{j,O_{jl},k} * latest_k \geq x_{j,O_{jl}} + p_{O_{jl}}$
$\forall j \in J, l \in \{1, ..., \mid O_j \mid\}$

C. Objective function:

Contrary to the usual job shop problem, where the objective is to minimize the makespan, the proposed objective to minimize is the sum of the starting times of the last operation of each job:

$$\min \sum_{j \in J} x_{j,O_{j,end}}$$

This is used in the purpose to minimize the production time of each job and limit the unnecessary waiting time of the jobs having no influence on the makespan.

IV. Evaluation

The proposed model has been implemented in the Julia programming language [9], [10] using the Gurobi solver [11]. Gurobi is considered as one of the state of the art commercial solver [2].

A. Numerical experiments

The experiments were conducted on three types of data: data generated specifically according to the constraint tested, randomly generated, as well as real data. The real data made it possible to ensure that the model proposed fits the real problem. The data presented here is an example, among the tests performed on real data. Four jobs have to be realized on the machines. The topology of the machines is shown in Figure 3, with the possible operations on each machine. Furthermore, five waiting machines, not represented in Figure

TABLE II
RECIPE FOR EACH JOB

job id	recipe (operation id)
1	1, 4, 4', 13, 9, 10
2	2, 5, 5', 14, 9, 11
3	7, 3, 6, 6', 15, 9, 11
4	12, 13, 8

TABLE III
INCOMPATIBILITY

machine id	machine id
D_1	E_3
D_3	E_1
D_3	E_2

3, are available. The manipulator is used for all displacements between machines that are not provided by the moving oven themselves. In order to maintain a reasonable computing time, the utilization of the manipulator has been neglected. The arrows on the figure symbolize the moving machines. Table II presents the recipes, which are defined as successions of operations to be accomplished for a given job. Data about the time-windows and the preparation times are not given here, but a full data set may be provided on request.

According to the arrangement of the machines, some displacements are not possible, whatever operations are carried out on them. This sequence of machine utilization is stored in Table III. Sometimes, the accessibility of a machine from another one is hindered by a specific operation on a third machine. These cases are presented in Table IV.

B. Solution

The solution returned by the model is presented as a Gantt chart, shown in Figure 4. The resolution time comprising the creation of the model and the resolution of the problem is less than five minutes. No better solution can be found since all jobs begin at time 0 and there are no waiting times between operations.

V. CONCLUSION

This study is a first part of a real-world industrial project. This production scheduling problem has the goal to minimize the production time of different jobs by optimizing the assignment of their operations on the machines, subject to several constraints. These constraints include limited availability of the machines, setup times for reconditioning of the machines, dynamic forbidden displacements and waiting machines. As a future work, manipulator constraints have to be integrated, and the objective function will also consider the monetary cost minimization, taking into account electricity tariffs. The use of heuristics will help in the design of a good real-time solution, able to adapt the solution to events occurring each minute, such as a breakdown of a machine or the arrival of new jobs.

TABLE IV
HINDERS

machineHinders	operationHinders	machineOrigin	machineDestination
E_1	4'	D_1	E_2
E_1	5'	D_1	E_2
E_1	6'	D_1	E_2
E_2	4'	D_2	E_1
E_2	5'	D_2	E_1
E_2	6'	D_2	E_1

REFERENCES

[1] J. Zhang, G. Ding, Y. Zou, S. Qin, and J. Fu, "Review of job shop scheduling research and its new perspectives under Industry 4.0," *Journal of Intelligent Manufacturing*, vol. 30, no. 4, pp. 1809–1830, Apr. 2019. doi: 10.1007/s10845-017-1350-2. [Online]. Available: https://doi.org/10.1007/s10845-017-1350-2

[2] W.-Y. Ku and J. C. Beck, "Mixed Integer Programming models for job shop scheduling: A computational analysis," *Computers & Operations Research*, vol. 73, pp. 165–173, Sep. 2016. doi: 10.1016/j.cor.2016.04.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054816300764

[3] S. Karimi, Z. Ardalan, B. Naderi, and M. Mohammadi, "Scheduling flexible job-shops with transportation times: Mathematical models and a hybrid imperialist competitive algorithm," *Applied Mathematical Modelling*, vol. 41, pp. 667–682, Jan. 2017. doi: 10.1016/j.apm.2016.09.022. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0307904X16304929

[4] M. Benttaleb, F. Hnaien, and F. Yalaoui, "Two-machine job shop problem under availability constraints on one machine: Makespan minimization," *Computers & Industrial Engineering*, vol. 117, pp. 138–151, Mar. 2018. doi: 10.1016/j.cie.2018.01.028. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835218300354

[5] H. Krim, R. Benmansour, and D. Duvivier, "On the single machine scheduling problem with sequence-dependent setup times and periodic maintenance," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*, Apr. 2018. doi: 10.1109/CoDIT.2018.8394827 pp. 374–378.

[6] S. A. Mansouri, E. Aktas, and U. Besikci, "Green scheduling of a two-machine flowshop: Trade-off between makespan and energy consumption," *European Journal of Operational Research*, vol. 248, no. 3, pp. 772–788, Feb. 2016. doi: 10.1016/j.ejor.2015.08.064. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221715008206

[7] S. Assia, E. A. Ikram, A. El Barkany, and E. Biyaali Ahmed, "Two-Machine Job Shop Scheduling Problem Under Availability Constraints on one machine: Total Energy Consumption (TEC) minimization," 2018.

[8] L. Shen, S. Dauzère-Pérès, and J. S. Neufeld, "Solving the flexible job shop scheduling problem with sequence-dependent setup times," *European Journal of Operational Research*, vol. 265, no. 2, pp. 503–516, Mar. 2018. doi: 10.1016/j.ejor.2017.08.021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037722171730752X

[9] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017. doi: 10.1137/141000671. [Online]. Available: http://julialang.org/publications/julia-fresh-approach-BEKS.pdf

[10] [Online]. Available: https://julialang.org

[11] [Online]. Available: http://www.gurobi.com/

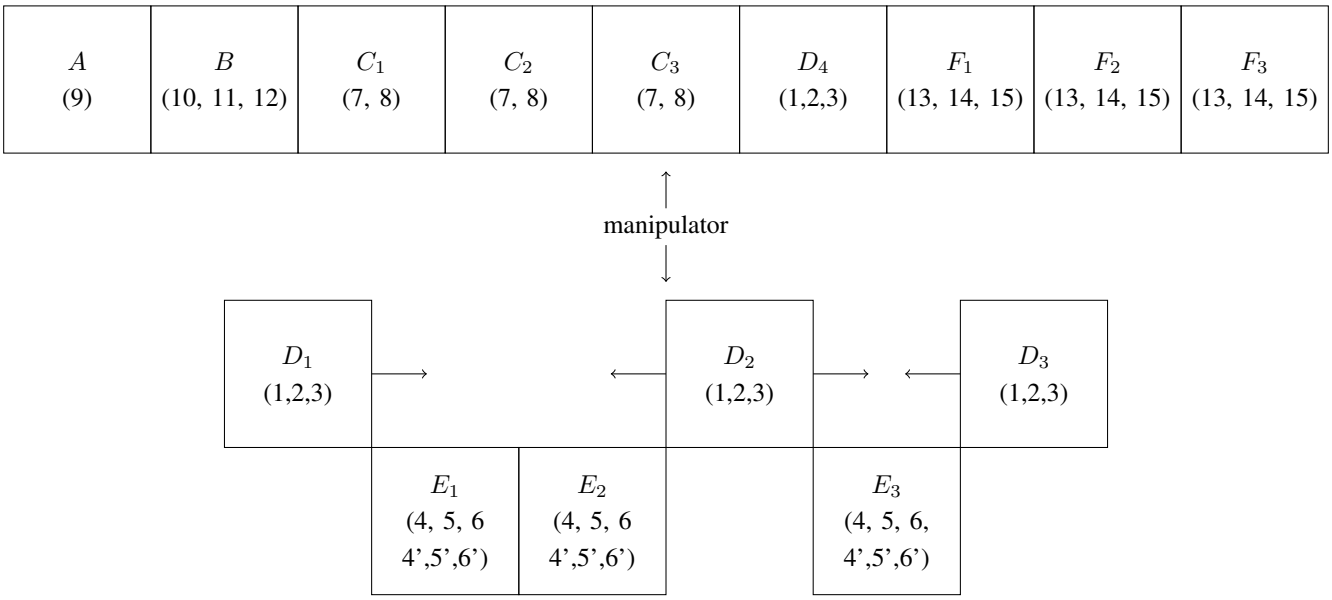


Fig. 3. Machine line topology with the name of each machine and in parenthesis the cycle that can be done on each machine (compatibility).

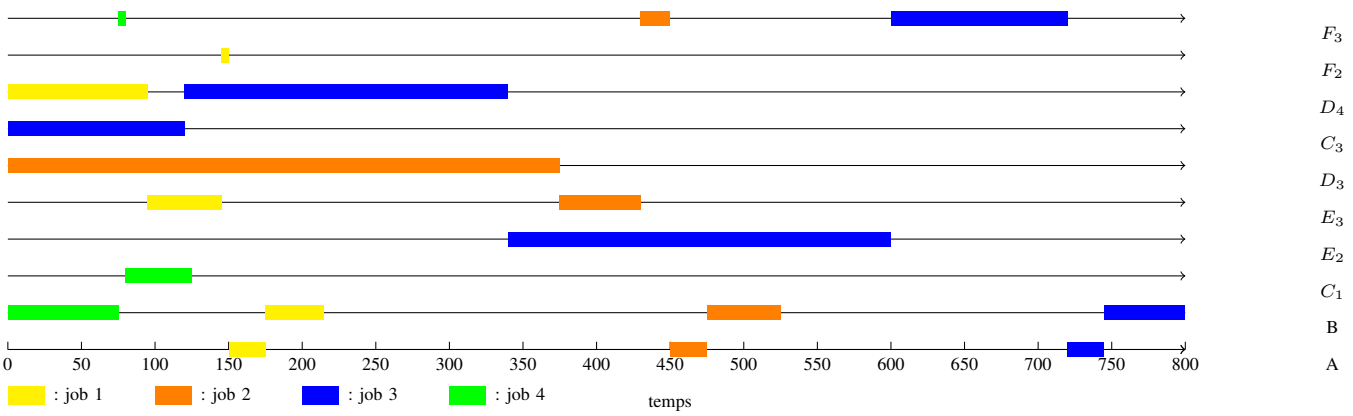


Fig. 4. time table representing the solution

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 103–111
DOI: 10.15439/2021F87
ISSN 2300-5963 ACSIS, Vol. 26

Alternatives for greedy discrete subsampling: various approaches including cluster subsampling of COVID-19 data with no response variable

Lubomír Štěpánek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

Filip Habarta
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
filip.habarta@vse.cz

Ivana Malá
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
malai@vse.cz

Luboš Marek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
marek@vse.cz

***Abstract*—An exhaustive selection of all possible combinations of $n = 400$ from $N = 698$ observations of the COVID-19 dataset was used as a benchmark. Building a random set of subsamples and choosing the one that minimized an averaged sum of squares of each variable's category frequency returned similar results as a "forward" subselection reducing the dataset one-by-one observation by the same metric's permanent lowering. That works similarly as k-means clustering (with a random clusters' number) over the original dataset's observations and choosing a subsample from each cluster proportionally to its size. However, the approaches differ significantly in asymptotic time complexity.**

I. Introduction

SUBSAMPLING is a method that reduces a size of a dataset by selecting a subset from the original dataset. However, in many areas, including biomedicine and many others, we often face a kind of opposite problem, i. e. we obtain a sample of only insufficient size and would need to enlarge its size. That can be done, e. g., by one of the resampling methods such as bootstrapping or others, or we need to use various inference methods to estimate properties of the entire population that our dataset comes from.

While such a data size reduction could not sound meaningful for the first impression, there are various situations where subsampling makes sense or is even necessary.

Usually, we can distinguish between two kinds of subsampling. Firstly, when we do the subsampling, we cannot even in theory collect all possible observations of an entire population. Or, secondly, we can gain all possible observations or, furthermore, we have already got them, but for some reason, we have to reduce the number of observations that will be utilized.

A typical example of the first subsampling kind is one of the large fields of statistics, called *sampling*, where subsampling as a method of choice deals with an idea of an entire population and its parameters but is limited to an option of gaining data of only a (small) subset coming from the population. Then, regardless of whether the population is more or less virtual, getting the sample that belongs to the population is still a problem fulfilling the subsampling definition.

The motivations for the subsampling could also be different and usually arise from any impossibility to utilize the entire original dataset, as may be true for the latter family of the subsampling problems. Thus, the rank of those motivations varies from the lack of (computational) power to analyze the entire original dataset to the lack of economic sources, making it impossible to collect all values for each observation of the original sample, e. g. populating a new (important) variable is considered to enrich the original dataset but can be done only for a limited number of observations (of the sub-selected dataset).

As a motivation for our study, using online surveys, we collected an original dataset of patients suffering from COVID-19 and undergoing anti-COVID-19 vaccination. To study a time development of COVID-19 antibodies after the vaccination, it is necessary to check the blood levels of the patients' antibodies from time to time. However, no matter how helpful would be the checking of antibodies for each patient, our financial sources were limited (and the antibody kits for laboratory serology tests are relatively expensive), so we had to select a subset of patients from the original dataset, no greater than a maximal number of laboratory tests funded by our financial sources. Furthermore, since the subsample can be done in many ways, we wanted to keep all categories of all categorical variables well balanced, i. e., to keep their frequencies in the final subsample equal or at least near-equal.

All the motivations share the demand on the quality by which the subsampling is done. As is naturally feasible, we usually want to avoid the "garbage in, garbage out", also known as the GIGO paradigm, which means that we cannot expect great outputs whenever the inputs are of low quality. The same logic applies to subsampling if followed by whatever kind of another analysis uses the subsample as an input. Thus, the authors suggest replacing the "garbage in, garbage out" paradigm more positively with "great in, great out".

However, regardless of the primary motivation why do subsampling, there is always a demand to keep the data homogeneity in the sub-selected sample, corresponding to the original data. More technically spoken, assuming the dataset contains only categorical variables, the homogeneity means that all categories of all categorical variables are near-equally represented in the final subsample.

In case there is a response variable included in the dataset, a popular and well-established method called propensity scoring (or propensity matching) is usually performed to identify the "best" subset of a given size that harmonizes effect sizes of individual explanatory variables [1].

Nevertheless, when a response variable is missing in the data because e. g. is planned to measure its values rather only for observations in the subsample than for the entire original sample, the logistic regression model behind the propensity scoring could not be built at all (since the response variable is not available). In such a case, the methodology that could be used for subsampling differ from naive approaches such as random sampling, even-odd sampling [2], to more intuitive, rather manual than automated sampling based on matching the observations so that they are balanced in pairs (or larger groups than pairs) [3]. In other words, when a response variable, commonly participating as a key part of the subsampling quality checking, is not available in the dataset, it could be "substituted" by a metric that might control for the quality of the subsampling process.

To check how balanced the subsample is, some metrics could be used [4]. They usually assume that numerical variables – if any – were prior transformed to categorical ones following more or less complex categorization rule. There are several commonly used metrics describing the rate of the categorical variables' levels balance in a final sample [5] such as entropy, mutability, Gini impurity, Simpson index, Shannon-Wiener index, and other diversity metrics. A sum of squares of categories' frequencies also becomes very popular; it is somewhat similar to Shannon entropy but is scaled, so it cannot be greater than 1.0 at maximum.

Based on the metric choice, the lower (or, the higher) is the metric's value; the better balanced is the subsample. Thus, for example, considering Shannon entropy or sum of squares of categories' frequencies, a lower value means better balancing the subsample; i. e. the frequencies of the categories in the subsample are equal or at least near-equal.

In fact, the subsampling itself is a discrete optimization task since the selection of a final subsample from the original sample may be made using a finite number of ways, but some of them are better than others, taking into account there is a given metric, checking the subsampling quality (categorical variables' levels well balancing) that is about to be minimized.

In this study, we selected a subpopulation ($n = 400$) from a COVID-19 dataset (or original size $N = 698$) with a missing response variable, which was up to be collected later. Whereas the response variable was not available, there were 18 more (explanatory) variables of interest. First, numerical variables were categorized. The quality of subpopulation selecting was measured using a sum of squares of each variable's category frequency and averaged over all variables. Minimizing the metric reflects the demand for keeping all the variables' categories numerically balanced, i. e. of similar sizes. Several subset-selecting strategies were applied. Besides a single random subsampling, an exhaustive method selecting all possible combinations of $n = 400$ observations from initial $N = 698$ observations was performed, choosing the subsample that grand totally minimized the metric. Similarly, a "forward" subselection, reducing the original dataset by one observation per each step, permanently lowering the metric, was done. A repeated random subsampling enabled to model a prior distribution of the metric and helped estimate its empirical minimum, determining one given subsample. Finally, k-means clustering (with a random number of clusters) of the original dataset's observations and choosing for a subsample from each cluster, proportionally to its size, and also based on a joint occurrence of each pair in one cluster, also lowered the metric compared to the random subsampling.

The aim of this study is to demonstrate that all the approaches except for a single random subsample offer a valid alternative to exhaustive sampling grant-totally minimizing the chosen metric.

II. Proposed research methodology

There are overall research methodology and the formal description of the dataset, the metric chosen for controlling the quality of the subsampling, and the proposed methods of the subsampling discussed in the following subsections.

A. Formal description of a dataset used for subsampling

The original dataset consists of N rows containing one observation per row and k categorical variable in columns.

The subsampling task means selecting a subset of n rows and k columns, where $n < N$. Thus, the sampling is applied on rows, not on columns.

For each $i \in \{1, 2, 3, \ldots, k\}$, the variable i contains exactly n_i categories and the frequencies of the category j is $n_{i,j}$. We can easily show that for the original dataset, and the subsample is

$$\sum_{j=1}^{n_i} n_{i,j} = N \qquad \text{and} \qquad \sum_{j=1}^{n_i} n_{i,j} = n,$$

respectively, so the sum of frequencies of a given variable i's categories is equal to N in the original dataset and is equal to n in the subsample, respectively, and based on the context.

B. A metric for controlling the quality of the subsampling

The Shannon entropy is defined as

$$H_i = -\sum_{j=1}^{n_i} p_{i,j} \log p_{i,j}$$

where $p_{i,j}$ is a probability of a category j for each $j \in \{1, 2, 3, \ldots, n_i\}$ in a sample of n_i categories of a variable i. We can easily prove by Jensen's inequality the upper bound of the entropy H_i defined by such formula is dependent on the probabilities $p_{i,j}$. Also, the formula may struggle with zero probabilities, i. e. when $\exists j \in \{1, 2, 3, \ldots, n_i\}$ such that $p_{i,j} = 0$ since the term $\log p_{i,j}$ is not defined for $p_{i,j} = 0$.

To overcome these difficulties, we rather used a sum of squares of each variable's category frequency. Using the finite samples, probabilities are only estimated by their frequencies, therefore we will replace the probability $p_{i,j}$ by its unbiased estimate $\pi_{i,j} = \frac{n_{i,j}}{n} = \hat{p}_{i,j}$, where $n_{i,j}$ is a number of occurrence of category j of variable i in the sample of size n. The sum of squares of the variable i's category frequencies then follows as

$$S_i = \sum_{j=1}^{n_i} \pi_{i,j}^2 = \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \sum_{j=1}^{n_i} \hat{p}_{i,j}^2. \tag{1}$$

Finally, when there is more than one variable, i. e. $i \in \{1, 2, 3, \ldots, k\}$ then in order to take into account for each variable's sum of squares given by formula (1), we can calculate the average value $\bar{S}$ of the sums of squares for individual variables, so

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2. \tag{2}$$

Let us derive the lower and upper bound of the sum of squares for the variable i.

(i) Firstly, let us consider one of the two possible extreme scenarios – the sample is populated by only one category. More technically, let us assume that $\exists j^* \in \{1, 2, 3, \ldots, n_i\}$ so that $n_{i,j^*} = n_i$. Then, $\forall j \in \{1, 2, 3, \ldots, n_i\} \setminus j^*$ is $n_{i,j} = 0$ and, eventually, $\frac{n_{i,j^*}}{n} = 1$ and $\frac{n_{i,j}}{n} = 0$.

The sum of squares S_i then follows the term

$$\begin{aligned} S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \\ &= \left(\frac{n_{i,j^*}}{n}\right)^2 + \sum_{j \in \{1,2,\ldots,n_i\} \setminus j^*} \left(\frac{n_{i,j}}{n}\right)^2 = \\ &= 1^2 + (n_i - 1) \cdot 0^2 = \\ &= 1. \end{aligned}$$

Thus, we derived the maximum value of the sum of squares S_i for the variable i is equal to 1.

(ii) Now suppose the other extreme scenario – all categories are equally populated in the sample and no one of the categories occurred more than once. So, in other words,

$$\frac{n_{i,1}}{n} = \frac{n_{i,2}}{n} = \cdots = \frac{n_{i,n_i}}{n} = \frac{1}{n}$$

and also $n_i = n$.

The sum of squares S_i then follows as

$$\begin{aligned} S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 = \sum_{j=1}^{n_i} \left(\frac{1}{n}\right)^2 = \\ &= \sum_{j=1}^{n} \left(\frac{1}{n}\right)^2 = \\ &= n \cdot \left(\frac{1}{n}\right)^2 = \\ &= \frac{1}{n}. \end{aligned}$$

So, we derived the minimum value of the sum of squares S_i for the variable i is equal to $\frac{1}{n}$, where n is the size of a sample containing only categories of the variable i.

Concluding this up, we derived that for each variable i and its sample size n is the sum of squares S_i of the variable's category frequencies lower then or equal to 1 and greater than or equal to $\frac{1}{n}$, more formally $\frac{1}{n} \leq S_i \leq 1$.

Going back to the idea of a well-balanced subsample, all category frequencies of all variables in the subsample should be of (near) equal sizes. That is a situation very close to scenario (ii) with balanced frequencies $\frac{1}{n}$ above – on the other hand, the frequencies in scenario (i) are imbalanced. Assuming this, the sum of squares S_i of the variable's category frequencies in the well-balanced subsample should be as low as possible and should approach the $\frac{1}{n}$. Finally, if all the variable would minimize their sums of squares, then also the average value $\bar{S}$ of all the sums of squares should be minimal.

Keeping the subsample well balanced, i. e. ensuring the categories of all the variables in the subsample are of (near) equal frequencies, means lowering the average value $\bar{S}$ of the sums of squares as much as possible. In theory, the minimal possible value of the average value $\bar{S}$ of the sums of squares is

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 \geq \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n} = \frac{1}{k} \cdot \frac{k}{n} = \frac{1}{n}.$$

In practise, assuming the categories are well balanced for each variable, i. e. for each $i \in \{1, 2, 3, \ldots, k\}$ is $n_{i,1} \approx n_{i,2} \approx \cdots \approx n_{i,n_i}$, then $\sum_{j=1}^{n_i} n_{i,j} = n \approx n_i \cdot n_{i,j}$ and so $\frac{n_{i,j}}{n} \approx \frac{n/n_i}{n} \approx \frac{1}{n_i}$, we can expect rather

$$\bar{S} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{n_{i,j}}{n}\right)^2 \gtrsim \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{1}{n_i}\right)^2 \approx$$
$$\approx \frac{1}{k}\sum_{i=1}^{k} n_i\left(\frac{1}{n_i}\right)^2 \approx \frac{1}{k}\sum_{i=1}^{k}\frac{n_i}{n_i^2} \approx \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}.$$

Eventually, what worth to be mentioned, is a comparison of each variable's sum of squares S_i given by formula (1) and Gini impurity. Using still the same mathematical notation, then Gini impurity is defined as

$$G_i = 1 - \sum_{j=1}^{n_i}\pi_{i,j}^2 = 1 - \sum_{j=1}^{n_i}\left(\frac{n_{i,j}}{n}\right)^2 = 1 - \sum_{j=1}^{n_i}\hat{p}_{i,j}^2,$$

which is obviously equal to $S_i = 1 - G_i$. That being written, using the Gini impurity G_i in this study instead of the sum of squares S_i would return exactly the same results (as far as the sign of Gini impurity is opposite than the one of the sum of squares and shifted by 1.0).

C. Single random subsampling without replacement

The term of random subsampling without replacement means that each observation of the original dataset has only one chance to be selected in the subsample.

If we subsample the original dataset of size N to a dataset of size n only once, there are in theory $\binom{N}{n}$ options how to do the random subsampling. Assuming one of the subsamples[1] minimizing the averaged sums of squares $\bar{S}$, the probability of randomly hitting such a subsample is about $\frac{1}{\binom{N}{n}} \simeq 0$ for large $N > n$.

An expected value of the averaged sums of squares $\bar{S}$, calculated using the obtained subsample, is in between the expected value of the worst-case scenario, 1, and the best-case scenario, $\frac{1}{n}$, so $\frac{1}{n} \leq \mathbb{E}(\bar{S}) \leq 1$.

The asymptotic time complexity is easy to derive, $\Theta(1)$, assuming the random subset generating costs 1 unit of complexity time.

D. Repeated random subsampling without replacement

Similarly to the previous approach, here we repeat the random subsampling $m > 1$ times.

The repetition of the random subsampling enables us to estimate an expected value $\hat{\mathbb{E}}(\bar{S})$ of the averaged sums of squares $\bar{S}$ and standard deviation $\sqrt{\hat{\text{var}}(\bar{S})}$, using the values of m obtained subsamples. Assuming the Ljapunov's version of the central limit theorem, the derived variable $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}}$ follows standard normal distribution, formally $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \sim \mathcal{N}(0, 1^2)$. That helps us to estimate the minimum value of the averaged sums of squares $\bar{S}$ following way. Supposing there $\frac{\bar{S}-\hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \sim \mathcal{N}(0, 1^2)$ holds, we know that

$$P\left(\frac{\bar{S} - \hat{\mathbb{E}}(\bar{S})}{\sqrt{\hat{\text{var}}(\bar{S})}} \leq u_{0.01}\right) = 0.01,$$

where $u_{0.01}$ is a 0.01-th quantile of the standard normal distribution. Continuing in the derivations, we get

$$P\left(\bar{S} \leq \hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}\right) = 0.01, \tag{3}$$

so approximately, the minimum value of $\bar{S}$ is very likely close to the term of $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$. Utilizing this piece of information, we can not only estimate the minimum value of the averaged sums of squares $\bar{S}$, but can also highlight the subsample approaching this minimum value (surely it is the subsample with minimal value – somewhat close to the subtraction $\hat{\mathbb{E}}(\bar{S}) - |u_{0.01}|\sqrt{\hat{\text{var}}(\bar{S})}$ from the positive direction – of the averaged sums of squares $\bar{S}$ in the set of all m generated subsamples).

The asymptotic time complexity of the (m times) repeated random subsampling without replacement is $\Theta(m)$, again assuming the random subset generating costs 1 unit of complexity time. The pseudocode of the repeated random subsampling process is in Algorithm 1.

E. Exhaustive subsampling

The method of exhaustive subsampling is based on greedy generating all possible subsamples of size n from the original dataset of size $N > n$.

In theory, there are $\binom{N}{n} = \frac{n!}{k!(n-k)!}$ ways how a subsample of size n could be sampled from the dataset of size N. It implies there is also $\binom{N}{n}$ values of the averaged sums of squares $\bar{S}$ (one value per each subsample), but the values are not necessarily different.

Regardless of that, this approach enables to convenient pick the subsample with a minimum possible value of the averaged sums of squares $\bar{S}$ (no other subsample could practically have the value of the averaged sums of squares $\bar{S}$ lower).

However, there is an obvious trade-off between the possibility to reach the practical minimum of the value of the averaged sums of squares $\bar{S}$ and asymptotic time complexity, which is enormous, $\Theta\left(\binom{N}{n}\right) = \Theta\left(\frac{n!}{k!(n-k)!}\right)$, assuming the random subset generating costs 1 unit of complexity time.

F. Subsampling by forwarding step-by-step size reduction of the original dataset

The logic of the step-by-step size reduction of the original dataset by permanent lowering a value of the averaged sums of squares $\bar{S}$ is based on random selection of such an observation that its removing from the original dataset tends to decrease (or at least not increase) a value of the averaged sums of squares $\bar{S}$. Thus, we also call this approach as *one-by-one* observation's sample reduction of also as *row-by-row* observation's sample reduction. Let's define a size of the dataset after τ steps, i. e. after removing of τ observations, as $n(\tau)$,

[1]Theoretically, there could be more than one subsample with the same but minimal value of the metric of the averaged sums of squares $\bar{S}$.

Algorithm 1: Repeated random subsampling without replacement and estimating of the minimum value of the averaged sums of squares $\bar{S}$, together with highlighting of the subsample minimizing the averaged sums of squares $\bar{S}$

Data: an original dataset of size N containing k variables
Result: a set of m random subsamples of size $n < N$, an estimate of the minimum value of the averaged sums of squares $\bar{S}$ and highlighting of the subsample minimizing the averaged sums of squares $\bar{S}$

```
N   // size of the original dataset ;
n   // size of the subsample;
m   // number of repetitions of;
    // subsampling;
S   // a tuple of subsamples of size
   n;
A   // a tuple of averaged sums of
   squares S̄;
for ℓ = 1 : m do
    generate a random subsample ∫ of size n without
     replacement from the original dataset of size N
     and calculate its averaged sums of squares S̄ ;
    S = {S, ∫};
    A = {A, S̄};
end
find the minimum of A and a corresponding
 subsample with S̄ = min{A} ;
calculate an estimate Ê(S̄) and vâr(S̄) ;
calculate the estimated minimum of S̄ as
 Ê(S̄) − |u_0.01| √vâr(S̄) ;
compare min{A} and Ê(S̄) − |u_0.01| √vâr(S̄) ;
```

and the averaged sums of squares $\bar{S}$ after τ steps as $\bar{S}(\tau)$. Evidently, $n(0) = N$, $n(1) = N - 1$, $n(2) = N - 2$, …, $n(N - n) = N - (N - n) = n$. Analogously, we demand on $\bar{S}(\tau + 1) \leq \bar{S}(\tau)$ for each $\tau \in \{0, 1, 2, \ldots, N - n - 1\}$.

It is easy to demonstrate that $\bar{S}(N - n) \leq \bar{S}(0)$, i. e. the averaged sums of squares $\bar{S}$ after $N - n$ steps (when dataset size is n) is lower than or equal to the value of the averaged sums of squares $\bar{S}$ in the beginning. Assuming the initial original dataset is not well balanced, then $\bar{S}(N - n) < \bar{S}(0)$ or even $\bar{S}(N - n) \ll \bar{S}(0)$. Based on the fact the selection of one observation per each step is random (until it leads to decreasing of the averaged sums of squares $\bar{S}$ value), a deterministic value of $\bar{S}(N - n)$ is not possible to calculate.

Let us suppose the random selection of the observation tending to reduce the averaged sums of squares $\bar{S}(\tau)$ in the $(\tau + 1)$-th step (so that $\bar{S}(\tau + 1) \leq \bar{S}(\tau)$), when the dataset contains exactly $N - \tau$ observations, would take averagely about $(N - \tau)/2$ samplings. Then the average asymptotic time complexity [6] of the row-by-row reduction of the original dataset by permanent lowering a value of the averaged sums of squares $\bar{S}$ is $\Theta(\bullet)$, so that

$$\begin{aligned}
\Theta(\bullet) &= \Theta\left(\sum_{\tau=0}^{N-n-1} (N-\tau)/2\right) = \\
&= \Theta\left(\frac{1}{2}\sum_{\tau=0}^{N-n-1} (N-\tau)\right) = \\
&= \Theta\left(\frac{1}{2}\left(\sum_{\tau=0}^{N-n-1} N - \sum_{\tau=0}^{N-n-1} \tau\right)\right) = \\
&= \Theta\left(\frac{1}{2}\left((N-n)N - \frac{(N-n-1)(N-n)}{2}\right)\right) = \\
&= \Theta\left(\frac{1}{2}\left(\frac{(N-n)(N+n+1)}{2}\right)\right) = \\
&= \Theta\left((N-n)(N+n+1)\right) \approx \\
&\approx \Theta(N^2).
\end{aligned}$$

The pseudocode of the subsampling by row-by-row reduction of the original dataset is in Algorithm 2.

Algorithm 2: Subsampling by row-by-row reduction of the original dataset, decreasing the value of the averaged sums of squares $\bar{S}$ per each step

Data: an original dataset of size N containing k variables
Result: a subsample minimizing the averaged sums of squares $\bar{S}$

```
N   // size of the original dataset;
n   // size of the subsample;
n_t // current size of the dataset;
S̄   // current averaged sums of
   squares;
n_t = N;
while n_t > n do
    while S̄ after removing the random observation
     ≥ S̄ do
        pick another random observation from the
         current dataset of size n_t (# of observations)
    end
    remove the picked observation from the dataset;
    n_t = n_t − 1;
    update S̄;
end
use the subsample of size n;
```

G. Subsampling using clustering

An idea behind the subsampling using unsupervised learning of clustering kind is to utilize the fact that observations within each cluster are similar enough, while observations between each cluster are different enough. Thus, when we require subsamples with well-balanced category frequencies for each variable, we should consider observations from different clusters when creating the final subsample. Thus, a big

question is *how* to pick the observations from different clusters to ensure the final subsample of a given size is well balanced.

The paper's authors suggest several ideas on how to use clusters for subsampling and, particularly, how to draw the observations from existing clusters when the final subsample is constructed.

Firstly, regardless of the fact the observations are picked randomly or following some pattern from d clusters of sizes $|c_1|$, $|c_2|$, ..., $|c_d|$, a number of observations picked from the cluster $\delta \in \{1, 2, \ldots, d\}$ should be proportional to its size, $|c_\delta|$.

Let us assume that a number of category frequencies of a variable i that are greater than zero is η_i in a given cluster δ. A total count of categories of a variable i is, following the previous notation, n_i, and a mean frequency for average category is about $\frac{|c_\delta|}{n_i}$. As we can see, the mean frequency is proportional to the cluster size $|c_\delta|$. In other words, the larger is the cluster (the larger is $|c_\delta|$), more categories would get non-zero frequency. Assuming the count of the variable i's categories with non-zero frequency is η_i in a given cluster δ, the η_i is proportional to $|c_\delta|$, $\eta_i \propto |c_\delta|$, and those frequencies are roughly similar, i. e. $n_{i,j} \approx \frac{n}{\eta_i} \approx \frac{|c_\delta|}{\eta_i}$, we can derive

$$
\begin{aligned}
S_i &= \sum_{j=1}^{n_i} \left(\frac{n_{i,j}}{n}\right)^2 \propto \sum_{j=1}^{\eta_i} \left(\frac{|c_\delta|/\eta_i}{|c_\delta|}\right)^2 \propto \\
S_i &\propto \sum_{j=1}^{\eta_i} \frac{1}{\eta_i^2} \propto \sum_{j=1}^{\eta_i} \frac{1}{|c_\delta|^2} \propto \\
&\propto \sum_{j=1}^{|c_\delta|} \frac{1}{|c_\delta|^2} \propto |c_\delta| \cdot \frac{1}{|c_\delta|^2} \propto \\
&\propto \frac{1}{|c_\delta|}, \qquad (4)
\end{aligned}
$$

that supports our suggestion to draw observations from the clusters proportionally to their sizes[2], i. e. the larger the cluster is, the more observations should be picked from the cluster towards the final subsample to minimize the sum of squares S_i.

Secondly, we also propose an experimental approach that requires another ongoing research. Considering the (not necessarily k-means) clustering is repeated m times, with a random number of clusters in each of m iterations, we can construct a symmetric square matrix T of dimensions $N \times N$, that for the p-th row and the q-th column describes a number of times that the p-th observation of the original dataset was together in the same cluster with the q-th observation of the original dataset. The matrix T follows a form of

$$
T = \begin{pmatrix}
t_{1,1} & t_{1,2} & \cdots & t_{1,N} \\
t_{2,1} & t_{2,2} & \cdots & t_{2,N} \\
\vdots & \vdots & \ddots & \vdots \\
t_{N,1} & t_{N,2} & \cdots & t_{N,N}
\end{pmatrix}, \qquad (5)
$$

where $t_{p,q}$ stands for a number of times both the p-th observation and q-th observation of the original dataset were together in the same cluster.

Once we want to construct a subsample of size n from the original dataset of size N, we demand on keeping all variables' each category frequency balanced with other frequencies, so the final subsample should include all categories of all variables with (near) similar frequencies, if possible. Drawing such observations that were many times together in the same clusters within the multiple clustering procedure would result in the final subsample containing too many similar observations, which would reduce the native variability of the variables.

Consequently, the final subsample should be constructed using observations that are mutually non-similar. The way of constructing such a subsample could be to pick two original observations with a minimum value of $t_{p,q}$ and then add to the subsample one by one new observation (until the subsample size is sufficient) such that each new one (the q-th) has the minimum value of

$$
\sum_{\forall p \in \{\text{observations in subsample}\}} t_{p,q},
$$

so that

$$
q = \operatorname{argmin}_{q \in \{1,2,\ldots,N\}} \sum_{\forall p \in \{\text{observations in subsample}\}} t_{p,q}, \qquad (6)
$$

that minimizes a chance of getting a subsample with too much similar observations. While this approach may look as completely deterministic, it contains a part that is based on randomness, namely the clustering part.

Adopting the time complexity of the m times repeated k-means clustering for small number of clusters is $\Theta(m \cdot N \cdot k)$ [7] and for the T matrix construction (5), the ongoing part using the formula (6) takes averagely $\Theta(n^2)$ complexity time units.

Whereas the clustering algorithm itself could vary (it is not necessary to apply only k-means algorithm), it is worth to be mentioned that – since the variables in the original dataset are categorical (or transformed into categorical ones) – the Gower distance was chosen for the clustering as it can handle categorical variables well within the clustering [8].

The pseudocode of the subsampling by clustering the original dataset is in Algorithm 3.

III. Results

We used COVID-19 survey data of our provenience for the application of the proposed methods. The original dataset contains $N = 698$ rows corresponding to observations and $k = 18$ columns related to variables. Since the dataset is of

[2]The proportional equation (4) might be confusingly understood as to pick a maximum of observations (towards the final subsample) from the larger cluster since this would lead to the minimization of the sum of squares S_i for the given variable. However, such a subsample would be constructed using almost only one of the clusters – the largest one – and thus, tends to include very similar observations, which could break the demand of well-balanced category frequencies over all variables.

Algorithm 3: Subsampling by clustering the original dataset using the matrix T of mutual occurrences in the same clusters as in (5)

Data: an original dataset of size N containing k variables and T matrix of mutual occurrences in the same clusters as in (5)
Result: a subsample minimizing the averaged sums of squares $\bar{S}$

n // size of the subsample;
n_t // current size of the dataset;
T // matrix of mutual occurrences in;
// the same clusters;
$\bar{S}$ // current averaged sums of squares;
$\mathcal{S}$ // current subsample;
populate the subsample $\mathcal{S}$ by the first two observations corresponding to row and column indices of minimum of T;
$n_t = 2$;
while $n_t < n$ **do**
  pick the q-th observation such that

$$q = \text{argmin}_{q \in \{1,2,\ldots,N\} \setminus \mathcal{S}} \sum_{\forall p \in \mathcal{S}} t_{p,q},$$

  where $t_{p,q}$ is the value of p-th row and q-th column of the matrix T ;
  $n_t = n_t + 1$;
  update $\bar{S}$;
end
use the subsample $\mathcal{S}$ of size n;

a questionnaire form including questions with the close format, the vast majority of the variables are categorical. A few of the numerical variables were categorized following experts' suggestions or natural logic, e. g. age was categorized into intervals of lengths ten years, starting and ending at an age divisible by a number 10, etc. Applying this approach, there are only categorical variables in the original dataset before the subsampling. The reason why the response variable, i. e. the serology levels of COVID-19 antibodies, is missing in the original dataset is that patients involved in the study were planned to undergo relatively expensive serology tests; thus, the original size ($N = 698$) had to be reduced significantly ($n = 400$) to keep the costs of the serology testing manageable.

The task was to get a subsample of $n = 400$ rows from the original dataset, containing the original number of $k = 18$ variables.

All the computations were performed using R programming language and environment [9]. There are more numerical applications of R language to various fields in [10]–[15].

We applied all the methods mentioned above to do the subsampling and compare the results using the metric controlling the quality of the subsampling in between the methods.

The metric of the subsampling quality, depicting particularly how well the category frequencies of all the variables are balanced, is the averaged sums of squared $\bar{S}$ as defined in (2).

Besides the single random subsampling without replacement, we started with the repeated random subsampling without replacement. Repeating the random subsampling multiple times ($m = 100$) enables modeling the prior distribution of the averaged sums of squared $\bar{S}$, and was also used for estimation of the minimum value of averaged sums of squares $\bar{S}$ using the formula (3).

Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 1. The minimum value of averaged sums of squares $\bar{S}$ was estimated following the (3) to be equal $\hat{\bar{S}} \doteq 0.247$.

The next method, subsampling by forwarding one-be-one reduction of the original dataset, was also performed $m = 100$ times. Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 2. The minimum value of averaged sums of squares $\bar{S}$ for the one-be-one size reduction that was obtained is equal to $\hat{\bar{S}} \doteq 0.242$.

The subsampling by clustering the original dataset was performed $m = 100$ times, too. The final subsample was designed using the T matrix (5) and creating the subsample from scratch using the logic of formula (6). Histogram of the prior distribution of the averaged sums of squared $\bar{S}$ is in figure 3. The minimum value of averaged sums of squares $\bar{S}$ for the one-be-one size reduction that was obtained is equal to $\hat{\bar{S}} \doteq 0.245$.

As we can see, all the three applied methods return similar accuracy considering the minimization of the averaged sums of squares. The formal comparison of statistical differences in between mean values of the averaged sums of squares $\bar{S}$ for the repeated ($m = 100$) random subsampling without replacement, repeated ($m = 100$) subsampling by forwarding one-be-one reduction of the original dataset, and the repeated ($m = 100$) subsampling by clustering the original dataset could be performed using one-way analysis of variance (ANOVA). However, considering the figures 1, 2 and 3, the practical differences are minimal. What practically differs is the asymptotic time complexity of the mentioned techniques, as discussed before.

IV. Conclusion

Subsampling may be an important task when the original dataset is larger than required. If there is a response variable available in the dataset, then the methodology used for the subsampling is well established; the popular propensity scoring is used to extract the subsample from the original data that harmonize size effects of all predictors using logistic regression model.

When the response variable from some reason or another is missing, e. g. is planned to be collected later, the methodology of the subsampling is not so straightforward. Many various

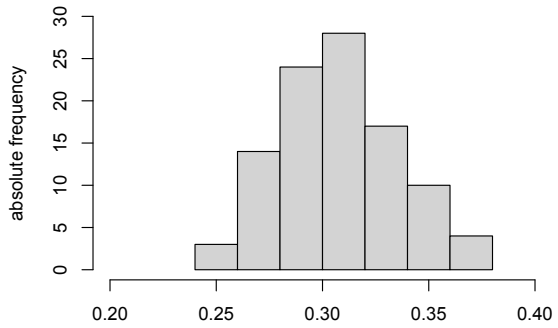


Fig. 1. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) random subsampling without replacement.

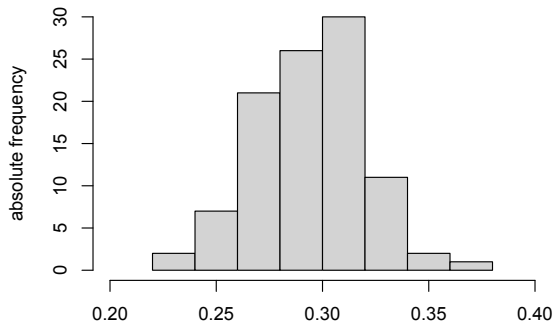


Fig. 2. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) subsampling by forwarding one-be-one reduction of the original dataset.

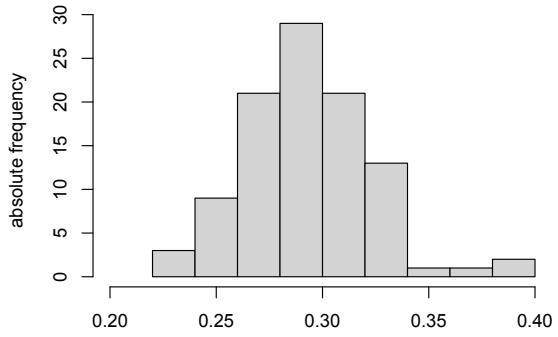


Fig. 3. Histogram of the prior distribution of the averaged sums of squares $\bar{S}$ calculated for the repeated ($m = 100$) subsampling by clustering the original dataset.

methods of low significance are used, based on different approaches – from totally random subsampling to manually matched pairs of observations with balanced all variables' category frequencies.

In this study, we proposed one metric – the averaged sums of squares – enabling to control a quality of the subsampling, including the fact the metric is in theory scaled to an interval not dependent on entry data, as was proven. Furthermore, we compared several methods; some of them are novel and proposed by this paper.

While the repeated random subsampling without replacement is relatively fast method, it can reach the minimum of the averaged sums of squares only approximately. The subsampling using one-by-one reduction of the original sample is a bit slower than the random multiple subsampling, but still feasibly applicable; it can approach the minimum of the averaged sums of squares only approximately, too. The exhaustive subsampling as the only one method can numerically calculate the exact value of the minimum of the averaged sums of squares; however, its executing time is enormously high. Finally, the subsampling by clustering is an innovative method that is relatively fast if implemented using standard algorithms and maturated computational environments, and furthermore, it offers a way to keep control over the mutual occurrences of each two observations from the same clusters, when the final subsample is constructed. Even the subsampling by clustering approached the minimum of the averaged sums of squares relatively closely.

All the proposed methods, i. e. repeated random subsampling without replacement, subsampling using one-by-one reduction of the original dataset and subsampling by clustering seem to be valid alternatives to exhaustive subsampling.

V. Acknowledgement

This paper is supported by the grant OP VVV IGA/A, CZ.02.2.69/0.0/0.0/19_073/0016936 with no. 18/2021, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] Peter C. Austin. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3 (May 2011), pp. 399–424. DOI: 10.1080/00273171.2011.568786. URL: https://doi.org/10.1080/00273171.2011.568786.

[2] Santhosh Pathical and Gursel Serpen. "Comparison of subsampling techniques for random subspace ensembles". In: *2010 International Conference on Machine Learning and Cybernetics*. IEEE, July 2010. DOI: 10.1109/icmlc.2010.5581032. URL: https://doi.org/10.1109/icmlc.2010.5581032.

[3] Elizabeth A. Stuart. "Matching Methods for Causal Inference: A Review and a Look Forward". In: *Statistical Science* 25.1 (Feb. 2010). DOI: 10.1214/09-sts313. URL: https://doi.org/10.1214/09-sts313.

[4] Sarda Sahney, Michael J. Benton, and Paul A. Ferry. "Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land". In: *Biology Letters* 6.4 (Jan. 2010), pp. 544–547. DOI: 10.1098/rsbl.2009.1024. URL: https://doi.org/10.1098/rsbl.2009.1024.

[5] David MacKay. *Information theory, inference, and learning algorithms*. Cambridge, UK New York: Cambridge University Press, 2003. ISBN: 0-521-64298-1.

[6] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. “Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test”. In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2020. DOI: 10.15439/2020f198. URL: https://doi.org/10.15439/2020f198.

[7] Malay K. Pakhira. “A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting”. In: *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE, Nov. 2014. DOI: 10.1109/cicn.2014.220. URL: https://doi.org/10.1109/cicn.2014.220.

[8] J. C. Gower. “A General Coefficient of Similarity and Some of Its Properties”. In: *Biometrics* 27.4 (Dec. 1971), p. 857. DOI: 10.2307/2528823. URL: https://doi.org/10.2307/2528823.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[10] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Evaluation of facial attractiveness for purposes of plastic surgery using machine-learning methods and image analysis”. In: *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Sept. 2018. DOI: 10.1109/healthcom.2018.8531195. URL: https://doi.org/10.1109/healthcom.2018.8531195.

[11] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language”. In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[12] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Machine-Learning and R in Plastic Surgery – Evaluation of Facial Attractiveness and Classification of Facial Emotions”. In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Sept. 2019, pp. 243–252. DOI: 10.1007/978-3-030-30604-5_22. URL: https://doi.org/10.1007/978-3-030-30604-5_22.

[13] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language”. In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[14] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Evaluation of Facial Attractiveness after Undergoing Rhinoplasty Using Tree-based and Regression Methods”. In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, Nov. 2019. DOI: 10.1109/ehb47216.2019.8969932. URL: https://doi.org/10.1109/ehb47216.2019.8969932.

[15] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. “A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer Data”. In: *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, Oct. 2020. DOI: 10.1109/ehb50910.2020.9280301. URL: https://doi.org/10.1109/ehb50910.2020.9280301.

Advances in Computer Science and Systems

ACSS is welcoming presentations of the scientific aspects related to applied sciences. The session is oriented on the research where the computer science meets the real world problems, real constraints, model objectives, etc. However the scope is not limited to applications, we all know that all of them were born from the innovative theory developed in laboratory. We want to show the fusion of these two worlds. Therefore one of the goals for the session is to show how the idea is transformed into application, since the history of modern science show that most of successful research experiments had their continuation in real world. ACSS session is going to give an international panel where researchers will have a chance to promote their recent advances in applied computer science both from theoretical and practical side.

Scope:

- Applied Artificial Intelligence
- Applied Parallel Computing
- Applied methods of multimodal, constrained and heuristic optimization
- Applied computer systems in technology, medicine, ecology, environment, economy, etc.
- Theoretical models of the above computer sciences developed into the practical use

Track Chairs

- **Dimov, Ivan,** Bulgarian Academy of Sciences, Institute of Information and Communication Technologies, Bulgaria
- **Wasielewska-Michniewska, Katarzyna,** Systems Research Institute, Polish Academy of Sciences, Poland

Program Chairs

- **Dimov, Ivan,** Bulgarian Academy of Sciences, Institute of Information and Communication Technologies, Bulgaria
- **Wasielewska-Michniewska. Katarzyna,** Systems Research Institute, Polish Academy of Sciences, Poland

Program Committee

- **Barbosa, Jorge,** University of Porto, Portugal
- **Braubach, Lars,** University of Hamburg, Germany
- **Cabri, Giacomo,** Università di Modena e Reggio Emilia, Italy
- **Fabijańska, Anna,** Technical University of Lodz, Poland
- **Georgiev, Krassimir,** Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- **Homles, Violeta,** University of Huddersfield, United Kingdom
- **Jezic, Gordan,** University of Zagreb, Croatia
- **Kotenko, Igor,** St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Russia
- **Lirkov, Ivan,** Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- **Mangioni, Giuseppe,** Dipartimento di Ingegneria Elettrica Elettronica e Informatica (DIEEI) - University of Catania, Italy
- **Millham, Richard,** Durban University of Technology, South Africa
- **Modoni, Gianfranco,** STIIMA-CNR, Italy
- **Pandey, Rajiv,** Amity University, India
- **Pawłowski, Wiesław,** University of Gdańsk, Poland
- **Petcu, Dana,** West University of Timisoara, Romania
- **Scherer, Rafał,** Częstochowa University of Technology, Poland
- **Schreiner, Wolfgang,** Research Institute for Symbolic Computation (RISC), Austria
- **Tudoroiu, Nicolae,** John Abbott College, Canada
- **Wyrzykowski, Roman,** Częstochowa University of Technology, Poland
- **Vardanega, Tullio,** University of Padua, Italy

14th Workshop on Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

TOPICS

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on coprocesors (GPU, Intel Xeon Phi, etc.)
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

TECHNICAL SESSION CHAIRS

- **Bylina, Beata,** Maria Curie-Sklodowska University, Poland
- **Bylina, Jaroslaw,** Maria Curie-Sklodowska University, Poland
- **Stpiczyński, Przemysław,** Maria Curie-Sklodowska University, Poland

PROGRAM COMMITTEE

- **Amodio, Pierluigi,** Universita' di Bari, Italy
- **Anastassi, Zacharias,** ASPETE School of Pedagogical and Technological Education, United Kingdom
- **Banaś, Krzysztof,** AGH University of Science and Technology, Poland
- **Bielecki, Włodzimierz,** West Pomeranian University of Technology in Szczecin, Poland
- **Brugnano, Luigi,** Università di Firenze, Italy
- **Burczynski, Tadeusz,** Polish Academy of Sciences, Poland
- **Czachórski, Tadeusz,** Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland
- **Domanska, Joanna,** Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland
- **Fialko, Sergiy,** Cracow University of Technology, Poland
- **Gemignani, Luca,** University of Pisa, Italy
- **Gepner, Paweł**
- **Giannoutakis, Konstantinos,** CERTH-ITI, Greece
- **Georgiev, Krassimir,** Bulgarian Academy of Sciences, Institute of Information and Communication Technologies, Bulgaria
- **Gravvanis, George,** Democritus University of Thrace, Greece
- **Kozielski, Stanisław,** Silesian University of Technology, Institute of Informatics, Poland
- **Krawczyk, Henryk,** Gdańsk University of Technology, Poland
- **Kucaba-Pietal, Anna,** Politechnika Rzeszowska, Poland
- **Lirkov, Ivan,** Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Poland
- **Ltaief, Hatem,** King Abdullah University of Science and Technology, Saudi Arabia
- **Luszczek, Piotr,** University of Tennessee Knoxville, United States
- **Marowka, Ami,** Bar-Ilan University, Isreal
- **Mehmood, Rashid,** King Abdulaziz University, Saudi Arabia
- **Mrozek, Dariusz,** Silesian University of Technology, Institute of Informatics, Poland
- **Palkowski, Marek,** Faculty of Computer Science, West Pomeranian University of Technology, Poland
- **Petcu, Dana,** West University of Timisoara, Romania
- **Rojek, Krzysztof,** Czestochowa University of Technology, Poland
- **Sawerwain, Marek,** University of Zielona Góra, Poland
- **Sidje, Roger B.,** University of Alabama, United States
- **Siminski, Krzysztof,** Silesian University of Technology, Poland
- **Skubalska-Rafajłowicz, Ewa,** Wrocław University of Science and Technology, Poland
- **Trivedi, Kishor S.,** Duke University, United States
- **Tudruj, Marek,** Polish-Japanese Institute of Information Technology, Institute of Computer Science, Polish Academy of Sciences, Poland
- **Ustimenko, Vasyl,** University of Maria Curie Sklodowska in Lublin, Poland

- **Wyrzykowski, Roman,** Czestochowa University of Technology, Poland
- **Vajtersic, Marian,** Deaprtment of Computer Sciences, University of Salzburg, Switzeland
- **Vardaneg, Tullio ,** University of Padova, Italy
- **Vazhenin, Alexander,** University of Aizu, Japan

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 117–121
DOI: 10.15439/2021F80
ISSN 2300-5963 ACSIS, Vol. 26

On new stream algorithms generating sensitive digests of computer files

Vasyl Ustimenko
University of Marie Curie-Sklodowska in Lublin, ul. Plac Marii Curie-Skłodowskiej 5, Lublin, 20-031, Poland
Email: vasulustimenko@yahoo.pl

Oleksandr Pustovit
Institute of Telecommunications and the Global Informatio the National Academy of Sciences of Ukraine, Chokolivsky 13, Kyiv, 02000, Ukraine
Email: pustovitoleksandr0709@gmail.com

The paper is dedicated to construction of new fast and flexible hash-based message authentication codes (HMACs) that will provide large files with cryptographically stable digestions in the Postquantum era.

These instruments can be used for detecting cyber-terrorist attacks, file audits and checking the integrity of messages during communication, We use algebraic properties of well known extremal graphs *D(n, q)* and *A(n,q)* with good expansion property for the construction of HMACS.

I. Introduction

We propose new fast algorithms for the creation of sensitive digests of electronic files to detect cyberattacks, computer viruses or other damages and check data integrirty. These tools can be used to defend virtual organization and conduct the audit of all files after a registered intervention. Cryptographic stability of new key-dependent hash functions is associated with complex algebraic problems, such as the study of systems of algebraic equations of large degree and the problem of decomposition of nonlinear transformation into the composition of given generators. These facts justify resistance of digests against adversary attacks with the usage of algorithms in terms of Turing machine or Quantum Computation Theory.

Algorithms of digests generation use idea of presentation of files in the form of sequences (words) of elements of finite commutative ring K , such as finite field or arithmetical ring modulo 2^m. Such words can be treated as elements of free semigroup or its modification S with the usage of additive ring operation. The presentation of k - regular tree (or other regular infinite graph) in the form of projective limit of finite graphs given by equations allows to define «compression homorphism» of S into group of polynomial transformations of affine space K^t (space of tigests of the chosen dimension (t). Affine transformations are used to hide the homomorphic map. This scheme is new.

Implemented accordingly to this scheme family of fast algorithms was investigated via computer simulations on the real data of large size. Change of single character of the document in binary alphabet causes the change of the majority of characters of produced digest (≥98%). This property and evaluation of time execution of software programs justify potential of practical usage of implemented algorithm for cybersecurity tasks.

II. On the verification of electronic documents

Protection of large data repositories against cyberattacks via creation of digests of both encoded and original electronic documents in the selected starting time is the impotent task. With a change of time new digests could be created and compared to the original ones. The presence of any changes indicates damage to the files (cyberattack, computer virus, hardware failure, staff error and more).

For the checking the integrity of electronic documents within transmission the correspondent creates a digest of the original file and the same file in an encrypted form. His/her partner creates a digest of the received decrypted document and the original encrypted document. Correspondents compare the digests and conclude whether or not document was damaged.

A simplified model of the global information space can be imagined as a large, growing network of registered virtual users (individuals or institutions) who exchange information and can store it in electronic repositories located on the network or isolated from them.

The size of files for sharing (electronic documents) tends to grow. An important category of information space is trustworthiness of documents.

Users can use a symmetric private key algorithm to encrypt documents and key exchange protocol to maintain encryption security. Certified public key algorithms may be also used to change the key. These methods ensure the security of the exchange channels.

It is easy to see that even if a reliable encryption is used it does not provide a complete trustworthines of the documents, because it is necessary to take into account the noise in the channels and the problems of safe storage of

files in electronic repositories, where documents can be tampered with, damaged by computer viruses, technical errors in the work of computing machinery, etc.

It should be noted that the threat of powerful cyberterrorist attacks on repositories of electronic information are recently increasing. There consequences are not only information leakage, but also a damage or a falsification of documents. It is clear that once a cyberattack is detected on a corporate information repository you need to audit all system files. Countering these threat require the development of a new software.

Other information security tasks require a general hash function that does not require a key or password (see for instance [1], [2] and further references).

III. Requirements to digests

The cryptographically stable hash function f must provide the practical impossibility of selecting a pair of links x and z with the same value of the hash function. The digest of a document created with a key-dependent hash function (MAC) uses the HMAC symbol. When users want to exchange correspondence securily, verifying who is the actual author of the letter, and the absence of changes when forwarding, they choose a shared MAC. Additionally they use a common symmetric encryption scheme.

In addition to cryptographical stability the execution speed and high indicator of avalanch effect are important. Avalanch effect can be measured in the following way. The HMAC of generated file has to be computed, after this step some chosen character of original file has to be changed for other symbol and HMAC for the new file has to be computed.

Finally bit to bit comparison of characters of two HMACs has to be done and persantage of changed characters has to be computed. For practical usage of HMAC is necessary to show that change of arbitrarily used character leads to the change of at least 40 persent of bites independently from the size of tested files.

Introduced approach of usage of special subgroups of endomorphisms from CSn(K) is useful for the development of stream ciphers of Symmetric Cryptography (see for instance [3], [4], [5] and further references) and constructions of HMACs (see [6] where special linear groups have been used). We use nonlinear subgroups of $CS_n(K)$. The method of generation of nonlinear transformations of free modules over commutative rings described in the terms of special graphs defined by algebraic equations (so called linguistic graphs) can be used instead of methods of generators and equations. Other applications of graph theory to Cryptography are considered in [7].

Studies of message authentification codes and HMACs is a hot topic.Complete list of all published papers within this direction is impossible to make, we only refer to some recent papers [8] - [17].

Recall that noncommutative cryptography is an active field of cryptology that explores cryptographic primitives and systems based on algebraic structures such as groups, semigroups, and non-commutative rings.

One of the earliest applications of noncommutative algebraic structure for cryptographic purposes was the usage of groups for the development of cryptographic protocols.

The method of usage of platform G which is a subgoup or subsemigroup of affine Cremona semigroup $CS(K^n)$ defined over finite commutative ring K under the condition that each element is presented in [19]. This is an attempt to merge methods of noncommutative cryptography and multivariate cryptography.

Studies of message authentification codes and HMACs is a hot topic. Noteworthy that arbitrary hash function such as MD5 or SHA-1 can be used for the composition of HMACs corresponding to MD5 and SHA-1 message authntication codes are known as HMAC-MD5 and HMAC-SHA-1 respectively. HMAC's cryptographic performance depends on the cryptographic performance of the underlying hash function, the size of its hash output, the size and quality of the key.

IV. Mathematical background of proposed hash functions

Let $F(K)$ be a space of potentially infinite texts in alphabit K which is the totality of all tuples of kind $(a_1, a_2, ..., a_k), a_i \in K$ of different case k. Assume that K is finite commutative ring and identify $F(K)$ with the semigroup with the following operation $(a_1, a_2, ..., a_k) \circ (b_1, b_2, ..., b_s) = (a_1, a_2, ..., a_k, b_1 + a_k, b_2 + a_k, b_s + a_k)$

Let $F'(K)$ be the subsemigroup of all words (tuples) of even length. We assume that CST(K n) stands for the semigroup of all polynomial maps of affine space K^n in itself.

Our algorithm is based on the following mathematical statement.

Theorem 1 (see [20]). For each natural integer $m \geq 2$ there exists homomorphism $\psi: F'(K) \rightarrow CS_m(K)$ such that its image $\psi(F'(K))$ is a group G of cubic polynomial transformations of degree 3.

Recall that the property of $\psi = \psi_m$ to be homomorphic map means that $\psi(a \circ b) = \psi(a) \circ \psi(b)$.

Transformations satisfying conditions of the theorem are defined in constructive way in terms of the theory of discrete dynamic systems defined via algebraic graphs with

extremal properties. These methods allow to get the lower bound $|G| \geq 2^{4n}$ of the order of *G*. Noteworthy that the proposition defines rare mathematical object. Superposition of two randomly chosen cubic maps will have degree 9 , in the case of 3 such maps resulting degree will be 27, composition of 4 such maps has degree 81, but in the constructed group all compositions of several maps will have degree ≤3.

It was not the *G* group itself that was used to create the MAC but the mapping ψ that defines it along with the affine *A* and *B* transformations of the Cremony group with the rule $g: x \to A\psi(x)B$. It is not hard to see that it's a natural data compression operator that maps an infinite set of all even-length words in the alphabet *K* to a finite set. The output is a list of coordinates *g (x)* to which the full differential operator is applied twice. Computer simulation made it possible to calculate a very high avalanche effect within 97-98 percents. For example, in MAC of Russian researchers the avalanche effect interval is estimated as 47-50% [18]. The constructive definition of compression homomorphisms is defined in terms of the theory of linguistic graphs. The known linguistic graphs *A (n, K)* and *D (n, K)* constructed for solving some problems of extreme graph theory are used (see [21] and furtherreferences).

V. On the option to speed up the algorithm

In this unit we present the modification of described above algorithm which allows to present (or even improve) the level of riched avalanche effect under essential increase of execution time. We have to admit that algorithm is described ''by modulo'' of computation of homomorphism value in a given point. Constructive definition of ψ were already described in the previous sections.

Let $(a_1, a_2, ..., a_n)$ be digital document presented in the alphabet *K* after the merge of file with some pseudorandom word of constant length.. We assume that parameter n is even. Users select the size of digest $m, m < n$ where $m = O(1)$ or $m = O(n)$ together with the key is formed by increasing sequence of positive integers $i(1), i(2), ..., i(m-1)$ and nonsingular matrix *M* with entries from commutative ring Z_{256} of residues modulo 256. Users form vector $u = (v_1, v_2, ..., v_m)$, where $v_1 = a_1 + a_2 + ... + a_n, ..., v_j = v_{j-1} - a_{i(j-1)}$. Secondly they compute the cubical map $F = \psi_m(a_1, a_2, ..., a_n)$ and its value on the vector u. Computed row vector $F(u)$ has to be multoplied on matrix M. Vector $w = F(u)M$ is the digest of document.

Note, that the value of $F(u)$ is calculated via recursive procedure, its complexity is approximated as $O(mn)$ and coincides with the complexity of digest generation.

This basic algorithm is easy to modify without the change of computational complexity. In particular the folloving variants can be used.

One can present the word $(a_1, a_2, ..., a_n)$ in a form of concatenation of finite number of words $z_1, z_2, ..., z_t$ of even length. Secondly he/she selects the sequence $u_1, u_2, ..., u_k$ where $u_i \in \langle z_1, z_2, ..., z_l \rangle$ such that each word z_i appears at least one time in this sequence. The next step is a computation of value of product of $u_1, u_2, ..., u_k$ in the presented above semigroup of words $F'(K)$. Algorithm is modified via the change of cubic map $\psi(a)$ for $\psi(y)$. In the case of open partition of file cryptographic stability such digest rests on the decomposition problem of $\psi(y)$ into the product of transformation $\psi(z_i)$ from affine Cremona group. Noteworthy that the polynomial postquantum algorithm for solving this problem is unknown.

In fact this problem appears under the condition of the incomplete knowledge because only the value $\psi(y)$ is known but not the cubical map itself. In this modification users have to understand that the partition of a on subwords z_i and the sequence u_j are considered as a part of common private key for correspondents.

2) Correspondents can compute v_1 as aproduct of expressions $2a_i + 1$ and obtain v_i by division of v_{i-1} on $2a_{i(j-1)} + 1$.

3) In the case 2 one can change v_i for its odd powers k, k<128. Then these degrees have to be counted as parameters of private key.

Implemented cases are convenient for their usage in blockchain technologies where digests in the form of sequence of bites 0 and 1 symbols are needed.

We have to note that good mixing properties of compression maps are based on the constructions of homomorphisms of infinite semigroup of words of even length in affine Cremona group defined via families of algebraic graphs with remarkable extremal properties.

VI. On the implementation of digest generation algorithms

Programs are inmplemented in C++ language. Time execution of a software depends on the parameters of a computer. We use ordinary personal computer with

Pentium 3.00 GHz processor, 2GB of RAM memory for Windows 7 system.

For the computer simulations with presented above basic algorithm on the base of group *GA(n,K)* we use sparse matrix M, computable in time *O(m)* where m is digest size.

Digests were presented in characters of binary alphabet to measure of avalanch effect. Time execution in seconds for files of various size is presented below.

Table 1 – Time execution of digests generation

Size of file (megabytes)	Size of digests (in bites)		
	256	512	1024
4,0	1,36	2,74	5,52
16,1	4,94	9,90	19,82
38,7	11,60	23,20	46,46
62,3	18,54	37,10	74,22
121,3	36,24	72,52	145,02
174,2	51,22	103,66	207,34

Computer simulation demonstrates that the change of a single character of an electronic document leads to the change of 98% of the corresponding digest.

VII. Conclusion

The routine work of an enterprise, corporation, financial institution requires a long-term work of specialists with a large number of electronic documents. Specialists must use proven information to make sound planning decisions. The validation tool for checking the documents can be large files compression algorithm producing a digest of a certain size, sensitive to any change in input characters.

New family of key-dependent fast algorithms for creating electronic documents digest is proposed. Computer simulation allows to investigate the high level of an emerging avalanche effect. Let K be a freely chosen finite commutative ring and m is a positive integer. The algorithms use the recently found homomorphic compression mapping of a semigroup of potentially infinite texts in the alphabet K to a finite group of cubic polynomial transformations of an affine space K^m.

The cryptographic stability of hashing functions is associated with complex algebraic problems, such as the investigation of large-scale algebraic equation systems and the problem of decomposition of a nonlinear mapping of a free module by given generators.

Algorithms are implemented in the cases of finite fields $F_2^8, F_2^{16}, F_2^{32}$, arithmetic ring Z_{256} and B(32) (Boolean ring of order 2^{32}).

Computer simulation demonstrates that the speed of the algorithm increases with the size of the base switching ring.

The proposed algorithms can handle data in the form of texts, video and audio files, movies , etc. The developed methods of creation of digests have flow character, the speed in the case of a constant size of digest depends linearly on size n of the file. The rise of parameter n increases the cryptographic stability. Block implementation is possible but not motivated, because fixed block size limits the number of variables of a system of nonlinear equations

The need for further research and technological development to create new key-dependent fast hash functions is linked to cybersecurity challenges, the growth of global information space, the expectation of a quantum computer, and the development of bitcoins technologies where we need to hash out arbitrary-sized inputs into the sequence of bits that is the digest of the so-called blockchains.

The proposed robust algorithms for creating sensitive digests of documents will now be practically used to detect cyberattacks and audit all system files after a logged-in intervention. This is the first successful attempt to implement the idea of non-commutative cryptography to create HMACs. We still believe that further work is needed to optimize the built algorithms, compare them with previously known HMACs and crypto-analytical studies.

References

[1] Oliynykov R., Gorbenko I., Kazymyrov O., Ruzhentsev V., Kuznetsov O., Gorbenko Yu., Dyrda O., Dolgov V., Pushkaryov A., Mordvinov R., Kaidalov D. Data Security. *Symmetric block transformation algorithm.* Ministry of Economical Development and Trade of Ukraine. DSTU 7624:2014. National Standard of Ukraine. Information technologies. Cryptographic. –2015.

[2] Aumasson J.Ph, Serious Cryptography: A Practical Introduction to Modern Encryption, No Starch Press. – 2017. – 312 pp.

[3] Pustovit O., Ustymenko V., Pro zastosuvannia alhebraichnoi kombinatoryky do problem koduvannia ta kryptohrafii [On the application of algebraic combinatorics to the problems of coding and cryptography] //Matematychne modeliuvannia v ekonomitsi, № 1-2. – Kyiv. – 2017. – s. 31-46.

[4] V. Ustimenko, U. Romanczuk-Polubiec, A. Wroblewska, M. Polak, E. Zhupa, On the constructions of new symmetric ciphers based on non-bijective multivariate maps of prescribed degree, Security and Communication Networks, 2019 . Volume 2019, Article ID 2137561, 15 pages

[5] V. Ustimenko, U. Roman'czuk-Polubiec, A. Wroblewska, M. Polak and E. Zhupa, *On the implementation of new symmetric ciphers based on non-bijective multivariate maps,* Proceedings of the 2018 Federated Conference on Computer Science and Informatics.
Proceedings of the Federated Conference on Computer Science and Information Systems pp. 397–405 DOI: 10.15439/2018F204 ISSN 2300-5963 ACSIS, Vol. 15, pp.397-405.

[6] Mathew Cary, Ramarathnam Venkatesam, *A Message Authentication Code Based on Unimodular Matrix Groups*, Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings, Lecture Notes in Computer Science.

[7] Priyadarsini P.L.K., *A Survey on some Applications of Graph Theory in Cryptography*, Journal of Discrete Mathematical Sciences and Cryptography, 18:3, 209-217 (2015).

[8] Mihir Bellare, Daniel J. Bernstein, and Stefano Tessaro, *Hash-function based PRFs:AMAC and its multi-user security,* LNCS, pages 566-595. Springer, Heidelberg, 2016.

[9] Kan Yasuda. A Double-Piped Mode of Operation for MACs, PRFs and PROs: *Security beyond the Birthday Barrier.* In Antoine Joux, editor, EUROCRYPT, volume 5479 of Lecture Notes in Computer Science, pages 242-259. Springer, 2009.

[10] Xiaoyun Wang, Hongbo Yu,WeiWang, Haina Zhang, and Tao Zhan. *Cryptanalysis on HMAC/NMACMD5 and MD5-MAC*. In Antoine Joux, editor, EUROCRYPT, volume 5479 of Lecture Notes in Computer Science, pages 121-133. Springer, 2009.

[11] Gaetan Leurent, Thomas Peyrin, and Lei Wang. *New Generic Attacks against Hash-Based MACs.* In Kazue Sako and Palash Sarkar, editors, Advances in Cryptology-ASIACRYPT 2013-1 volume 8270, pages 11-20. 2013.

[12] Neal Koblitz and Alfred Menezes. Another look at HMAC. Cryptology ePrint Archive, Report 2012/074, 2012.

[13] Yevgeniy Dodis, Eike Kiltz, Krzysztof Pietrzak, and Daniel Wichs. Message authentication, revisited.In David Pointcheval and Thomas Johansson, editors, EUROCRYPT 2012, volume 7237 of LNCS, pages 355-374. Springer, Heidelberg, April 2012

[14] Yevgeniy Dodis and John P. Steinberger, Domain Extension for MACs Beyond the Birthday Barrier, In Kenneth G. Paterson, editor, EUROCRYPT, volume 6632 of Lecture Notes in Computer Science,pages 323-342. Springer, 2011.

[15] Yevgeniy Dodis, Thomas Ristenpart, John P. Steinberger, and Stefano Tessaro. To Hash or Not to Hash Again? ,(In) Difererentiability Results for H2 and HMAC. In Reihaneh Safavi-Naini and Ran Canetti,editors, CRYPTO, volume 7417 of Lecture Notes in Computer Science, pages 348-366. Springer, 2012.

[16] Pierre-Alain Fouque, Gaetan Leurent, and Phong Q. Nguyen. Full Key-Recovery Attacks on HMAC/NMAC-MD4 and NMAC-MD5, In Alfred Menezes, editor, CRYPTO, volume 4622 of Lecture Notes in Computer Science, pages 13-30. Springer, 2007.

[17] Jongsung Kim, Alex Biryukov, Bart Preneel, and Seokhie Hong. On the Security of HMAC and NMAC Based on HAVAL, MD4, MD5, SHA-0 and SHA-1 (Extended Abstract). In Roberto De Prisco and Moti Yung, editors, SCN, volume 4116 of Lecture Notes in Computer Science. Springer, 2006.

[18] Krendelev S., Sazonova P., *Parametric Hash Function Resistant to Attack by Quantum Computer*, Based on Problem of Solving a System of Polynomial Equations in Integers, Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS. – Vol. 15. –pp. 387 – 390 (2018)

[19] V. A. Ustimenko, On the Families of Stable Multivariate Transformations of Large Order and Their Cryptographical Applications, Tatra Mountains Mathematical Publications,2O17, 70(1), pp 107-117.

[20] V. A. Ustimenko, On multivariate public keys based on the pair of transformations with the density gap. Доповіді НАН У, 2018. 9, с. 21-27.

[21] V.Ustimenko, On the usage of postquantum protocols defined in terms of transformation semi-groups and their homomorphisma, Theoretical and Applied Cybersecurity, National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute", vol 2, 2020, pp. 32-44.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 123–130
DOI: 10.15439/2021F67
ISSN 2300-5963 ACSIS, Vol. 26

On computations with Double Schubert Automaton and stable maps of Multivariate Cryptography

Vasyl Ustimenko
Faculty of Mathematics, Physics and Computer Science
Maria Curie-Skłodowska University
Pl. Maria Curie Skłodowska 1, 20-031 Lublin, Poland
vasyl@hektor.umcs.lublin.pl

***Abstract*—The families of bijective transformations G_n of affine space K^n over general commutative ring K of increasing order with the property of stability will be constructed. Stability means that maximal degree of elements of cyclic subgroup generated by the transformation of degree d is bounded by d. In the case $K = F_q$ these transformations of K^n can be of an exponential order. We introduce large groups formed by quadratic transformations and numerical encryption algorithm protected by secure protocol of Noncommutative Cryptography. The construction of transformations is presented in terms of walks on Double Schubert Graphs.**

***Index Terms*—Affine Cremona Group, Double Schubert Automaton, Multivariate Cryptography, Noncommutative Cryptography, Post Quantum Cryptography**

I. Introduction

In 2017 the international tender of the National Institute of Standartisation Technology (NIST) of the USA for the selection of public key based on postquantum algorithms was announced. It has been considering algorithms for the encryption task and for the procedure of digital signature. The last third round of this competition started in summer time of 2020. Only one candidate from the multivariare cryptography area remains. This is a special case of ''Rainbow like unbalanced oil and vinegar" digital scheme.The final list does not contain algorithms of Multivariate Cryptography for the encryption task. This outcome stimulates alternative research on numerical encryption asymmetrical postquantum algorithms of Multivariate cryptography such as algorithms which are not public keys and use the composition of several nonlinear maps of bounded degree. Our paper is dedicated to new postquantum secure cryptosystem with the encryption process based on bijective quadratic maps of large order. Postquantum status of these encryption is justified by recent results of Noncommutative Cryptography.

In March 2021 it was announced that prestigious Abel prize will be shared by A. Wigderson an L. Lovasz. They contribute valuable applications of theory of Extremal graphs (see [20]) and Expanding graphs [21] to Theoretical Computer Science. We have been working on applications of these graphs to Cryptography (see [22], [23], [24], [25] and further references). This paper is dedicated to the problem of postquantum secure encryption of rather large files in terms of Multivariate Cryptography but with usage of ideas of Noncommutative Cryptography. We will use Double Schobert graphs which belong to class of geometrical expanders introduced in [26]. Remarkable symbiotic combination of absolutely secure one time pad with Diffie-Hellman protocol in terms of groups F_p^*, p is prime, can not be used in our postquantum times because classical discrete logarithm problem can be solved in polynomial time with usage of quantum computer. The proof of this fact waspublished by Peter Shor in 1995. We present a possible substitutor of mentioned above symbiotic combination.

Classical encryption tools of Multivariate Cryptography are nonlinear polynomial maps F of affine space K^n over finite commutative ring K into itself. Traditionally a map F is presented in the form T_1GT_2, where T_1 and T_2 are representatives of affine general group $AGL_n(K)$ of all polynomial bijective transformations of K^n of degree 1 and central G is a nonlinear polynomial map. We refer to F as linear deformation of G. Popular computer tools for the generation of G are packages for symbolic computations ("Mathematica", "Maple", "Sage", "Magma" and special symbolic tools for professionals). Alternative approach to the construction of core maps G via *numerical* computations with sparse algebraic graphs was presented at some talks at CANA conference of FedSCIS [16], [17], [18]. The idea is to convert albebraic graphs into special automata for computations in polynomial ring $K[x_1, x_2, \ldots, x_n]$ in terms of "arithmetical" operations of addition and multiplication in the ring. It allows to use standard C^{++} or Java languages for the construction of polynomial maps over finite fields, arithmetical and Boolean rings. It is interesting that automata was constructed from bipartite algebraic graphs defined by systems of equations $x_i - y_i = x_l y_k$, some properties of graphs (stability, degree, in particular) were obtained theoretically but other properties (orders, density) were investigated via computer simulation.

This paper is a theoretical one, we present theoretical results which demonstrate potential of the graph based approach. It turns out that the method allows to generate a stable nonlinear polynomial maps of chosen degree with a prescribed density and exponentially growing order. Results are obtained via explicit constructions of automaton maps based on bipartite graph $DS(n, K)$ over general commutative ring K such that point $(x_1, x_2, \ldots, x_n, x_{11}, x_{12}, \ldots, x_{nn})$ is incident to line $[y_1, y_2, \ldots, y_n, y_{11}, y_{12}, \ldots, y_{nn}]$ if and only if $x_{ij} - y_{ij} =$

x_iy_j, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$. Special walks on this graph of even length induce nonlinear map of affine space $K^{n(n+1)}$ to itself. The graph has geometrical nature, in case of $K = F_q$ it is induced subgraph of the incidence relation of finite projective geometry.

The approach was motivated by cryptographical applications. That is why explicit constructions lead to some new cryptosystems. In Section 2 we discuss the concepts of postquantum security and multivariate cryptography (MC), some references on usage of cryptographical properties of algebraic graphs are given. Next section is devoted to the concept of stable transformation connected with the investigation of discrete logarithm problem in the affine Cremona group $C(K^n)$ of all bijective polynomial transformations of affine space K^n such that their inverses are also polynomial maps. This problem is motivated by related multivariate Diffie - Hellman key exchange protocol and corresponding El Gamal cryptosystem. In Section 4 we modify El Gamal algorithm, one can use high non commutativity of $C(K^n)$ and conjugate the inverse of the generator of large cyclic group. In the next section we state theorems on the existence of families of nonlinear stable multivariate maps over finite fields of exponentially growing order with prescribed degree and density. The existence of corresponding explicit construction is also formulated. The impact of such theorems is an option of implementation of multivariate key exchange protocols and related shifted El Gamal cryptosystems in case of family of cyclic subgroups of exponentially growing order in affine Cremona group over F_q. In Section 7 we discuss natural restriction on parameters for such algorithm. The Double Schubert graph and related automaton are introduced in Section 8. The sketch of proof of the main theorem on the existence of stable maps of exponentially growing order is given as a chain of lemmas.

The last section of the paper present new encryption algorithm of Post Quantum Cryptography. Encryption is descrribed in terms of Quadratic Multivariate Cryptography. The speed of encryption is standard for this area. Suggested cryptosystem is Post Quantum Secure. It is not a public key. So it differs from candidates investigated by well known NIST competition. Cryptosystem combines secure protocol with quadraticsable platform of polynomial transformations with flexible encryption procedure.

II. On Post Quantum and Multivariate Cryptography

Post Quantum Cryptography serves for the research of asymmetrical cryptographical algorithms which can be potentially resistant against attacks based on the use of quantum computer.

The security of currently popular algorithms is based on the complexity of the following three well known hard problems: integer factorisation, discrete logarithm problem, discrete logarithm for elliptic curves. Each of these problems can be solved in polynomial time by Peter Shor's algorithm for theoretical quantum computer. So cryptographers already started research on postquantum security. They also have to investigate the impact of the new results on general complexity theory such as complexity estimates of graph isomorphism problem obtained by L. Babai [11].

We have to notice that Post Quantum Cryptography (PQC) differs from Quantum Cryptography, which is based on the idea of usage of quantum phenomena to reach better security.

Modern PQC is divided into several directions such as Multivariate Cryptography, Lattice based Cryptography, Hash based Cryptography, Code based Cryptography, studies of isogenies for superelliptic curves.

The oldest direction is Multivariate Cryptography which uses polynomial maps of affine space K^n defined over a finite commutative ring into itself as an encryption tool. It exploits the complexity of finding solution of a system of nonlinear equations from many variables.

This is still young promising research area with the current lack of known cryptosystems with the proven resistance against attacks with the use of ordinary Turing machines. Studies of attacks based on Turing machine and Quantum computer have to be investigated separately because of different nature of two machines, deterministic and probabilistic respectively. Multivariate cryptography started from the studies of potential for the special quadratic encryption multivariate bijective map of K^n, where K is an extention of finite field F_q of characteristic 2. One of the first such cryptosystems was proposed by Imai and Matsumoto, cryptanalysis for this system was invented by J. Patarin. The survey on various modifications of this algorithm and corresponding cryptanalysis the reader can find in [1]. Various attempts to build secure multivariate public key were unsuccessful, but the research of the development of new candidates for secure multivariate public keys is going on (see for instance [2] and further references).

Applications of Algebraic Graph Theory to Multivariate Cryptography were recently observed in [3]. This survey is devoted to algorithms based on bijective maps of affine spaces into itself.

III. On stable multivariate transformations for the key exchange protocols

It is widely known that Diffie - Hellman key exchange protocol can be formally considered for the generator g of a finite group or semigroup G. Users need a large set $\{g^k | k = 1, 2, \ldots\}$ to make it practical. One can see that security of the method depends not only on abstract group or semigroup G but on the way of its representation. If G is a multiplicative group F_p^* of finite field F_p than we have a case of classical Diffie - Hellman algorithm. If we change ${F_p}^*$ for isomorphic group Z_{p-1} then the security will be completely lost. We get a problem of solving linear equation instead of a discrete logarithm problem to measure the security level.

Let K be a commutative ring. $S(K^n)$ stands for the affine Cremona semigroup of all bijective polynomial transformations of affine space K^n.

Let us consider a multivariate Diffie - Hellman key exchange algorithm for the generator $g(n)$ semigroup G_n of affine Cremona semigroup. Correspondents (Alice and Bob) agree on the generator $g(n)$ of group of kind

$x_1 \to f_1(x_1, x_2, \ldots, x_n)$, $x_2 \to f_2(x_1, x_2, \ldots, x_n), \ldots$, $x_n \to f_n(x_1, x_2, \ldots, x_n)$ acting on the affine space K^n, where $f_i \in K[x_1, x_2, \ldots, x_n]$, $i = 1, 2, \ldots, n$ are multivariate polynomials. Alice chooses a positive integer k_A as her private key and computes the transformation $g(n)^{k_A}$ (multiple iteration of $g(n)$ with itself).

Similarly Bob chooses k_B and gets $g(n)^{k_B}$. Correspondents complete the exchange: Alice sends $g(n)^{k_A}$ to Bob and receives $g(n)^{k_B}$ from him. Now Alice and Bob computes independently common key h as $(g(n)^{k_B})^{k_A}$ and $(g(n)^{k_A})^{k_B}$ respectively. So they can use coefficients of multivariate map $h = g(n)^{k_B k_A}$ from G_n written in the standard form.

There are obvious problems preventing the implementation of this general method in practice. In case $n = 1$ the degree $\deg(fg)$ of composition fg of elements $f, g \in S(K)$ is simply the product of $\deg(f)$ and $\deg(g)$. Such effect can happen in multidimensional case: $(\deg(g))^x = \deg(g^x) = b$. It causes the reduction of discrete logarithm problem for multivariate polynomials to a number theoretical problem. If g is a bijection of degree d and order m then $d^x = b$ in cyclic group Z_m. Similar reduction can appear in the case of other degree functions $s(x) = \deg(g^x)$. If $s(x)$ is a linear function then multivariate discrete logarithm problem with base g is reducible to the solution of linear equation. The degenerate case $\deg(g^x) = \text{const}$ is an interesting one because in such situation the degree function does not help to investigate multivariate discrete logarithm.

We refer to the sequence of multivariate transformations $f(n) \in S(K^n)$ as stable maps of degree d if $\deg(f(n))$ is a constant $d, d > 2$ and $\deg(f(n)^k) \leq d$ for $k = 1, 2, \ldots$ (see [3]). If τ_n is a bijective affine transformation of K^n, i. e. a bijective transformation of degree 1, then the sequence of stable maps $f(n)$ can be changed for other sequence of stable maps $\tau f(n) \tau^{-1}$ of the same degree d.

The first families of special bijective transformations of K^n of bounded degree were generated by *discrete dynamical systems* defined in [4] in terms of graphs $D(n, K)$. In the paper [5] the fact that each transformation from these families of maps is cubic was proven. In [6] authors notice that this family is a stable one, the order of its members grows with the increase of parameter n and suggests key exchange protocols with generators from these families. Other results on the usage of algebraic graphs for construction of families of nonlinear multivariate maps of degree ≤ 3 the reader can find in [7], [8].

Recall that the other important property for the generator $g(n)$ in the described above protocol is a large cardinality of $\{g(n)^k | k = 1, 2, \ldots\}$. Let us assume that $g(n)$ is bijection.

The famous family of linear bijections of ${F_q}^n$ of exponential order is formed by Singer cycles $s(n)$, they have order $q^n - 1$ (see [11], [12] and further references). Statements on the existence of explicit construction of families of nonlinear maps of exponential order are formulated in the section 5 of this paper.

The above mentioned key exchange protocol can be used for the design of the following multivariate El Gamal cryptosystem (see [9], [10]).

Alice takes generator $g(n)$ of the group G_n together with its inverse $g(n)^{-1}$. She sends the transformation $g(n)^{-1}$ to Bob. He will work with the plainspace K^n as public user.

At the beginning of each session Alice chooses her private key k_A. She computes $f = g(n)^{k_A}$ and sends it to Bob.

Bob writes his text $(p_1, p_2, \ldots, p_n)$, chooses his private key k_B and creates the ciphertext $f^{k_B}((p_1, p_2, \ldots, p_n)) = \text{c}$.

Additionally he computes the map $g(n)^{-1} k_B = h(n)$ and sends the pair $(c_1, c_2, \ldots, c_n), h(n)(\text{x})$ to Alice.

Alice computes $h(n)^{k_A}(\text{c}) = (p_1, p_2, \ldots, p_n)$.

REMARK 1. *It is proven (see [9]) that the security level of above multivariate Diffie - Hellman and El Gamal algorithms is the same. It is based on the multivariate discrete logarithm problem on solving the equation $g^x = d$, where g and d are elements of special cyclic subgroup G_n of affine Cremona group.*

IV. On the shifted multivariate El Gamal cryptosystem

We suggest here the following modification of above described algorithm. Alice takes generator $g(n)$ of the group G_n together with its inverse $g(n)^{-1}$. At the beginning of each session Alice chooses her private key k_A and pair $(h(n), h(n)^{-1})$, where $h(n)$ is an element of affine Cremona group. She computes $f = g(n)^{k_A}$ and sends it to Bob together with transformation $m(n) = h(n) g(n)^{-1} h(n)^{-1}$. Public user Bob writes his text $(p_1, p_2, \ldots, p_n)$, chooses his private key k_B and creates the ciphertext $f^{k_B}((p_1, p_2, \ldots, p_n)) = \text{c}$.

Additionally he computes the map $m(n)^{k_B} = a(n)$ and sends the pair $(c_1, c_2, \ldots, c_n), a(n)(\text{x})$ to Alice.

Alice computes $h(n)^{-1} a(n)^{k_A} h(n)(\text{c}) = (p_1, p_2, \ldots, p_n)$.

The shifted algorithm can have better protection against attacks by adversary. One can choose $h(n)$ to make the discrete logarithm problem in affine Cremona group with base $m(n)$ harder than one in a case of base $g(n)^{-1}$. Additionally the adversary has to compute the inverse of $f = g(n)^{k_A}$.

Alice can work with a stable map $g(n)$ of a large polynomial degree and a polynomial density of a large order such that its inverse conjugate with stable map $m(n)$ of prescribed small constant degree d.

REMARK 2. *It is clear, that the algorithm above can be formally considered for the general pair of bijective nonlinear polynomial transformations $g(n)$ and $h(n)$ of affine Cremona group of the free module K^n. But the best computational complexity will be achieved in the case of quadratic stable elements $g(n)$ and $m(n) = h(n) g(n)^{-1} h(n)^{-1}$. In the case of a family $m(n)$ of exponential order corresponding discrete logarithm problem looks as a hard one.*

V. Results on existence of families with prescribed properties

Recall that the density of multivariate polynomial $f \in K[x_1, x_2, \ldots, x_n]$ is its number of monomial terms.

The density of a transformation F of K^n given by rules $x_1 \to f_1(x_1, x_2, \ldots, x_n)$, $x_2 \to f_2(x_1, x_2, \ldots, x_n)$, $\ldots$,

$x_n \to f_n(x_1, x_2, \ldots, x_n)$ is defined as a maximum of densities of f_i, $i = 1, 2, \ldots, n$.

We refer to $F(n) : K^n \to K^n$ as a family of density d if a density of $F(n)$ is estimated by Cn^d, where C is a positive constant. If each transformation $F(n)$ of a density d has constant degree t, then $d \leq t$.

We refer to a family of bijective linear transformation $\tau(n)$ given by rule $(x_1, x_2, \ldots, x_n) \to (x_1, x_2, \ldots, x_n)A$ of affine space K^n as sparse transformations if each row and column of matrix $A(n)$ contains only finite number of nonzero entries and this number is bounded by some positive constant. We refer to $G(n) = \tau(n)F(n)\tau(n)^{-1}$ as a sparse deformation of a family $F(n)$. if one of the families $F(n)$ and $G(n)$ has density d then the density of the other is also d.

THEOREM 1

For each pair (d, T), where $d \leq T$ there is a family $F(n)$ of stable transformations of K^n, $n = k(k+1)$ of degree T and density d of order bounded below by $\sqrt{n} - 1$.

In the case of finite fields we get the following statement.

THEOREM 2.

Let F_q be a finite field. For each pair (d, T), where $1/2 \leq d \leq T$ there is a family $F(n)$ of stable transformations of K^n of degree t, density d and order at least $q^{\sqrt{n}-1} - 1$.

Sketches of proofs ot the theorems 1 and 2 are presented in the section 6. This technique is used also for the proof of the following statement.

THEOREM 3.

Let F_q be a finite field. For each pair (d, T), where $1/2 \leq d \leq T$ there is a family $F(n)$ of stable transformations of ${F_q}^n$, $n = (k+1)^2 - 1$ of degree T, density d and order at least $q^{2k} - 1$.

Each proposition is proven via explicit construction of $F(n)$.

VI. Remarks on the size of numerical parameters and choice of generators

We propose a constructive method of generation of bijective families $F(n)$ of stable bijective multivariate maps of vector space F_q of dimension $n = k(k+1)$ or $n = (k+1)^2 - 1$ of prescribed polynomial degree s, polynomial density $d \geq 1/2$ and exponential order $\geq q^k - 1$. It allows to generate $F(n)$ and its inverse in polynomial time.

We suggest the described above algorithms with the usage of such $F(n)$ defined over F_q and integer parameters k_A and k_B are of size $O(n^{d_A})$ and $O(n^{d_B})$. Such choice insures the polynomial time for the computation of ${F(n)}^{k_A}$, ${F(n)}^{k_B}$ and ${F(n)}^{k_A k_B}$ in the case of key exchange protocol. Notice that density of ${F(n)}^{K_A}$ can be higher than d, it is bounded from above by s. Alice sends the generator $F(n)$, ${F(n)}^{k_A}$ in a standard form to Bob together with parameter d_B which restricts his choice of the private key.

The precise computation of the order of $F(n)$ is a difficult task. In the case of El Gamal algorithm our scheme below allows Alice to compute ${F(n)}^{-1}$ without the knowledge of order of $F(n)$.

Shifted El Gamal cryptosystem requires additional transformation $h(n)$. Notice that the map ${F(n)}^{-1}$ is hidden.

The known function is $M(n) = h(n)F(n)h(n)^{-1}$. The adversary has to solve for k_B the discrete logarithm problem with the base $M(n)$ and given $M(n)^{k_B}$. Our method allows to generate both $F(n)$ and $M(n)$ as stable bijective transformations with prescribed degrees n_F and n_M and densities d_F and d_M.

VII. Double Schubert graphs and automata for the generation of stable maps

We define Double Schubert Graph $DS(k, K)$ over commutative ring K as incidence structure defined as disjoint union of points from $PS = \{(\mathrm{x}) = (x_1, x_2, \ldots, x_k, x_{1,1}, x_{1,2}, \ldots x_{k,k})| \in (\mathrm{x}) \in K^{(k+1)k}\}$ and lines from $LS = \{[\mathrm{y}] = [y_1, y_2, \ldots, y_k, y_{1,1}, y_{1,2}, \ldots y_{k,k}]| \in (\mathrm{y}) \in K^{(k+1)k}\}$ where (x) is incident to [y] if and only if $x_{i,j} - y_{i,j} = x_i y_j$ for $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, k$. It is convenient to assume that indices of kind i, j are placed in lexicographical order.

REMARK. *The term* Double Schubert Graphs *is chosen because points and lines of $DS(k, F_q)$ can be treated as subspaces of ${F_q}^{2k+1}$ of dimensions $k+1$ and k which form two largest Schubert cells. Recall that the largest Schubert cell is the largest orbit of group of unitriangular matrices acting on the variety of subsets of given dimensions.* (see [12] and further references).

We define the colour of point $(\mathrm{x}) = (x_1, x_2, \ldots, x_k, x_{1,1}, x_{1,2}, \ldots x_{k,k})$ from PS as tuple $(x_1, x_2, \ldots, x_k)$ and the colour of line $[\mathrm{y}] = [y_1, y_2, \ldots, y_k, y_{1,1}, y_{1,2}, \ldots y_{k,k}]$ as tuple $(y_1, y_2, \ldots, y_k)$. For each vertex v of $DS(k, K)$ there is a unique neighbour $N_\alpha(\mathrm{v})$ of given colour $\alpha = (a_1, a_2, \ldots, a_k)$, $a_i \in K$, $i = 1, 2, \ldots, k$.

The symbolic colour g from $K[z_1, z_2, \ldots, z_k]^k$ of kind $f_1(z_1, z_2, \ldots, z_k)$, $f_2(z_1, z_2, \ldots, z_k)$, ..., $f_k(z_1, z_2, \ldots, z_k)$, where f_i are polynomials from $K[z_1, z_2, \ldots, z_k]$ defines the neighbouring line of point (x) with colour $(f_1(x_1, x_2, , \ldots, x_k), f_2(x_1, x_2, \ldots, x_k), \ldots, f_k(x_1, x_2, \ldots, x_k)$.

Let us consider a tuple of symbolic colours $(g_1, g_2, \ldots, g_{2t}) \in K[z_1, z_2, \ldots, z_k]^k$ and the map F of PS to itself which sends the point (x) to the end v_{2t} of the chain v_0, v_1, ..., v_{2t}, where $(\mathrm{x}) = \mathrm{v}_0$, $\mathrm{v}_i I \mathrm{v}_{i+1}$, $i = 0, 1, \ldots, 2t-1$ and $\rho(\mathrm{v}_j) = g_j(x_1, x_2, \ldots, x_k)$, $j = 1, 2, \ldots, 2t$. We refer to F as closed point to point computation with the symbolic key $(g_1, g_2, \ldots, g_{2t})$. As it follows from definitions $F = F_{g_1, g_2, \ldots, g_{2t}}$ is a multivariate map of $K^{k(k+1)}$ to itself. When symbolic key is given F can be computed in a standard form via elementary operations of addition and multiplication of the ring $K[x_1, x_2, \ldots, x_k, x_{11}, x_{12}, \ldots, x_{kk}]$. Recall that $(x_1, x_2, \ldots, x_k, x_{11}, x_{12}, \ldots, x_{kk})$ is our plaintext treated as symbolic point of the graph. Let $\mathrm{Sk}(k, K)$ be the totality of all symbolic keys. We define product $(g_1, g_2, \ldots, g_{2t})(h_1, h_2, \ldots, h_{2s})$ of symbolic keys $(g_1, g_2, \ldots, g_{2t})$ and $(h_1, h_2, \ldots, h_{2s})$ as $(g_1, g_2, \ldots, g_{2t}, h_1(g_{2t}), h_2(g_{2t}), \ldots, h_{2s}((g_{2t}))$.This product converts $\mathrm{Sk}(k, K)$ to a semigroup. It is easy to check that the map ${}^k\eta$ sending $(g_1, g_2, \ldots, g_{2t})$ to $F_{g_1, g_2, \ldots, g_{2t}}$ is the

homomorphism of $\mathrm{Sk}(k,K)$ onto $C(K^n)$ where $n=k(k+1)$. We refer to ${}^k\eta$ as *linguistic retraction morphism*.

We write fg to the composition $g(f(\mathrm{x}))$. If (g_1, g_2, ... , g_k) are elements of affine Cremona group $C(K^k)$ then $F_{g_1,g_2,\ldots,g_{2t}}=F_{g_1}F_{{g_1}^{-1}g_2}F_{{g_2}^{-1}g_3}\ldots F_{g_{2t-1}^{-1}g_{2t}}$.

We refer for expression $F_{g_1,g_2,\ldots,g_{2t}}$ as automaton presentation of F with the symbolic key $g_1,g_2,\ldots,g_{2t}$. Notice that if g_{2t} is an element of affine Cremona group $C(K^k)$ then $F_{g_1,g_2,\ldots,g_{2t}}\in C(K^{k(k+1)})$ and automaton presentation of its inverse is $F_{g_{2t}^{-1}g_{2t-1},{g_{2t}}^{-1}g_{2t-2},\ldots,{g_{2t}}^{-1}g_1,{g_{2t}}^{-1}}$.

The restrictions on degrees and densities of multivariate maps g_i of K^k to K^k and size of parameter t allow to define a polynomial map F of polynomial degree and density.

Let us assume that $g_i=({h_1}^i,{h_2}^i,\ldots,{h_k}^i)$, $i=1,2,\ldots,2t$ is the symbolic key of the closed point to point computation $F=F(k)$ of the symbolic automaton $DS(k,K)$. We set that $g_0=({h_1}^0,{h_2}^0,\ldots,{h_k}^0)=(x_1,x_2,\ldots,x_k)$. We set that ${h_1}^0,{h_2}^0,\ldots,{h_k}^0)=(z_1,z_2,\ldots,z_k)$. Then F is a transformation of kind

$z_1\to {h_1}^{2t}(z_1,z_2,\ldots,z_k)$, $z_2\to {h_2}^{2t}(z_1,z_2,\ldots,z_k)$, ..., $z_k\to {h_k}^{2t}(z_1,z_2,\ldots,z_k))$

$z_{11}\to z_{11}-{h_1}^1z_1+{h_1}^1{h_1}^2-{h_1}^3{h_1}^2+{h_1}^3{h_1}^4+\ldots+{h_1}^{2t-1}{h_1}^{2t}$

$z_{12}\to z_{12}-{h_1}^1z_2+{h_1}^1{h_2}^2-{h_1}^3{h_2}^2+{h_1}^3{h_1}^4+\ldots+{h_2}^{2t-1}{h_1}^{2t}$

...

$z_{kk}\to z_{kk}-{h_k}^1z_k+{h_k}^1{h_k}^2-{h_k}^3{h_k}^2+{h_k}^3{h_k}^4+\ldots+{h_k}^{2t-1}{h_k}^{2t}$

LEMMA 1.

The degree of F is bounded by a maximum M of $\gamma_{r,s,i}(n)=\deg({h_r}^i)+\deg({h_s}^{i+1})$, $0\le i\le 2t$, $1\le r\le k$, $1\le s\le k$. The density of F is at most a maximum of $d(r,s)$, where $d(r,s)-1$ is the sum of parameters $\mathrm{den}({h_r}^i)\times\mathrm{den}({h_s}^{i+1})$ for $i=0,1,\ldots,2t$.

We say that closed point to point computation F is balanced if its degree coincides with parameter M of the previous lemma.

LEMMA 2.

If the map $g_{2t}:K^k\to K^k$ is a bijection then the presentation defines one to one transformation of $PS=K^{k(k+1)}$ to itself. The order of F is bounded below by the order of g_{2t}.

LEMMA 3.

If the map $g_{2t}:K^k\to K^k$ is an affine bijective transformation ($\deg(g_{2t})=1$) and the computation is balanced then the map F is stable one two one transformation of $PS=K^{k(k+1)}$ to itself.

PROOF OF THEOREM 1 and 2.

Proof. Theorem 1 and 2 can be deduced from lemmas 1,2 and 3. We assume that parameter t is a constant and $n=(k+1)k$. Let us choose F as $F_{g_1,g_2,\ldots,g_{2t}}$ such that $(g_{2t})\in AFL_n(K))$ and parameter M of Lemma 1 equals T. Other maps g_i, $1\le i\le 2k-1$ can be chosen to keep the density of balanced F in the interval C_1n^d and C_2n^d where C_1 and C_2 are constants, $C_1\le C_2$. In the case of Theorem 1 we can choose g_{2t} as linear permutation map corresponding to cycle of length k. This parameter k gives the lower bound for the order of bijective map F. In the case of theorem 2 we can take Singer cycle in ${F_q}^k$ as g_{2t}. So $|F|\ge g^k-1$.

□

PROOF OF THEOREM 3.

Proof. Let us consider the edge of the graph $DS(2k,F_q)$. It contains a point $(\mathrm{p})=(x_1,x_2,\ldots,x_k,x_{11},x_{12},\ldots,x_{kk})$ and incident line [l] of colour $(y_1,y_2,\ldots,y_k)$. We consider a chain of kind (p), [l], (p_1), [l_1], (p_2), [l_2], ..., (p_s), [l_s] of odd length $2s+1$ such that $\rho(\mathrm{p}_i)=g_i\in F_q[x_1,x_2,\ldots,x_k,y_1,y_2,\ldots,y_k]^k$, $\rho(\mathrm{l}_i)=h_i\in F_q[x_1,x_2,\ldots,x_k,y_1,y_2,\ldots,y_k]^k$, $i=1,2,\ldots,s$. If pair $(x_1,x_2,\ldots,x_k)\to g_{2s}(x_1,x_2,\ldots,x_k,y_1,y_2,\ldots,y_k)$, $(y_1,y_2,\ldots,y_k)\to h_{2s}(x_1,x_2,\ldots,x_k,y_1,y_2,\ldots,y_k)$ defines bijective map Δ of ${F_q}^{2k}$, then the map F sending edge p,l to the edge p_{2s},l_{2s} is bijective transformation of edge set $E(2k,q)$ of $DS(2k,F_q)$. Notice, that $E(2k,q)$ is isomorphic to ${F_q}^{k(k+1)+k}={F_q}^{(k+1)^2-1}$. If we chose Δ as Singer transformation of the vector space ${F_q}^{2k}$, then the order of F will be bounded below by $q^{((k+1)^2-1)}-1$. Similarly to lemma 1 the degree and density of F are maximum of parameters $\deg(g_i)+\deg(h_i)$, $i=0,1,2,\ldots,s$, $\deg(h_i)+\deg(g_{i+1})$, $i=0,1,2,\ldots,s$ ($\deg(g_0)=\deg(h_0)=1$). So appropriate choice of the symbolic key insure that degree of transformation F is T and density d.

□

VIII. Double Schubert automaton as a stable groups generator

We refer to a subgroup G in $S(K^n)$ as a stable subgroup of degree d if the maximal degree for its representative g equals d.

Let $AGS_n(K)$ be the semigroup of affine transformations of K^n, i. e. the group of all transformations of degree 1.

It is easy to see that symbolic keys of kind (g_1, g_2, ..., g_r) of even length from the semigroup $\mathrm{Sk}(k,K)$ with $g_i\in AGS_k(K)$, $i=1,2,\ldots,r-1$ and $g_r\in AGL_k(K)$ form a subgroup. We denote it as $\mathrm{Lk}(k,K)$. The degree of the transformation $F_{g_1,g_2,\ldots,g_r}$ for $<g_1,g_2,\ldots,g_r>$ from $\mathrm{Lk}(k,K)$ is bounded by 2. Let us consider the group $E_k(K)=^k\eta(\mathrm{Lk}(k,K))$. As it follows from lemma 1 group $E_k(K)$ is stable subgroup of degree 2 in $C(K^{n(n+1)})$. The family of groups $E_k(K)$ can be used for the following cryptosystem which can process rather large file. It consists on following protocol, step of exchange of encryption rules and encryption process.

PROTOCOL.

Correspondents use family of group $E_k(F_q)$ for chosen papameters k and q. Alice computes $n=k(k+1)$ and selects affine transformation T from $AGL_n(F_q)$. She computes T^{-1}. Alice selects positive integers t and r together with two strings $\mathrm{a}=(g_1,g_2,\ldots,g_k)$ and $\mathrm{b}=h_1,h_2,\ldots,h_r$ of elements g_i and h_j from $AGS_k(F_q)$ such that g_t and h_r are Singer cycles from $GL_k(F_q)$, i. e. elements of order q^k-1.

She uses the homomotphism $\eta=^k\eta$ from the semigroup $\mathrm{Sk}(k,K)$ onto affine Cremona group $C(K^n)$ and computes $\eta(a)$ and $\eta(b)$. Alice forms elements $G=T\eta(a)T^{-1}$ and $H=$

$T\eta(a)T^{-1}$ of orders at most q^k-1. In fact high orders of these elements are insured by the choice of linear transformations g_t and h_r.

Alice sends to Bob the transformations G and H presented in their standard forms $x_i \to^i g(x_1, x_2, \ldots, x_n)$, $x_i \to^i h(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$ where monomial terms of polynomials ${}^ig$ and j^h are listed in the lexicographical order.

Secondly Alice selects positive constant integers $k_A < q^k - 1$ and $r_A < q^k - 1$. She computes standard form $G_A = H^{r_A}G^{k_A}H^{-r_A}$ and sends it to Bob.

In his turn Bob selects parameters $k_B < q^k - 1$ and $r_A < q^k - 1$. He computes standard form of $Z_B = H^{r_B}{G_A}^{k_B}H^{-r_B}$ and keeps it safely.

Secondly Bob form $G_B = H^{r_B}G^{k_B}H^{-r_B}$ and sends its standard form to Alice. She computes $Z_A = H^{r_A}{G_B}^{k_A}H^{-r_A}$.

Noteworthy that $Z = Z_A = Z_B$ is a collision element of the protocol.

In fact Alice and Bob share an element Z from the stable group $Y(k, F_q) = T^k\eta(\mathrm{Sk}(k, F_q))T^{-1}$ of degree 2.

STEP 2. ENCRYPTION TOOLS EXCHANGE.

Alice takes different from T affine transformation T' such that $TT' \neq T'T$.

She forms $G' = T\eta(a')T'^{-1}$ and $H' = T\eta(b')T'^{-1}$ of orders at least q^k-1. Correspondents again execute the protocol and elaborate another collision map Z'.

Secondly Alice (or Bob) selects extra two elements T_1 and T_2 from $AGL_n(F_q)$ such that elements of each pair selected from $\{T, T', T_1, T_2\}$ does not commute. She takes two strings b and c from $\mathrm{Sk}(k, F_q)$ of length $O(1)$, computes $P = T_1\eta(b){T_1}^{-1}$ and $Q = T_2\eta(b){T_2}^{-1}$ together with their inverses. Finally Alice sends $Z + P$ and $Z' + Q$ to Bob via open channel. He restores P and Q.

ASYMMETRIC ENCRYPTION PROCEDURE.

Before the start of exchange session Alicia and Bob agree on the tuple of integers $(\alpha_1, \alpha_2, \ldots, \alpha_s)$ of length $s = O(1)$ (*password of the session*).

Bob takes his plaintext $\mathrm{p} = (p_1, p_2, \ldots, p_n)$ and applies transformation P to it α_1 times and gets $P^{\alpha_1}(\mathrm{p} =^1 \mathrm{p}$. In similar way he constructs $Q^{\alpha_2}({}^1\mathrm{p}) =^2 \mathrm{p}$. Bob gets ${}^3\mathrm{p}$ via the multiple application of P to ${}^2\mathrm{p}$. Let us assume for simplicity that s is even. Then continuation of the process of recurrent applications of P and Q forms the output $P^{\alpha_1}Q^{\alpha_2}\ldots P^{\alpha_s}(\mathrm{p}) = \mathrm{y}$. So Bob sends the ciphertext y to his partner.

Alice uses natural decryption procedure. She takes reverse word $(\alpha_s, \alpha_{s-1}, \ldots, \alpha_1)$. She applies P^{-1} to the ciphertext y with multiplicity α_s and gets ${}^1\mathrm{y}$, applies Q^{-1} to ${}^1\mathrm{y}$, ... Continuation of this process gives her the plaintext p.

COMPLEXITY ESTIMATES.

Straightforward computation of the number of elementary operations shows that Alice can construct multivariate map G (as well as H, G', H' in $O(k^7) = O(n^{3.5})$. This bound can be used for time evaluation of Step 2.

The execution time of presented above key agreement protocol is determined by the hardest operation to compute the composition of two quadratic maps of dimension $n = k(k+1)$ given in their standard forms. This operation requires $O(n^5)$.

It is easy to see that encryption of single message costs standard for Multivariate Cryptography time $O(n^3)$

ELEMENTS OF CRYPTANALISIS.

In the case of agstract finite group X twisted key agreement protocol with input elements $G \in X$, $H \in X$ and output $Z \in X$ is known instrument of NONCOMMUTATIVE CRYPTOGRAPHY (see [27]-[38]). It based on complexity of Power Conjugacy Problem. Adversary can intercept $G_A = H^{r_A}G^{k_A}H^{-r_A}$ (or $G_B = H^{r_B}G^{k_B}H^{-r_B}$). To break the protocol he / she hast find the presentation of G_A in the form of word of kind $H^xG^yG^{-x}$

Currently algorithm with the joint usage of Turing machine and Quantum computer for breaking this problem in the case of affine Cremona group $X = C(K^n)$ are unknown. So adversary has no chances to break the protocol. During single session of exchanges with the string $(\alpha_1, \alpha_2, \ldots, \alpha_s)$ adversary can estimate the degree of encryption map $U = P^{\alpha_1}Q^{\alpha_2}\ldots P^{\alpha_s}$ as $D = 2^d$, $d = \alpha_1 + \alpha_2 + \ldots + \alpha_s$. He/she can execute interception of n^D pairs plaintext-ciphertext and try to approximate U via costly linearisation attack. (the cost is at least $\geq O(n^{2D+1})$. To prevent this option correspondents can agree on the maximal number $M = cn^{D-1}$ of messages during the session. They can start new session with other selected string without repetition of the protocol.

Noteworthy that increase of s to $O(log_2(k))$ makes linearisation attacks unfeasible. So correspondents can work with unlimited session for which encryption costs $O(n^3log_2(n^{1/2}))$.

REMARK 1. *Alice can increase the number of protocols sessions from 2 to chosen l, $l \geq 2$. It gives her opportunity of safe delivery of noncomputing transformations G_1, G_2, ..., G_l and use this larger set of generators in the described above encryption algorithm.*

REMARK 2. *Alice can take $l = 4$ and $G_3 = {G_1}^{-1}$ and $G_4 = {G_2}^{-1}$. In this case both sides have option to decrypt. So we have symmetric encryption algorithm.*

IX. Conclusion

Algebraic system on $K[x_1, x_2, \ldots, x_n]$, where K is a commutative ring with operations of addition, multiplication and composition is the core part of Computer Algebra. Let $\deg(f)$ be the degree of polynomial $f \in K[x_1, x_2, \ldots, x_n]$, then $\deg(f) + \deg(g) = \max(\deg(f), \deg(g))$. The general formula for $\deg(f(g))$ does not exist, only inequality $\deg(f(g)) \leq \deg(f)\deg(g)$ holds. The addition and multiplication of n polynomials from $K[x_1, x_2, \ldots, x_n]$ of bounded degree can be computed in polynomial time but there is no polynomial algorithm for the execution of the computation of n elements from $K[x_1, x_2, \ldots, x_n]$. It means that in Cremona semigroup $S(K^n)$ of all endomorphisms of $K[x_1, x_2, \ldots, x_n]$ the computation of the product of n representatives is unfeasible task. Noteworthy that each endomorphism $F \in S(K^n)$ is defined by its values f_i on x_i and can be identified with the rule $x_i \to f_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$, where f_i is given via the list of its monomial terms written in the lexicographical order. Noteworthy that the semigroup $S(K^n)$ and its subgroup $C(K^n)$ of all automorphisms of $K[x_1, x_2, \ldots, x_n]$

are core objects of Multivariate Cryptography (MC). Classical Multivariate Cryptography considers only compositions of kind T_1FT_2 of single nonlinear element F of small degree (2 or 3) with linear bijective endomorphisms T_1 and T_2 of degree 1 because of the heavy complexity for the computation of compositions.

Discovery of large stable subsemigroups X of $S(K^n)$ of degree bounded by constant degree d gives new option. One can compute the composition of n representatives of X in polynomial time. So Diffie-Hellman protocol or its modifications with generator from X are possible. Security of them requires further investigations. The cases $d = 2, 3$ has computational advantage because the composition of two nonlinear map can be computed in time $O(n^5)$ and $O(n^{13})$. The existence of implemented models of quantum computers even with restricted number of cubits stimulates studies of analogs of Diffie - Hellman protocol with at least two generators $g_1, g_2, \ldots, g_s$ and noncommutative semigroup $X =< g_1, g_2, \ldots, g_s >$. Current paper contains description of such protocol in the case of stable subgroups X of degree 2 in $S(K^n)$. The security of this algorithms rests on hard Power Conjugacy Problem. In the case $K = F_q$ one can select generators of exponential order. For the construction of Postquantum Secure Cryptosystem we combine this protocol with asymmetrical encryption algorithm , which allows execution of encryption for Bob and decryption for Alice in times $O(n^3)$ and $O(n^2)$ respectively. We consider the way to convert encryption procedure into symmetrical algorithm in previous section.

Other cryptosystems with the same platform ot its expansion are presented in [39], [40]. They use Word Decomposition Problem instead Power Conjugacy Problem. The implementations of such protocol of Noncommutative Cryptography for stable subgroups X of degree 3 is described in [24].

REFERENCES

[1] J. Ding, J. E. Gower, D. S. Schmidt, *Multivariate Public Key Cryptosystems*. Springer, Advances in Information Security, V. 25, 2006.

[2] J. Porras, J. Baena, J. Ding *New Candidates for Multivariate Trapdoor Functions*, Revista Colombiana de Matematicas 49(1):57-76 (November 2015).

[3] V. A. Ustimenko, *Explicit constructions of extremal graphs and new multivariate cryptosystems*, Studia Scientiarum Mathematicarum Hungarica, Special issue "Proceedings of The Central European Conference, 2014, Budapest", volume 52, issue 2, June 2015, pp. 185-204.

[4] V. Ustimenko, *Linguistic Dynamical Systems, Graphs of Large Girth and Cryptography*, Journal of Mathematical Sciences, Springer, vol.140, N3 (2007) pp. 412-434.

[5] A. Wróblewska, On some properties of graph based public keys, Albanian Journal of Mathematics, Volume 2, Number 3, 2008, 229-234, NATO Advanced Studies Institute: "New challenges in digital communications".

[6] V. Ustimenko, A. Wróblewska, *On the key exchange with nonlinear polynomial maps of stable degree*, Annalles UMCS Informatica AI XI, 2 (2011), 81-93.

[7] V Ustimenko, U. Romańczuk, *On Dynamical Systems of Large Girth or Cycle Indicator and their applications to Multivariate Cryptography*, in "Artificial Intelligence, Evolutionary Computing and Metaheuristics ", In the footsteps of Alan Turing Series: Studies in Computational Intelligence, Vol. 427, Springer, January , 2013, 257-285.

[8] V. Ustimenko, A. Wróblewska, *Dynamical systems as the main instrument for the constructions of new quadratic families and their usage in cryptography*, Annales UMCS Informatica AI, ISSN 1732-1360, vol.12, N3 (2012), 65-74.

[9] M. Klisowski, *Zwiększenie bezpieczeństwa kryptograficznych algorytmów wielu zmiennych bazujących na algebraicznej teorii grafów*, PhD thesis, Częstochowa, 2014.

[10] M. Klisowski, V. Ustimenko, *Graph based cubical multivariate maps and their cryptographical applications*, in " Advances on Superelliptic curves and their Applications", IOS Press, NATO Science for Peace and Security series - D: Information and Communication Security, vol 41, 2015 , pp. 305 -327.

[11] L. Babai, *Graph Isomorphism in Quasipolinomial Time*, arXive: 1512 03547v1 [cs. DS], 11 Dec 2015.

[12] V. Ustimenko, *On Schubert cells in Grassmanians and new algorithms of multivariate cryptography*, Proceedings of the Institite of Mathematics, Minsk, 2015, Volume 23, N 2, pp. 137-148 (Proceedings of international conference "Discrete Mathematics, algebra and their applications", Minsk, Belarus, September 14-18, 2015, dedicated to the 100th anniversary of Dmitrii Alexeevich Suprunenko).

[13] A. Cossidente, M. J. de Ressmine, *Remarks on Singer Cycle Groups and Their Normalizers*, Designs, Codes and Cryptography, 32, 97-102, 2004.

[14] W. Kantor, *Linear groups containing a Singer cycle*, J. of Algebra 62, 1982, 232-234.

[15] V. A. Ustimenko, *On the Families of Stable Multivariate Transformations of Large Order and Their Cryptographical Applications*, Tatra Mountains Mathematical Publications,2O17, 70(1), pp 107-117.

[16] V. A. Ustimenko, *On algebraic graph theory and non–bijective maps in cryptography*, Algebra and Discrete Mathematics, Volume 20 (2015). Number 1, pp. 152-170.

[17] V. A. Ustimenko, *On the hidden discrete logarithm for some polynomial stream ciphers*, International Multiconference on Computer Science and Informational Technology, 20-22 October 2008, Wisla, Poland, CANA Proceedings

[18] M. Klisowski, V. Ustimenko, *On the public keys based on the extremal graphs and digraphs*, International Multiconference on Computer Science and Informational Technology, October 2010, Wisla, Poland, CANA Proceedings, 12 pp.

[19] J. Kotorowicz, U. Romańczuk, V. Ustimenko, *Implementation of stream ciphers based on a new family of algebraic graphs*, Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, 13.

[20] L A. Grzesik, D. Kr´al', L. M. Lov´asz, *Elusive extremal graphs*, preprint (2018), ar-Xiv:1807.01141.

[21] N Hoory, A. Linial, A.Wigderson. *Expander graphs and their applications*, Bull. Amer. Math Soc., 43, pp 439-561, 2006.

[22] M. Polak, U. Romańczuk, V. Ustimenko, A. Wróblewska, *On the applications of Extremal Graph Theory to Coding Theory and Cryptography*, Electronic Notes in Discrete Mathe-matics, N 43, p. 329-342.

[23] V. Ustimenko, U. Romanczuk-Polubiec, A. Wroblewska, M. Polak, E. Zhupa, *On the constructions of new symmetric ciphers based on non-bijective multivariate maps of prescribed degree*, Security and Communication Networks, Volume 2019, Article ID 2137561, 15pages https://doi.org/10.1155/2019/2137561

[24] V. Ustimenko, M. Klisowski, *On Noncommutative Cryptography with cubical multivariate maps of predictable density*, Proceedings of ''Computing 2019" conference, London, 16-17, July , Volume 2, Part of Advances in Intelligent Systems and Computing (AISC), volume 99, pp, 654-674.

[25] V. Ustimenko, U. Romanczuk-Polubiec, A. Wroblewska, *Expanding Graph of the Extremal Graph Theory and expanded platforms of post-quantum cryptography*, Annals of Computer Science and information Systems 2019, vol.19, pp 41-46.

[26] N.Alon, *Eigenvalues, geometric expanders, sorting in rounds, and Ramsey Theory*, Combinatorica, 6 (3), 1986, pp.207-219.

[27] D. N. Moldovyan, N. A. Moldovyan, *A New Hard Problem over Non-commutative Finite Groups for Cryptographic Protocols*, International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2010: Computer Network Security pp 183-194.

[28] V. Shpilrain, A. Ushakov, *The conjugacy search problem in public key cryptography: un-necessary and insufficient*, Applicable Algebra in Engineering, Communication and Computing, August 2006, Volume 17, Issue 3–4, pp 285–289.

[29] Delaram Kahrobaei, Bilal Khan, *A non-commutative generalization of ElGamal key exchange using polycyclic groups*, In IEEE GLOBECOM 2006 - 2006 Global Telecommu-nications Conference [4150920] DOI: 10.1109/GLOCOM.2006.

[30] Alexei Myasnikov, Vladimir Shpilrain, Alexander Ushakov (2008). *Group-based Cryptography*. Berlin: Birkhäuser Verlag.
[31] Zhenfu Cao, *New Directions of Modern Cryptography*. Boca Raton: CRC Press, Taylor & Francis Group (2012). ISBN 978-1-4665-0140-9.
[32] G.Maze., C. Monico, J. Rosenthal, *Public key cryptography based on semigroup actions*. Adv.Math. Commun. 1(4), 489–507 (2007).
[33] P.H. Kropholler, S.J. Pride , W.A.M. Othman K.B. Wong, P.C. Wong, *Properties of cer-tain semigroups and their potential as platforms for cryptosystems*, Semigroup Forum (2010) 81: 172–186.
[34] Gautam Kumar and Hemraj Saini, *Novel Noncommutative Cryptography Scheme Using Extra Special Group*, Security and Communication Networks , Volume 2017, Article ID 9036382, 21 pages, https://doi.org/10.1155/2017/9036382.
[35] V. A. Roman'kov, *A nonlinear decomposition attack*, Groups Complex. Cryptol. 8, No. 2 (2016), 197-207.
[36] V. Roman'kov, *An improved version of the AAG cryptographic protocol*, Groups, Complex., Cryptol, 11, No. 1 (2019), 35-42.
[37] A. Ben-Zvi, A. Kalka and B. Tsaban, *Cryptanalysis via algebraic span*, In: Shacham H. and Boldyreva A. (eds.) Advances in Cryptology - CRYPTO 2018 - 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2018, Proceedings, Part I, Vol. 10991, 255-274, Springer, Cham (2018).
[38] B. Tsaban, *Polynomial-time solutions of computational problems in noncommutative-algebraic cryptography*, J. Cryptol. , 28, No. 3 (2015), 601-622.
[39] V. Ustimenko, *On desynchronised multivariate algorithms of El Gamal type for stable semigroups of affine Cremona group*, Theoretical and Applied Cybersecurity, National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute", vol 1, 2019, pp. 22-30.
[40] V. Ustimenko, *On the usage of postquantum protocols defined in terms of transformation semi-groups and their homomorphisma*, Theoretical and Applied Cybersecurity, National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute", vol 2, 2020, pp. 32-44.

14^{th} International Symposium on Multimedia Applications and Processing

SOFTWARE Engineering Department, Faculty of Automation, Computers and Electronics, University of Craiova, Romania "Multimedia Applications Development" Research Centre

Background and Goals

Multimedia and information have become ubiquitous on the web and communication services, creating new challenges for detection, recognition, indexing, access, search, retrieval, automated understanding, processing and generation of several applications which are using image, signal or various multimedia technologies.

Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices—such as laptops, iPods, personal digital assistants (PDA), and smartphones—have stimulated the rapid development of intelligent applications. These key technologies by using Virtual Reality, Augmented Reality and Computational Intelligenceare creating a recent multimedia revolution which will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain.

We welcome papers covering innovative applications, practical usage but also theoretical aspects of the above mentioned trends. The key objective of this sessionis to gather results from academia and industry partners working in all subfields of multimedia and language: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications—including middleware—along with the related security issues. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages including home-networked entertainment and games. However, innovative contributions which don't exactly fit into these areas are also welcomed to this session.

The Multimedia Applications and Processing (MMAP) will provide an opportunity for researchers and professionals to discuss present and future challenges as well as potential collaboration for future progress in the field. The MMAP Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference. Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

Call for Papers

MMAP 2020 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAP 2020 Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP 2020 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference. Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

Topics

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Big Data Science and Multimedia Systems
- Cloud Computing and Multimedia Applications
- Machine Learning, Fuzzy Systems, Neural Networks and Computational Intelligence for Information Retrieval in Multimedia Applications
- Data Mining, Warehousing and Knowledge Extraction
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Multimedia in Internet and Web Based Systems
- E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications and Computational biology
- Entertainment, Personalized Systems and Games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture and Fuzzy Logic Systems
- Intelligent Multimedia Network Applications
- Future Trends in Computing System Technologies and Applications
- Trends in Processing Multimedia Information

BEST PAPER AWARD

A best paper award will be made for work of high quality presented at the MMAP Symposium. Award comprises a certificate for the authors and will be announced on time of conference. Selected papers will be invited to high IF journals organized for the participants of MMAP.

- Authors should submit draft papers (as Postscript, PDF or MSWord file).
- The total length of a paper should not exceed 10 pages IEEE style (including tables, figures and references). IEEE style templates are available here.
- Papers will be refereed and accepted on the basis of their scientific merit and relevance to the workshop.
- Preprints containing accepted papers will be published on a USB memory stick provided to the FedCSIS participants.
- Only papers presented at the conference will be published in Conference Proceedings and submitted for inclusion in the IEEE Xplore®database.
- Conference proceedings will be published in a volume with ISBN, ISSN and DOI numbers and posted at the conference WWW site.
- Conference proceedings will be indexed in BazEkon and submitted for indexation in: Thomson Reuters—Conference Proceedings Citation Index, SciVerse Scopus, Inspec, Index Copernicus, DBLP Computer Science Bibliography and Google Scholar
- Extended versions of selected papers presented during the conference will be published as Special Issue(s).
- Organizers reserve right to move accepted papers between FedCSIS events.

ADVISORY BOARD

- **Neustein, Amy,** Boston University, USA, Editor of Speech Technology
- **Jain, Lakhmi C.,** University of South Australia and University of Canberra, Australia
- **Zurada, Jacek,** University of Louisville, United States
- **Ioannis, Pitas,** University of Thessaloniki, Greece
- **Badica, Costin,** University of Craiova, Romania
- **Borko, Furht,** Florida Atlantic University, USA
- **Kosch, Harald,** University of Passau, Germany
- **Uskov, Vladimir,** Bradley University, USA
- **Deserno, Thomas M.,** Aachen University, Germany
- **Burdescu, Dumitru Dan,** University of Craiova, Romania

TECHNICAL SESSION CHAIR

- **Korzhik, Valery,** State University of Telecommunications, Russia
- **Schiopoiu Burlea, Adriana,** University of Craiova, Romania

PROGRAM COMMITTEE

- **Azevedo, Ana,** CEOS.PP-ISCAP/IPP, Portugal
- **Badica, Amelia,** University of Craiova, Romania
- **Cano, Alberto,** Virginia Commonwealth University, United States
- **Cordeiro, Jose,** EST Setúbal/I.P.S.
- **Cretu, Vladimir,** Politehnica University of Timisoara, Romania
- **Debono, Carl James,** University of Malta, Malta
- **Fabijańska, Anna,** Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland
- **Fomichov, Vladimir,** National Research University Higher School of Economics, Moscow, Russia., Russia
- **Giurca, Adrian,** Brandenburg University of Technology, Germany
- **Grosu, Daniel,** Wayne State University, United States
- **Kabranov, Ognian,** Cisco Systems, United States
- **Keswani, Dr. Bright,** Suresh Gyan Vihar University, Mahal, Jagatpura, Jaipur
- **Kotenko, Igor,** St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia
- **Logofatu, Bogdan,** University of Bucharest, Romania
- **Mangioni, Giuseppe,** DIEEI - University of Catania, Italy
- **Marghitu, Daniela,** Auburn University
- **Mihaescu, Cristian,** University of Craiova, Reunion
- **Mocanu, Mihai,** University of Craiova, Romania
- **Ohzeki, Kazuo,** Professor Emeritus at Shibaura Institute of Technology, Japan
- **Pohl, Daniel,** Intel, Germany
- **Popescu, Dan,** CSIRO, Sydney, Australia, Australia
- **Popescu, Daniela E.,** Integrated IT Management Service, University of Oradea
- **Querini, Marco,** Department of Civil Engineering and Computer Science Engineering
- **Radulescu, Florin,** University "Politehnica" of Bucharest
- **Romansky, Radi,** Professor at Technical University of Sofia
- **Rutkauskiene, Danguole,** Kaunas University of Technology
- **Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt
- **Sari, Riri Fitri,** University of Indonesia, Indonesia
- **Scherer, Rafał,** Częstochowa University of Technology, Poland
- **Sousa Pinto, Agostinho,** Instituto Politécnico do Porto, Portugal
- **Stanescu, Liana,** University of Craiova, Romania
- **Stoicu-Tivadar, Vasile,** University Politehnica Timisoara
- **Trausan-Matu, Stefan,** Politehnica University of Bucharest, Romania
- **Trzcielinski, Stefan,** Poznan University of Technology, Poland
- **Tsihrintzis, George,** University of Piraeus, Greece
- **Tudoroiu, Nicolae,** John Abbott College, Canada
- **Vega-Rodríguez, Miguel A.,** University of Extremadura, Spain
- **Virvou, Maria,** University of Piraeus, Greece
- **Watanabe, Toyohide,** University of Nagoya

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 133–139
DOI: 10.15439/2021F33
ISSN 2300-5963 ACSIS, Vol. 26

A Comparison between a Relational and a Graph Database in the Context of a Recommendation System

Liana Stanescu
University of Craiova, Faculty of Automation, Computers and Electronics, Craiova, Romania
Email: stanescu@software.ucv.ro

***Abstract*—This paper presents a comparison between relational and graph-based database systems' performance in a modern web application recommendation system. The comparison is conducted on five different queries starting with simple ones, leading up to more complex queries, that are performed in a typical web social application. The implementation is done in C# using .NET framework and the database systems used are SQL Server and Neo4j. For the comparative study we used a database designed in the context of a recommendation system for a culinary application. In order to effectively test the performance of both graph and relational database systems, tests were performed on four data sets that contain 350.000, 700.000, 1.400.000 and 2.100.000 entries. The tests imply performing five different retrieval queries taken in order of difficulty both in SQL and Neo4J.**

I Introduction

WEB applications and mobile applications have gained popularity in recent years, being user-friendly, offering commodity and an easy to use environment for research, reading, buying and so on.

Considering the fact that users often prefer the use of a mobile or a web application, over physical resources, for quick information search, the creation of web applications that rapidly deliver information based on user filtering seems natural and has become well spread.

However, this might not be enough for users, who are eager to rapidly learn about an item based on their preferences, without having to search for particular criteria lead the path for development of more complex recommendation systems.

For a majority of people, especially individuals living in an urban environment, time-consuming activities retain them from spending time researching. This can be avoided by the development of online web application that offers fast and innovative ideas based on users' habits. The recommendations, in the form of responses to users, need to be delivered fast, no matter how complex the application becomes, as it is a requirement implied by the fast-paced living era.

A critical aspect to consider is the database where the information will be stored. There are classical solutions like relational databases or the more recent NoSQL solutions, as graph databases. A well conducted research is mandatory when choosing the database fit for the application [1].

Different aspects such as what type of database will fit the application, what kind of structure the database will have and how fast will it be able to deliver data to the users depending on its structure are important to be taken into account.

Because in the literature we found few similar studies, we decided to experiment with the use of both a relational database and a graph type to analysis which is in this context the appropriate solution.

The paper is thus organized: section 2 presents related work, section 3 briefly introduces graph databases, section 4 contains the experiments and section 5 shows the conclusions.

II Related Work

In the last few years, the focus shifted from typical applications to ones that are focused on the users and their preferences. Clearly, the most used applications nowadays are social media platforms, so the attention shifted from relational databases to more appropriate database systems. There is research done is this area, since graph databases are gaining more and more ground every day and choosing the proper system has an enormous impact on the application's functionality and response time [6].

Surajit Medhi and Hemanta K. Baruah in [7] create a similar comparison between a relational and graph database performance on a simple Cricket application reaching similar results in favor of the Neo4J system. However, their tests are performed on only 400 objects and 3 queries with a decreased difficulty.

In [8] the authors present a similar comparative study in the context of a cancer treatment application. They compared the performance of a relational database implemented in MySQL and a graph database implemented in Neo4J. The comparison was made on twelve queries and three datasets: 1000, 10.000

and 100.000 records. The results indicate that MySQL performs better than Neo4J in most cases, but Neo4J is better when the queries involve multiple joins between tables and the number of records is 100.000.

In [11] the authors review the literature of recent years that have analyzed in detail the NoSQL databases and relational ones in order to highlight the characteristics of each type of database, especially for NoSQL technology that appears as a new solution over relational databases.

Another article [10] presents the results of the comparison between Oracle relational database and NoSQL graph database. The comparison was made in two directions: the first direction aimed at executing queries in the types of databases and the second direction involved performing a predictive analysis on the experimental results.

Given that both relational and graph databases can manage both relational and graph data, other researchers have tried to establish the limitations of these two technologies. In [9] they present their conclusions of the experiments that involved a unified benchmark for relational and graph databases over the same datasets using the same queries and the same metrics.

In a more recent article, the authors compared the performance of MySQL and Neo4J databases regarding the memory usage and execution time. The results highlighted the following: MySQL has a faster execution time than Neo4J, both these databases have the same time complexity, Neo4J has a higher memory usage than MySQL and Neo4J has better flexibility than MySQL [12].

This activity of comparing the relational and NoSQL databases is a current concern, as it is clear that there are applications for which relational databases are the best solution, while for other types of applications, new NoSQL technologies are preferred.

III. RELATIONAL DATABASES VS GRAPH DATABASES

Relational databases have been the basis of software applications since the 1980s, and still are [5]. Relational databases store data in a well-structured format within tables consisting of columns of certain data types and rows of those defined data types [5]. Relational databases require designers and applications to strictly structure the data used in their applications. Relational data is stored in tables, and the relationships between them are made simply through referential integrity which involves the presence of the external key that refers to a primary key [5]. To retrieve the data from several linked tables, the JOIN operation is used at query time by matching primary and foreign keys of all rows in the connected tables. These operations involve large processing capacity and memory usage, having an exponential cost [5]. If the data modeling implies the existence of many to many relationships, in the relational model will appear an additional table, a so called join table with two, or more external keys, to the initial participating tables, which further increases the cost of the JOIN operation [5].

NoSQL databases have appeared, aiming to cover certain requirements of users and applications, but many of them still did not satisfy the data links optimally. Hence the need for graph databases, that are the best choice for modeling the modern world we live in [1], [5].

In the graph data model, the relationships are as important as the data themselves. As a result, there is no need to implement the links between the entities using additional concepts, such as external keys [1], [5].

Graph databases allow the creation of models that map well to the problems to be solved. In this type of database, the data looks very similar to those in the modeled mini-world, small, normalized and keeping connections. The user can query and view the data from any point of view [1], [5].

In the graph database model, each node, either entity or attribute, has a list of link records that model the links to other nodes. These relationship records are organized by type and direction and may have additional attributes. [1], [5].

This list is used by graph databases, when running an operation equivalent to the JOIN operation in the relational model, to access the connected nodes, eliminating expensive computations. In graph databases, traversing the joins or relationships is very fast because they are not calculated at query time as they are persistent [1], [5].

Neo4J is a graph database system implemented in Java and the access to data is done with Cypher Query [3], [4]. It is an ACID-compliant transactional database with native graph storage and processing [1], [5]. The relationships are materialized at creation time, which results in no penalties for complex runtime queries.

Neo4J implements the Property Graph Model in an efficient way [3] (figure 1). The property graph model is an extension of the graphs from mathematics. The following concepts are used to model a property graph:

- Nodes that are the entities in the graph
- Labels that are used to represent the role of the node; a node can have multiple labels at the same time
- Relationships that describe directed, semantically connections between two nodes
- Properties that are key-value pairs that contain information about the node or relationship.

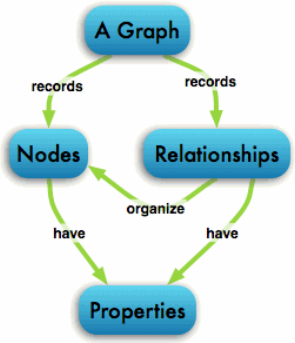


Fig. 1. Block Diagram of System Modules

The query of the relational databases is done with the help of SQL – a declarative query language. SQL commands can be used within the interfaces provided by relational database management systems, or they can be nested in an application and sent for execution to the database engine [5].

Cypher is also a declarative graph query language which is based on the basic concepts and clauses of SQL but which added a multitude of additional features specific to graphs to make it easier to work with the graph model. For describing visual patterns in graphs it uses ASCII-Art syntax. Using Cypher users can build expressive and efficient queries on graph databases [2], [5].

IV. EXPERIMENTS AND RESULTS

In order to efficiently test the performances of both graph and relational database systems, there were performed tests on four data sets as in Table I. The tests represent performing five different retrieval queries taken in order of difficulty both in SQL and Neo4J.

TABLE I. Data Sets

Set Number	Number of entries
1	350.000
2	700.000
3	1.400.000
4	2.100.000

The dataset on which the tests were performed is represented by a culinary web application and its structure can be seen in the figures below in both database systems: MS SQL Server and Neo4J.

The implementation of the application began with the development of the SQL database (figure 2) that was later exported as CSV files and imported into Neo4J (figure 3).

The database contains tables that store data about different types of ingredients, culinary recipes, and join tables that resulted from many to many relationships between data (figure 2).

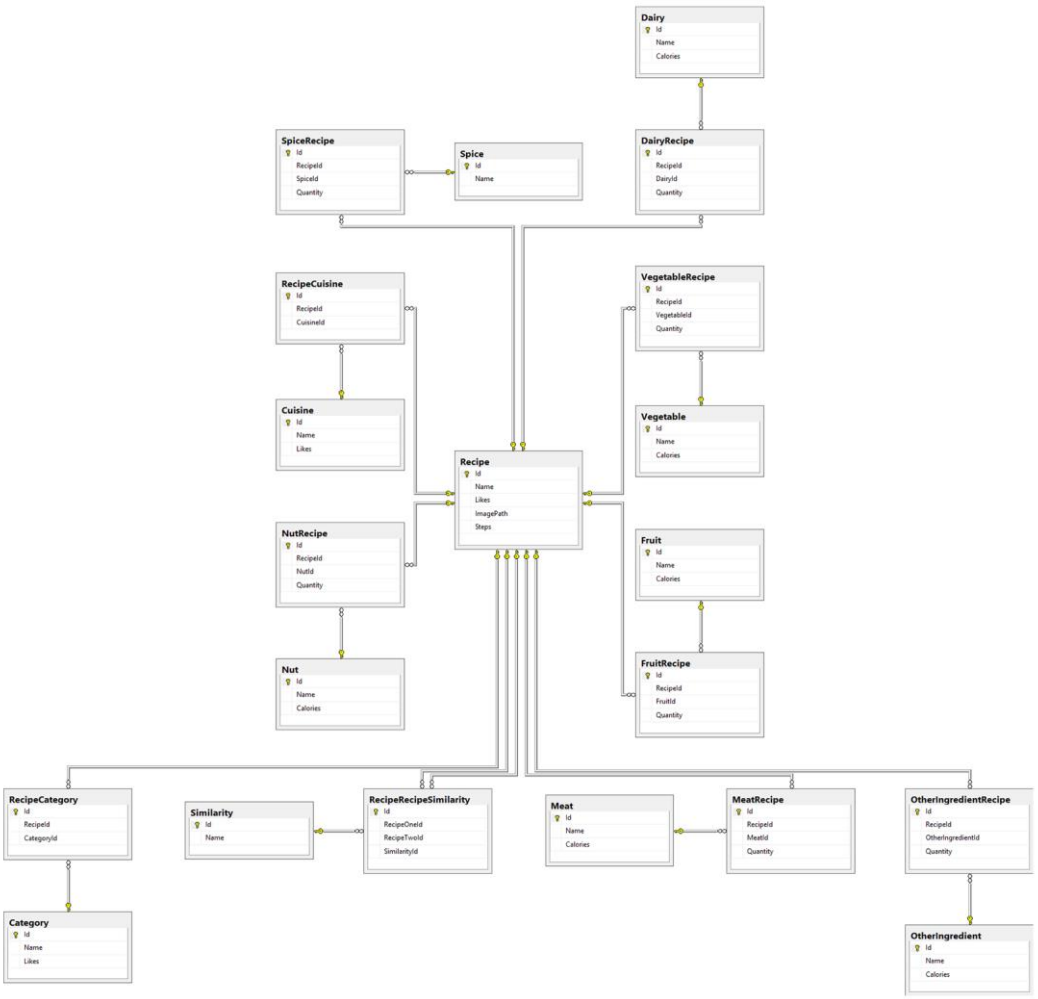


Fig. 2. Culinary App database - MS SQL Server

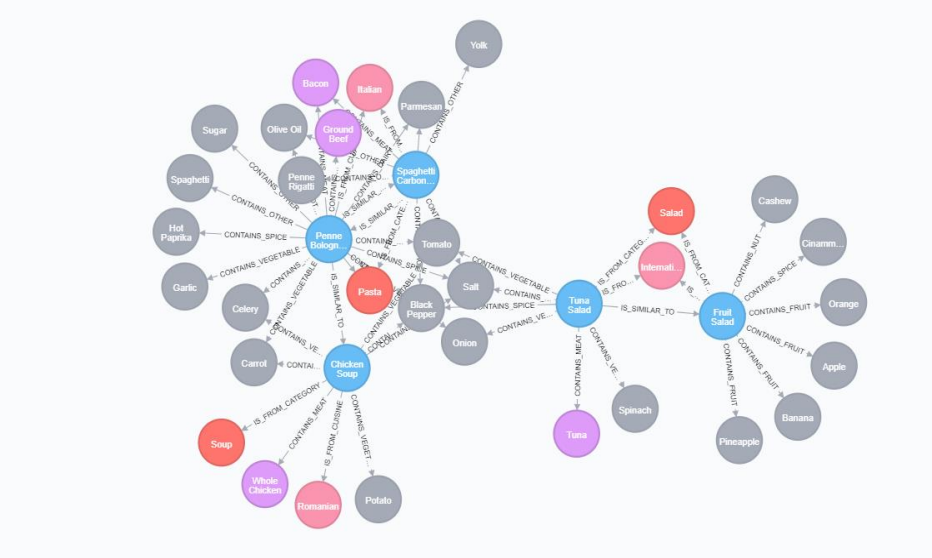


Fig. 3. A sample of Culinary App database - Neo4J

In this sample (figure 3) we can see a number of nodes and edges that represent relationships. For example the node „Chicken Soup" is related by node „Soup" with an edge called „is_from_category", by node „Romanian" with the relationship "is_from_cuisine", etc.

The tests were performed on a personal computer, in the application's solution developed in Microsoft Visual Studio 2017. The running time of the methods was measured using the Stopwatch class from the System Diagnostics namespace in the .NET Framework.

PC Configuration:

- CPU: Intel I5 @ 3.40GHz
- RAM: 8.00 GB
- OS: Windows 10 x64
- SQL Database System – SQL SERVER 2019
- Graph Database System – Neo4J Database 4.0

The connection to the SQL Database was made using Entity Framework and the connection to the Neo4J Database was possible using Neo4J Driver and Neo4J Client libraries.

Experiment 1:

Query: Get all recipes containing "Bacon"

SQL Syntax

SELECT DISTINCT recipes
FROM Recipes IN Recipe TABLE
JOIN MeatRecipe IN MeatRecipe TABLE
ON RecipesId EQUALS MeatRecipe.RecipeId
WHERE MeatRecipe.MeatId EQUALS "Bacon"

Neo4j Syntax

MATCH (Recipe)- [CONTAINS MEAT]-> (Meat {"Bacon"})
return Distinct Recipe

This query involves the join of two tables and a condition. The results of the comparison appear in figure 4. It can be seen that for all four data sets, the query execution was faster in MS SQL server than in Neo4J.

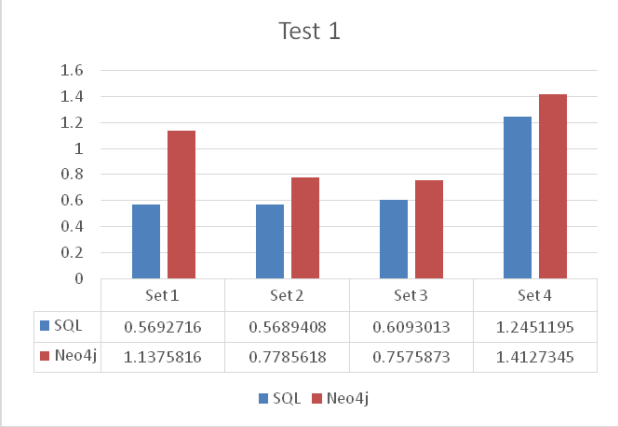


Fig. 4. Experiment 1 results

Experiment 2:

Query: Get all recipes containing "Bacon" from the "Italian" Cuisine

SQL Syntax

SELECT DISTINCT recipes
FROM Recipes IN Recipe TABLE
JOIN MeatRecipe IN MeatRecipe TABLE
ON RecipesId EQUALS MeatRecipe.RecipeId
WHERE MeatRecipe.MeatId EQUALS "Bacon"
JOIN RecipeCuisine IN RecipeCuisine TABLE
ON RecipesId EQUALS RecipeCuisine.RecipeId
WHERE RecipeCuisine.CuisineId EQUALS "Italian"

Neo4j Syntax

MATCH (Recipe)- [CONTAINS MEAT]-> (Meat {"Bacon"}),
(Recipe)- [IS_FROM_CUISINE]-> (Cuisine {"Italian"})
return Distinct Recipe

Query number 2 involves the join of three tables and two conditions on data. The results for this experiment appear in figure 5. In this case, in which we increased the number of joined tables, the execution time in Neo4j decreased a lot. Neo4j outperformed MS SQL server. The gap between the two systems grew larger as the number of records increased.

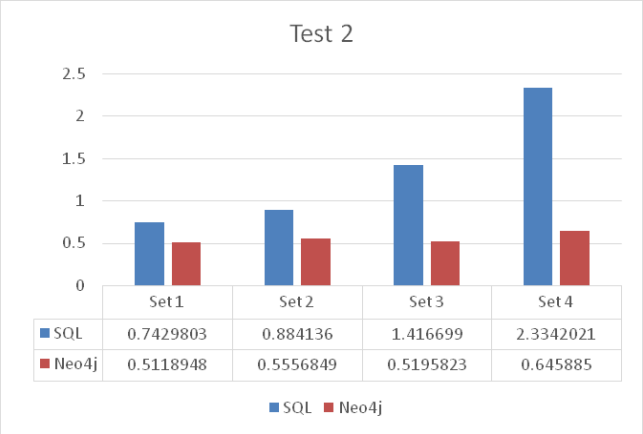


Fig. 5. Experiment 2 results

Experiment 3:

Query: Get all recipes containing "Bacon" from the "Italian" Cuisine from the "Pasta" Category

SQL Syntax

SELECT DISTINCT recipes
FROM Recipes in Recipe TABLE
JOIN MeatRecipe in MeatRecipe TABLE
ON Recipes.Id equals MeatRecipe.RecipeId
WHERE MeatRecipe.MeatId EQUALS "Bacon"
JOIN RecipeCuisine in RecipeCuisine TABLE
ON Recipes.Id equals RecipeCuisine.RecipeId
WHERE RecipeCuisine.CuisineId EQUALS "Italian"
JOIN RecipeCategory in RecipeCategory TABLE
ON Recipes.Id equals RecipeCategory.RecipeId
WHERE RecipeCategory.CategoryId EQUALS "Pasta"

Neo4j Syntax

MATCH (Recipe)- [CONTAINS_MEAT]-> (Meat {"Bacon"}),
(Recipe)- [IS_FROM_CUISINE]-> (Cuisine {"Italian"}),
(Recipe)-[:IS_FROM_CATEGORY]-> (Category {"Pasta"})
return Distinct Recipe

Experiment number 3 represents an even more complex query defined on four tables (three joins) and two conditions. The results for experiment 3 appear in figure 6. The same observation can be made as in experiment 2. Neo4j executes the query much faster than MS SQL Server, and moreover, the execution time decreases drastically for the graph database system.

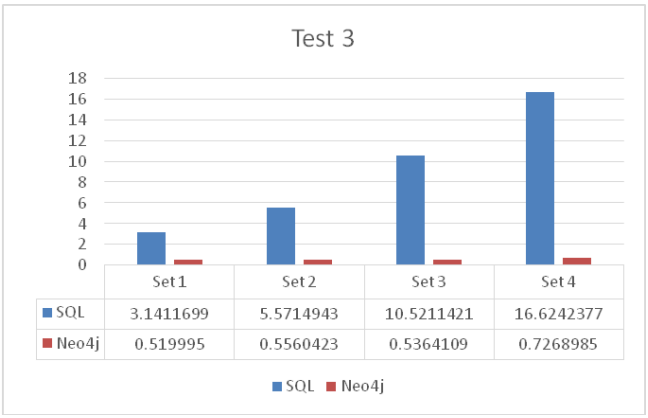


Fig. 6. Experiment 3 results

Experiment 4:

Query: Get all recipes similar to "Example Recipe" containing "Tomato"

SQL Syntax

SELECT DISTINCT recipes
FROM Recipes IN Recipe TABLE
JOIN RecipesSimilarity IN RecipeRecipeSimilarity TABLE
ON Recipes.Id EQUALS RecipesSimilarity.RecipeTwoId
WHERE RecipesSimilarity.SimilarityId EQUALS "Strong" AND RecipesSimilarity.RecipeOneId EQUALS "Example Recipe" JOIN VegetableRecipe IN VegetableRecipe TABLE ON Recipes.Id EQUALS VegetableRecipe.RecipeId WHERE VegetableRecipe.VegetableId EQUALS "Tomato"

Neo4j Syntax

MATCH (Recipe {"Example Recipe"})- [similarity: IS_SIMILAR_TO {Similarity: "Strong"}]-> (Other Recipe),
(Other Recipe)-[CONTAINS_VEGETABLE]-> (Vegetable {"Tomato"})
return Distinct Other Recipe

The results for experiment 4 appear in figure 7. Again, Neo4J is superior to MS SQL server for all four data sets.

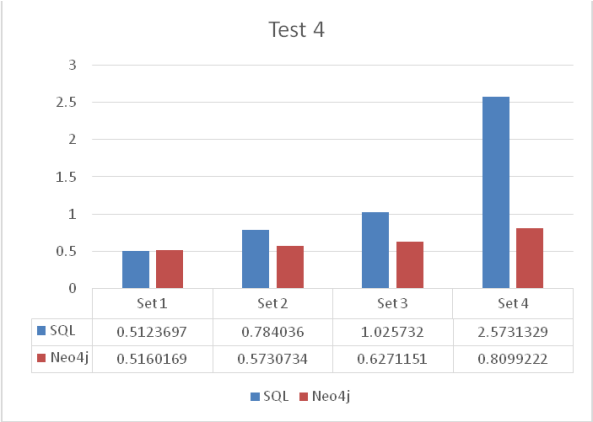


Fig. 7. Experiment 4 results

Experiment 5:

Query: Get all recipes similar to "Example Recipe" that have more than 100 likes containing "Tomato" from the "Italian" or from the "Romanian" Cuisine

SQL Syntax

```
SELECT DISTINCT recipes
FROM Recipes IN Recipe TABLE
WHERE Recipes.Likes BIGGER THAN 100
JOIN RecipesSimilarity IN RecipeRecipeSimilarity TABLE
ON Recipes.Id EQUALS RecipesSimilarity.RecipeTwoId
WHERE RecipesSimilarity.SimilarityId EQUALS
"Strong" AND RecipesSimilarity.RecipeOneId EQUALS
"Example Recipe"
JOIN VegetableRecipe IN VegetableRecipe TABLE
ON Recipes.Id EQUALS VegetableRecipe.RecipeId
WHERE VegetableRecipe.VegetableId EQUALS
"Tomato" JOIN RecipeCuisine in RecipeCuisine TABLE
ON Recipes.Id equals RecipeCuisine.RecipeId
WHERE RecipeCuisine.CuisineId EQUALS "Italian" OR
RecipeCuisine.CuisineId EQUALS "Romanian"
```

Neo4j Syntax

```
MATCH (Recipe {"Example Recipe"})- [similarity:
IS_SIMILAR_TO {Similarity: "Strong"}] -> (Other
Recipe),
(Other Recipe)- [CONTAINS_VEGETABLE]->
(Vegetable {"Tomato"}),
 (Other Recipe)- [IS FROM CUISINE]-> (Cuisine)
where Cuisine EQUALS "Italian" OR Cuisine EQUALS
"Romanian" AND Other Recipe Likes BIGGER THAN 100
return Distinct Other Recipe
```

This query also involves four tables (three joins) and many conditions expressed with "and" or "or" operators.

The results for experiment 5 appear in figure 8. The same observation can be made. The query execution time that involves many junctions between tables and multiple conditions is much shorter in Neo4J than in the relational system and has also very little value for all four datasets.

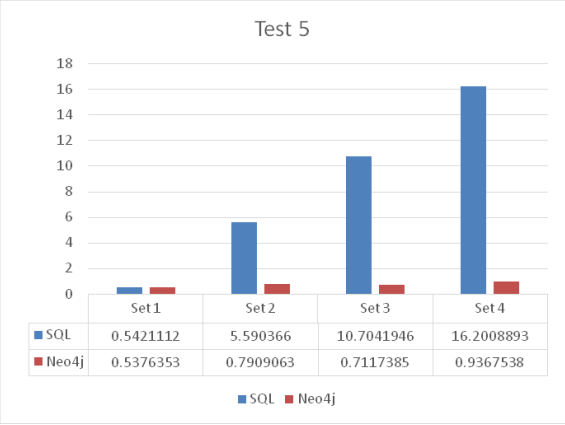


Fig. 8. Experiment 5 results

V. CONCLUSIONS

The idea of this experimental study started as a Web Application that interacts with the end-user and quickly delivers responses based on the requests performed. However, nowadays, as the web applications are extensively popular and often used as opposed to physical items like books or magazines, they must adapt and be able to handle a large number of data. As the trend is to use SQL databases, that are the most popular databases worldwide, the development started with such a database as data storage 'device'.

When considering this from the future's perspective, it is interesting to analyze the manners in which they respond in such an application type, where there are many relationships between items and also, the recommending engine and how it will respond.

As described before, the same structure of the database was exported to a graph database, and for multiple sets of data, tests were performed. These tests were designed based on how users tend to interact with such a system.

As seen, for a slightly large number of records, where the query has to perform search based on a limited number of relationships SQL tend to perform better than the Neo4J database. However, as the numbers get bigger the Neo4J database appears to be superior when it comes to computing time. For a low number of JOINs, the SQL database doesn't fall back so much even with large numbers of records, but for this type of application, where items are strongly related through relationships the graph database, it is safe to say, it is clearly superior.

In conclusion, for applications that involve large number of relationships between data, the graph databases are a suitable choice. Such projects could be social media applications, collaborative systems or libraries of any kind, books, music or videos. Even though, relational databases are strong and well-performing, so, there are cases where there is a (slightly) better alternative for data storage.

Nowadays, many large companies world-wide have migrated to NoSQL alternatives and the results are astonishing. The performance of their applications is keeping users interested and satisfied everyday by providing fast responses to their requests and that is generating success.

When it comes to choosing what type of database should be used, one must first perform a type of research activity, read and most important perform tests on their applications ahead of time, with large numbers of records to predict how they will perform in the future.

Designing the application with a well-researched and well-chosen alternative is a critical step in the early stages of development. Performing changes along different development cycles and stages, when the application has already become complex, delivered and in use for users, implies migrating data from one database to another, which is a complex, time-consuming and high-risk task.

REFERENCES

[1] J. J. Miller, "Graph Database Applications and Concepts with Neo4j", *in Proceedings of the Southern Association for Information Systems Conference*, Atlanta, GA, USA. Vol. 2324, No. 36, 2013

[2] http://neo4j.com/developer/cypher/

[3] https://linkurio.us/using-neo4j-to-build-a-recommendation-engine-based-on-collaborative-filtering/

[4] http://graphaware.com/neo4j/2013/10/11/neo4j-bidirectional-relationships.html

[5] https://neo4j.com/developer/graph-db-vs-rdbms/

[6] https://sdtimes.com/databases/guest-view-relational-vs-graph-databases-use/

[7] S. Medhi, and H. K. Baruah, "Relational Database And Graph Database: A Comparative Analysis", New Technologies, International Vol. 5, No 2, 2017

[8] A. Martinez, R. Mora, D. Alvarado, G. Lopez, and S. Quiros, " A Comparison between a Relational Databases and a Graph Database in the Context of a Personalized Cancer Treatment Application", *in CEUR Workshop Proceedings*, Vol. 1644, 2016, http://ceur-ws.org/Vol-1644/paper37.pdf

[9] Y. Cheng, P. Ding,T. Wang, et al., "Which Category Is Better: Benchmarking Relational and Graph Database Management Systems", *Data Sci. Eng*., vol.4, pp. 309–322, 2019 https://doi.org/10.1007/s41019-019-00110-3

[10] W. Khan, E. Ahmed, and W. Shahzad, "Predictive Performance Comparison Analysis of Relational & NoSQL Graph Databases", *International Journal of Advanced Computer Science and Applications*, vol. 8, no.5, 2017

[11] K. Sahatqija, J. Ajdari, X. Zenuni, B. Raufi, and F. Ismaili, "Comparison between relational and NOSQL databases", in *Proceedings of MIPRO*, pp. 0216-0221, 2018

[12] R. J. Sholichah, M. Imrona, and A. Alamsyah, "Performance Analysis of Neo4j and MySQL Databases using Public Policies Decision Making Data", in *Proceedings of ICITACEE*, pp. 152-157, 2020

Advances in Network Systems and Applications

THE rapid development of computer networks including wired and wireless networks observed today is very evolving, dynamic, and multidimensional. On the one hand, network technologies are used in virtually several areas that make human life easier and more comfortable. On the other hand, the rapid need for network deployment brings new challenges in network management and network design, which are reflected in hardware, software, services, and security-related problems. Every day, a new solution in the field of technology and applications of computer networks is released. The ANSA technical session is devoted to emphasizing up-to-date topics in networking systems and technologies by covering problems and challenges related to the intensive multidimensional network developments. This session covers not only the technological side but also the societal and social impacts of network developments. The session is inclusive and spans a wide spectrum of networking-related topics.

The ANSA technical session is a great place to exchange ideas, conduct discussions, introduce new ideas and integrate scientists, practitioners, and scientific communities working in networking research themes.

TOPICS

- Networks architecture
- Networks management
- Quality-of-Service enhancement
- Performance modeling and analysis
- Fault-tolerant challenges and solutions
- 5G developments and applications
- Traffic identification and classification
- Switching and routing technologies
- Protocols design and implementation
- Wireless sensor networks
- Future Internet architectures
- Networked operating systems
- Industrial networks deployment
- Software-defined networks
- Self-organizing and self-healing networks
- Mulimedia in Computer Networks
- Communication quality and reliability
- Emerging aspects of networking systems

TRACK CHAIRS

- **Armando, Alessandro,** University of Genova, Italy
- **Awad, Ali Ismail,** Luleå University of Technology, Sweden
- **Furtak, Janusz,** Military University of Technology, Poland
- **Suri, Niranjan,** Institute of Human and Machine Cognition, United States

PROGRAM CHAIRS

- **Awad, Ali Ismail,** Luleå University of Technology, Sweden
- **Furtak, Janusz,** Military University of Technology
- **Hodoň, Michal,** University of Žilina, Slovakia

PROGRAM COMMITTEE

- **Ahad, M**ohd Abdul, Department of Computer Science and Engineering, Jamia Hamdard, New Delhi
- **Ajlouni, Naim,** Istanbul Aydin University, Turkey
- **Antkiewicz, Ryszard,** Military University of Technology, Poland
- **Brida, Peter,** University of Zilina, Slovakia
- **Bridova, Ivana,** University of Zilina, Slovakia
- **Brzoza-Woch, Ada,** AGH University of Science and Technology, Poland
- **Chaganti, Raj,** ExpediaGroup Inc, Seattle, USA
- **Chmielewski, Mariusz,** National Cyber Security Centre, Poland
- **Chumachenko, Igor,** Kharkiv National University of Municipal Economy named after Beketov, Ukraine
- **Cui, Huanqing,** Shandong University of Science and Technology, China
- **Davidsson, Paul,** Malmö University, Sweden
- **Dotsenko, Sergii,** Ukrainian State University of Railway Transport, Ukraine
- **Długosz, Rafał,** UTP University of Science and Technology, Poland
- **Elmougy, Samir,** Mansoura University, Egypt
- **Faria, Lincoln,** Department of Computer Science, Fluminense Federal University, Brazil
- **Farooq, Ali,** University of Turku, Finland
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Gheisari, ,** Mehdi, Southern University of Science and Technology, China
- **Karpiš, Ondrej,** University of Žilina, Slovakia
- **Kochláň, Michal,** University of Žilina, Slovakia
- **Lavrov, Eugeniy,** Sumy State University, Ukraine
- **Molenda, Karol,** National Cyber Security Centre, Poland
- **Monov, Vladimir V.,** Bulgarian Academy of Sciences, Bulgaria
- **Murawski, Krzysztof,** Military University of Technology, Poland
- **Niewiadomska-Szynkiewicz, Ewa,** Research and Academic Computer Network (NASK), Institute of Control and Computation, Poland
- **Papaj, Jan,** Technical university of Košice, Slovakia
- **Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt

5th Workshop on Internet of Things—Enablers, Challenges and Applications

THE Internet of Things is a technology which is rapidly emerging the world. IoT applications include: smart city initiatives, wearable devices aimed to real-time health monitoring, smart homes and buildings, smart vehicles, environment monitoring, intelligent border protection, logistics support. The Internet of Things is a paradigm that assumes a pervasive presence in the environment of many smart things, including sensors, actuators, embedded systems and other similar devices. Widespread connectivity, getting cheaper smart devices and a great demand for data, testify to that the IoT will continue to grow by leaps and bounds. The business models of various industries are being redesigned on basis of the IoT paradigm. But the successful deployment of the IoT is conditioned by the progress in solving many problems. These issues are as the following:

- The integration of heterogeneous sensors and systems with different technologies taking account environmental constraints, and data confidentiality levels;
- Big challenges on information management for the applications of IoT in different fields (trustworthiness, provenance, privacy);
- Security challenges related to co-existence and interconnection of many IoT networks;
- Challenges related to reliability and dependability, especially when the IoT becomes the mission critical component;
- Zero-configuration or other convenient approaches to simplify the deployment and configuration of IoT and self-healing of IoT networks;
- Knowledge discovery, especially semantic and syntactical discovering of the information from data provided by IoT.

The IoT technical session is seeking original, high quality research papers related to such topics. The session will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. The focus areas will be, but not limited to, the challenges on networking and information management, security and ensuring privacy, logistics, situation awareness, and medical care.

TOPICS

The IoT session is seeking original, high quality research papers related to following topics:

- Future communication technologies (Future Internet; Wireless Sensor Networks; Web-services, 5G, 4G, LTE, LTE-Advanced; WLAN, WPAN; Small cell Networks. . .) for IoT,
- Intelligent Internet Communication,
- IoT Standards,
- Networking Technologies for IoT,
- Protocols and Algorithms for IoT,
- Self-Organization and Self-Healing of IoT Networks,
- Object Naming, Security and Privacy in the IoT Environment,
- Security Issues of IoT,
- Integration of Heterogeneous Networks, Sensors and Systems,
- Context Modeling, Reasoning and Context-aware Computing,
- Fault-Tolerant Networking for Content Dissemination,
- IoT Architecture Design, Interoperability and Technologies,
- Data or Power Management for IoT,
- Fog—Cloud Interactions and Enabling Protocols,
- Reliability and Dependability of mission critical IoT,
- Unmanned-Aerial-Vehicles (UAV) Platforms, Swarms and Networking,
- Data Analytics for IoT,
- Artificial Intelligence and IoT,
- Applications of IoT (Healthcare, Military, Logistics, Supply Chains, Agriculture, ...),
- E-commerce and IoT.

The session will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. Focus areas will be, but not limited to above mentioned topics.

TECHNICAL SESSION CHAIRS

- **Cao, Ning,** College of Information Engineering, Qingdao Binhai University
- **Chudzikiewicz, Jan,** Military University of Technology, Poland
- **Zieliński, Zbigniew,** Military University of Technology, Poland

PROGRAM COMMITTEE

- **Al-Anbuky, Adnan,** Auckland University of Technology, New Zealand
- **Antkiewicz, Ryszard,** Military University of Technology, Poland
- **Brida, Peter,** University of Zilina, Slovakia
- **Chudzikiewicz, Jan,** Military University of Technology in Warsaw, Poland

- **Cui, Huanqing,** Shandong University of Science and Technology, China
- **Ding, Jianrui,** Harbin Institute of Technology, China
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Fuchs, Christoph,** Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
- **Hodoň, Michal,** University of Žilina, Slovakia
- **Johnsen, Frank T.,** Norwegian Defence Research Establishment (FFI), Norway
- **Karpiš, Ondrej,** University of Žilina, Slovakia
- **Krco, Srdjan,** DunavNET
- **Laqua, Daniel,** Technische Universit"at Ilmenau, Germany
- **Lenk, Peter,** NATO Communications and Information Agency, Other
- **Li, Guofu,** University of Shanghai for Science and Technology, China
- **Marks, Michał,** NASK - Research and Academic Computer Network, Poland
- **MURAWSKI, Krzysztof,** Military University of Technology, Poland
- **Papaj, Jan,** Technical university of Košice, Slovakia
- **Savaglio, Claudio,** University of Calabria, Italy
- **Ševčík, Peter,** University of Žilina, Slovakia
- **Shaaban, Eman,** Ain-Shams university, Egypt
- **Staub, Thomas,** Data Fusion Research Center (DFRC) AG, Switzerland
- **Suri, Niranjan,** Institute of Human and Machine Cognition, United States
- **Wrona, Konrad,** NATO Communications and Information Agency

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 145–149
DOI: 10.15439/2021F102
ISSN 2300-5963 ACSIS, Vol. 26

Connectivity Maintenance in IoT-based Mobile Networks: Approaches and Challenges

Vahid Khalilpour Akram
International Computer Institute
Ege University
Izmir, Bornova
vahid.akram@ege.edu.tr

Moharram Challenger
University of Antwerp
and Flanders Make
Flanders, Belgium
moharram.challenger@uantwerpen.be

***Abstract*—Connectivity is an important requirement in almost all IoT-based wireless networks. The multi-hop networks use intermediate nodes to create a communication path between other nodes. Hence losing some nodes may cut off all communication paths between other active nodes. Generally, the connectivity of a partitioned network can be restored by adding new or activating redundant nodes, moving available nodes to the new location, and increasing the wireless communication range of nodes. The restoration problem may have many constraints and sub-problems. The network may initially be disconnected, the nodes may be heterogeneous, reliable connections may be required between the nodes, we may have unreachable locations in the network area to put the new nodes or move exciting nodes, more than one node may fail at the same time and the expected coverage area may complicate the connectivity restoration problem. In this paper, we study the main challenges and methods of connectivity restoration in IoT-based wireless networks.**

***Index Terms*—Internet of Things, Connectivity, Multi-hop Wireless Network, Mobile Networks.**

I. Introduction

INTERNET of Things (IoT) is one of the fastest-growing and promising technologies that already formed a revolution in daily human life. In recent years, the new generation of smart buildings, structures, vehicles, clothes and almost all types of objects that every day are used by people benefit from IoT technologies [1], [2]. Technically, IoT is a set of small, low-energy electronic devices that can connect to the Internet over wired or wireless communication platforms [3], [4]. These devices may have different types of capabilities such as processing, sensing, and data storage. Recent advances in electronic and hardware technologies allow the generation of a wide range of tiny, low-cost, low-energy devices that support local processing, sensing, and various communication methods. The diversity and capabilities of IoT devices grow exponentially day by day which allows people to use them in different application areas. Tracking the status and location of patients and health care devices in hospitals [5], automation of activities and increasing the quality and efficiency of products in agriculture [6], tracking a mobile object in indoor or outdoor environments [7], controlling the objects in smart homes [8], automation of fabrication in factories [9], fast and efficient rescue systems [10], real-time monitoring systems of critical infrastructures [11], and providing ad-hoc or mobile communication platforms [12] are a few samples of IoT applications.

Connectivity is a critical necessity in all sorts of networks, including wired local area networks, wireless ad-hoc networks, mobile networks, and the Internet of Things. Ideally, all available devices in a network should be able to communicate with other devices in the network. In other words, the network must keep the connectivity between all available devices. In some types of networks, such as wired local area networks, preserving the connectivity between the nodes is almost straightforward. As long as the routers, switches, and cables work properly, any connected device may communicate with other devices under predefined security policies. In these networks, the status of endpoint devices has no effect on the connectivity of the network. For example, if a device stop working, the connectivity of other nodes will not be affected. However, preserving the connectivity in ad-hoc wireless networks may be much more complicated. In a wireless ad-hoc network, the nodes communicate with other remote nodes over multi-hop links. Using the ad-hoc routing protocols, each node forwards the received message to its neighbors which allows the nodes to remote nodes which are outside of their communication range. Therefore, the connectivity of nodes relies on the proper working of available intermediate nodes in the network. Consequently, if a node stop working, we may lose the connectivity between other working nodes. The problem will be much more complicated if the nodes are mobile. If a node changes its initial location, the connectivity between some other nodes may be completely destroyed. In a vehicle or drone network, if a mobile node changes its location, the communication paths between its neighbors will be changed. In the worst case, if there is no other redundant path, moving or losing a node may cut the communication paths to a large set of working nodes and waste many active resources.

The diversity of device and communication technologies allows establishing ad-hoc networks almost everywhere even in harsh environments such as mountains, sea-bed, and forests. In these networks, the nodes may use hybrid communication technologies such as Bluetooth, WiFi, GSM, LTE, LoRa, and Zigbee. Also, some nodes may be static with a fixed location and some other nodes may be mobile. For example, for real-

time monitoring of an environment, we may distribute some sensor nodes in the environment and collect their sensing data over multi-hop links, mobile drones, or mobile vehicles (Fig. 1).

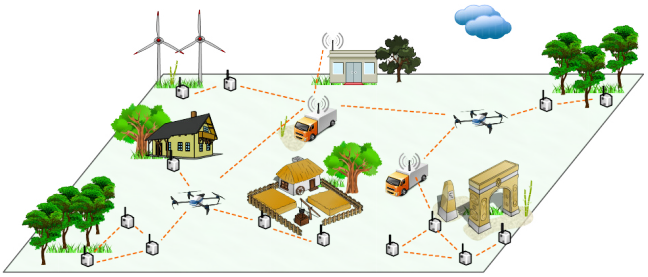


Fig. 1: Sample network for collecting sensed data from environment

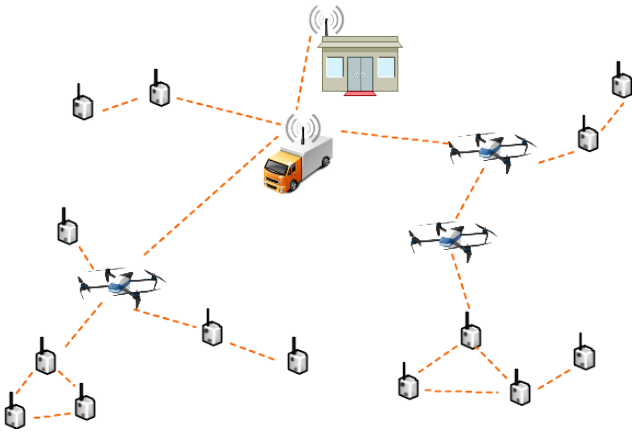
A wide range of hardware and sensors are available for establishing a network similar to Fig. 1. For example, an ESP32 device support both WiFi, low energy Bluetooth communication technologies and have enough memory and processing power for most of monitoring applications. This device may be equipped with different types of sensors to gather various data from the environment. The new generation of drones [13] have more than one hour fly time and wide communication range which allows them to reach far locations miles away from the base station. However, preserving continues and reliable connectivity in wireless ad-hoc networks still is a challenging problem. In this paper we, focus on the applications and different challenges of connectivity maintenance in IoT based mobile ad-hoc networks. The remaining parts of this paper has been organized as follow; Section II provides a formal definition for connectivity problem and its different variants. Section III focuses on the open challenges and research problems on the efficient connectivity maintenance in mobile networks. Finally, Section IV provides the conclusion and future works.

II. Problem Formulation

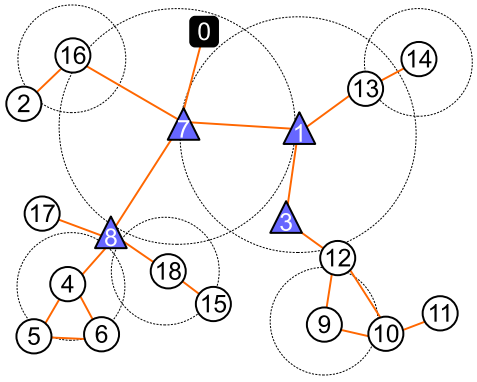
We can model an ad-hoc network as graph $G(V, E)$ where V is the set of nodes and E is the set of edges between the nodes. For example, Fig. 2a shows a sample mobile ad-hoc network with 4 mobile nodes and 15 static nodes. Fig. 2b shows the graph model of this network where $V = \{0, 1, ..18\}$ and $E = \{(0, 7), (1, 3), (1, 7), ...\}$ is the set of links between the nodes. In Fig 2b triangles show the mobile nodes and circles show the static nodes. We assume that node 0, (the filled black node) is the base station of the network. The dashed big circles in Fig. 2b shows the communication range of the node which may differ based on the node types.

Generally, a network is called connected if there is at least a communication path between every pair of nodes. Connectivity is one of the most important requirements in all networks. In wireless ad-hoc networks, where the network connectivity relies on the proper working of nodes, different strategies have been developed to increase connectivity robustness. Placing redundant nodes, creating alternate paths between the nodes, and increasing the radio range of nodes are some of these strategies which have their own advantages and disadvantages. Placing redundant nodes in the environment is a simple and feasible approach but increases the network cost. Increasing the radio power of node allows them to connect more nodes but at the same time increase the energy consumption of nodes which are not desirable in the battery-powered networks. Creating and maintaining alternate paths between the nodes needs complex algorithms and real-time topology control which may be hard to implement.

Formally, a network is called k-connected if there is at least k path between every pair of nodes. Therefore in a 1-connected, there is at least one path and in a 3-connected network, there are at least 3 disjoint paths between every pair of nodes. Higher k values increase the reliability of the network but need precise nodes deployment and restoration strategies. Generally, challenges and problems on network connectivity can be classified into 2 groups as connectivity detection and connectivity restoration problems which are discussed in more detail in the following subsections.



(a)



(b)

Fig. 2: a) Sample mobile network, b) Graph model of the network.

A. Connectivity Detection

Connectivity detection is the problem of finding the connectivity status and reliability of connections between the nodes. In the simplest case of the connectivity detection problem, the aim is to determine whether all nodes in the network are connected. In most applications, we need to ensure that all nodes have at least one communication path to each other which leads to the simplest form of connectivity detection problem. There are many central and distributed algorithms for the connectivity detection problem [14]. The central connectivity detection algorithms may use different methods such as depth-first search, network flow, path traversal, and matching to find the connectivity of the network.

Existing of a communication path between all nodes is a required condition in most applications, but in most cases is not enough. In wireless ad-hoc networks, 1-connectivity usually is considered unreliable because losing some nodes or links may disconnect a large number of nodes from the others. For example, Fig. 3a shows a sample 2-connected network that can tolerate any node or links failure without losing its connectivity. In contrast, Fig. 3b shows a 1-connected network with many critical links (orange color) and nodes (filled with orange) that losing each one destroy the network connectivity. A node whose failure destroys the network connectivity is called a critical node. Similarly, a link whose failure destroys the network connectivity is called a critical link or bridge. Detecting the critical nodes and bridges may help to improve connectivity reliability. For example, Fig. 3a and Fig. 3b show that adding only two links to the graph can resolve all critical nodes and links.

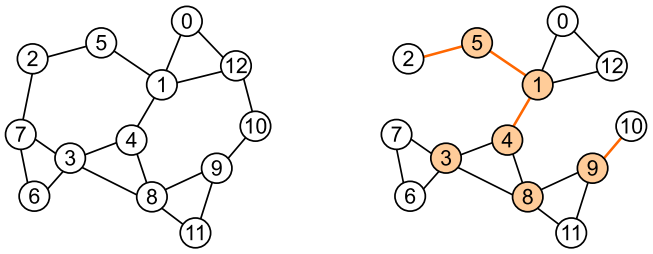


Fig. 3: a) a sample 2-connected network, b) a sample 1-connected network with critical nodes and critical bridges.

Besides the bridges and critical nodes, we may find the minimum cut edges and minimum cut vertex of a network to measure its connectivity reliability. The minimum edge cut of a network is the smallest set of edges whose removal, destroys the connectivity of the network. For example, in Fig. 3a a minimum edge cut of the network is $\{(1,5)(2,7)\}$ which their removal disconnects node $\{2,5\}$ from the other nodes. Similarly, a minimum vertex cut of presented network in Fig. 3a is $\{1,9\}$. A network may have more than one minimum edge or minimum vertex cut. Finding the minimum vertex and edge cuts reveals the weak points and connectivity robustness of the network.

B. Connectivity Restoration

Network connectivity restoration is the process of increasing the reliability of network connectivity by reconnecting the disconnected nodes [15]. In some applications, the connectivity restoration is started after failure in some nodes that disconnect some working nodes from the others. However, some applications require continuous and reliable connectivity. In these applications, the connectivity restoration process must be started before complete disconnection to reinforce the unreliable connections. So, the connectivity restoration strategies can be classified into proactive and reactive groups. The proactive methods start after each node or links failure and reinforce the connectivity if required. For example, in the k-connectivity restoration methods [16], if a node failure reduces the k value, the restoration algorithm tries to increase the k value by moving other nodes or activating redundant nodes. The reactive methods start after network disconnection and try to reconnect the disconnected parts.

The connectivity restoration algorithms usually rely on the connectivity detection algorithms to determine the current connectivity status and decide about the required actions. Generally, the main approaches for connectivity restoration are moving the available mobile nodes to the new locations, activating or placing new nodes in the network environment, and increasing the radio communication of the nodes. Each approach has its own advantages and disadvantages. The movement-based methods use available resources in the network but require mobile nodes which are not feasible in some applications. Also moving the nodes from their initial location may disconnect some other links which complicate the connectivity restoration process.

Placing new nodes or activating redundant nodes simplifies the connectivity restoration process but requires additional resources. Also placing new nodes in the desired locations may not be possible in some harsh environments. Increasing the radio communication range of reaming nodes is another solution that may reconnect the disconnected parts. But increasing the radio communication range increases the energy consumption of nodes and may reduce the network lifetime. Besides these issues and constraints, the connectivity restoration problem has some other difficulties and challenges which are discussed in the next section.

III. Challenges

In this section we discuss about the main challenges of connectivty restoration in mobile ad-hoc networks.

A. Initial Connectivity

A network can be initially connected or it can be disconnected after deployment. For example, after distributing a large set of sensor nodes to a forest using an airplane, with a high probability the resulting network will be disconnected. Some researchers assume that the network is initially connected and

the connectivity restoration may start after failure or moving of nodes. This assumption simplifies the restoration problem as we ensure that restoring the disconnected links is enough for establishing the network connection. Connecting all nodes in a network that is initially disconnected is a hard problem because the set of possible solutions is very large. In the movement-based restoration, selecting the candidate nodes for moving, selecting the direction of movement, and calculating the movement distance is a hard problem because usually, the optimal solution needs a combination of different movements. For example, Fig. 4a shows the movement-based connectivity restoration in a network that is initially connected and Fig 4b shows another network that is initially disconnected. Similarly, connectivity restoration by placing new nodes or activating redundant nodes is much harder in the networks which are initially disconnected.

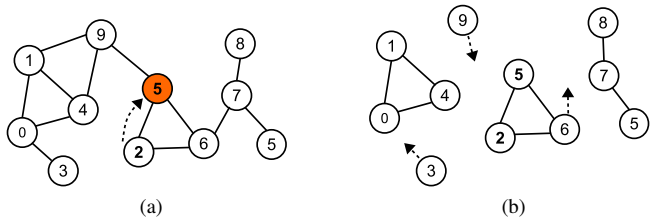


Fig. 4: a) Connectivity restoration when network is initially connected, b) Connectivity restoration when network is initially disconnected.

B. Heterogeneity

An IoT network may include a set of similar nodes with the same hardware and software properties. In such a homogeneous network all nodes have almost the same communication range, processing power, memory capacity, moving capability, etc. In contrast, in a heterogeneous network, the nodes may have different hardware and communication ranges. When the nodes have different communication ranges, some nodes may connect to a large number of nodes and some nodes may only have a limited set of neighbors. Also in a heterogeneous network, we may have uni-directed links which only allow one-way communications. Connectivity restoration in heterogeneous networks is much harder than homogeneous networks because the communication range of each node and the direction of links should be considered in graph model [17]. Most of the existing researches in connectivity restoration assume that all nodes have the same communication range.

C. k-connectivity

The aim of k-connectivity restoration is preserving the k value of a given network [14]. For example, in a 3-connected network, we want to preserve the 3-connectivity after losing some nodes. For $k = 1$ the problem is converted to the traditional connectivity restoration but for higher k values the problem will be much more complicated because moving every node in the network may affect the k value. In a 1-connected network, moving most of the nodes have no effect on the connectivity. For example in Fig. 4a moving each of the nodes $\{1, 2, 4\}$ does not affect the connectivity. However, in a k-connected network the set of candidate nodes that can leave their position without affecting k is limited, and finding these nodes needs some computation.

D. Target Positions

In the movement or deployment-based methods, we may assume that any position in the network area can be selected as a target position for moving the nodes or placing new nodes. Most of the existing research assumes that all nodes can move to their desired location or we may put the redundant or new nodes to the desired location. However, this assumption is not true for most real-world applications. Due to environmental conditions and obstacles, the nodes may not move to some location or we may not put the new nodes in the desired locations. To simplifies the restoration problem, some researchers assume that the new nodes can be only added to the location of exciting nodes or the nodes can only move to the location of existing nodes. This assumption simplifies the problem and converts it to a polynomial-time problem.

E. Single vs. Multiple Failure

Restoring the connectivity after a single node failure is generally simpler than the multiple nodes failure. After the failure of a single node its neighbor nodes may change their location to restore the connectivity because all of them may know the exact location of the failed node. However, in multiple nodes failure, a node and it's all neighbors may stop working at the same time. In this case, some of the failed nodes may be undetectable, or moving multiple nodes is impossible. Despite that the multiple node failures can happen in most real-world application, the researches that consider this case is limited and the number of proposed solutions is restricted [18].

F. Coverage

In some applications, the IoT nodes collect various data from enshrinement using different sensors. Losing a node in an IoT-based network or moving a node to a new location may lead to some coverage lost in the network. The coverage lost is not acceptable in some applications hence during the connectivity restoration we should preserve the maximal coverage. Restoring the connectivity and preserving the maximal coverage at the same time complicate the restoration process [19]. Especially in movement-based connectivity restoration, the nodes which have the minimal effect of total coverage area should be selected for movement. Generally, the coverage-aware connectivity restoration methods try to find the nodes which their covered are is also covered by the other nodes.

IV. Conclusion

Connectivity is one of the most important properties in most IoT-based wireless networks and robust connectivity is a vital requirement in most applications. In multi-hop networks, the

connectivity of the network relies on the proper working of the nodes, and losing some nodes may destroy the connectivity.

In this paper, we surveyed the main challenges and methods of connectivity restoration in IoT-based wireless networks. Generally, the connectivity of a partitioned network can be restored by adding new or activating redundant nodes, moving available nodes to new locations, and increasing the wireless communication range of nodes. The restoration problem may have many constraints and sub-problems. Restoring the connectivity of a network that is initially connected is much simpler than connectivity all nodes in a network that is initially disconnected.

In a homogeneous network in which all nodes have the same hardware and software capabilities, the connectivity restoration is simpler than a heterogeneous network. In a heterogeneous network, the communication range and moving capabilities of each node may be different from the other nodes which complicate the restoration process. While the 1-connectivity allows the nodes to communicate with each other, the 1-connected networks are usually considered unreliable because losing a single node may destroy the connectivity. The k-connectivity restoration process tries to preserve k disjoint paths between every pair of nodes.

In some applications, the nodes in the network may go to every desired location or we may add new nodes to the desired location. However, in some other networks, the environmental conditions do not allow to put the new nodes or move the existing nodes to the desired locations. The connectivity restoration after a single failure can be simpler than the connectivity restoration after multiple failures because losing a node and its neighbors may complicate the restoration process. Finally losing a node in the network may lead to some coverage loss which may be not acceptable in some applications. Hence coverage-aware connectivity restoration algorithm tries to reconnect the connectivity while preserving the maximal coverage.

As future works, we will focus on the discussed challenges of the restoration problem to find efficient approaches that consider more than one criteria at the same time. For example, proposing a comprehensive approach that can handle multiple failures, maximize the coverage, preserve the k-connectivity, support heterogeneous nodes, and allow flexible target position selection can be very useful in many real-world applications. Also developing platform-specific languages and frameworks to support the deployment and connectivity restoration of different mobile and flying nodes under the discussed constraints can simplify the development and maintaining of complex IoT-based applications [20], [21].

REFERENCES

[1] S. Arslan, M. Challenger, and O. Dagdeviren, "Wireless Sensor Network based Fire Detection System for Libraries," in *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017, pp. 271–276.

[2] L. Özgür, V. K. Akram, M. Challenger, and O. Dağdeviren, "An IoT based Smart Thermostat," in *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*. IEEE, 2018, pp. 252–256.

[3] B. Karaduman, T. Aşıcı, M. Challenger, and R. Eslampanah, "A cloud and Contiki based Fire Detection System using Multi-hop Wireless Sensor Networks," in *Proceedings of the Fourth International Conference on Engineering & MIS 2018*, 2018, pp. 1–5.

[4] B. Karaduman, M. Challenger, and R. Eslampanah, "ContikiOS based Library Fire Detection System," in *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*. IEEE, 2018, pp. 247–251.

[5] N. Karimpour, B. Karaduman, A. Ural, M. Challenger, and O. Dagdeviren, "IoT based Hand Hygiene Compliance Monitoring," in *2019 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2019, pp. 1–6.

[6] M. S. Mekala and P. Viswanathan, "A survey: Smart agriculture iot with cloud computing," in *2017 international conference on microelectronic devices, circuits and systems (ICMDCS)*. IEEE, 2017, pp. 1–7.

[7] S. Shao, A. Khreishah, and I. Khalil, "Enabling real-time indoor tracking of iot devices through visible light retroreflection," *IEEE Transactions on Mobile Computing*, vol. 19, no. 4, pp. 836–851, 2019.

[8] Y. Jie, J. Y. Pei, L. Jun, G. Yun, and X. Wei, "Smart home system based on iot technologies," in *2013 International conference on computational and information sciences*. IEEE, 2013, pp. 1789–1791.

[9] I. E. Etim and J. Lota, "Power control in cognitive radios, internet-of things (iot) for factories and industrial automation," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 4701–4705.

[10] T. Ahn, J. Seok, I. Lee, and J. Han, "Reliable flying iot networks for uav disaster rescue operations," *Mobile Information Systems*, vol. 2018, 2018.

[11] S. L. Ullo and G. Sinha, "Advances in smart environment monitoring systems using iot and sensors," *Sensors*, vol. 20, no. 11, p. 3113, 2020.

[12] N. H. Motlagh, M. Bagaa, and T. Taleb, "Uav-based iot platform: A crowd surveillance use case," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 128–134, 2017.

[13] M. Hassanalian and A. Abdelkefi, "Classifications, applications, and design challenges of drones: A review," *Progress in Aerospace Sciences*, vol. 91, pp. 99–131, 2017.

[14] V. K. Akram and O. Dagdeviren, "Deck: A distributed, asynchronous and exact k-connectivity detection algorithm for wireless sensor networks," *Computer Communications*, vol. 116, pp. 9–20, 2018.

[15] Y. Zhang, J. Wang, and G. Hao, "An autonomous connectivity restoration algorithm based on finite state machine for wireless sensor-actor networks," *Sensors*, vol. 18, no. 1, p. 153, 2018.

[16] V. K. Akram and O. DAĞDEVİREN, "Tapu: Test and pick up-based k-connectivity restoration algorithm for wireless sensor networks," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, no. 2, pp. 985–997, 2019.

[17] Y. Zeng, L. Xu, and Z. Chen, "Fault-tolerant algorithms for connectivity restoration in wireless sensor networks," *Sensors*, vol. 16, no. 1, p. 3, 2016.

[18] M. Imran, M. Younis, A. M. Said, and H. Hasbullah, "Localized motion-based connectivity restoration algorithms for wireless sensor and actor networks," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 844–856, 2012.

[19] N. Tamboli and M. Younis, "Coverage-aware connectivity restoration in mobile sensor networks," *Journal of network and computer applications*, vol. 33, no. 4, pp. 363–374, 2010.

[20] H. M. Marah, R. Eslampanah, and M. Challenger, "DSML4TinyOS: Code Generation for Wireless Devices," in *2nd International Workshop on Model-Driven Engineering for the Internet-of-Things (MDE4IoT), 21st International Conference on Model Driven Engineering Languages and Systems (MODELS2018)*. Copenhagen, Denmark, 2018.

[21] T. Z. Asici, B. Karaduman, R. Eslampanah, M. Challenger, J. Denil, and H. Vangheluwe, "Applying Model Driven Engineering Techniques to the Development of Contiki-based IoT Systems," in *2019 IEEE/ACM 1st International Workshop on Software Engineering Research & Practices for the Internet of Things (SERP4IoT)*. IEEE, 2019, pp. 25–32.

Advances in Information Systems and Technology

AIST is a FedCSIS conference track aiming at integrating and creating synergy between disciplines of information technology, information systems, and social sciences. The track addresses the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This track takes a socio-technical view on information systems and, at the same time, relates to ethical, social and political issues raised by information systems.

AIST provides a forum for academics and professionals to share the latest developments and advances in the knowledge and practice of these fields. It seeks new studies in many disciplines to foster a growing body of conceptual, theoretical, experimental, and applied research that could inform design, deployment and usage choices for information systems and technology within business and public organizations as well as households.

We call for papers covering a broad spectrum of topics which bring together sciences of information systems, information technologies, and social sciences, i.e., economics, management, business, finance, and education. The track bridges the diversity of approaches that contributors bring to the conference. The main topics covered are:

- Advances in information systems and technologies for business;
- Advances in information systems and technologies for governments;
- Advances in information systems and technologies for education;
- Advances in information systems and technologies for healthcare;
- Advances in information systems and technologies for smart cities; and
- Advances in information systems and technologies for sustainable development.

AIST invites papers covering the most recent innovations, current trends, professional experiences and new challenges in the several perspectives of information systems and technologies, i.e. design, implementation, stabilization, continuous improvement, and transformation. It seeks new works from researchers and practitioners in business intelligence, big data, data mining, machine learning, cloud computing, mobile applications, social networks, internet of thing, sustainable technologies and systems, blockchain, etc.

Extended versions of high-marked papers presented at technical sessions of AIST 2015-2020 have been published with Springer in volumes of Lecture Notes in Business Information Processing: LNBIP 243, LNBIP 277, LNBIP 311, LNBIP 346, and LNBIP 380.

Extended versions of selected papers presented during AIST 2021 will be published in Lecture Notes in Business Information Processing series(LNBIP, Springer).

- Data Science in Health, Ecology and Commerce (3rd Special Session DSH'21)
- Information Systems Management (16th Conference ISM'21)
- Knowledge Acquisition and Management (27^{th} Conference KAM'21)

Track Chairs

- **Ziemba, Ewa,** University of Economics in Katowice, Poland
- **Chmielarz, Witold,** University of Warsaw, Poland
- **Cano, Alberto,** Virginia Commonwealth University, Richmond, United States

Program Chairs

- **Chmielarz, Witold,** University of Warsaw, Poland
- **Raban, Daphne,** University of Haifa, Israel
- **Wątróbski, Jarosław,** University of Szczecin, Poland
- **Ziemba, Ewa,** University of Economics in Katowice, Poland

Program Committee

- **Anton Agafonov,** Samara National Research University, Russia
- **Andrzej Białas,** Institute of Innovative Technologies EMAG, Poland
- **Ofir Ben-Assuli,** Ono Academic College, Israel
- **Robertas Damasevicius,** Silesian University of Technology, Poland
- **Gonçalo Dias,** University of Aveiro, Portugal
- **Rafal Drezewski,** AGH University of Science and Technology, Poland
- **Leila Halawi,** Embry-Riddle Aeronautical University, United States
- **Ralf Haerting,** Hochschule Aalen, Germany
- **Adrian Kapczyński,** Silesian University of Technology, Poland
- **Wojciech Kempa,** Silesian University of Technology, Poland
- **Agnieszka Konys,** West Pomeranian University of Technology, Szczecin, Poland
- **Eugenia Kovatcheva,** University of Library Studies and Information Technologies, Bulgaria
- **Jan Kozak,** University of Economics in Katowice, Poland
- **Marcin Lawnik,** Silesian University of Technology, Faculty of Applied Mathematics, Poland

- **Antoni Ligeza,** AGH University of Science and Technology, Poland
- **Amit Rechavi,** Ruppin Academic Center, Israel
- **Nina Rizun,** Gdansk University of Technology, Poland
- **Joanna Santiago,** Universidade de Lisboa - ISEG, Portugal
- **Wojciech Sałabun,** West Pomeranian University of Technology in Szczecin, Poland
- **Marcin Sikorski,** Gdank University of Technology, Poland
- **Francesco Taglino,** IASI-CNR, Italy
- **Łukasz Tomczyk,** Pedagogical University of Cracow, Poland
- **Gerhard-Wilhelm Weber,** Poznan University of Technology, Poland
- **Paweł Ziemba,** University of Szczecin, Poland

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 153–158
DOI: 10.15439/2021F2
ISSN 2300-5963 ACSIS, Vol. 26

MEDICAL STEEL FAULT PREDICTION USING DEEP LEARNING TECHNIQUES

A.Sheik Abdullah,
Assistant Professor, Department of Information Technology, Thiagarajar college of Engineering, Madurai, Tamilnadu, India.

Karthikeyan Jothikumar,
Research Associate, Department of Information Technology, Thiagarajar college of Engineering, Madurai, Tamilnadu, India.

K.R.A.Bhubesh,
Department of Information Technology, Thiagarajar college of Engineering, Madurai, Tamilnadu, India.

Abstract— **Fault detection and analysis is considered to be an important factor in industrial production for medical applications. As industrial era has evolved tons, new fault identification ways are required to differentiate faults with totally good distinctions. The superior best a production is wanted to possess, the higher fault identification methodology the factories should apply. This research work focus on the assessment and evaluation of fault detection using deep learning techniques. The evaluation is made accordingly using Deep CNN with the variants corresponding to simple CNN, Resnet, Alexnet and Vgg_16. Besides, classification accuracy is improved by parameter optimizing and sample size equalization strategy. Experimental results shows that evaluation using the proposed methods with Vgg_16 gives an improved training accuracy of about 90% and validation accuracy of about 87%. This proves that fault detection and analysis in medical equipments and transplanting devices can be efficiently identified for better treatment and device management.**

Keywords—Deep learning, Image Augmentation, Neural Networks, Fault Detection, Medical devices.

I. INTRODUCTION

Fault detection and analysis is becoming an important phenomenon in the forth coming days. Medical fault prediction and analysis is needed in day-to-day surgical incorporation and its applications [1]. The classification of fault and its mechanism need to be considered more important because the materials used for incorporation varies accordingly with the treatment concerned. Managing and maintaining the fault types with materials, cost and its waste level we can improve the quality of the device or the equipment that is to be used [2]. Similar measures corresponding to the recycling of materials will also happen corresponding to the medical or surgical fault diagnosis [3].

The operational environments should be made flexible with regard to the fault and its diagnosing principles. Industrial production is also threatening to the medical environment for fault methods and makes things fine distinctions [4]. The quality should not be compromised at any situations with regard to the production and analysis methods [5]. Parameter estimation and optimizing the quality of the materials used makes the medical steel plate production to have a good and gatherable environment for the industry to produce the material [6].

Traditional learning algorithms in machine learning platform have its restrictions in different domains of real-time engineering applications. When considering the medical domain algorithms focusing on deep learning models has a good impact and predictive nature with the application considered [7]. If this is made in novel practice for the prediction of fault rate in steel prediction then the exact relevance and its coordinates for the material can be determined in advance [8]. The dimension, shape, size, and quality of the material are the most important parameters for any material to test for evaluation. This if made clear and concise then the production of good quality material can be adhered at all the stages of usage for medical practice [9].

Deep learning models are mainly used in medical applications in order to improve the prediction and to make diagnosis in a better way [10]. Researchers and clinical investigators use the concept of deep learning to predict and enhance the nature of disease, risk factors, and medical diagnostic compliances [11]. At certain stages, the realm of heart failure rates and its detective models can be made accordingly with the deep learning techniques. Algorithms such as

convolutional neural network, vector classifiers, and tree-based evaluation methods are the most used algorithmic models for prediction and classification in medical informatics [12]. This research work focus on the assessment and evaluation of efficiency of medical devices specifically on steel plate that is used for surgical practice. The entire process has been modelled using Deep learning techniques upon statistical evaluation and analysis.

II. LITERATURE REVIEW

The work by the authors Isermann [1] strategies are based, e.g., on boundary assessment, equality conditions or state eyewitnesses. Additionally signal model methodologies were created. The objective is to produce a few manifestations demonstrating the distinction among ostensible and defective status. In view of various side effects deficiency analysis techniques follow, deciding the flaw by applying grouping or derivation strategies. This commitment gives a short presentation into the field and gives a few applications for an actuator, a traveler vehicle and an ignition motor.

The work by the authors Dong [2] lot of channels is planned where each channel intends to mutually appraise the framework states and a particular conceivable issue. Upper limits of the assessment blunder covariance's are acquired in the concurrent presence of the linearization mistakes and decentralized occasion set off transmissions, and afterward the channel gains are determined to limit such limits. The channels are planned in a recursive manner and subsequently the calculation is relevant for online execution. At the point when a shortcoming is identified, the channel with the least remaining is viewed as the one comparing to the real flaw and its yield can be viewed as the states and issue assessment. The adequacy of the proposed strategy is represented by a reenactment model.

The work by the authors Good et al [17] introduces an algorithm that greatly reduces the overall size of the PCA problem by breaking the analysis of a large number of variables into multiple analyses of smaller uncorrelated blocks of variables. From the statistical summary it has been observed that the compatibility of PCA found to be good at the implementation level of all variables.

Similarly, the authors Yin et al [3] proposed an approach which is different from the standard PCA- and PLS-based techniques which rely on mean-extraction for residual generation; the proposed CVA-based scheme takes process dynamics into account as well. Also, the significance and the corresponding property provide an improvement when compared to that of the benchmark process.

The work by the authors Breiman et al [15] tests on genuine and mimicked informational collections utilizing arrangement and relapse trees and subset determination in direct relapse show that stowing can give significant increases in exactness. The crucial component is the shakiness of the forecast strategy. In the event that annoying the learning set can cause huge changes in the indicator developed, at that point stowing can improve precision.

The work by the authors Bewick et al [18] presents strategic relapse, which is a strategy for demonstrating the reliance of a paired reaction variable on at least one illustrative factors. Constant and clear cut illustrative factors are thought of.

The work made by the author Widodo [4] manages the utilization of the previously mentioned classifiers for flaw finding of a concoction cycle containing a ceaseless mixed tank reactor and a warmth exchanger. The outcomes show a predominant characterization execution of the help vector machine versus the chose fake neural organization. Also, the help vector machine classifier is extremely delicate to the correct determination of the preparation boundaries. It is demonstrated that the use of hereditary calculation for ideal determination of these boundaries is doable and can assist with improving the help vector machine classifier execution. Similarly, Terzi et al [20] provided a scientific categorization for arranging classifiers is introduced. Another meta-classifier, Meta-Consensus, with a establishment in both agreement hypothesis and the hypothesis of autonomous appointed authorities, is presented.

The work by the author Basheer [5] manages the utilization of the previously mentioned classifiers for flaw finding of a concoction cycle containing a ceaseless mixed tank reactor and a warmth exchanger. The outcomes show a predominant characterization execution of the help vector machine versus the chose

fake neural organization. Also, the help vector machine classifier is extremely delicate to the correct determination of the preparation boundaries. It is demonstrated that the use of hereditary calculation for ideal determination of these boundaries is doable and can assist with improving the help vector machine classifier execution. In this research work we explicitly focus on the evaluation of the steel plate which is specifically used in surgical process. The implementation is preceded with the deep learning techniques upon statistical evaluation.

III. PROPOSED METHODOLOGY

A. Data Collection

Data corresponding to the Northeastern University (NEU) is collected which specifically corresponds to the surface defect focusing on six different surfaces. The defect is analyzed with regard to the hot-rolled strip of steel with rolled-in scale, crazing, pitted, scratches, patches and inclusion surfaces. The database corresponds to 1800 grayscale images and 300 samples of each of the six forms of defects on each of the surface. The observed images clearly depict the intra-class defects focusing on horizontal, vertical and slanting scratch surfaces. But, the inter-class defects have similar aspect of defect focusing on scale, crazing and regions of pitted surface. From the surface defect database there exists two challenges such as the appearance of the defect and the influence of enlightened patterns when the quality of the material changes.

B. Data Preprocessing

The dataset used for the fault detection is the preprocessed image dataset. So we applied some noise reduction alone to use it in our algorithm. The image is in bmp format and image is preprocessed as grey scale image. The reason to use bmp format and grey scale image is bit map format identifies the spot of faults in steel images more precisely than jpg format. The grey scale is more precise for fault identification than coloured images.

III.B.1 Basic Image Data Analysis: Basic image data analysis show the image basic properties like image size, RGB value calculation and heuristic natures of image to train without much discrepancy.

```
Type of the image :  <class 'imageio.core.util.Array'>
Shape of the image : (200, 200)
Image Hight 200
Image Width 200
Dimension of Image 2
```

Figure 1. Data Preprocessing

C. Image Augmentation

The data provided from the source was only in limited numbers. For efficient training and classification of faults we need more images for training and testing. So we used some automation on image augmentation which will generate images based upon given needs like image size shape and format etc.

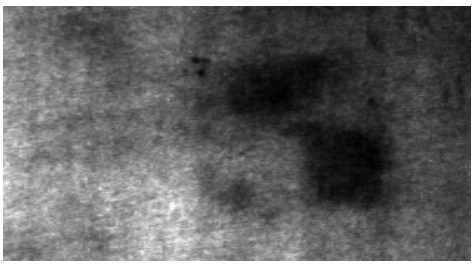


Figure 2. Original Image

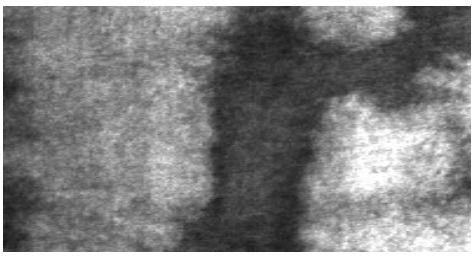


Figure 3. Augmented Image

From above 2 images it is clear that the augmented images are matched upto 95 % with original image. So we augmented up to 1000 images for each class of fault in the given dataset. The original and augmented images are structured in Figure 2 and Figure 3.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

a. Simple Deep CNN

In this the architecture is pretty old architecture that is Deep CNN. It has 6 Convolution layer and 1 output layer, each convolution layers uses a filters of 30 X 30 which is better of viewing the fault section in grey scale training images. The kernel size used in here is 3 X 3 and activation function in all layers is tanh, which responds well for minor pixel changes it does not show many deviations but with major steep pixel changes the activation performs well. The following Figure 4 signifies the processed image for steel fault rate.

The image data is converted to numpy arrays and the numpy arrays fed into these convolution layers for training. The pixels values of fault spots are trained inside the architectures and the model is validated against the validation data.

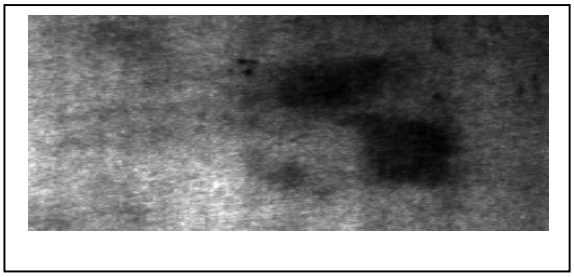


Figure 4. Processed Image (Steel fault detection)

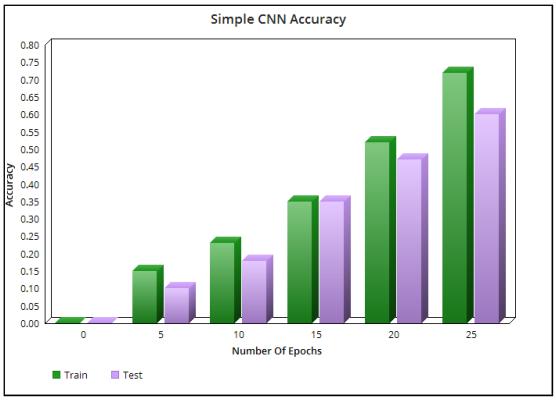


Figure 5. Accuracy Graph for Simple Deep Cnn

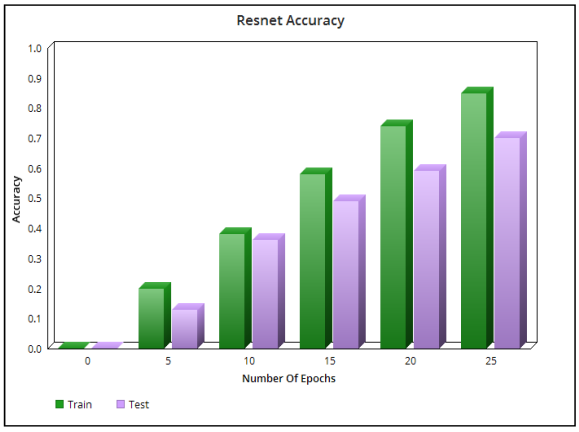


Figure 6. Accuracy Graph for Resnet

b. Resnet

In this 16 convolution layers are used for more precise insights of pixel values each layer uses 512 filters with kernel size of 3x3. Here the image is trained as it is without conversion of numpy arrays. Since resnet Architecture will take for pixels with 2 of its convolution layers. The activation function used in each layer was rectified linear unit which was most efficient activation function for the resent architecture. The output layer uses the softmax as output function. Every layer is trained with minute value of pixels so for 16 convolution layers the pixel of faults will be well trained and provide greater accuracy than the simple deep cnn. The Accuracy evaluation is given in Figure 6.

c. Alexnet

In this 12 convolution architectures is used with 256 filters with kernel size of 3x3. Dropout values are used in this architecture here also the image is trained as it is. Alexnet uses softmax activation function in its convolution layers in this recurrent unit is attached so that some pixel conditions are maintained for the future remembrance which makes this architecture as most powerful one.

The output layer uses the relu activation function which converts and gives the normalised for higher grey scale values. The accuracy is depicted in Figure 7.

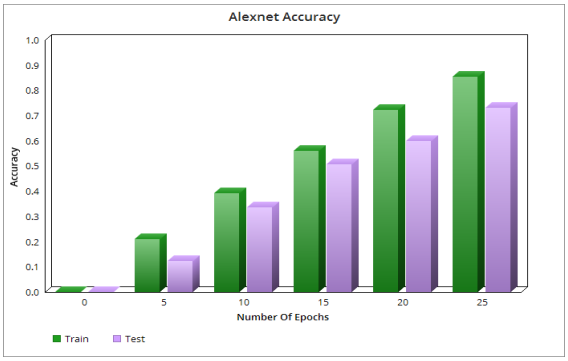


Figure 7. Accuracy Graph for Alexnet

d. Vgg

Vgg uses the 16 and 32 convolution layers for greater accuracies. Here the vgg 16 Architecture is used so 16 convolution layers are used the activation function used is softmax at each layer. The image is converted into numpy arrays and then trained and validated with the model. The accuracy is depicted in Figure 8. The output layer uses the relu activation function with some regularization to avoid unnecessary pixels. The comparison among the accuracy of the model is depicted in Table 1.

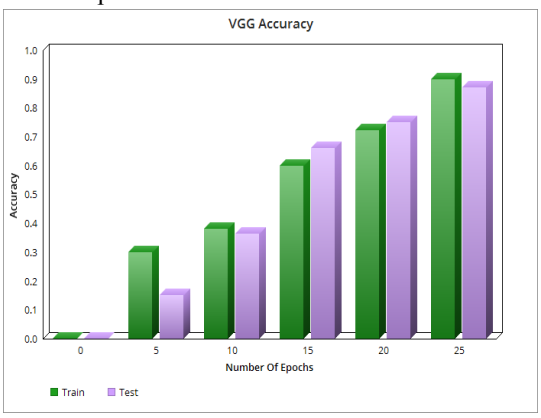


Figure 8. Accuracy Graph for Vgg

Table 1. Experimental results

Architecture	Training Accuracy	Validation Accuracy
Simple CNN	72%	60%
RESNET	85%	70%
ALEXNET	85.4%	73%
VGG_16	90%	87%

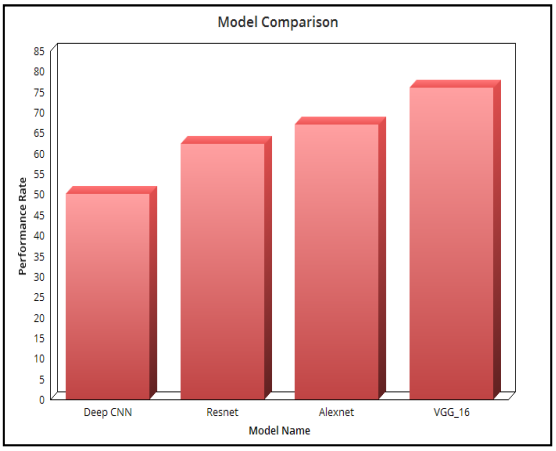


Figure 9. Model comparison and analysis

V. CONCLUSION AND FUTURE WORK

Fault identification and analysis in devices and equipments is one of the most frequently occurring phenomenon's in production technology. This research work focus on the assessment and evaluation of fault rate in steel plates using the variants of deep CNN. The variants involved are simple CNN, Resnet, Alexnet and Vgg_16. Among the four different variants of Deep CNN vgg_16 provided a good impact in fault identification and validation accuracy. This if deployed in production engineering can fix the problems and issues related to fault detection and analysis. Moreover manual intervention collection of data for classification algorithms is literally reduced since the image is directly trained with help of CNN. In future works we will enhance the operational facilities of this idea to all metal sheets with greater accuracy and less time complexity.

REFERENCES

[1] Isermann, R. Model-based fault-detection and diagnosis—status and applications. Annu. Rev. Control 29 (1) pp. 71–85, 2005.

[2] Dong, H., Wang, Z., & Gao, H. Fault detection for markovian jump systems with sensor saturations and randomly varying nonlinearities, Circuits and Systems I: Regular Papers. IEEE Transactions on 59 (10) pp. 2354–2362, 2012.

[3] Yin, S., Ding, S.X., Haghani, A., Hao, H., & P. Zhang. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. J. Process Control 22 (9) pp. 1567–1581, 2012.

[4] Widodo, A., & Yang, B.S. Support vector machine in machine condition monitoring and fault diagnosis. Mech. Syst. Signal Process. 21 (6) pp. 2560–2574, 2007.

[5] Basheer, I., & Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application, J. Microbiol. Methods 43 (1) pp. 3–31, 2000.
[6] Yin, S., Ding, S., Xie, X., & Luo, H. A review on basic data-driven approaches for industrial process monitoring, IEEE Trans. Ind. Electron. 61 (11) 6418–6428, 2014.
[7] Du, W., & Zhan, Z. Building decision treeclassifieronprivatedata, in: Proceedings of the IEEE International Conference on Privacy, Security and Data Mining, vol. 14, Australian Computer Society, Inc., pp. 1–8, 2002.
[8] Zou, H., Hastie, T., & Tibshirani, R. Sparse principal component analysis, J. Comput. Graphical Stat. 15 (2) pp. 265–286, 2006.
[9] Braga, J., Heuze, Y., Chabadel, O., Sonan, N., & Gueramy, A. Non-adult dental age assessment: correspondence analysis and linear regression versus Bayesian predictions, Int. J. Legal Med. 119 (5) pp. 260–274, 2005.
[10] Russell, E.L., Chiang, L.H., & Braatz, R.D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemom. Intell. Lab. Syst. 51 (1) pp. 81–93, 2000.
[11] Bach, F.R., & Jordan, M.I. Kernel independent component analysis, J. Mach. Learn. Res. 3 pp. 1–48, 2003.
[12] Yin, S., Zhu, X., & Kaynak, O. Improved pls focused on key performance indictor related fault diagnosis, IEEE Trans. Ind. Electron. 2014.
[13] Amit ,Y., & Geman, D. "Shape quantization and recognition with randomized trees," Neural Comput., vol. 9, no. 7, pp. 1545–1588, 1997.
[14] Breiman, L. "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[15] Breiman, L. "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123– 140, 1996.
[16] Dong, Y., Du, B., & Zhang ,L."Target Detection Based on Random Forest Metric Learning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 4, pp. 1830-1838, April 2015.
[17] Good, R.P., Kost, D., & Cherry, G.A. "Introducing a Unified PCA Algorithm for Model Size Reduction," in IEEE Transactions on Semiconductor Manufacturing, vol. 23, no. 2, pp. 201-209, May 2010.
[18] Bewick, V., L. Cheek andJ. Ball, Statistics review 14: Logistic regression. Crit Care 9: 112-118. DOI: 10.1186/cc3045, 2005.
[19] Breiman, L., Classification and Regression Trees. 1st Edn., Wadsworth International Group, Belmont, ISBN-10: 0534980538 pp: 358, 1984.
[20] Buscema, M., S. Terzi and W. Tastle, A new meta-classifier. Proceedings of the North American Fuzzy Inform Processing Society, Jul. 12-14, IEEE Xplore Press, Toronto, pp: 1-7. DOI: 10.1109/NAFIPS.2010.5548298, 2010.
[21] Chaudhuri, B.B. and U. Bhattacharya, Efficient training and improved performance of multilayer perceptron in pattern classification. Neurocomputing, 34: 11-27. DOI: 10.1016/S0925-2312(00)00305-2, 2000.
[22] Dong, L., D. Xiao, Y. Liang and Y. Liu, Rough set and fuzzy wavelet neural network integrated with least square weighted fusion algorithm based fault diagnosis research for power transformers. Elec. Power Syst. Res., 78: 129-136. DOI: 10.1016/J.EPSR.2006.12.013, 2008.
[23] Eslamloueyan, R., Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process. Applied Soft Comput., 11: 1407-1415. DOI: 10.1016/J.ASOC.2010.04.012, 2011.
[24] Haykin, S.S., Neural Networks: A Comprehensive Foundation. 1st Edn., Macmillan, New York, ISBN-10: 0023527617, pp: 69, 1994.
[25] Zhang, Y., Jiang, J, Bibliographical review on reconfigurable fault-tolerant control systems. Annual Rev. Control, 32: 229-252. DOI: 10.1016/J.ARCONTROL.2008.03.008, 2008.
[26] Maurya, M.R., Rengaswamya, R., Venkatasubramanian, V., Fault diagnosis using dynamic trend analysis: A review and recent developments. Eng. Appli. Artif. Intell., 20: 133-146. DOI: 10.1016/j.engappai.2006.06.020, 2007.
[27] Lo, C.H., Wong, Y.K., Rad, A.B., Chow, K.M., Fusion of qualitative bond graph and genetic algorithms: A fault diagnosis application. ISA Trans., 41: 445-456. 10.1016/S0019-0578(07)60101-3, 2002.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 159–166
DOI: 10.15439/2021F41
ISSN 2300-5963 ACSIS, Vol. 26

Conception of 4-Component Architecture of Information Systems on Example of Artificial Neural Networks

Dmitriy Gakh
Institute of Control Systems,
Bakhtiyar Vahabzadeh str. 68,
AZ1141, Baku, Azerbaijan
Email: dgakh@sinam.net

***Abstract*—Nowadays Information Systems (IS) become more and more distributed, complex, and heterogeneous. Such nature of IS make them or their components a Black Box. Although classical software operates according understandable logic, modern complex software often shows non-determinism in its operation. Artificial Intelligence (AI) based on Artificial Neural Networks (ANN) is an example of such systems.**

This paper considers IS architecture consisting of 4 components, one of which represents non-determinism as an "Machine Intuition". The architecture is derived from 3-tier computer architecture and based on psychological findings. This approach allowed building a simple and user/developer friendly model.

Practical value of the architecture is concluded in ability to better understand, design, and develop the IS containing units with non-deterministic behavior, deal with AI overfitting, underfitting, and threat problems. Architecture and principles represented in this paper can be applied not only to AI/ANN but different IS types.

I. Introduction

ARTIFICIAL Intelligence (AI) is one of the most intensively developing technology. Design of simple AI systems is concluded in composing the Artificial Neural Network (ANN) layout (in this article, unless otherwise stated, the abbreviation AI refers to AI based on ANN). However, at this stage there is no clear understanding of how the ANN will operate. This fact demonstrates a difference of AI development and that of classical software. So, while classical software is a series of commands and its behavior is deterministic, AI introduces some kind of non-deterministic behavior. Another example of non-deterministic calculations is quantum computing (QC / A QC is not a deterministic machine; in other words, there is no single solution for which any other result would be an error [1]). To simplify the readability, the further text will discuss AI, but many statements can be also applied to other technologies. The article mentions AI non-deterministic behavior and units providing "Machine Intuition" (there is also a term "Artificial Intuition").

Since ANN are originated from observation of real life processes, life simulation solutions are very interesting examples where results of their operations can be compared to life. Genetic and Machine Learning Algorithms should be also considered because in most cases they provide non-deterministic results. A very interesting example is the life simulation where wolfs preferred suicide over eating sheep [2]. There are also other examples showing unintended consequences of Black Boxes [2-5]. Complex systems are intrinsically hazardous systems. Complex systems contain changing mixtures of failures latent within them, change introduces new forms of failure, human operators have dual roles: as producers and defenders against failure, human practitioners are the adaptable elements of complex systems, etc. [6].

The AI Failures Incident Database provides a publicly accessible view of AI failures [7, 8]. The classification schema detailing AI failures has been developed [7, 9]. The methods of avoiding AI failures that provides a balance between being excessively rigid (which would make its use difficult and brittle) and overly subjective (which would render the framework useless) have been elaborated [7].

Many problems presented by a super intelligence resemble exercises in international diplomacy more than computer software challenges; for instance, the value alignment problem of aligning AI values with humans'. Failure is defined as 'the nonperformance or inability of the system or component to perform its expected function for a specified time under specified environmental conditions'. Intelligence definitions converge towards the idea that it 'measures an agent's ability to achieve goals in a wide range of environments' [9]. These definitions imply such requirements as specified time and specified environmental conditions. However, actual AI solutions can be applied for undefined period of time and in unpredictable environment. Thus, failures of AI systems are non-deterministic nature. We can say here about something like "Machine intuition" instead of failures. We can as well say that there is a kind of "Machine intuition" failures. "Machine Intuition" can be presented in computer architecture as a separate specific component.

Architecture of solution is an important conception that to be formulated at the very beginning of the development. Although there are papers describing AI or QC solutions architecture, this description is specific and quite complex (One example is AI Infrastructure Reference Architecture from IBM [10]. Other examples are described in [11, 12] representing QC hardware architectures). Literature analysis shows that there is lack of papers describing general architecture of AI solutions reflecting their generic features, the most important of which can be non-determinism (at least in

comparison with classic software). Absence of such literature is a gap between understanding of AI solutions and well known classical (3-tier or N-tier) solutions. As a result it prevents smooth integration of AI and the classical solutions (implementation of AI in the classical solutions become popular method of software upgrade).

This paper considers simple 4-component architecture of IS solution derived from 3-tier architecture. This approach allows describing AI architecture as understandable for the classical programmers (those who write software on base of 3-tier or N-tier architectures) and integrate AI with the existing solution more smoothly. The research considers psychological findings that make the model closer to the human thinking and better align AI values with humans. Some phenomena that previously considered as failures can now be considered as behavior of "Machine Intuition". Introduction of computer architecture that simplifies development of software with complex unpredictable behavior is the main motivation to carry out this research. The main contribution of the article is rising the question about presence in modern IS components with non-deterministic behavior and necessity to study it. 4-component IS architecture and some problems that can be solved by its means are presented.

II. Background and Research Questions

AI has its own advantages and disadvantages. It differs from classical determined software. One requirement of classical software is providing determined output for determined input. This requirement is assured by tests. However, size and complexity of input data grow exponentially and it leads to impossibility to test the software for each case. This fact in its turn leads to non-determinism when input data or IS architecture become large and complex. Software Development methodologies do not reflect this non-determinism explicitly. However, it should be mentioned that quality assurance methods are effective in many cases. Nevertheless, these methods regard this non-deterministic behavior as a disadvantage. As a result, this can reduce the flexibility of AI solutions.

Software Quality itself is not a deterministic conception. The definitions of "quality" shows that it is not an objective index. Software Quality Models introduce metrics to make quality measurable. Due to interrelation of quality attributes a trade-offs must take place. There are also stakeholder's expectations that needs trade-offs between them [13].

In practice, a gap exists between abstract quality definitions provided in common quality taxonomies, such as ISO 25010, and concrete quality assessment techniques and measurements [14, 15]. Company-specific quality models are widely used. Quality models typically are adapted. ISO standards are not well accepted. Quality model users are moderately satisfied with their models [16].

Architecture of IS solutions impacts its quality. Analyses of literature demonstrates lack of description of AI architectures that will be simple and close to the classic software architectures. Because classic software architectures do not contain a component reflecting non-deterministic behavior, designers cannot pay significant attention to this IS character. Introduction of separate component reflecting "Machine Intuition" allowed designing IS with possible non-deterministic behavior in mind. Thus, it makes sense to design an architecture that takes this feature into account.

Meanwhile, AI originated from discovery of human brain's neural structure. This fact allowed to look for application of psychological findings to the AI solutions. Design of AI solution architecture on base of psychological findings has two benefits:

• Basic primitive rules that are true for the human brain most likely will be true for AI, because AI is originated from human brain's neural structures (proving of this hypothesis is out of scope of current research);

• A solution based on psychological findings will be likely close to human understanding (there are papers about research of psychological approaches to the software development processes).

A. Three-Tier Architecture

Three-tier architecture was developed by John J. Donovan in Open Environment Corporation (OEC), a tools company he founded in Cambridge, Massachusetts [17-19]. Fowler describes three principal layers of computer architecture as the following [20]:

1. Data Source - Databases, messaging systems, transaction managers, etc.;
2. Domain - Logic that is the real point of the system;
3. Presentation - Provision of services, display of information, user interface, HTTP requests, command-line invocations, batch API.

Three-tier computer architecture is the classical architecture that can be used to design complex IS or development of small applications. Division of IS into three tiers allows not only to simplify the design, but also distribute tasks between different developers. So, for example, the data tier can be developed and maintained by database specialists, logic tier developed by programmers, and presentation designed by user interface designers.

B. Literature about Psychology in Computer Science

Computer science can be applied to many aspects of human life. One of the them is psychology. This article does not consider application of computers in psychology. It rather considers application of psychology in computer science.

Proper design of interfaces between humans and machines humans wish to control requires cooperation of engineers and psychologists. Such cooperation allows dealing with so called "Human Factors" [21]. Considering psychological issues in human-computer interaction is not a new approach [22]. References to the synthesis of psychological knowledge and computer science are also mentioned in later sources [23]. In his research Prabhaker Panditi concluded that Software Engineering should consider the latest scientific discoveries in psychology, social psychology and be-

havioral economics. There is a need to conduct experiments to identify how the psychological discoveries apply to various phases, processes and practices in Software Engineering [24]. Kam Hou VAT considers so called "Software Psychology" as the domain of human behavior study in software engineering [25].

C. McWhinney's Realities

McWhinney's model is selected due to its simplicity and prove in practice (Young and Kovalev [26, 27]). Three-tier computer architecture can be mapped to 3 of 4 of his realities. Additional "Machine Intuition" component can be mapped to the 4th McWhinney's reality. This simple approach allowed building an effective computer architecture model.

Will McWhinney proposes to consider phenomena through a prism of four realities. He has drawn the coordinate system where axis X represented monistic - pluralistic quality and axis Y represented free will - determined quality. By this way, quarter "monistic + determined" represents unitary (U) reality, quarter "pluralistic + determined" represents sensory (SE) reality, quarter "pluralistic + free will" represents social (SO) reality, and quarter "monistic + free will" represents social (M) reality (See Fig. 1, Fig. 2) [28, 29]. Young and Kovalev have expanded this model and applied it to solve psychological problems [26, 27].

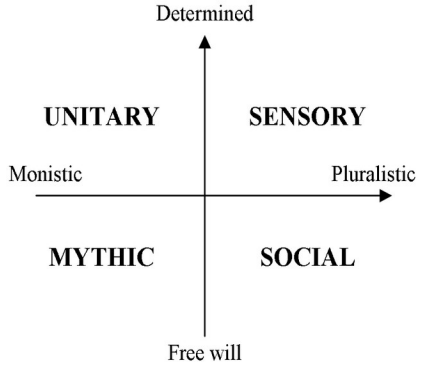


Fig 1. McWhinney's Realities

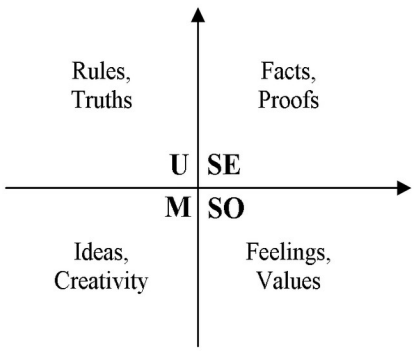


Fig 2. Characteristics of McWhinney's Realities

Phenomena situated in the unitary reality are characterized by rules and truths (See Fig. 2). They represent something stable and unchangeable. Phenomena situated in the sensory reality are characterized by facts and proofs. They represent something useful and profitable. Phenomena situated in the social reality are characterized by feelings and values. They represent something pleasant and comfortable. Phenomena situated in the mythic reality are characterized by ideas and creativity. They represent something synergetic and intuitive.

D. Research Questions

Building a 4-component IS architecture on base of 3-tier architecture is not a trivial problem. We cannot say that we just add 4th tier into architecture. Even the structure of the architecture model is not complex, the model changes dramatically. So, the research questions to be answered are:

RQ1: What is the 4-component IS architecture and what is its structure?

RQ2: How the 4-component IS architecture relates to 3-tier architecture?

RQ3: Which advantages/disadvantages does the 4-component IS architecture have?

E. Research Methodology

This research is based on the following main steps:

1. Select appropriate psychological model (McWhinney's realities in this research;
2. Juxtapose components of 3-tier architecture with the McWhinney's realities;
3. Juxtapose AI component with the McWhinney's realities;
4. Build the 4-component IS architecture model on base of identified patterns;
5. Prove the model theoretically.

Scope of this research allows presenting only basic findings and perform simple theoretical proof. Full theoretical proof of the model requires studying many implicit factors and interrelations. Practical proof of the 4-component IS architecture requires attempts to implement it in actual projects. Thus, this research implies in descriptive design and is based on literature analyses, in 28-years author's experience in software development, in experience in psychological and AI studies.

According to abstraction hierarchy [30] this paper describes the following phenomena:

- Theory. Presented 4-component architecture is a theory;
- Concepts. The concepts are presented by tiers of 3-tier architecture, realities of McWhinney's model, and components of the 4-component architecture;
- Indicators. Indicators are identified patterns of relations and interrelations of Software Components and realities of McWhinney's model;
- Variables. Variables are presented by components of the models (4-component AI architecture and McWhinney's

model should be considered together until the theory matures).

• Values. Values are actual software modules, pieces of code, and even hardware devices.

III. Findings

IS non-determined behavior occupies more and more spaciousness. Quality Assurance is aimed to mitigate and eliminate its negative effects. However, number of Quality Assurance methods are non-determined themselves because they are subjective.

Synthesis of computer science and psychology is not a new conception. Psychology is considered mainly for human-computer or human-robot interaction. Behavioral psychology is the main direction of psychology discussed in previous studies. Human factors can be considered as psychological phenomena.

Model of McWhinney's realities shows the 4 types of human vision of the world. Its practical value is proven in management and psychotherapy and has a simple basic structure.

As a result of considering of 3-tier architecture through the prism of McWhinney's realities the following parallels can be drawn (See Fig. 3):

1. Data Source relates to unitary reality because it represents rules (relations, constraints…) and truths (data). The statement "Deterministic systems of truths, assumptions, and propositions. Logics, morality, and spiritual oneness." [28] most likely relates to the data source;

2. Domain relates to sensory reality because facts and proofs can be provided by the logic. The statement "Raw characteristics and atomistic objects are derived from the senses. Empiricism." [28] most likely relates to the data processing by the functions. These functions in their turn represent the programming logic;

3. Presentation relates to social reality because feelings and values are the result of presentation abilities. Statement "Emotions and group values associated with distinct individuals and groups. Ethics and human relations." [28] most likely relates to the presentation. In other words, presentation of a computer system determines emotions of its users and their group values.

IV. Discussion of Findings

ANN are underlying structures of AI systems in most cases. AI solutions in their turn show non-deterministic behavior to a greater extent. In other words, they show "Machine Intuition". So, ANN are the best example to discuss 4-component architecture. Other types of solutions can be discussed in the same manner. Indeed, solutions with determined behavior can be presented without "Machine Intuition" component. Bugs and errors can be considered as belonging to "Machine Intuition" component where they can be considered as "wrong decision of the intuition". This statement is a subject for further discussions. Complex IS could be considered as set of interconnected components

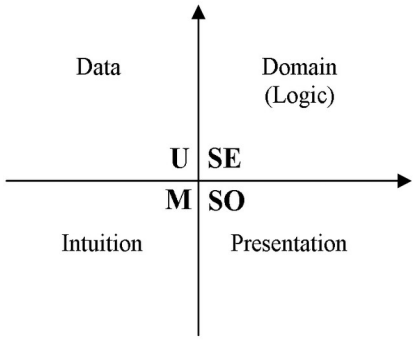


Fig 3. Components of the Model in McWhinney's Realities

forming network structures like ANN where each component can be considered as an artificial neuron. Testing and debugging of such systems could be considered as supervised Machine Learning (ML). Non-deterministic behavior of such systems could have the same origin as ANN. These facts confirm that although 4-component architecture is build on example of ANN, it can be applied for IS of different types.

ANN provide generalization of many data at a time. AI systems give a result with some uncertainty. Even if the system gives 10000 results with confidence 99.99%, there is no assurance that the next result will be of such high confidence. On the one hand, the deviation can be considered as an error, on the other hand it can be considered as an AI's intuition, idea, or creativity. Most often AI systems are too complex to make any assumption about their operation and represent Black Box. They require learning and testing after building. All these features of AI systems show that according to McWhinney's model it belongs to the mythic reality.

One can say that AI relates to the sensory reality. But AI systems are built as one whole construction processing large data entirely in one operation (by one call). Sensory reality is specified by "atomistic objects" [28] that corresponds more to number of functions processing number of data by number of operations (iterations, calls). One can also say that AI could be assigned to unitary reality because it is represented by the structure of neural network. But this structure is active and cannot be considered as the static data. Some kind of uncertainty does not allow considering AI as an element of unitary reality.

A. Four Components vs. Three Tiers

Three tier model contains 3 computer components connected sequently. Thus, these components can be called "tiers". But addition of AI as the fourth component leads to change of the system orderly structure then the components cannot be named "tiers" any more. The new structure is presented by Fig. 4. It should be marked that AI component within the 4-component architecture is presented as "Intuition". It is because the whole model can represent an AI solution. Indeed, an AI solution contains data, presentation,

and logic (this is also described in the text below). To follow technical language, one can say that "Machine Intuition" is better term for naming computer components relating to the 4th component. But "Intuition" is laconic and simple to use in text and diagrams.

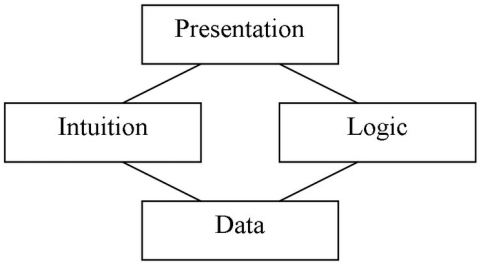


Fig 4. Four-Component Architecture

This layout is the most correct one by the following reasons:

• It could be considered as derived from 3-tier model where tier "Domain" is divided into two components "Logic" and "Intuition";

• According to McWhinney's realities "Data" locates beside "Logic" and "Intuition", "Logic" locates beside "Data" and "Presentation", "Presentation" locates beside "Logic" and "Intuition", and "Intuition" locates beside "Data" and "Presentation". Such closeness determines the connections between the components;

• AI systems can be even represented by 3-tier model where AI locates in tier "Domain" (but after introduction the 4-component model, use of 3-tier model for IS could be considered as deprecated).

McWhinney's model considers all relations between the realities. According to this principle, "Presentation" can be connected to "Data" and "Intuition" can be connected to "Logic" (See Fig. 5). Considering the "Presentation-Data" connection is not interested in current research because it is equal to corresponding connection between tiers in 3-tier model. But considering the "Logic-Intuition" connection should be discussed within this research because "Intuition" is a new component (relating to 3-tier model). At the same time this connection cannot be implemented easily because "Logic" relates to determined conception while "Intuition" relates to undetermined one. Any data flow from "Intuition" to "Logic" will make "Logic" undetermined and entered into "Intuition" component as a result. Any data flow from "Logic" to "Intuition" can change behavior of "Intuition" in undetermined way (because Intuition is undetermined) that most possible will require retraining in case of AI solution. In this case we can say that "Logic" will be entered into "Intuition".

Epstein showed that two systems of human brain are operating simultaneously: experiential/intuitive and rational/analytic. Both systems are adaptive, but in different ways, and neither system is generally superior to the other as each has unique strengths and limitations [31]. This fact is an additional prove of the model, represented in Fig. 5 and shows that "Intuition" and "Logic" components can be considered as ones originated from "Domain" level of 3-tier architecture.

It should be mentioned that actual trained and error-free AI components can contain a logical part. An evidence of a logical part is a fact that some AI components are determined on the training set (for example overfitting problem). Another example may be concluded in fact that AI unit may be designed as logically joined neural subnets. Thus, the 4-component architecture is an idealistic model.

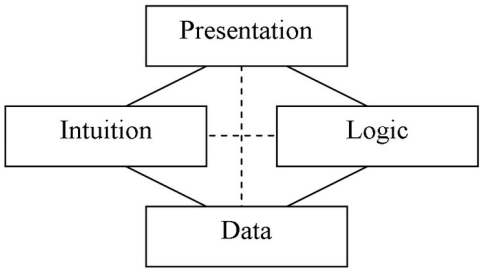


Fig 5. Relations in Four-Component Architecture

In real projects "Intuition" components can be ones where undetermined calculations are performed. This assumption allows considering different intelligent components performing undetermined calculations to be considered as "Intuition" components. Speaking strongly detailed description and requirements aimed to understand which component should be specific code or device should be related. But for this study components performing determined calculations should be considered as belonging to "Logic". Components performing undetermined calculations in their turn should be considered as belonging to "Intuition". Such division allows using the 4-component IS architecture at the beginning stages of design.

The 4-component IS architecture can be presented in more details where "Intuition" is presented as 4 levels (See Fig. 6). This case shows architecture where "Intuition" contains its own data source, logic and presentation serving for integration purposes. An example of such architectures can be a quantum computing unit that needs to transfer data to/from the quantum gates according to specific logic. Besides all "Intuition" accumulates data during the training, that also can be indicated as a data component and separated as a "gene" (that is useful for genetic algorithms).

B. Underfitting and Overfitting Problems

Underfitting and overfitting problems could be considered as classic in the ML. The problems are quietly often discussed in literature at different levels – from theory to prac-

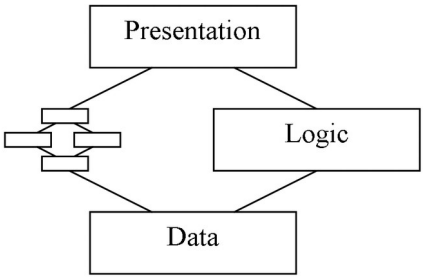


Fig 6. Nesting in Four-Component Architecture

tice [32-36]. These problems can be analyzed through the prism of the 4-component model. AI can be considered here as consisting of all the four components considered. The problems will be discussed at high abstraction level and is presented as a hypothesis. Proof of this hypothesis needs actual experiments that is beyond of the scope of this paper.

The model enables having a new look to the underfitting and overfitting problems in the ML. "Intuition" component relates to the Mythic reality of McWhinney's model that in its turn relates to the free will. "Logic" component relates to the Sensory reality of McWhinney's model that consequently relates to the deterministic realities. These facts allow one to draw the following conclusions:

• Underfitting issues relate to cases where "Intuition" component dominates over "Logic" component;

• Overfitting issues relate to cases where "Logic" component dominates over "Intuition" component.

Dominance of "Intuition" can also be due to dominant data/command flows through "Data" - "Intuition" and/or "Presentation" -"Intuition" links. Dominance of "Logic" in its turn can be due to dominant data/command flows through "Data"-"Logic" and/or "Presentation"-"Logic" links accordingly. Thus, data and interface ("Presentation") can contribute underfitting and overfitting problems. For the Artificial Neural Networks (ANN) "Presentation" can relate to input and output neurons.

Concerning to the modern methods of solving the underfitting and overfitting problems one can say the following. Network reduction allows decreasing AI capability to memorize weak relations in the training dataset that in its turn reduces dominance of Logic. At the same time, it can relate to dominance of the links to "Presentation" and "Data" if reduces input or output neurons (Neighbor and deeper neurons can also have influence. This influence should be studied).

Penalty method, expansion of the training data method, regularization method relate to "Data". Regularization method can relate also to "Presentation". Early stopping method prevents growth of AI dominance over Logic by preventing strengthening one part over another.

Techniques against the underfitting includes expansion of training dataset ("Data"), increase the number or size of parameters in the model ("Data", "Presentation"), increase the complexity of the model ("AI", "Logic"), increase the training time (strengthening "Logic").

C. AI Threat Problem

There is a concern that AI can get out of human control and significantly harm humanity. Implementation of technological inventions without an in-depth laboratory analysis of the consequences is real precondition for this threat [37]. Analyses of AI threat with means of the 4-component model leads to considering the problem sourced in two spheres:

• Logical – in case if AI system trained to do harm the people;

• Intuition - in case if AI system not trained not to harm people.

Although it is just a hypothesis, it shows another possible application of the 4-component model.

D. Value alignment problem

The 4-component architecture can help solve the value alignment problem of aligning AI values with humans. McWhinney's model can be used as a mediator between human values and the 4-component architecture. Description of human values in terms of McWhinney's model is psychologists' problem. Nevertheless, alignment of described human values with the 4-component architecture is problem of programmers. Because 3 of 4 components of the architecture are well known, studied, and used in practice, while a new component "Machine Intuition" is the only subject for such research.

It should be mentioned that the value alignment problem should be considered not only for AI, ANN, QC solution, but also for classical solutions. One example of aligning values with humans is a user interface, that relates to visual aesthetics experience. Another example is the software quality (there are many quality assurance methodologies that could be helpful).

E. Business Processes Modeling and Moral Decision Making

Business Processes Modeling (BPM) widely uses diagrams and formulae. It shows its orientation to strong logic and determined calculations. Intuition in classic BPM is the prerogative of a human. As this research shows, complex systems and AI introduce new "Intuition" component which is used more and more in IS. So, BPM can include (or may be separate to specific technique) modeling of intuitive processes. These processes can be based on experience [31].

Miller selected success groups and success factors in moral decision making and algorithms [38]. But it should be mentioned that morality of the decision is based on evaluation of consequences of the decision and originally this evaluation can be made only by human. As a result automatic evaluation of morality by the computer can be only based on

experience in similar cases and corresponds to “Intuition” component of the architecture.

F. Sustainable Development and Smart Cities

There are four dimensions of Sustainable Development (SD), i.e. ecological, economic, socio-cultural, and political [39]. These dimensions introduce behavior with some non determinism. AI solutions are very useful to handle issues with such behavior. Presented 4-component architecture seems helpful to analyze SD issues and develop solutions.

Smart Culture is the specific component of Smart Cities. Current understanding of Smart Culture is concluded in provision of information [40]. AI can handle issues in culture, related to non deterministic calculations, such as assessment of music, paintings, and other artifacts. The 4-component architecture can help to develop such solutions.

V. Conclusion

AI and QC introduce new kind of IS where calculations significantly differ on classical ones. These calculations contain some kind of uncertainty and non-determinism and introduce new problems. The 4-component model is a simple tool allowing design, develop, and analyze of such IS. Moreover, it allows better understanding and learning these systems and solve related problems. Application of the 4-component model to Smart City and related issues seem to be very promising. There is also interest to use the model for quality assurance and evaluation of ability of IS use in critical systems.

Presentation of the 4-component AI architecture in lectures “Programming technology” is approved by Baku State University. It is implicit proof of interest to the model and chance to discuss it with the students.

A “Machine Intuition” component is the point where software developers and psychologists can cooperate to create human friendly solutions. Such cooperation could also be helpful to better understand the human psyche.

It is also hypothesis that growing impact of non-deterministic behavior will prevent growth of complexity and functionality of IS. In other words, a “Machine Intuition” component should not remain without attention of researchers.

A. Disadvantages and Further Research

Disadvantages of the 4-component IS architecture are concluded in fact that there is no obvious boundary between Intuition and Logic components. Often even classical software has bugs (it could be acceptable levels of bugs) introducing some kind of uncertainty and non-determinism. But determination of “boundary” between solid logic and uncertainty is a difficult task. This makes the 4-component IS architecture more theoretical, rather than a practical tool.

There is huge interest to study Quality Assurance methodologies with application to 4-component architecture. Integration of psychological approaches could improve quality of complex IS.

There is also the fact that the advantages of the architecture have not been practically proved yet. Further research of the 4-component IS architecture can include:

- proof of benefits of the model in practical use including teaching of students;
- a deeper and wider research of the model;
- the way the model can help solve underfitting and overfitting problems;
- the way the model can help solve value alignment problem;
- possibility of integration with programming languages and code constructions;
- possibility of use as a design pattern in IS development;
- application in Quantum Computing;
- IS quality assurance;
- application in Business Process Modeling;
- SD, Smart Cities, Smart Culture;
- AI threat problem.

It should be mentioned that the discussion was carried out for AI. QC is now considered sufficient. There is an assumption that the model is useful for QC as well. Although AI is a young conception, QC is much younger. Further development of QC can require revision of abilities of 4-component model to apply to QC and implement addition research.

Acknowledgment

I wish to express my sincere gratitude to SINAM Ltd. and Baku State University for their kind support and inspiration in the research.

References

[1] S. Fulton, “What is quantum computing today? The how, why, and when of a paradigm shift”, 2020, https://www.zdnet.com/article/what-is-quantum-computing-understanding-the-how-why-and-when-of-quantum-computers/, accessed on 19th April 2021.

[2] L. Ng, “The AI Wolf That Preferred Suicide Over Eating Sheep”, https://onezero.medium.com/the-ai-wolf-that-preferred-suicide-over-eating-sheep-49edced3c710, accessed in July, 2021.

[3] J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. Bentley, S. Bernard, G. Beslon, D. Bryson, N. Cheney, P. Chrabaszcz, A. Cully, S. Doncieux, F. Dyer, K. Ellefsen, R. Feldt, S. Fischer, S. Forrest, A. Fŕenoy, C. Gagńe, L. Goff, L. Grabowski, B. Hodjat, F. Hutter, L. Keller, C. Knibbe, P. Krcah, R. Lenski, H. Lipson, R. MacCurdy, C. Maestre, R. Miikkulainen, S. Mitri, D. Moriarty, J. Mouret, A. Nguyen, C. Ofria, M. Parizeau, D. Parsons, R. Pennock, W. Punch, T. Ray, M. Schoenauer, E. Schulte, K. Sims, K. Stanley, F. Taddei, D. Tarapore, S. Thibault, R. Watson, W. Weimer, J. Yosinski, “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities”, Artif Life, 26 (2), pp. 274–306, 2020, https://doi.org/10.1162/artl_a_00319.

[4] L. Yaeger, “Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or PolyWorld: Life in a new context”, Artificial Life III, vol. XVII, pp. 263-298, Addison-Wesley, 1993, https://doi.org/10.1.1.38.6719.

[5] R. Feldt, "Generating multiple diverse software versions with genetic programming", Proc. 24th EUROMICRO Conference, vol.1 pp. 387-394, 1998, https://doi.org/10.1109/EURMIC.1998.711831.

[6] R. Cook, “How Complex Systems Fail”, Cognitive Technologies Laboratory, University of Chicago, Chicago, IL, USA, 1998.

[7] R. Williams, R. Yampolskiy, “Understanding and Avoiding AI Failures: A Practical Guide”, Philosophies, 6(3):53, 2021, https://doi.org/10.3390/philosophies6030053.
[8] S. McGregor, C. Custis, J. Yang, J. McHorse, S. Reid, S. McGregor, S. Yoon, C Olsson, R. Yampolskiy, “AI Incident Database”, 2021, https://incidentdatabase.ai/, accessed in July, 2021.
[9] P. Scott, R. Yampolskiy, “Classification Schemas for Artificial Intelligence Failures”, Delphi - Interdisciplinary Review of Emerging Technologies, vol. 2, iss. 4, pp. 186–199, 2019, https://doi.org/10.21552/delphi/2019/4/8.
[10] K. Lui, J. Karmiol, “AI Infrastructure Reference Architecture”, IBM Systems, 2018.
[11] K. Bertels, A. Sarkar, T. Hubregtsen, M. Serrao, A. A. Mouedenne, A. Yadav, A. Krol, I. Ashraf, “Quantum computer architecture: towards full-stack quantum accelerators”, Design, Automation and Test in Europe, pp. 1-6, 2020, https://doi.org/10.23919/DATE48585.2020.9116502.
[12] N. Jones, R. Meter, A. Fowler, P. McMahon, J. Kim, T. Ladd, Y. Yamamoto, “Layered Architecture for Quantum Computing”, Phys. Rev. X 2, 031007, 2012, https://doi.org/10.1103/PhysRevX.2.031007.
[13] P. Berander, L. Damm, J. Eriksson, T. Gorschek, K. Henningsson, P. Jönsson, S. Kågström, D. Milicic, F. Mårtensson, K. Rönkkö, P. Tomaszewski, “Software quality attributes and trade-offs”, Blekinge Institute of Technology, June 2005.
[14] S. Wagner, K. Lochmann, S. Winter, A. Goeb, M. Klaes, “Quality Models in Practice: A Preliminary Analysis”, ESEM’09, 2009.
[15] S. Wagner, K. Lochmann, L. Heinemann, M. Kläs, A. Trendowicz, R. Plösch, A. Seidl, A. Goeb, J. Streit, “The Quamoco Product Quality Modelling and Assessment Approach”, 34th Int. Conf. on Software Engineering (ICSE), pp. 1133-1142, 2012, https://doi.org/10.1109/ICSE.2012.6227106.
[16] S. Wagner, K. Lochmann, S. Winter, A. Goeb, M. Kläs, S. Nunnenmacher, “Software Quality Models in Practice. Survey Results”, http://mediatum.ub.tum.de/doc/1110601/274701.pdf, accessed in July, 2021.
[17] Wikipedia, the free encyclopedia, http://en.wikipedia.org, accessed in April 2021.
[18] A. Tafti, S. Janosepah, N. Modiri, A. Noudeh, H. Alizadeh, “Development of a Framework for Applying ASYCUDA System with N-Tier Application Architecture”, Comm. in Comp. and Inf. Science, vol. 181 pp. 533-541, 2011, https://doi.org/10.1007/978-3-642-22203-0_46.
[19] Z. Durdik, “Architectural Design Decision Documentation through Reuse of Design Patterns”, KIT Scientific Publishing, 2016, https://doi.org/10.5445/KSP/1000043807.
[20] M. Fowler, “Patterns of Enterprise Application Architecture”, Addison-Wesley, 2011
[21] D. Brée, “Artificial Intelligence and Cognitive Psychology: A New Look at Human Factors”, Human-Computer Interaction, 1988, https://doi.org/10.1007/978-3-642-73402-1_17.
[22] T. Moran, S. Card, “Applying Cognitive Psychology to Computer Systems”, A Graduate Seminar in Psychology, 1980.
[23] A. E. Bolock, J. Salah, Y. Abdelrahman, C. Herbert, S. Abdennadher, “Character Computing: Computer Science meets Psychology”, MUM’18, pp. 557-562, 2018, https://doi.org/10.1145/3282894.3286060.
[24] P. Panditi, “Psychology – The Land That Software Engineering Forgot”, Proc. of Innovations in Software Engineering Conference, 2018, https://doi.org/10.1145/3172871.3172889.
[25] K. VAT, “Teaching Software Psychology: Expanding the Perspective”, Proc. of the 31st SIGCSE Technical Symposium on Comp. Sci. Ed., 2000, https://doi.org/10.1145/331795.331892.
[26] P. Young, “Understanding NLP Principles & Practice”, Crown House Publishing 2nd ed, 2004.
[27] С. Ковалёв, “Психотерапия человеческой жизни”, Москва, 2018.
[28] W. McWhinney, “Growing Into the Canopy”, Journal of Transformative Education, vol. 5 no. 3 pp. 206-220, 2007, https://doi.org/10.1177/1541344607307023.
[29] W. McWhinney, J. Webber, D. Smith, B. Novokowsky, ”Creating Paths of Change: Managing Issues and Resolving Problems in Organizations”, SAGE Publications, 1997.
[30] N. Walliman, “Research Methods. The Basics”, Routledge, London, 2011.
[31] S. Epstein, “Demystifying Intuition: What it is, What it Does, and How it Does it”, Psychological Inquiry, 21, 295-312, 2010.
[32] H. Zhang, L. Zhang, Y. Jiang, "Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems", 11th Int. Conf. on Wireless Comm. and Sign. Processing (WCSP), pp. 1-6, 2019, https://doi.org/10.1109/WCSP.2019.8927876.
[33] H. Gabbar, R. Khan, “Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)”, CompSci Comm. Instr. Dev., pp. 163–172, 2014.
[34] F. Yusufi, A. Ahmed, J. Ahmad, “Modelling and developing diabetic retinopathy risk scores on Indian type 2 diabetes patients”, Int. J. Diabetes Dev Ctries 39, pp. 29–38, 2019, https://doi.org/10.1007/s13410-018-0652-z.
[35] E.A. Martínez-García, N. Rodríguez, R. Rodríguez-Jorge, J. Mizera-Pietraszko, J. Sheba, R. Mohan, E. Magid, “Non Linear Fitting Methods for Machine Learning”, Lecture Notes on Data Engineering and Communications Technologies, vol 13 pp. 807-818, 2018, https://doi.org/10.1007/978-3-319-69835-9_76.
[36] X. Ying, “An Overview of Overfitting and its Solutions”, Journal of Physics: Conference Series, vol. 1168, issue 2, 2019, https://doi.org/10.1088/1742-6596/1168/2/022022.
[37] M. Tegmark, “An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence, https://futureoflife.org/ai-open-letter. accessed in April 2021.
[38] G. Miller, “Artificial Intelligence Project Success Factors: Moral Decision-Making with Algorithms”, Proceedings of the 16th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS, Vol. 25, pp. 379–390, 2021, http://dx.doi.org/10.15439/2021F26
[39] E. Ziemba, “The ICT adoption in enterprises in the context of the sustainable information society”, Proceedings of the Federated Conference on Computer Science and Information Systems, vol. 11 pp. 1031–1038, 2017, https://doi.org/10.15439/2017F89.
[40] M. Fanea-Ivanovici, M. Pană, "From Culture to Smart Culture. How Digital Transformations Enhance Citizens’ Well-Being Through Better Cultural Accessibility and Inclusion", IEEE Access, vol. 8, pp. 37988-38000, 2020, https://doi.org/10.1109/ACCESS.2020.2975542.

3rd Special Session on Data Science in Health, Ecology and Commerce

DATA Science in Health, Ecology and Commerce is a forum on all forms of data analysis, data economics, information systems and data based research, focusing on the interaction of those four fields. Here, data-driven solutions can be generated by understanding complex real-world (health) related problems, critical thinking and analytics to derive knowledge from (big) data. The past years have shown a forthcoming interest on innovative data technology and analytics solutions that link and utilize large amounts of data across individual digital ecosystems. First applications scenarios in the field of health, smart cities or agriculture merge data from various IoT devices, social media or application systems and demonstrate the great potential for gaining new insights, supporting decisions or providing smarter services. Together with inexpensive sensors and computing power we are ahead of a world that bases its decisions on data. However, we are only at the beginning of this journey and we need to further explore the required methods and technologies as well as the potential application fields and the impact on society and economy. This endeavor needs the knowledge of researchers from different fields applying diverse perspectives and using different methodological directions to find a way to grasp and fully understand the power and opportunities of data science.

This is a joint track by WIG2, the Scientific Institute for health economics and health service research, the Information Systems Institute of Leipzig University and the Helmholtz Environmental Research Institute.

Topics

We embrace a rich array of issues on data science and offer a platform for research from diverse methodological directions, including quantitative empirical research as well as qualitative contributions. We welcome research from a medical, technological, economic, political and societal perspective. The topics of interest therefore include but are not limited to:

- Data analysis in health, ecology and commerce
- (Health) Data management
- Health economics
- Data economics
- Data integration
- Semantic data analysis
- AI based data analysis
- Data based health service research
- Smart Service Engineering
- Integrating data in integrated care
- AI in integrated care
- Spatial health economics
- Risk adjustment and Predictive modelling
- Privacy in data science

Technical Session Chairs

- **Franczyk, Bogdan,** University of Leipzig, Germany
- **Militzer-Horstmann, Carsta,** WIG2 Institute for health economics and health service research, Leipzig, Germany
- **Häckl, Dennis,** WIG2 Institute for health economics and health service research, Leipzig, Germany
- **Bumberger, Jan,** Helmholtz-Centre for Environmental Research – UFZ, Germany
- **Reinhold, Olaf,** University of Leipzig / Social CRM Research Center, Germany

Program Committee

- **Alpkoçak, Adil,** Dokuz Eylul University
- **Cirqueira, Douglas,** Dublin City University
- **Dey, Nilanjan,** Techno India College of Technology, India
- **Kossack, Nils,** Head Mathematics and Statistics, WIG2 Institute for Health Economics and Health Service Research
- **Kozak, Karol,** Fraunhofer and Uniklinikum Dresden, Germany
- **Müller, Marco,** WIG2 Institute for Health Economics and Health Service Research
- **Popowski, Piotr,** Medical University of Gdańsk, Poland
- **Sachdeva, Shelly,** National Institute of Technology Delhi, India
- **Viehbahn, Malte,** WIG2 Institute for Health Economics and Health Service Research
- **Wasielewska-Michniewska, Katarzyna,** Systems Research Institute of the Polish Academy of Sciences, Poland

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 169–173
DOI: 10.15439/2021F115
ISSN 2300-5963 ACSIS, Vol. 26

Shorter Length of Stay Keeps the Doctor Away?

About the Influence of the Length of Hospital Stay on the Recovery

Felix Krüger, Tobias Schäffer and Gerrit Stahn

Martin Luther University Halle-Wittenberg
Große Steinstraße 73, D-06112 Halle (Saale), Germany
Corresponding Email: felix.krueger@wiwi.uni-halle.de

***Abstract*—Since at least the 1960s, the average length of stay in German hospitals has been declined. Early discharge can cause health risks for the patient and incurs cost risks for health insurers. Otherwise, a shorter length of stay can also indicate more efficient and better care in hospitals. The aim of this research project is therefore to investigate whether the decreasing length of stay has an effect on the quality of care provided by hospitals, and whether a shorter length of stay in inpatient care results in an increase in follow-up outpatient care. Routine data will be used.**

I. Research question and motivation

SINCE at least the 1960s, the average length of stay (ALOS) of patients in German hospitals has been declined (see [1]), and almost halved since the early 1990s (see [2] and [3]). It is assumed that the introduction of Diagnosis Related Groups (DRGs) in hospital reimbursement in 2003 was an additional driver of this development (see [4] and [5]). Compared to previously used equal daily reimbursement rates, the DRG system reduces a hospital's profit if a patient stays longer. The upper bound of the length of stay in a DRG determines up to which length of hospital stay a flat rate is paid (see [6]). As soon as the duration of stay in an individual case exceeds the upper bound, additional payments are made. However, these additional payments are unprofitable for the hospital (see [7]). This creates an economic incentive for hospitals to discharge patients as early as possible (see [8] and [9]). In order to counteract early and premature discharges for cost reasons, hospitals have to accept reductions in the per-case flat rates if the length of stay falls below the lower bound due to early discharge or transfer to another hospital (see [10]). As a result, hospitals generate the greatest profit per case at the lower bound. This is illustrated in Fig. 1. The difference between the amount of the flat rate payment per case **P** and the costs of the hospital stay **K** is greatest at the point of the lower bound **DRG_L**. Depending on the slope of the hospital's cost curve, profit is generated up to the point of the upper bound **DRG_U**.

Early discharge from the hospital can cause health risks for the patient and incurs cost risks for health insurers (see [11]). If, for example, more outpatient treatment, nursing care or readmission to hospital becomes necessary (*revolving door effect*), this can increase the total costs for the payer (see [12]).

This work was not supported by any organization

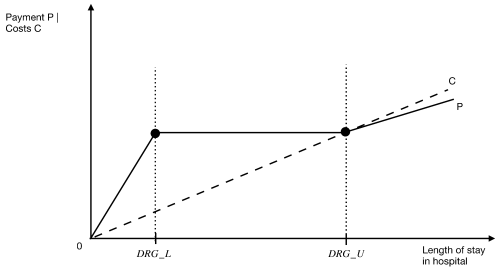


Fig. 1. Hypothetical relation between costs and returns

Source: Own representation

The implementation of the length of stay boundaries follows the objective of more efficient care in hospitals in terms of the *pay for performance* principle (see [13]). On the other hand, there is the concern that an excessively shortened length of stay will lead to underuse and misuse of care and an increased workload for medical staff (see [14] and [15]). Whether the DRG system leads to changes in the quality of care due to earlier discharge of patients is therefore still controversial and not clearly evident (see [16]).

The aim of this research project is therefore to investigate whether the decreasing length of stay has an effect on the quality of care provided by hospitals in Germany and whether a shorter length of stay in inpatient care results in an increase in follow-up outpatient care. More specifically, we aim to answer the following research questions:

1) Does the decreasing length of hospital stay have an effect on the quality of care or on the patient's health status after discharge?
2) Is a shorter length of stay substituted for increased follow-up outpatient care?

II. Brief literature review

Current scientific literature provides little clarity on the potential relationship between individual length of stay and a patient's quality of care. Previous studies could only show associations for specific areas of health care. If the length of stay is reduced solely on grounds of economic considerations, premature discharge can have negative health effects on the patient. Cases like these are also referred to as *bloody discharge* (see [11]). In the REDIA study by von Eiff *et al.* ([11]), the authors examined the effects of the introduction of the G-DRG system on rehabilitation. The authors argue that because of shorter hospital stays, orthopedic patients start rehab earlier and in poorer health. Nevertheless, the treatment goals of the rehab process could be achieved. The authors explain this with the increased treatment effort by the medical staff in the rehab facilities.

Other studies assess a decreasing length of stay less critically, as this can be interpreted as an indicator of an increase in process quality (see [17]). The patient's desired state of health is achieved more quickly through better treatment. Thus, improved treatment quality could also lead to a decrease in length of stay. That a decreasing length of stay could have no negative effects, but even positive effects, is suggested by the studies of Kehlet and Wilmore ([18]), of Husted *et al.* ([19]) & of Barmer GEK ([20]).

As already mentioned, since the introduction of the DRG-based reimbursement system in Germany, the length of stay has continued to decrease each year (see [3]). It should be noted that the length of stay was already declining before the introduction of the DRG-based system (see [16] and [1]). Therefore, it is still unclear whether the introduction of the DRG-based reimbursement system has an impact on length of stay. This makes research on the effects of this system on quality of care also interesting with regard to length of stay, since the reduction of length of stay is an objective of the introduction of the DRG-based system that has not yet been sufficiently evaluated.

In principle, too little empirical research has been done to investigate this relationship. Therefore, no reliable statements can be made yet. Research to date has essentially been based on structured quality reports. For Germany, for example, Fürstenberg *et al.* (see [21] as well as [22]) observed a general decline in post-hospital mortality in the period between 2004 and 2010. The effect of the DRG-based system and the length of stay on the quality of care remained unclear. The overall picture among international research is currently similarly unclear (see [23]). An overview of international literature is listed in Table I.

International research concerning other countries with DRG systems there also indicates a given shift from inpatient care to outpatient care structures (see [24]) and [25]). A shift from inpatient to outpatient care influenced by decreasing length of stay has not yet been observed in Germany. Therefore, it is still unclear whether patients show increased use of outpatient services as a result of earlier hospital discharges.

III. Data

Routine data from the research database of the WIG2 Scientific Institute for Health Economics and Health Systems Research [35] will be used primarily to answer the research questions. Routine data are the accounting data of the statutory health insurances. The pseudonymized personal reference of the data is of central importance for answering the research question of this thesis. This makes it possible to trace the individual treatment paths of the insured and thus analyze the influence of the length of stay. Since the methodology used requires the largest possible sample, the entire available scope of the research database will be used for the estimation. The observation of the individuals should take place on a monthly basis. Alternatively, the observation can be done quarterly, as this corresponds to the rhythm of ambulatory care.

These data are to be supplemented with the publicly available data on DRGs from the German Institute for the Hospital Remuneration System (InEK).

IV. Methodology

At the center of the empirical analysis is the investigation of the potential relationship between hospital length of stay and quality of care, as well as variables related to outpatient follow-up treatment. Because quality of care is not directly reflected in the data, recovery indicators such as mortality, medication use, complications, or comorbidities will be used as proxies. These indicators can provide information on whether and to what extent the patient's state of health has changed after hospitalization, depending on the length of stay. In addition, variables on further treatment, such as outpatient follow-up treatment (e.g., physician visits) and hospital readmission, can provide information on whether a shorter length of stay results in a shift in the care structure (e.g., from inpatient care to outpatient care). Quality indicators and variables for follow-up treatment are summarized below as outcome indicators. Simple OLS regressions of the length of stay on the indicators, as stated in model 1, would likely be biased as differences in characteristics between hospitals as well as seasonal variations most likely have an impact on outcome indicators and on the individual length of stay.

$$\mathbf{E_i} = \alpha_0 + \alpha_1 \cdot length_i + \theta_{\mathbf{X}} \cdot \mathbf{X_i} + \eta_i \tag{1}$$

with

$\mathbf{E_i}$	Vector of indicators for post stationary recovery or for quality of medical treatment
$length_i$	Length of stay for patient i
X_i	Vector of different control variables
η_i	Error term

Source	Country	Findings
[26]	USA	No effect of the DRG system on quality of care detected.
[27]	USA	No effect of the DRG system on quality of care detected.
[28]	USA	Shift from inpatient hospital care to lower-cost providers.
[29]	USA	Unclear whether the DRG system leads to a reduction in quality of care. Suggestive evidence for premature discharge.
[30]	USA	Unclear whether the DRG system leads to a reduction in quality of care. Suggestive evidence for premature discharge. (Rate of unstably discharged patients increased from 10% to 15% within 3 years after introduction of the DRG-based system).
[24]	Norway	Suspicion of treatment preference for patients with milder orthopedic diagnoses. Also, evidence of a shift from inpatient care to outpatient care.
[31]	Great Britain	No effect of the DRG system on quality of care detected.
[32]	Japan	DRG system introduction is associated with lower mortality and higher readmissions.
[25]	Great Britain	Expansion of better reimbursed hip TEP procedures compared to less highly reimbursed procedures. Also, evidence of a shift from inpatient care to outpatient care.
[33]	France	No effect of DRG system on readmissions after surgical procedures.
[34]	Switzerland	DRG system is associated with lower mortality and higher readmissions.

TABLE I
INTERNATIONAL LITERATURE

We expect that the model behind the structural equation 1 would be still biased by unobserved factors and that the exogeneity assumption is thus violated even if we control for hospital and time fixed effects. An example of an uncontrolled influencing factor of this kind is the varying adaptation of new treatment methods between hospitals, as well as the varying adaptation of technical innovations in medical care. Also, unobservable variables (such as the actual health status of patients, actual quality of care or the cost structure of hospitals) or measurement errors (incorrectly or incompletely maintained database) could lead to biases and violation of the exogeneity assumption. By means of an instrument variable estimation, an attempt can be made to counteract this problem. The upper and lower bounds of stay of the billed DRGs will be used as instruments for this purpose. As Figure 1 already illustrates, these boundaries are expected to have a relevant influence on the individual length of stay, since they determine the area of the greatest profit for the hospital. The boundaries applicable for a particular year are specified externally by the InEK in the respective previous year. The actual length of stay of the calculation hospitals from the respective previous year is used as the basis for this determination. Therefore, these calculation hospitals potentially have the opportunity to influence the length of stay boundaries in the next year with their discharge and transfer behavior. However, we do not assume that this potential influence is intentional or particularly high. For this to be the case, the calculation hospitals would have to behave strategically in a coordinated manner to increase the length of stay and the costs per case in the same way. However, such behavior seems rather unlikely. Furthermore, the overall trend towards decreasing length of stay and case costs do not suggest such behavior. It can therefore be assumed that the length-of-stay boundaries are set externally by InEK and that the calculation hospitals have little or no influence. Accordingly, these quasi-experimental circumstances result in the following stages of the IV-regression:

$$\begin{aligned} \text{First stage: } Length_i = &\beta_0 + \beta_2 \cdot DRG\_L_i + \\ &\beta_3 \cdot DRG\_U_i + \theta_{\mathbf{X}} \cdot \mathbf{X_i} + u_i \end{aligned} \tag{2}$$

$$\text{Second stage: } \mathbf{E_i} = \gamma_0 + \gamma_1 \cdot \widehat{Length}_i + \theta_{\mathbf{X}} \cdot \mathbf{X_i} + \epsilon_i \tag{3}$$

with

DRG_L_i	Lower limit length of stay per billed DRG
DRG_U_i	Upper limit length of stay per billed DRG
u_i, ϵ_i	Error terms

V. Next Steps

We consider the following points as the next main steps for our research project. First, quality indicators for individual diseases and procedures will be identified by means of a literature search. At the time of writing, the scope of research includes quality indicators for procedures such as appendectomies, transcatheter aortic valve implantation and the insertion of artificial hip joints. We are currently concentrating on these three medical procedures, since they are very common, and therefore we expect to have a high number of observations in the database. In a further step, the relevant DRGs will be derived from the relevant procedures and diagnoses to be included in the regression. Based on this, the dataset will be compiled and validated. After compiling the data, the described analyses can be performed and the results will be described.

References

[1] Statistisches Bundesamt, *Data about hospitals and benefit or rehabilitation facilities 1992 (in German)*. Wiesbaden: Fachserie 12: Gesundheitswesen, Reihe 6.1, 1992.

[2] B. Augurzky and A. Beivers, "Digitization and investment financing (in german)," in *Krankenhaus-Report 2019: Das digitale Krankenhaus*, J. Klauber, M. Geraedts, J. Friedrich, and J. Wasem, Eds. Berlin, Heidelberg: Springer, 2019, pp. 67–82. doi: 10.1007/978-3-662-58225-1_5

[3] Statistisches Bundesamt, *Hospital data 2017 (in German)*. Wiesbaden: Fachserie 12: Gesundheitswesen, Reihe 6.1.1, 2018.

[4] A. Beivers and L. Waehlert, "Control of the quantity dynamics according to the KHSG: Implications for hospitals (in German)," in *Entrepreneurship im Gesundheitswesen II: Geschäftsmodelle – Prozesse – Funktionen*, M. A. Pfannstiel, P. Da-Cruz, and C. Rasche, Eds. Wiesbaden: Springer Fachmedien, 2018, pp. 123–137. doi: 10.1007/978-3-658-14781-5

[5] T. Reinhold, K. Thierfelder, F. Müller-Riemenschneider, and S. N. Willich, "Health economic effects of the drg introduction in germany - a systematic overview (in german)," *Das Gesundheitswesen*, vol. 71(05), pp. 406—312, 2009. doi: 10.1055/s-0028-1119399

[6] InEK GmbH, *Proposal procedure for the integration of medical, scientific and other expertise in the further development of the G-DRG system (in German)*, 2019. [Online]. Available: www.g-drg.de/G-DRG-Vorschlagsverfahren2

[7] W. Fiori, H.-J. Lakomek, K. Buscham, H. Lehmann, A. Fuchs, F. Bessler, and N. Roeder, "Hospital financing 2015 - important aspects for internal rheumatology (in german)," *Zeitschrift für Rheumatologie*, vol. 74, pp. 447—455, 2015. doi: 10.1007/s00393-015-1605-2

[8] A. Geissler, M. Wörz, and R. Busse, "German hospital capacity in an international comparison (in german)," in *Krankenhaus-Report 2010: Krankenhausversorgung in der Krise?*, J. F. Jürgen Klauber, Max Geraedts, Ed. Stuttgart: Schattauer, 2010, pp. 25–40.

[9] G. Neubauer, "On the economic control of hospital care under drg flat rates (in german)," in *Krankenhaus-Report 2003: G-DRGs im Jahre 1*, H. S. Jürgen Klauber, Bernt-Peter Robra, Ed. Stuttgart: Schattauer, 2003, pp. 101–119.

[10] InEK GmbH, *Final report 2006 (in German)*, 2006. [Online]. Available: https://www.g-drg.de/Archiv/DRG_Systemjahr_2006_Datenjahr_2004#sm7

[11] W. Von Eiff, S. Schüring, B. Greitemann, and M. Karoff, "Redia - effects of the drg introduction on rehabilitation (in german)," *Die Rehabilitation*, vol. 50, no. 04, pp. 214–221, 2011. doi: 10.1055/s-0031-1275720

[12] Deutscher Ethikrat, "Patient welfare as an ethical benchmark for the hospitals (in german)," *Stellungnahme vom 5.April 2016*, 2016. [Online]. Available: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-patientenwohl-als-ethischer-massstab-fuer-das-krankenhaus.pdf

[13] J. Friedrich, W.-D. Leber, and J. Wolff, "Base case values - on the price and productivity development of stationary services (in german)," in *Krankenhaus-Report 2010: Krankenhausversorgung in der Krise?*, J. F. Jürgen Klauber, Max Geraedts, Ed. Stuttgart: Schattauer, 2010, pp. 127–47.

[14] Alliance for hospitals instead of factories (in German), *Fakten und Argumente zum DRG-System und gegen die Kommerzialisierung der Krankenhäuser*, Maintal, 2019. [Online]. Available: https://www.krankenhaus-stattfabrik.de/196

[15] M.-L. Müller, "Mid-term evaluation from the perspective of the german nursing council (in german)," in *Auswirkungen der DRG-Einführung in Deutschland. Standortbestimmung und Perspektiven*, F. Rau, N. Roeder, and P. Hensen, Eds. Stuttgart: W. Kohlhammer Verlag, 2009, pp. 32–36.

[16] R. Milstein and J. Schreyögg, "Empirical evidence on the effects of the introduction of the g-drg system (in german)," in *Krankenhaus-Report 2020: Finanzierung und Vergütung am Scheideweg*, J. Klauber, M. Geraedts, J. Friedrich, J. Wasem, and A. Beivers, Eds. Berlin, Heidelberg: Springer, 2020, pp. 25–39. doi: 10.1007/978-3-662-60487-8_2

[17] S. Hilgers, *DRG reimbursement in German hospitals: effects on length of stay and quality of treatment (in German)*. Springer-Verlag, 2011. doi: 10.1007/978-3-8349-6242-3

[18] H. Kehlet and D. W. Wilmore, "Fast-track surgery," *British Journal of Surgery*, vol. 92, no. 1, pp. 3–4, 2005. doi: 10.1002/bjs.4841

[19] H. Husted, G. Holm, and S. Jacobsen, "Predictors of length of stay and patient satisfaction after hip and knee replacement surgery: fast-track experience in 712 patients," *Acta orthopaedica*, vol. 79, no. 2, pp. 168–173, 2008. doi: 10.1080/17453670710014941

[20] Barmer GEK, "Hospital report 2010 (in german)," *Schwerpunktthema: Trends in der Endoprothetik des Hüft-und Kniegelenks. Schriftenreihe zur Gesundheitsanalyse*, vol. 3, 2010.

[21] T. Fürstenberg, M. Laschat, K. Zich, S. Klein, P. Gierling, H.-D. Nolting, and T. Schmidt, "G-drg accompanying research: final report of the second research cycle (2006–2008) (in german)," *Berlin: IGES*, 2011.

[22] T. Fürstenberg, M. Laschat, K. Zich, S. Klein, P. Gierling, T. Schmidt, and H.-D. Nolting, "G-drg accompanying research: final report of the third research cycle (2008–2010) (in german)," *Deutsche Krankenhaus-Verlagsgesellschaft, Düsseldorf*, 2013.

[23] J. O'Reilly, R. Busse, U. Häkkinen, Z. Or, A. Street, and M. Wiley, "Paying for hospital care: the experience with implementing activity-based funding in five european countries," *Health economics, policy and law*, vol. 7, no. 1, pp. 73–101, 2012. doi: 10.1017/S1744133111000314

[24] P. E. Martinussen and T. P. Hagen, "Reimbursement systems, organisational forms and patient selection: evidence from day surgery in norway," *Health Economics, Policy and Law*, vol. 4, no. 2, pp. 139–158, 2009. doi: 10.1017/S1744133109004812

[25] I. Papanicolas and A. McGuire, "Do financial incentives trump clinical guidance? hip replacement in england and scotland," *Journal of health economics*, vol. 44, pp. 25–36, 2015. doi: 10.1016/j.jhealeco.2015.08.001

[26] C. K. Davis and D. J. Rhodes, "The impact of drgs on the cost and quality of health care in the united states," *Health Policy*, vol. 9, no. 2, pp. 117–131, 1988. doi: 10.1016/0168-8510(88)90029-2

[27] M. W. Rich and K. E. Freedland, "Effect of drgs on three-month readmission rate of geriatric patients with congestive heart failure." *American Journal of Public Health*, vol. 78, no. 6, pp. 680–682, 1988. doi: 10.2105/ajph.78.6.680

[28] M. A. Sager, D. V. Easterling, D. A. Kindig, and O. W. Anderson, "Changes in the location of death after passage of medicare's prospective payment system," *New England Journal of Medicine*, vol. 320, no. 7, pp. 433–439, 1989. doi: 10.1056/NEJM198902163200705

[29] W. H. Rogers, D. Draper, K. L. Kahn, E. B. Keeler, L. V. Rubenstein, J. Kosecoff, and R. H. Brook, "Quality of care before and after implementation of the drg-based prospective payment system: a summary of effects," *Jama*, vol. 264, no. 15, pp. 1989–1994, 1990.

[30] J. Kosecoff, K. L. Kahn, W. H. Rogers, E. J. Reinisch, M. J. Sherwood, L. V. Rubenstein, D. Draper, C. P. Roth, C. Chew, and R. H. Brook, "Prospective payment system and impairment at discharge: the quicker-and-sicker'story revisited," *Jama*, vol. 264, no. 15, pp. 1980–1983, 1990.

[31] S. Farrar, D. Yi, M. Sutton, M. Chalkley, J. Sussex, and A. Scott, "Has payment by results affected the way that english hospitals provide care? difference-in-differences analysis," *Bmj*, vol. 339, 2009. doi: 10.1136/bmj.b3047

[32] H. Hamada, M. Sekimoto, and Y. Imanaka, "Effects of the per diem prospective payment system with drg-like grouping system (dpc/pdps) on resource usage and healthcare quality in japan," *Health Policy*, vol. 107, no. 2-3, pp. 194–201, 2012. doi: 10.1016/j.healthpol.2012.01.002

[33] A. Vuagnat, E. Yilmaz, A. Roussot, V. Rodwin, M. Gadreau, A. Bernard, C. Creuzot-Garcher, and C. Quantin, "Did case-based payment influence surgical readmission rates in france? a retrospective study," *BMJ open*, vol. 8, no. 2, pp. 1–9, 2018. doi: 10.1136/bmjopen-2017-018164

[34] A. Kutz, L. Gut, F. Ebrahimi, U. Wagner, P. Schuetz, and B. Mueller, "Association of the swiss diagnosis-related group reimbursement system with length of stay, mortality, and readmission rates in hospitalized adult patients," *JAMA network open*, vol. 2, no. 2, pp. 1–12, 2019. doi: 10.1001/jamanetworkopen.2018.8332

[35] WIG2 Scientific Institute for Health Economics and Health Systems Research. (2020) WIG2 Research database. [Online]. Available: https://www.wig2.de/analysetools/wig2-forschungsdatenbank.html

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 175–180
DOI: 10.15439/2021F113
ISSN 2300-5963 ACSIS, Vol. 26

Maximum Simulated Likelihood: Don't Stop Believin'?

Christopher Schrey
Lipsiusstrasse 44, 04317 Leipzig, Germany
Email: christopher.schrey@outlook.de

***Abstract*—Unobserved heterogeneity may complicate model estimation in econometrics. To integrate out the effect of unobserved heterogeneity via maximum simulated likelihood (MSL) estimation, assumptions regarding the underlying distribution need to be made. Researchers seldomly discuss these assumptions. This raises the question, to what extent estimation results in the MSL-context are robust to potential distributional mismatch. This work-in-progress derives the research question from the literature. A simulation study is conducted that underpins the relevance of this matter, where results imply that mismatch may introduce significant bias. Intended future work to properly address and answer this question is defined and discussed.**

I. Introduction

UNOBSERVED heterogeneity may complicate model estimation in (health) econometrics. When modelling discrete choice, such as patients decisions regarding health insurance plans, unobserved heterogeneity may come in the form of private information regarding awareness of and attitudes towards an individuals health risks, resulting in self-selection into healthcare plans [1], [2], [3]. Similarly, unobserved heterogeneity may occur in every aspect of commerce, such as when consumers choose among alternatively-fuelled vehicles [4], among energy efficient refrigerators [5] or among modes of transportation [6], while their preferences (i.e., coefficients) are allowed to vary randomly among their choices. Generally speaking, unobserved heterogeneity may be considered whenever researchers cannot measure patient or consumer characteristics that determine preferences or equivalently, whenever features of the alternatives that are chosen from remain unrecorded [4].

Econometricians need to address unobserved heterogeneity, that materialises either though self-selection or varying preferences among alternatives. When researchers make an assumption regarding the distribution of these unobservable factors, their effect can be integrated out. This can be achieved, among others, by conducting *maximum simulated likelihood* (MSL) estimation. Simulation refers to the fact that integration over a density is but a form of averaging [7]. By averaging the likelihood function over a sufficiently large number of draws from the assumed distribution, MSL-estimation becomes feasible. Put differently, researchers need to make an as assumption, which distribution to choose, herein after referred to as *assumed distribution*, to approximate the *true distribution* which is unknown. While several distributional forms may be assumed, researchers most frequently assume that their unobserved heterogeneity follows a normal distribution [8], [9].

Accordingly, the researchers' assumption regarding the assumed distribution seems to be a critical one. MSL-estimation may be sensitive to poor approximations of the simulated probabilities [10] and even the wrong amount (i.e., too little) or quality of random draws may jeopardise the reliability of the results [11]. But what if researchers choose the assumed distribution incorrectly, resulting in *distributional mismatch*? The consequences of such distributional mismatch do not seem to be adequately addressed within the relevant literature. Many [12], [13], [3], [14], [9], [15], [6], state they assume unobserved heterogWas folgt eneity to follow a normal distribution without any justification or further elaboration. Some [1], [2], [16] provide little context regarding their choice.

[1] state that they obtained similar results with the uniform and beta as assumed distribution as with choosing the standard normal distribution. [2] justify their assumption regarding the standard normal distribution to handle location invariance. The readers are informed by [16] that distributional mismatch within their model *" (. . .) would potentially lead to biased parameter estimates"*.

As such, the research question of this piece of work-in-progress is to investigate bias in parameter estimates due to distributional mismatch between assumed and true distribution. Specifically, the mismatch will be limited to mismatch within the normal distribution, i.e., mean and standard deviation. Addressing this research problem will be beneficial to both econometricians conducting analysis with MSL-estimation as well as the research community interpreting the respective results. Further tools and methods to detect such biases and to potentially correct them may follow.

To this end, the MSL-method and its features will be introduced and a simulation study conducted, which aims at identifying bias due to distributional mismatch. The results of the simulation study will be discussed and interpreted. The bias introduced by the mismatch, i.e., through mismatch in mean and standard deviation, will be approximated by two equations, that will serve as basis for further discussion. Intended future work to properly address and answer this question is defined and discussed.

[0]This work was not supported by any organisation

II. Unobserved Heterogeneity

An example of unobserved heterogeneity that can be addressed with MSL-estimation is provided by [8]. Their example will serve as basis and will be enhanced to serve as a simulation study subsequently.

Let y_i be individual i's (with $i = 1 \dots N$) outcome of a sample with size N. Here, y_i depends on the observable variable x_i times its coefficient α, which is additionally be influenced by unobservable heterogeneity u_i with coefficient β and a standard normally distributed error term ε, such that [8]

$$y_i = \alpha x_i + \beta u_i + \varepsilon_i. \tag{1}$$

While the standard normally distributed error terms ε might similarly be viewed as a source of unobserved heterogeneity, their effect could simply be taken into account by OLS-regression or regular maximum likelihood estimation.

The density of *y* conditional on *u* is given by [8]

$$f(y_i|x_i, u_i) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y_i - \alpha x_i - \beta u_i)^2}{2}\}. \tag{2}$$

Inference on *x* is based on the marginal density $f(y|u)$, which requires to integrate out the effect of *u* [8]. In the original case study by [8], the u's (true) distribution is the extreme value type 1 distribution. Here, for simplicity u's true distribution will be the normal distribution in different settings (regarding mean and standard deviation, as will be explained later). By drawing a number of S random draws from the distribution of *u*, their effect can be integrated out via simulation, hence the name maximum *simulated* likelihood. Given that the number of simulation draws S and sample size N both $S, N \to \infty$ while S increases faster than $\sqrt{N}$, such that $\sqrt{N}/S \to 0$, MSL is asymptotically normal, efficient and equivalent to maximum likelihood estimation [17], [7].[1] Here, MSL-estimation is achieved by drawing *S* random draws from the assumed distribution $\hat{u}$ of the unobserved heterogeneity *u* for each individual and averaging over each individual, such that [8]:

$$\ln \mathrm{L}_N = \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y_i - \alpha x_i - \beta \hat{u}_i^s)^2}{2}\} \right) \tag{3}$$

Put differently, an assumption regarding the true distribution of the unobserved heterogeneity needs to be made, so that it can be approximated by this assumed distribution. In this case study, the true distribution of the unobserved heterogeneity is known, such that the assumed distribution can be chosen correctly. This is a crucial point and the main focus of the study at hand: What if the assumption by the researcher does not match the true distribution, i.e., distributional mismatch occurs? Only a few of the before mentioned pieces of research offer a theoretical or practical justification for choosing the (standard) normal as the assumed distribution to match unobserved heterogeneity. Similarly, only few make the reader aware that their assumption may have consequences on the estimation results.

To this end, the here introduced unobserved heterogeneity example will be employed and modified to gain insights on the consequences of mismatching true and assumed distribution in the MSL-context. Although many distributional forms of unobservable heterogeneity seem plausible, e.g., extreme value or the uniform distribution, within this example the mismatch will be achieved by mismatching mean, i.e., μ vs. $\hat{\mu}$ (0 vs. 1), and standard deviation, i.e., $\hat{\sigma}$ vs. σ (1 vs. 2) across the normal distribution, as summarised in Table I. The underlying parameter choice is purely for experimental purposes and is not justified by any other reference. Each of the four constellations will serve as the true data-generating (i.e., unobserved heterogeneity) distribution and will be benchmarked against each of the other four as an assumed distribution which will be employed in MSL-estimation. This will result in sixteen cases, of which four times true and assumed distribution match, whereas in twelve scenarios a mismatch will occur. Table II provides an overview.

Within the simulation study, the α and β coefficients (cf. Equation 1) are to be estimated. Each time the assumed and true distribution match one another, the estimates for α and β, i.e., $\hat{\alpha}$ and $\hat{\beta}$, are hypothesised to be fairly close to their true values, i.e., $\alpha = \frac{1}{2}$ and $\beta = 1$. Yet, interest lies in the situation when a mismatch between assumed and true distribution occurs. It is unclear beforehand whether or not results will be biased and if so how much. This is the central question of this piece of research.

Due to the study design, mismatches will occur along two dimensions: Firstly, there will be four mismatches only among the mean of the assumed and true distribution. Secondly, there will be four mismatches only among the standard deviation of the assumed and true distribution. Also, there will be four mismatches along both dimensions. These twelve mismatches will be exploited for further analysis. Interest lies in the bias of the estimated $\hat{\alpha}$ vs. the true α, as well as the estimated $\hat{\beta}$ vs. the true β. If possible, the bias will be explained by the deviation in μ vs. $\hat{\mu}$ and $\hat{\sigma}$ vs. σ.

III. Preliminary findings

Each of the sixteen scenarios, as described in Table II was estimated 500 times, using [20], [21], [22]. Results are summarised in Figure 1, where the upper part displays the results for the estimates $\hat{\alpha}$, whereas the bottom presents results for $\hat{\beta}$. For each of the two coefficients the diagonal from the top left to the bottom right displays the four scenarios, in which the distributional parameters of u, i.e., $\sim \mathcal{N}(\mu, \sigma)$ and $\hat{u}$, i.e., $\sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$ match one another. As was expected, the observed values are fairly close to their respective true values, i.e, $\alpha = \frac{1}{2}$ and $\beta = 1$, which are represented by a grey vertical line in Figure 1.

Surprisingly, $\hat{\alpha}$ seems to respond differently to mismatches in mean and standard deviation of u than $\hat{\beta}$ does, which was

[1] How to know whether or not one has employed a sufficient amount of simulation draws is subject to another discussion [18], [19].

TABLE I
PARAMETER SUMMARY

Variable	Value	Description
α	.5	true coefficient of x
β	1	true coefficient of u
$\hat{\alpha}$		estimated coefficient of x
$\hat{\beta}$		estimated coefficient of u
x	1	observable characteristics
u	$\sim \mathcal{N}(\mu, \sigma)$	true unobservable heterogeneity
μ	$\{0, 1\}$	true mean
σ	$\{1, 2\}$	true standard deviation
$\hat{u}$	$\sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$	assumed unobservable heterogeneity
$\hat{\mu}$	$\{0, 1\}$	assumed mean
$\hat{\sigma}$	$\{1, 2\}$	assumed standard deviation
ε	$\sim \mathcal{N}(0, 1)$	error term
S	1,000	Number of simulation draws
N	1,000	Sample size
R	500	Number of repetitions

not anticipated. Yet, in hindsight, it makes sense, as $\hat{\beta}$ belongs to the unobservable u variable that is incorrectly approximated, whereas $\hat{\alpha}$ belongs to the x variable which can be observed. $\hat{\alpha}$ seems to be shifted away from the true value of α by the difference in true mean and assumed mean, amplified by the relation in mismatch of the standard deviation. The reaction of $\hat{\alpha}$ seems to be described by:

$$\hat{\alpha} = \alpha(1 + \mu - \hat{\mu}\frac{\sigma}{\hat{\sigma}}). \tag{4}$$

For each of the sixteen scenarios in the upper part of Figure 1, this Equation 4 is represented by a blue vertical line.

The reaction of $\hat{\beta}$ on the other hand does not seem to be influenced by any difference in true mean and assumed mean. Nevertheless, it seems to be shifted away from the true value of β by the relation in mismatch of the standard deviation. The reaction of $\hat{\beta}$ can be approximated by:

$$\hat{\beta} = \beta\frac{\sigma}{\hat{\sigma}}. \tag{5}$$

For each of the sixteen scenarios in the bottom part of Figure 1, this Equation 5 is represented by a red vertical line. One notable exception for the latter Equation 5 is the behaviour of $\hat{\beta}$ where the true $u \sim \mathcal{N}(\mu = 1, \sigma = 1)$ and the assumed $\hat{u} \sim \mathcal{N}(\hat{\mu} = 0, \hat{\sigma} = 2)$ (second row from the top, third column from the left, bottom part of Figure 1). In this case, $\hat{\beta}$ seems to be represented both as implied by Equation 5 as well as its negative, even though the former occurred more often than the latter.

IV. DISCUSSION AND OUTLOOK

The lack in guidance regarding potential bias due to mismatch in true and assumed distribution in MSL-estimation motivated this simulation study. It seemed unclear, to what extent the estimation coefficients may be biased from distributional mismatch of mean and standard deviation within the normal distribution. This lead to an back-of-the-envelope calculation, resulting in Equation 4 and Equation 5. These two equations were deduced from the underlying results and seem to approximate the bias in $\hat{\alpha}$ vs. α and $\hat{\beta}$ vs. β fairly well, except for one notable exception, as mentioned in section III. Nevertheless, they are only trial-and-error approximations of the observed results.

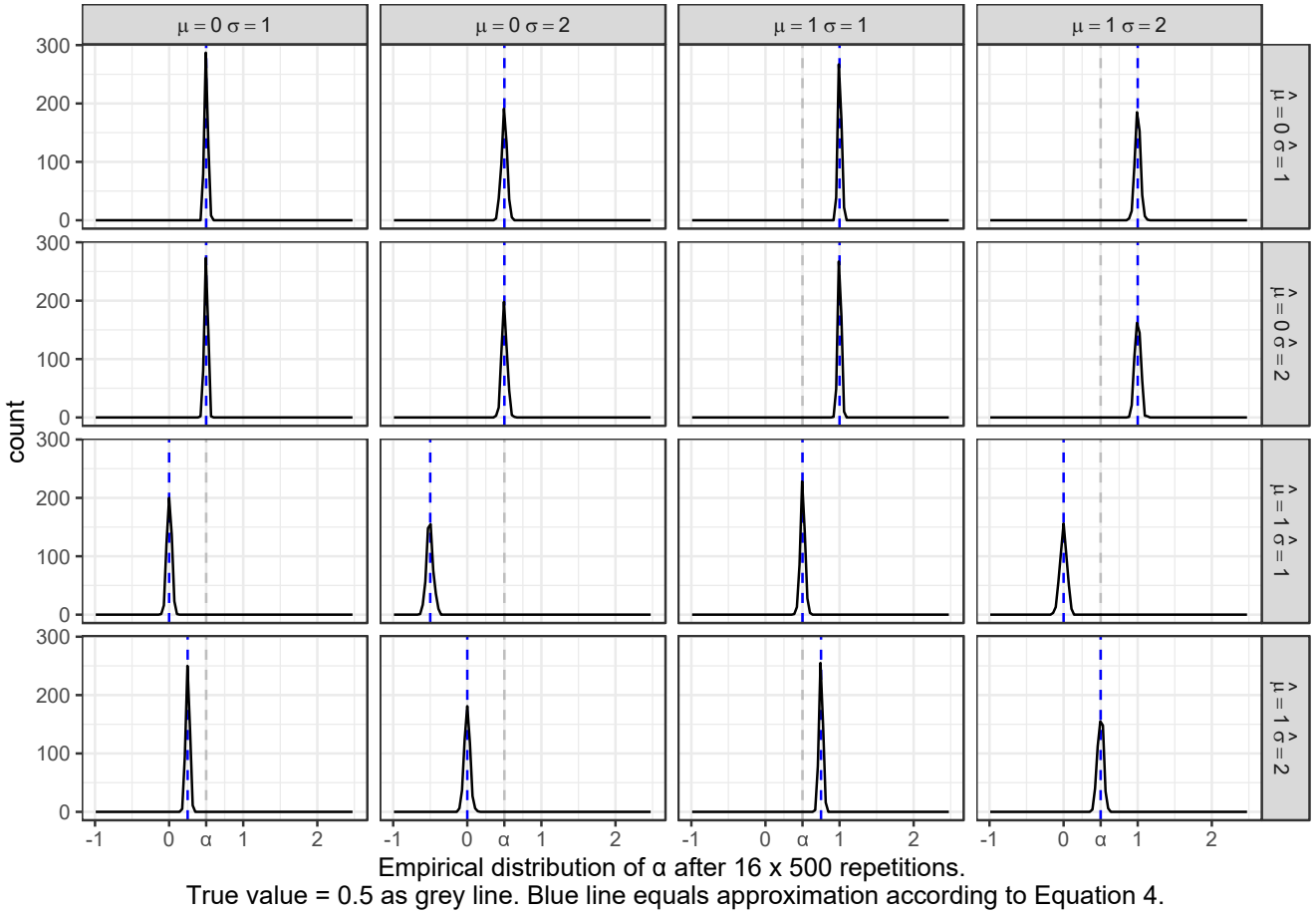
While the lack of guidance, such as provided by Equation 4 and Equation 5, was the motivation to looking for it in the first place, it needs to be assumed that such relation were found and discussed earlier. Yet, this would similarly raise the question why, if it was already common knowledge, none of the found pieces of research that apply MSL-estimation pointed out to this direction when discussing limitations of their models and findings?

Future intended work is motivated by this question: A more quantitatively comprehensive and qualitatively structured literature research will be conducted in the realm of what is described by [23] as *Maximum Approximated Likelihood*, i.e., MSL-estimation, Gaussian-quadrature and integration on sparse grids. The main focus will be placed on the *distributional assumption* regarding the assumed distributions, its theoretical materialisation, i.e., whether it is applied to varying preferences or endogeneity. Variation in the latter findings will then be structured among the dimensions:

- **scope**: theoretical vs. applied papers,
- **estimation method**: e.g., MSL-estimation, Gaussian-quadrature, integration on sparse grids,
- **models**: e.g., mixed multinomial, multinomial treatment regression [24] and
- **field of research**: e.g., healthcare, commerce, transportation.

Additional interest lies in finding pieces of applied research that already had similar findings as given by Equation 4 and Equation 5, as it is assumed that these findings were made already earlier.

Additionally, and equivalently important, remains the further exploration of the bias induced by distributional mismatch between assumed and true distribution in the simulation-context. Depending on the findings of the literature review, Equation 4 and Equation 5 may be further explored, as especially Equation 5 could not approximate all of the sixteen scenarios. As of now it remains unclear, whether or not the findings of Equation 4 and Equation 5 may be applicable to any other situation than the underlying (toy) example. To potentially detect distributional mismatch, consequences regarding the log-Likelihood seems promising with respect to diagnostic tests, such as the Likelihood-ratio test. Similarly, consequences of variance reduction techniques will be discussed.



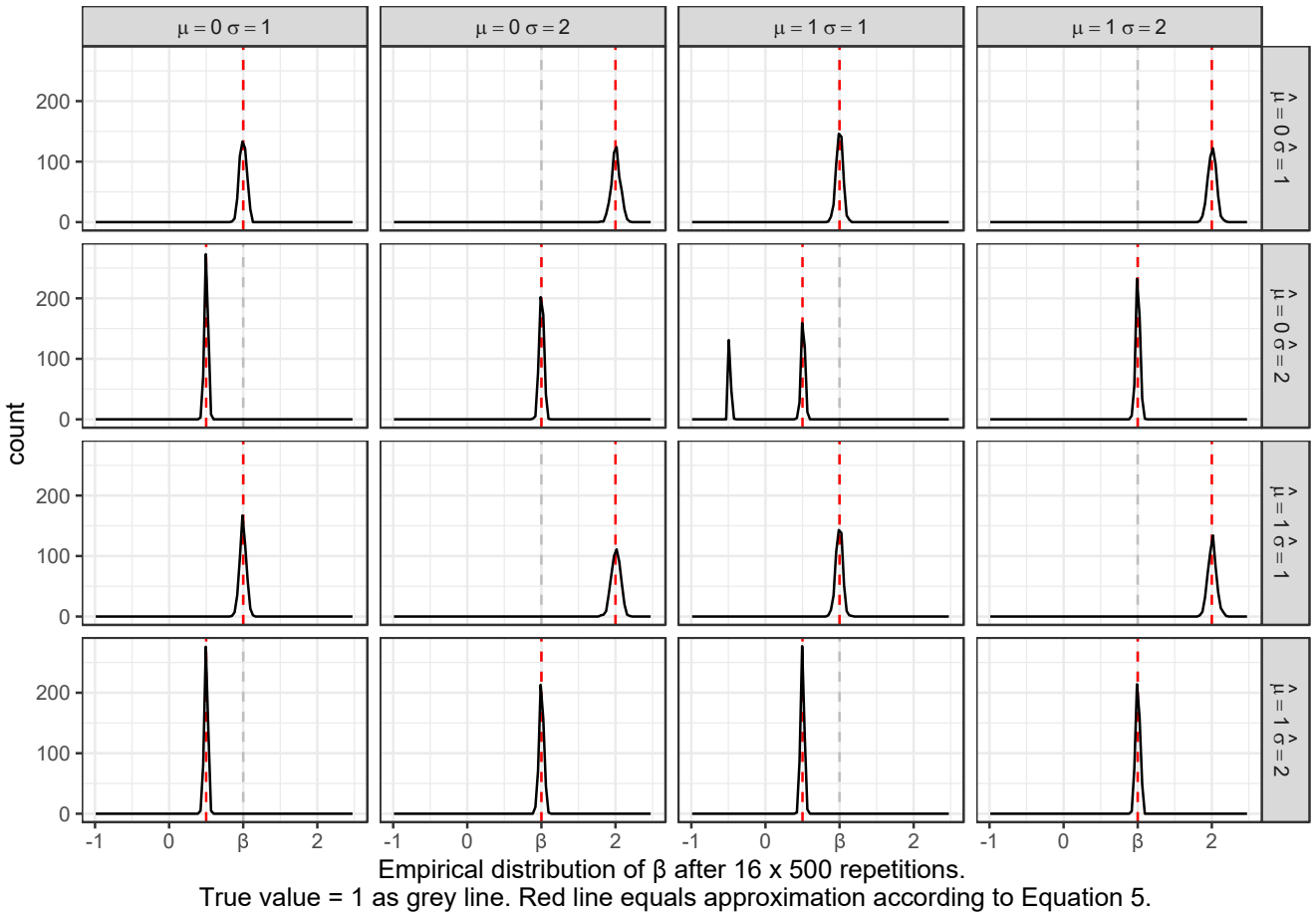


Fig. 1: RESULTS OF THE SIMULATION STUDY: EMPIRICAL DISTRIBUTION OF THE $\hat{\alpha}$ (TOP) AND $\hat{\beta}$ (BOTTOM) COEFFICIENTS.

TABLE II
SETUP OF SIMULATION STUDY: DISTRIBUTIONAL MISMATCH

		True			
		$\mu = 0, \sigma = 1$	$\mu = 0, \sigma = 2$	$\mu = 1, \sigma = 1$	$\mu = 1, \sigma = 2$
Assumed	$\hat{\mu} = 0, \hat{\sigma} = 1$	match	mismatch ($\hat{\sigma}$)	mismatch ($\hat{\mu}$)	mismatch ($\hat{\mu}, \hat{\sigma}$)
	$\hat{\mu} = 0, \hat{\sigma} = 2$	mismatch ($\hat{\sigma}$)	match	mismatch ($\hat{\mu}, \hat{\sigma}$)	mismatch($\hat{\mu}$)
	$\hat{\mu} = 1, \hat{\sigma} = 1$	mismatch ($\hat{\mu}$)	mismatch ($\hat{\mu}, \hat{\sigma}$)	match	mismatch ($\hat{\sigma}$)
	$\hat{\mu} = 1, \hat{\sigma} = 2$	mismatch ($\hat{\mu}, \hat{\sigma}$)	mismatch ($\hat{\mu}$)	mismatch ($\hat{\sigma}$)	match

REFERENCES

[1] P. Deb, C. Li, P. K. Trivedi, and D. M. Zimmer. "The effect of managed care on use of health care services: results from two contemporaneous household surveys". In: *Health economics* 15.7 2006, pp. 743–760. DOI: http://dx.doi.org/10.1002/hec.1096.

[2] P. Deb and P. K. Trivedi. "Specification and simulated likelihood estimation of a non–normal treatment–outcome model with selection: Application to health care utilization". In: *The Econometrics Journal* 9.2 2006, pp. 307–331. DOI: http://dx.doi.org/10.1111/j.1368-423X.2006.00187.x.

[3] D. Shane and P. K. Trivedi. "What drives differences in health care eemand? The role of health insurance and selection bias". In: *Health, Econometrics and Data Group (HEDG) Working Papers* 2012. URL: https://www.york.ac.uk/media/economics/documents/herc/wp/12_09.pdf.

[4] D. McFadden and K. Train. "Mixed MNL models for discrete response". In: *Journal of Applied Econometrics* 15.5 2000, pp. 447–470.

[5] D. Revelt and K. Train. "Mixed logit with repeated choices: households' choices of appliance efficiency level". In: *The Review of Economics and Statistics* 80.4 1998, pp. 647–657. DOI: http://dx.doi.org/10.1162/003465398557735. URL: https://direct.mit.edu/rest/article/80/4/647/57083/Mixed-Logit-with-Repeated-Choices-Households.

[6] D. Munger, P. L'Ecuyer, F. Bastin, C. Cirillo, and B. Tuffin. "Estimation of the mixed logit likelihood function by randomized quasi-Monte Carlo". In: *Transportation Research Part B: Methodological* 46.2 2012, pp. 305–320. DOI: http://dx.doi.org/10.1016/j.trb.2011.10.005.

[7] K. E. Train. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press, 2009. DOI: http://dx.doi.org/10.1017/CBO9780511805271.

[8] A. Cameron and P. K. Trivedi. *Microeconometrics: Methods and applications*. New York, NY: Cambridge University Press, 2005.

[9] C. R. Bhat. "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model". In: *Transportation Research Part B: Methodological* 35.7 2001, pp. 677–693. DOI: http://dx.doi.org/10.1016/S0191-2615(00)00014-X.

[10] L. Chiou and J. L. Walker. "Masking identification of discrete choice models under simulation methods". In: *Journal of Econometrics* 141.2 2007, pp. 683–703. DOI: http://dx.doi.org/10.1016/j.jeconom.2006.10.012.

[11] L. M. Andersen. "Obtaining reliable likelihood ratio tests from simulated likelihood functions". In: *PloS one* 9.10 2014, e106136. DOI: http://dx.doi.org/10.1371/journal.pone.0106136.

[12] M. Bratti and A. Miranda. "Endogenous treatment effects for count data models with endogenous participation or sample selection". In: *Health economics* 20.9 2011, pp. 1090–1109. DOI: http://dx.doi.org/10.1002/hec.1764.

[13] M. B. Buntin, C. H. Colla, P. Deb, N. Sood, and J. J. Escarce. "Medicare spending and outcomes after postacute care for stroke and hip fracture". In: *Medical care* 48.9 2010, pp. 776–784. DOI: http://dx.doi.org/10.1097/MLR.0b013e3181e359df.

[14] M. M. Garrido, P. Deb, J. F. Burgess, and J. D. Penrod. "Choosing models for health care cost analyses: issues of nonlinearity and endogeneity". In: *Health services research* 47.6 2012, pp. 2377–2397. DOI: http://dx.doi.org/10.1111/j.1475-6773.2012.01414.x.

[15] Z. Sándor and K. Train. "Quasi-random simulation of discrete choice models". In: *Transportation Research Part B: Methodological* 38.4 2004, pp. 313–327. DOI: http://dx.doi.org/10.1016/S0191-2615(03)00014-6.

[16] D. Brunner, F. Heiss, A. Romahn, and C. Weiser. *Reliable estimation of random coefficient logit demand models: DICE Discussion Papers*. 2017. URL: https://EconPapers.repec.org/RePEc:zbw:dicedp:267.

[17] C. Gouriéroux and A. Monfort. *Simulation-based econometric methods*. Oxford University Press, 1997. DOI: http://dx.doi.org/10.1093/0198774753.001.0001.

[18] M. Czajkowski and W. Budziski. "Simulation error in maximum likelihood estimation of discrete choice models". In: *Journal of Choice Modelling* 31 2019, pp. 73–85. DOI: http://dx.doi.org/10.1016/j.jocm.2019.04.003.

[19] C. Schrey, T. Schäffer, C. Militzer-Horstmann, and N. Kossack. "Maximum Simulated Likelihood: Don't stop 'til you get enough?" In: *Position Papers of the 2019 Federated Conference on Computer Science and Information Systems*. Annals of Computer Science and Information Systems. PTI, 2019, pp. 79–82. DOI: http://dx.doi.org/10.15439/2019F354.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. URL: https://www.R-project.org/.

[21] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York, 2016. URL: https://ggplot2.tidyverse.org.

[22] Lionel Henry and Hadley Wickham. *purrr: functional programming tools*. 2020. URL: https://CRAN.R-project.org/package=purrr.

[23] M. Griebel, F. Heiss, J. Oettershagen, and C. Weiser. "Maximum approximated likelihood estimation". In: INS Preprint No. 1905 2019. URL: https://ins.uni-bonn.de/media/public/publication-media/INSPreprint1905.pdf?pk=1424.

[24] P. Deb and P. K. Trivedi. "Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment". In: *Stata Journal* 6.2 2006, pp. 246–255.

16th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from three complimentary directions: management of information systems in an organization, uses of information systems to empower managers, and information ssytems for sutainable development. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in organizations. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome. Papers about the influence of information systems on sustainability are also expected.

TOPICS

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles
- Information Systems for Sustainability
  - Sustainable business models, financial sustainability, sustainable marketing
  - Qualitative and quantitative approaches to digital sustainability
  - Decision support methods for sustainable management

TECHNICAL SESSION CHAIRS

- **Arogyaswami, Bernard,** Le Moyne University, USA
- **Chmielarz, Witold,** University of Warsaw, Poland
- **Jankowski, Jarosław,** West Pomeranian University of Technology in Szczecin, Poland
- **Karagiannis, Dimitris,** University of Vienna, Austria
- **Kisielnicki, Jerzy,** University of Warsaw, Poland
- **Ziemba, Ewa,** University of Economics in Katowice, Poland

PROGRAM COMMITTEE

- **Janis Bicevskis,** University of Latvia, Latvia
- **Alberto Cano,** Virginia Commonwealth University, United States
- **Vincenza Carchiolo,** Dipartimento di Matematica e Informatica - Universita di Catania, Italy
- **Beata Czarnacka-Chrobot,** Warsaw School of Economics, Poland
- **Pankaj Deshwal,** Netaji Subhas Institute of Technology, India
- **Robertas Damasevicius,** Silesian University of Technology, Poland
- **Monika Eisenbardt,** Univeristy of Economics Katowice, Poland
- **Marcelo Fantinato,** University of São Paulo, Brazil
- **Renata Gabryelczyk,** University of Warsaw, Poland
- **Nitza Geri,** The Open University of Israel, Israel
- **Dariusz Grabara,** University, Economics in Katowice, Poland
- **Jarosław Jankowski,** West Pomeranian University of Technology in Szczecin, Poland
- **Andrzej Kobylinski,** Warsaw School of Economics, Poland
- **Christian Leyh,** Technische Universität Dresden, Germany

- **Karolina Muszyńska,** University of Szczecin, Poland
- **Tomasz Parys,** University of Warsaw, Poland
- **Uldis Rozevskis,** University of Latvia, Latvia
- **Nina Rizun,** Gdansk University of Technology, Poland
- **Andrzej Sobczak,** Warsaw School of Economics, Poland
- **Jakub Swacha,** University of Szczecin, Poland
- **Symeon Symeonidis,** Democritus Univesity of Thrace, Greece
- **Oskar Szumski,** University of Warsaw, Poland
- **Jarosław Wątróbski,** University of Szczecin, Poland
- **Janusz Wielki,** Opole University of Technology, Poland
- **Dmitry Zaitsev,** Odessa State Environmental University, Ukraine
- **Marek Zborowski,** University of Warsaw, Poland

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 183–190
DOI: 10.15439/2021F39
ISSN 2300-5963 ACSIS, Vol. 26

Assessing Enterprise Governance of Information Technology Maturity Models in Middle East and North Africa Region

Mostafa M. AlShamy, Walid M. Abdelmoez, Essam Eldean Elfakharany
College of Computing and Information Technology,
Arab Academy for Science, Technology and Maritime Transport, Egypt
Email: mostafa.alshamy@egybyte.net, walid.abdelmoez@aast.edu, essam.elfakharany@aast.edu

Hany H. Ammar
Lane Department of Computer Science and Electrical Engineering,
West Virginia University, WV, USA
Email: hammar@wvu.edu

***Abstract*—Enterprise Governance of IT (EGIT) is an important topic for academics and practitioners in the context of achieving enterprise goals while optimizing resource utilization and risk management. EGIT is playing a critical role in developing countries as resources are rare and risk levels are higher. There is a need for EGIT Maturity Models (MMs) in Middle East and North Africa (MENA) region detected by delivering and analyzing two questionnaires which were shared with a group of participants working in the field. The obtained results have been generalized and consolidated into a generally applicable requirement list covering the specific needs of MENA region. The results of this paper reveal that although there are some global EGIT MMs used in MENA region which cover some maturity dimensions, there is a lack of easy-to-use integrated multi-dimensional EGIT MMs specific for the region needs.**

I. Introduction

MATURITY Models (MMs) are techniques developed and used to determine the level of performance, capability or maturity of process or organization [1]. They are used to discover organizations strength and weakness points to enable them to define respective opportunities for improvement. They are also used to determine maturity targets and how to reach them. We tried to get Middle East and North Africa (MENA) region respective stakeholders involved by developing and sharing two questionnaires to know more about how their organizations select, use, and integrate Enterprise Governance of IT (EGIT) MMs. The analysis of these two questionnaires revealed a demand in the MENA region for having an EGIT MM with special characteristics covering the local needs and context.

MENA region has some specific needs based on the nature of its member countries as they all are developing countries with emerging economies based on the natural and human resources they have and great opportunities for improvement. At the same time, the Arab Peninsula countries are emerging with their eagerness to achieve great improvements over short times due to their new economic strategies which include many dimensions include information technology, cybersecurity, and data management among others. The region is starting to believe in the importance of corporate governance and EGIT in achieving the national and organizational goals and objectives effectively. Many counties like Egypt, Kingdom of Saudi Arabia (KSA), United Arab Emirates (UAE) among others begin to have strategic vision for 2030 and many respective initiatives with clear goals, roles and responsibilities and actual measurement techniques.

Organizations in MENA region are becoming more interested in EGIT as we discovered that around 80% of organizations are trying to implement or have already implemented an EGIT MM based on the conducted questionnaires. Those organizations need a MM that can enable them to measure their EGIT maturity and guide them to improve their performance to achieve their goals and comply with emerging regulations while optimizing resources and risk. None of MMs examined in this research uses stage-based and multi-dimensional maturity measurement methodology as they are just using separate dimensions and maturity levels except for COBIT 2019 [2] which uses different dimensions but does not have stage-based maturity measurement methodology. Therefore, these MMs enforce the interested organizations to use more than one of them together. At the same time, to measure all respective processes/aspects of the organization against each maturity level is specially considered huge effort for small and medium organizations. Therefore, many organizations cannot implement EGIT measurement and improvement easily due to the lack of a single easily integrated MM. We could not find any information about any EGIT MM which was developed in or for the MENA region to cover its context and maturity level. Therefore, we will assess the MMs in MENA region and define the needs of its organizations.

In this paper we present the result of assessing EGIT MMs in the MENA region and the specific needs of organizations working in it which are interested in measuring and improving their EGIT maturity. It should be noted that EGIT here has governance stands for Governance, Risk and Compliance (GRC) combined. We aim to define the needs of the organizations working in the MENA region to know if there is a need for a new EGIT MM, or the existing MMs are effective and efficient. If a new EGIT MM is needed, we target to identify its characteristics to guide researchers who may be interested in developing one. The objective is to design questionnaires to cover EGIT management and usage

and provide them to respective representatives from some organizations working in different fields and representing different sized organizations. Their answers will be analyzed properly. This will lead us to know the MENA region EGIT MM needs and customization.

This paper is organized as the following. Section 2 covers the literature review. Section 3 covers assessing organizations using EGIT MM in the MENA. Section 4 covers conclusions and future work.

While maturity itself is defined by Rosemann and de Bruin [3] as "a measure to evaluate the capabilities of an organisation in regard to a certain discipline", Becker et al [1] define MM as "conceptual models that outline anticipated, typical, logical, and desired evolution paths towards maturity". They are also used to determine maturity targets and how to reach them. MMs can have three purposes [3]:

- Descriptive which measures the current state (AS-IS) of an entity,
- Prescriptive which determines the desired state (TO-BE) of an entity and
- Comparative which allows entities to benchmark.

Descriptive MMs measures the current existing maturity levels in organization against predefined maturity levels to enable organizations to know their actual achievements. This enables organizations to understand their capabilities and weaknesses based on neutral assessment and analysis techniques.

Prescriptive MMs enable organizations to determine which future maturity level suites their goals and objectives that can be achievable too. They help organizations in defining maturity targets to follow by initiating improvement initiatives and assign needed competent resources.

Comparative MMs enable different organizations to compare their maturity achievements in a benchmark style. It is a great type of MMs, but it needs many arrangements to guarantee its effectiveness and efficiency. It enables organizations to rank their maturity in a specific market or field. Participating organizations must accept to share specific information with other entities including the other participating organizations to enable the MM to measure the actual maturity level and there shall be an external neutral assessor to manage the whole process professionally.

II. Literature Review

There are more than 150 MMs developed and published in the last few years as stated by de Bruin et al. [3] to support IT management. and in a research conducted by Becker et al [1] they found more than one thousand academic articles probably dealing with MMs published during the period of 1994 to 2009 when they applied a maturity model keyword search in ten scientific databases. When they tried to extend their analysis in 19 pure IS journals their search resulted in 20 articles that focus on MM. They discovered that there is no clear guidance on how to develop a MM using a scientific methodology. Becker's procedure model [4] is considered the greatest source of guidance for developing any EGIT MM due to the simple and scientific eight requirements provided.

We conducted a search for Maturity Model and IT Governance Maturity Model key words in three major publishers indexed in Scopus with good Cite Score which are IEEE, Springer Nature and Elsevier. The search covered MM in two geographical locations which are worldwide and MENA region. The result of the search is depicted in Table 1. There are no IT Governance MMs in the MENA region based on IEEE and Elsevier while there is a few found on Springer Nature. After examining those found on Springer Nature, we found them not related to IT Governance by any means.

The existent MMs are belonging to one of two different approaches, the first one of them is the commercial approach which is based on the efforts of big service providers and bodies of knowledge. The other approach is the academic one with many researchers who attempted to develop MMs while they do not have enough resources and capabilities like the first approach. This part will cover the existent MMs and compare among them.

Although there are many existing MMs in the field of information and technology, all of them lack one or more needed EGIT dimensions and some of them are not targeting EGIT. There is a need to assess the existing MMs from the organizations and stakeholders' perspectives. Therefore, we developed and shared two questionnaires to collect and analyze stakeholders' feedback to have general overview of the current situation of EGIT MMs in the MENA region. The need for a new MM with specific characteristics which suite more organizations in the region has be detected.

A. Maturity Models Classification

The first approach which is commercial EGIT MMs will be covered here by three of the most famous maturity/capability models in the market and two ISO

TABLE 1.
MM AND IT GOVERNANCE MM IN LEADING SCIENTIFIC JOURNALS.

Search Keyword	**IEEE**		**Elsevier**		**Springer Nature**	
	Worldwide	**MENA**	**Worldwide**	**MENA**	**Worldwide**	**MENA**
MM	3,165	N/A	27,452	N/A	7,115	39
IT Governance MM	86	N/A	340	N/A	71	3

standard which many companies were trying to use to know their maturity level and how to improve it. IT Infrastructure Library (ITIL) framework [5] is considered the most famous public framework for IT Service Management (ITSM) for the last thirty years and it has so many practitioners in the MENA who have attended its training courses and took certification exams. It has a lifecycle for any IT service which includes five stages empowered by twenty-six processes, four main IT functions and many techniques for managing IT services and increasing the customer satisfaction in a measurable manner. ITIL v3 and its 2011 update included a MM called Process Maturity Framework (PMF) [5] which is an easy-to-use multi-purpose ITSM MM that measures the maturity of ITSM processes using five maturity levels from 1 to 5 which are called initial; repeatable; Defined; Managed and Optimizing. Each level measures five areas which are considered dimensions and they are Vision and steering; Process; People; Technology and Culture. It is an ITSM MM while it can be used to measure any other domain. It assesses all the processes against each maturity level.

Control Objectives for Information and Related Technology (COBIT 5) [6] was considered the most famous public framework for Governance of Enterprise Information and Related Technology from 2012 to 2018 when it was replaced by COBIT 2019, and it has a Process Capability Model [7]. It has many practitioners in the MENA who have attended its training courses and took certification exams. It differentiated between governance and management and the processes of each. It was created based on other best practices which are governance principles from ISO/IEC 38500 [8], risk management from ISO 31000 [9], enterprise architecture from TOGAF [10], project management methodology from PRINCE2 [11] and PMBOK [12], information security management from ISO/IEC 27001 [13], application capability measurement from CMMI [14], IT service management processes from ITIL 2011 [4] and ISO/IEC 20000 [15]. COBIT5 has thirty-two management processes in four domains and five governance processes in one domain. COBIT5 has capability model called Process Capability Model which is built on the internationally recognized ISO/IEC 15504-2 standard [16] for Software Engineering - Process Assessment standard. ISACA which is the owner and developer of COBIT5 refuses to use the term maturity as it assumed that maturity can be used for measuring many dimensions of an organization and not only one as COBIT5 which was just measuring processes, so it will be reasonable to use capability instead of maturity. It measures the capability of IT governance and management processes using six capability levels from 0 to 5 which are Incomplete Process; Performed Process; Managed Process; Established Process; Predictable Process and Optimising Process. Each capability level has four ratings which are Fully (> 85 % to 100 % achievement), Largely (> 50 % to 85% achievement), Partially (> 15 % to 50 % achievement) and Not achieved (0 to 15 % achievement). It has nine attributes within the second to the sixth capability levels. The attributes found in a specific capability level shall be fully achieved so that the assessment can go for the next level. It is not an easy-to-use multi-purpose EGIT MM for MENA as many organizations do not have many of its processes and do not have enough resources to conduct its complex assessment. COBIT5 has only one dimension which is Process and therefore it measures capability and not maturity, other dimensions are still needed like information security, business continuity and compliance. It assesses all the processes against each maturity level and a simpler version tailored for the needs of MENA region is needed to cover its specific needs.

COBIT 2019 [2] is the new version of COBIT5 which was released by ISACA at the end of 2018. Now it has more processes as it has 35 processes for IT Service Management and 5 processes for Governance. It is not an easy-to-use multi-purpose EGIT MM [17]. It uses CMMI® Development 2.0 [14] process capability scheme. It has four dimensions which are Process, Organizational Structures, Information Items and Culture and Behavior. It has the same six maturity levels and four ratings like COBIT5. It is not easy to be used as it has six maturity levels including nine attributes and four ratings per each. It needs training, experience, and more resources to be implemented. It covers four dimensions, and therefore it measures capability for each dimension and maturity for all of them combined. It covers ITSM, information security, continuity, and compliance as processes and not as dimensions. It assesses all the processes against each maturity level.

ISO/IEC 15504-2 [16] is a guidance ISO standard created for process improvement and process capability determination. It is not an easy-to-use multi-purpose process MM. It has only one dimension which is Process. It has six capability levels which are Incomplete, Performed, Managed, Established, Predictable and Optimizing. It has nine attributes covering the second to the sixth capability levels. It is dedicated to process measurement and provided an exemplar software life cycle process assessment model. It is not easy to be used in MENA region due to its complexity and resource consuming style. It needs training and experience to be implemented. It covers one dimension, which is process, and therefore it measures capability and not maturity, other dimensions are still needed. It assesses all the processes against each maturity level.

ISO 19600 [18] is a guidance ISO standard published in 2014 and was created to provide organizations with guidance on how to comply with regulations and avoid fines by having a compliance management system. It is not a MM nor provides a maturity measurement like the other MMs mentioned above. Like many ISO standards it can measure compliance to its requirements by having one of two states which are conformity or non-conformity. It has only one dimension which is compliance which was missing in all the other mentioned MMs except for COBIT. It could be easy to be used in MENA region due to its straightforward requirements and maturity measurement technique and the increasing number of emerging regulations in the region. It covers one dimension and therefore other dimensions are

still needed. It was replaced by ISO 37301 [19] which was released in 2021.

For MENA region, ITIL is considered the best one for ITSM while COBIT is considered the best one for EGIT. But they still need to be customized to cover MENA region specific requirements.

The second approach covers relevant academic governance of IT maturity/capability models that represent researchers' participation which does not reach to proper audience in many cases. Although, de Bruin stated that there are more than 150 MMs in the last few years, the related work here represents the most related MMs or their development guidance. The related works can be divided into three categories which are the first category proposing MMs, the second category comparing among the developed MMs and the third category providing guidance on how to develop a scientific MM.

GoCoMM: A Governance and Compliance Maturity Model [17] by G. Gheorghe et al, Toward an IT Governance Maturity self-assessment Model Using EFQM and CobiT [20] by S. Arezki et al, Maturity Model Architect A Tool for Maturity Assessment Support [21] by Diogo Proença et al, Using Enterprise Architecture Model Analysis and Description Logics for Maturity Assessment [22] by D. Proença et al, Software process improvement and capability determination [23] by A. Mas et al and An Overview of the Business Process Maturity Model (BPMM) [24] by Jihyun Lee et al, are representing the first category which propose MMs. All these researchers tried to develop and propose a MM related to one aspect or more of EGIT.

Comparing among the developed MMs which is the second category is represented by MATURITY MODELS IN IS RESEARCH [1] by J. Becker et al, The Maturity Models for Information Systems - A State of the Art [25] by D. Proença et al and Understanding maturity models Results of a Structured Content Analysis [26] by Kohlegger, M., Maier, R., & Thalmann, S. The researchers are comparing among a group of proposed and released MMs and trying to discover their respective shortcomings.

Providing guidance on how to develop a scientific MM which is the third category is represented by Information Governance Maturity Model - Final Development Iteration [27] by Proença et al, What makes a useful maturity model? a framework of general design principles for maturity models and its demonstration in business process management [28] by J. Pöppelbuß et al, Maturity assessment models: a design science research approach [29] by T. Mettler et al and Developing Maturity Models for IT Management – A Procedure Model and its Application [30] by J. Becker et al, Assessing Organizational Capabilities: Reviewing and Guiding the Development of Maturity Grids [31] by Anja M. Maier et al, What Makes A Useful Maturity Model? A Framework Of General Design Principles For Maturity Models And Its Demonstration In Business Process Management [32] by Jens Pöppelbuß and Maximilian Röglinger, IT Evaluation in Business Groups: A Maturity Model [33] by Hamel, F., Ph, T., Falk, H., & Walter, U, Understanding the Main Phases of Developing a Maturity Assessment Model, (December) [34] by Bruin, T. De, Freeze, R., & Rosemann, M and A Design Science Research Perspective on Maturity Models in Information Systems [35] by Mettler, T. In this category researchers tried to provide other researchers who are interested in developing MMs with guidance on how to develop and evaluate MMs properly.

Although the first category which proposes new MMs and the second category which compares among the already developed MMs are important, the last category which is providing guidance on how to develop a MM is very important as we will use its provided guidance in understanding how to develop a scientific MM if needed for the MENA region. The following Table 2 gives a summary of the covered dimensions of the MMs used in the market

TABLE 2.
MM COMPONENTS COVERED BY EXISTING MMS AND ISO STANDARDS.

References	ITSM	Information Security	Business Continuity	Compliance	Process	People	Technology
ITIL PMF	*	*	*		*	*	*
COBIT 5/2019	*	*	*	*	*	*	*
ISO/IEC 15504-2					*		
M_o_R MM							
P3M3 MM							
ISO/IEC 20000-1:2018	*	*	*		*	*	
ISO/IEC 27001:2013		*	*		*	*	*
ISO 2230-1:2019			*		*	*	*
ISO 19600:2014/37301:2021				*	*	*	
GoCoMM					*		
Toward an IT Governance Maturity self-assessment Model Using EFQM and CobiT					*		
Maturity Model Architect A Tool for Maturity Assessment Support					*		
Using Enterprise Architecture Model Analysis and Description Logics for Maturity Assessment					*		
A Formalization of the ISO/IEC 15504 Enabling Automatic Inference of Capability Levels					*		

TABLE 3.
EXISTING MMS AND ISO STANDARDS FEATURES AND MAIN AREAS OF APPLICATION IN ORGANIZATIONS.

References	Features			Main Areas of Application
	Public/Proprietary	Easy to use	Descriptive (D)/ prescriptive (P)/ comparative (C)	
ITIL PMF	Public	Yes	(D)/(P)/(C)	ITSM
COBIT 5/2019	Public	No	(D)/(P)/(C)	EGIT
ISO/IEC 15504-2	Public	No	(D)/(P)/(C)	Process Measurement
M_o_R MM	Public	Yes	(D)/(P)/(C)	Risk Management
P3M3 MM	Public	Yes	(D)/(P)/(C)	Portfolio, Program, and Project Management Maturity Model
ISO/IEC 20000-1:2018	Public	Yes	(D)/(P)/(C)	ITSM
ISO/IEC 27001:2013	Public	Yes	(D)/(P)/(C)	Information Security
ISO 2230-1:2019	Public	Yes	(D)/(P)/(C)	Business Continuity
ISO 19600:2014/37301:2021	Public	Yes	(D)/(P)/(C)	Compliance Management

whether they are commercial or academic to depict the differences among them. It will start with the commercial MMs and then the academic MMs. No MM covers all EGIT dimensions with stage-based maturity levels.

Table 3 covers the most famous MMs used currently in the market and all of them are not academic ones. It compares among these MMs based on their features and main areas of application in organizations. The features show whether they are public or proprietary, ease of use and whether they are descriptive, prescriptive, or comparative.

B. MENA Region Evaluation

We searched Elsevier, IEEE and Springer Nature for MENA EGIT MM and we could not find any MMs which were developed in MENA region or specially developed for it. Although we have many EGIT MMs and regulations developed out of MENA region and used worldwide, we can find only regulations developed and enforced in the MENA region. These regulations are like Saudi National Cybersecurity Authority (NCA) cybersecurity controls and SAMA business continuity management and cybersecurity frameworks in the Kingdom of Saudi Arabia or the Egyptian Personal Information Protection Act among others. We could not find any EGIT MMs developed in or for MENA region that care about its context and special needs.

III. ASSESSING ORGANIZATIONS USING EGIT MM IN THE MENA REGION

The study aims to understand the existing EGIT MMs in the MENA region which should enable organizations to improve their enterprise governance of IT in an easy and affordable manner while helping them to comply with emerging regulations. The quantitative approach is used in this research by developing and distributing two questionnaires, which were developed and published using Google Docs, to 118 participants who are working in EGIT and related functions including IT, Information Security, GRC, QA, Business Analysis, IT Service Management, IT Project Management, Infrastructure, IT Networks, IT Operations, Performance Management and Audit. The first questionnaire [36], which was published in 2019, asks the participants about their organizations' behavior regarding EGIT using eighteen questions while the second one [37], which was published after a few months in 2020, asks them about their organizations' behavior towards EGIT and regulations using thirty questions. The number of participants in the first questionnaire is eighty-three and the number of the second questionnaire participants is thirty-five and they provided valuable feedback which enriched our research with market needs. The first questionnaire is dedicated to EGIT MMs while the second one is concentrating on compliance MMs in addition to EGIT MMs. The number of participants in the second questionnaire is less than of those of the first one as the number of those who are interested in compliance is less than the number of those who are generally interested in EGIT. The participants of the two questionnaires are working in micro, small, medium, and large enterprises operating in Egypt, Saudi Arabia, UAE, Sudan, Jordan, Yemen, among other countries in MENA.

The main objective of these two questionnaires is to define how organizations in the MENA region manage the EGIT maturity measurement using MMs. Both are structured questionnaires with a group of sequence questions leading the participants to describe how their organizations measure their EGIT maturity levels by using MMs. They start with asking the participants about their organizations type, size, location, and sector in addition to the position of participants. Then, a group of multiple-choice questions is provided to participants with the ability to choose one or more options that match their environment.

The two questionnaires were trying to get answers for questions like:

- Whether organizations define and internally publish their strategies, goals, applicable regulations, and the implications of not complying with them?
- Whether organizations define their EGIT goals and map them to applicable regulations with respective annual improvement initiatives and how to measure their achievements?
- Whether organizations measure the maturity of their EGIT and how?
- Whether organizations use one or more MMs to measure the maturity of their EGIT and for which purpose (Comparative, Prescriptive and Comparative)?
- The ease of use of the used MMs and cost of implementation in addition to the need for professional training?
- Whether the used MMs are scientifically developed, and which dimensions are included?
- The first-, second- and third-party assessments applicability of the used MMs and the type of assessors?
- The preferred EGIT MM dimensions and language?
- The need for one integrated MM or different ones to measure organizations' EGIT maturity and compliance to respective regulations?

The results gained after analyzing the answers of participants confirmed our point of view that MENA region needs an easy-to-use integrated multi-dimensional EGIT MM that suites the specific context of the region.

Table 4 gives a small set of the most important preferences of users of the MMs that depict the characteristics of the MMs used in MENA market based on the answers of the first questionnaire participants. It is clear that some of the participants' organizations just measure the achievement of goals and objectives instead of using market well-known MMs. About third of the participants confirmed that they use more than one MM while half of participants stated that the used MMs are developed using a scientific methodology and are easy to be used. More than half of the participants confirmed that they use MMs for descriptive, prescriptive, or comparative purposes, while a higher percentage of participants stated that they need training to implement these MMs. Some of the participants stated that their MMs are expensive. Few participants stated that their MMs are easy to be used for a self-assessment.

These statistics show that MENA market needs to have a scientifically developed EGIT MM that integrates other market well-known MMs while having descriptive, prescriptive, and comparative purposes. It should be easy-to-understand and easy-to-use. If the MM will be publicly free, it will increase the number of its users and specially the organizations which do not have enough resources. The analysis of the first questionnaire also depicts the needs of MENA market for all types of MMs and if there is one MM covering First, Second and Third-party assessments it will cover different segments of organizations. Self-assessment and easy-to-use capabilities are also needed to enable small organization to measure their EGIT too.

Table 5 gives a summary of the dimensions of the MMs used in MENA region based on the answers of the first questionnaire participants. Not less than half of the participants confirmed that their organizations have one or more of the measured dimensions in their used EGIT MMs. Based on their use, the dimensions are ordered in a descending manner from Process, Technology, Risk, Projects, Programs, and Portfolio, collectively, Compliance, Management Commitment to People.

All the provided percentages show big demand for all these EGIT dimensions in MENA region, and it will be great having them combined in just one integrated MM.

Table 6 gives a summary of the preferences of the users of the existing MMs in MENA market based on the answers of the second questionnaire participants. A high percentage of participants stated that their organizations need to use MM for measuring EGIT and compliance and more than half of participants preferred to use one integrated MM instead of using many MMs in addition to the ability of having internal and external assessors. This means that there is a big need for an integrated EGIT MM which can be used by internal and external assessors.

Based on the analysis of the second questionnaire, Table 6 gives a summary of how the organizations started to define applicable regulations and measure their compliance. A high percentage of organizations define applicable regulations and the implications of not complying with them and map their goals to applicable regulations. Measuring compliance

TABLE 4.
CHARACTERISTICS AND TYPES OF MMS IN MENA REGION BASED ON 1ST QUESTIONNAIRE.

Aspect	Percentage
Organizations use market well known MMs	23.5
Organizations just measure goals/objectives	28.4
Organizations use more than one MM	37.3
Scientifically developed MMs	50.8
Easy to use MMs	47.5
MMs that need training for implementation	78
Expensive MMs	35.6

TABLE 5.
DIMENSIONS OF MMS IN THE MENA REGION BASED ON 1ST QUESTIONNAIRE.

Aspect	Percentage
Management commitment	50.8
Process	72.9
People	50.8
Technology	71.2
Risk	59.3
Projects, programs, and portfolio	55.9
Compliance	55.9

TABLE 6.
PREFERENCES OF USERS OF MMS IN THE MENA REGION BASED ON 2ND QUESTIONNAIRE.

Aspect	Percentage
Organizations need to use EGIT MM	79.8
Organizations prefer to use one integrated MM	57.1
Organizations prefer internal and external assessors	63.1
Organizations need compliance MM	77.1
Defined applicable regulations and the implications of non-compliance	74.3
Organizations map their goals to applicable regulations	62.9
Internal auditors measure compliance to respective regulations	40
External auditors measure compliance to respective regulations	42.8
Organizations prefer Arabic/English MM	51.4
Organizations prefer English MM	37.1
Organizations prefer Arabic MM	11.4

with respective regulations by internal and external auditors is almost equal.

For language preference, the highest percentage is for Arabic/English MM, then English alone and the least percentage is for Arabic alone. It should be noted that all Arab Gulf Region countries have a high percentage of foreign labor who do not speak Arabic and their second language is English. The International Labour Organization (ILO) stated on its 2021 published Global Estimates on International Migrant Workers report [38] that the Arab States are considered the third highest sub-region hosting the majority of international migrant workers with 14.3% after Northern, Southern and Western Europe with 24.2% and Northern America with 22.1%.

The current increase in the IT and Personal Indefinable Information (PII) regulations worldwide and in MENA region affects the EGIT in all organizations which do not like to breach these regulations and have many risks like losing reputation, having to pay huge fines and penalties or legal implications.

After analyzing the results of the conducted two questionnaires we discovered that there is a great need in the MENA region for an EGIT MM that suites the characteristics of the region. These statistics can be used in the future to develop a customized EGIT MM for the MENA region.

IV. CONCLUSIONS AND FUTURE WORK

MMs importance is increasing every day due to the increasing dependance in our world on IT services and their respective regulations. There are many IT and EGIT MMs in the market which were developed globally. At the beginning there was a belief that there is a demand for measuring EGIT maturity in MENA region organizations due to the increase of EGIT awareness level and the increase in emerging cybersecurity and PII regulations. A literature review was conducted leading us to know that there is a lack of scientifically developed EGIT MMs worldwide and that famous MMs are concentrating on limited number of dimensions to measure. Two dedicated questionnaires were developed and shared with 118 participants who are working in EGIT and related functions in MENA region to understand if the currently used EGIT MMs are effective and efficient or there is a need to develop other MMs.

Based on the literature review, our published two questionnaires and the analysis of the answers of the participated individuals, we found that many organizations use worldwide known best practices frameworks and standards to measure their EGIT maturity. Based on 57.1% of participating individuals, there is a need to develop an integrated MM for measuring the EGIT that suites the MENA region specific context instead of using more than one MM. An inexpensive, easy-to-use, and scientifically developed multi-purpose, multi-dimensional and stage-based maturity levels MM is needed to cover organizations' needs.

The results of this paper can provide guidance to other researchers who are interested in developing EGIT MM for developing countries and especially those in MENA region. Those researchers can use this guidance in developing our recommended research topics:

- Develop MMs and all their components to enable organizations to use them for assessing their EGIT maturity with the capability of multi-purpose, multi-dimension and stage-based maturity levels maturity assessment.
- Develop guidance for those who would like to be assessors on how to develop their knowledge about any newly developed MM, what are the needed skills and how to use the developed MMs.
- Choose some of the interested organizations and support them in implementing the developed MMs to check the validity of the developed MMs and how organizations can accept the idea of having just one multi-dimensional stage-based MM.
- Develop awareness program and gain participating organizations top management commitment from the beginning to make all stakeholders comply and reduce change resistance as much as possible.

REFERENCES

[1] Becker, Joerg; Niehaves, Bjoern; Poeppelbuss, Jens; and Simons, Alexander (2010), "Maturity Models in IS Research". ECIS 2010 Proceedings. Paper 42.

[2] ISACA, COBIT® 2019 FRAMEWORK: INTRODUCTION & METHODOLOGY, (2018).

[3] de Bruin, Tonia & Rosemann, Michael. (2005). Towards a Business Process Management Maturity Model. Proceedings of the 13th European Conference on Information Systems. 521–532.

[4] J. Becker, R. Knackstedt, and J. Pöppelbuß, (2009) "Developing Maturity Models for IT Management," Bus. Inf. Syst. Eng., vol. 1, no. 3, pp. 213–222

[5] AXELOS ITIL Homepage, https://www.axelos.com/best-practice-solutions/itil. Accessed 20/7/20.

[6] Office of Government Commerce, Service Design, UK, TSO (2007), pp. 263.

[7] ISACA, A Business Framework for the Governance and Management of Enterprise IT (2012).
[8] ISO organization, ISO-IEC 38500:2008 Corporate governance of information technology.
[9] ISO organization, ISO 31000:2009 Risk management — Principles and guidelines.
[10] Open Group Homepage, https://www.opengroup.org/togaf. Accessed 20/7/20.
[11] AXELOS PRINCE2 Homepage, https://www.axelos.com/best-practice-solutions/prince2. Accessed 20/7/20.
[12] PMI PMBOK Homepage, https://www.pmi.org/pmbok-guide-standards. Accessed 20/7/20.
[13] ISO organization, ISO-IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements.
[14] CMMIINSTITUTE Homepage, https://cmmiinstitute.com. Accessed 20/7/20.
[15] ISO organization, ISO-IEC 20000-1:2018 Information technology — Service management —
[16] ISO organization, ISO-IEC 15504-2:2003 Information technology — Process assessment — Part 2- Performing an assessment.
[17] G. Gheorghe, F. Massacci, and A. Pretschner, (2009) "GoCoMM : A Governance and Compliance Maturity Model ∗," pp. 33–37.
[18] ISO organization, ISO 19600:2014 Compliance management systems - Guidelines.
[19] ISO organization, ISO 37301:2021 Compliance management systems — Requirements with guidance for use.
[20] S. Arezki and Y. Elhissi, (2018) "Toward an IT governance maturity self-assessment model using EFQM and CobiT," Proc. Int. Conf. Geoinformatics Data Anal. - ICGDA '18, pp. 198–202.
[21] Diogo Proença and José Borbinha, (2018) "Maturity Model Architect: A Tool for Maturity Assessment Support" - 10.1109/CBI.2018.10045
[22] D. Proença and J. Borbinha, (2018) "Using enterprise architecture model analysis and description logics for maturity assessment," Proc. 33rd Annu. ACM Symp. Appl. Comput. - SAC '18, no. November, pp. 102–109.
[23] A. Mas and A. Mesquida, (2018) "Software process improvement and capability determination conference 2017," Comput. Stand. Interfaces, vol. 60, no. November 2018, pp. 1–2.
[24] Lee, J., Lee, D., & Kang, S. (2007). An Overview of the Business Process Maturity Model (BPMM) BT - Advances in Web and Network Technologies, and Information Management. In K. C.-C. Chang, W. Wang, L. Chen, C. A. Ellis, C.-H. Hsu, A. C. Tsoi, & H. Wang (Eds.) (pp. 384–395). Berlin, Heidelberg: Springer Berlin Heidelberg.
[25] D. Proença and J. Borbinha, (2016) "Maturity Models for Information Systems - A State of the Art," Procedia Comput. Sci., vol. 100, no. December, pp. 1042–1049.
[26] Kohlegger, M., Maier, R., & Thalmann, S. (2009). Understanding maturity models Results of a Structured Content Analysis, (December 2016).
[27] Proença, Diogo & Vieira, Ricardo & Borbinha, José. (2017). Information Governance Maturity Model Final Development Iteration. 128-139. 10.1007/978-3-319-67008-9_11.
[28] J. Pöppelbuß and M. Röglinger, (2011) "What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management," Proc. IEEE Int. Eng. Management Conf., no. August 2014, pp. 244–249.
[29] T. Mettler, (2011) "Maturity assessment models: a design science research approach," Int. J. Soc. Syst. Sci., vol. 3, no. 1/2, p. 81.
[30] J. Becker, R. Knackstedt, and J. Pöppelbuß, (2009) "Developing Maturity Models for IT Management," Bus. Inf. Syst. Eng., vol. 1, no. 3, pp. 213–222.
[31] Maier, A. M., Moultrie, J., & Clarkson, P. J. (2012). Assessing Organizational Capabilities: Reviewing and Guiding the Development of Maturity Grids.
[32] Poeppelbuss, J., & Roeglinger, M. (2011). What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management
[33] Hamel, F., Ph, T., Falk, H., & Walter, U. (2013). IT Evaluation in Business Groups: A Maturity Model.
[34] Bruin, T. De, Freeze, R., & Rosemann, M. (2005). Understanding the Main Phases of Developing a Maturity Assessment Model, (December).
[35] Mettler, T. (2009). A Design Science Research Perspective on Maturity Models in Information Systems.
[36] M. Alshamy, (2019) "Multi-Purpose and Dimension Enterprise Governance of IT (EGIT) Maturity Model (MM). Internet: https://docs.google.com/forms/d/e/1FAIpQLSf0ForiyeJ46qVeew5MYdynYAedwIaRXXyjGpk63SJmUfaRZg/viewform?usp=sf_link
[37] M. Alshamy, (2020) "Multi-Purpose and Multi-Dimension Enterprise Governance of IT (EGIT) Maturity Model (MM) for MEA. Internet: https://docs.google.com/forms/d/e/1FAIpQLSe8p_ZmTl13wQo6gFrHyDHfqE9o0JYViZ86luosWq8csHSnIA/viewform?usp=sf_link.
[38] ILO Global Estimates on International Migrant Workers, https://www.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---migrant/documents/publication/wcms_808946.pdf. Accessed 16/7/21.

27^{th} Conference on Knowledge Acquisition and Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference “Knowledge Acquisition and Management". The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we've joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

TOPICS

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

TECHNICAL SESSION CHAIRS

- **Hauke, Krzysztof,** Wroclaw University of Economics, Poland
- **Nycz, Malgorzata,** Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pondel, Maciej,** Wroclaw University of Economics, Poland

PROGRAM COMMITTEE

- **Abramowicz, Witold,** Poznan University of Economics, Poland
- **Andres, Frederic,** National Institute of Informatics, Tokyo, Japan
- **Bodyanskiy, Yevgeniy,** Kharkiv National University of Radio Electronics, Ukraine
- **Chmielarz, Witold,** Warsaw University, Poland
- **Christozov, Dimitar,** American University in Bulgaria, Bulgaria
- **Jan, Vanthienen,** Katholike Universiteit Leuven, Belgium
- **Mercier-Laurent, Eunika,** University Jean Moulin Lyon3, France
- **Sobińska, Małgorzata,** Wroclaw University of Economics, Poland
- **Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Vasiliev, Julian,** University of Economics in Varna, Bulgaria
- **Zhu, Yungang,** College of Computer Science and Technology, Jilin University, China

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 193–200
DOI: 10.15439/2021F139
ISSN 2300-5963 ACSIS, Vol. 26

Characteristic and comparison of UML, BPMN and EPC based on process models of a training company

Marcin Nizioł, Piotr Wiśniewski, Krzysztof Kluza and Antoni Ligęza
AGH University of Science and Technology
al. A. Mickiewicza 30, 30-059 Krakow
Email: {wpiotr,kluza,ligeza}@agh.edu.pl

***Abstract*—We describe, characterize and compare three selected modeling notations of business processes: Unified Modeling Language, Business Process Model and Notation, as well as Event-Driven Process Chain. Using processes implemented in a training company, the selected notations were discussed in detail. We compare various aspects, such as modeling notation origin, the number of graphical elements included. Moreover, notations were analyzed using the 4+1 architectural view model. Justified results of the survey conducted among employees of above-mentioned organization let us conclude that there exist notation differences. Both BPMN and EPC allow the process architects to prepare more precise and legible models than UML.**

***Index Terms*—Business Process Management, BPMN, EPC, UML, process modeling**

I. Introduction

BUSINESS process modeling is a graphical representation of processes taking place in organizations. Process models are most often developed by process analysts. They present how the organization and its structures work. At the same time, they provide information that helps define the way the organization should act and indicates the direction of change. Such models may be further automated, as growing interest in the robotic process automation might be observed [1].

Modeling of business processes allows observation of their implementation, and thus the optimization of processes (simplification, increased transparency) or duration. It indicates which employee will be responsible for implementation at a given stage, and also allows to determine who is responsible for a given fragment of the implemented process. The most popular notation of business process modeling is Business Process Model and Notation (BPMN) while Unified Modeling Language (UML) or Event-Driven Process Chain (EPC) are also used successfully.

In this paper, the aforementioned notations have been thoroughly characterized with the help of original drawings, which present graphic elements, fragments or models of the entire processes of a training company.

The paper is organized as follows: Section II presents basic information about the UML language. Section III characterizes the BPMN notation while Section IV describes the EPC language together with the ARIS methodology [2]. We also compare these three ways of modeling business processes (Section V). The list is based on existing sources and proprietary models. The two processes of the above-mentioned organization are presented graphically – each in three variants. This section also presents the results of the survey of employees of an enterprise dealing with the organization of training. The last section (numbered as Section VIII) presents a summary containing the overall conclusions of the conducted comparison and presents ideas for future work.

II. Unified Modeling Language (UML)

The Unified Modeling Language (UML) [3] from the Object Management Group (OMG) is a standardized notation for modeling object-oriented software applications [4]. This multipurpose modeling language offers a variety of notations to capture different aspects of software [5], [6]. UML has become the dominant notation among software engineers and attempts to be a universal visual notation for software design.

UML is a quite complex notations, which makes it hard to understand by non-technicians [7] and is not suitable for all aspects of modelling [8]. Although it was created for modeling IT systems and is constantly developed in this area, it can be successfully used as a notation of business process modeling [9]–[14].

Due to the fact that UML is very popular and widely known, we decided to use it also at the business level. An analogy was also noted between an IT system and a business process – both are modeled from two perspectives – the structure and course of the process (dynamic structure). However, it should be remembered that Unified Modeling Language is not dedicated to business processes, and using it for this purpose may carry some risk – for example, ambiguous interpretation of the model caused by inconsistent understanding of the presented elements in the context of the created business process [15].

Although the UML offers over a dozen of diagram types, in the case of business process models, activity diagrams are used the most often [16]. A UML activity diagram is responsible for presenting the system dynamics. This is the type that is used to prepare material for business analysts in organizations. Activity diagrams are also used in modeling systems, algorithms or use case scenarios. The following graphic elements are used to construct these types of diagrams:

- **Activity** – behavior of the actor of the modeled process. Presented as a rectangle with rounded edges. One activity can consist of more than one subactivity. Action names are formulated in imperative mode, most often placed inside the element;
- **Action** – specifies the activity. The graphic notation of this element is the same as for activities;
- **Control Flow** – represents the relationship between actions and activities, as well as the sequence of flow between them. Represented by an arrow;
- **Start node** – the point that initiates the start of the process. Represented as a filled black circle. Most often there is one beginning for one diagram, while there may be more in modeling complex processes and systems;
- **End node** – the point where the process ends. In the diagram it will be an empty circle with a black dot inside. Activity diagrams may appear more than once;
- **Flow End** – the moment at which the selected control flow is stopped. It can occur repeatedly. Presented using unfilled crossed with two lines circle;
- **Decision node** – represented by a diamond. This is the place where the decision is made determining further control flows. A logical condition is placed next to the decision block (in square brackets, in the form of an infinitive). The number of outgoing flows from the block depends on the number of results of the logical condition. These also receive their names. For the decision block to make sense, the output results must be mutually exclusive.

Figure 1 shows an example UML activity diagram prepared for the administration department of the training company. The diagram represents a fragment of the large onboarding process of a new employee who needs to have a valid certificate after training in OHS rules to start working in a new position. The model consists of a small number of elements. The first step is checking by an employee of the administration department whether the person being onboarded does not have valid documents. The flow then goes into the decision block (*Valid Certificate*), from which two control flows are outgoing. One of them (*Yes*) leads into the graphic element *end*. The new employee has provided current documents, organization of OHS training will not be necessary. The other result of the decision block (*No*) directs the flow to the next step followed by *end*, where the process ends.

III. Business Process Model and Notation (BPMN)

Business Process Model and Notation [17], [18] was created and provided by the Business Process Management Initiative (BPMI) [19]. Its current version – 2.0.2 (introduced in January 2012) and the standard are maintained by the Object Management Group (OMG). The goal that guided the creators of BPMN notation was to create a language to describe processes taking place in the enterprise that would be understandable for all business users. The notation was to be universal and unambiguous enough to make graphic representations legible

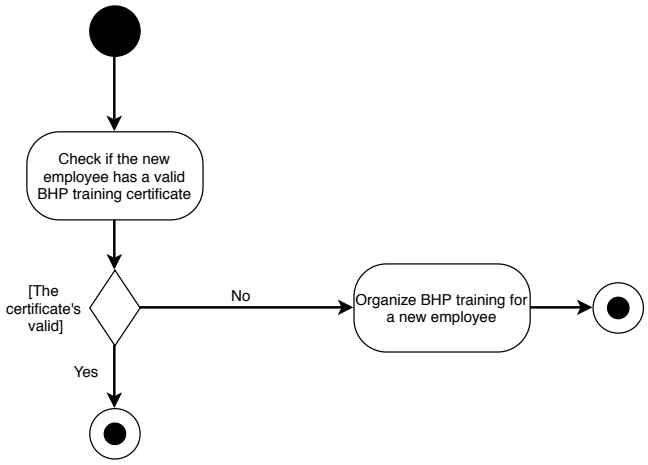


Figure 1. Example UML activity diagram.

and understandable for analysts responsible for creating models, technical persons dealing with the implementation of processes, not forgetting about the business representatives who manage and monitor workflows [20], [21]. Preparation of such models may also support process optimization, which usually has a positive impact on the efficiency of the organization and the time users spend on getting familiar with processes.

A. BPMN Process Diagrams

Process diagrams constitute the main type of models within the BPMN notation. The basic graphical elements of a process diagram are [20], [21]:

- activities;
- events;
- logical gateways;
- sequence flows;
- message flows
- pools and lanes;
- data objects;
- artifacts.

The BPMN standard alone does not specify the level of detail in modeling. This means that not all graphic elements need to be used in the final process model. It is its intended use that determines how accurate the prepared diagram should be. Drejewicz [20] lists three levels of detail of a model prepared using the BPMN notation:

1) **Illustrative model** – intended to present only general assumptions in the process. In this case, there is no description of technical issues, penetration into details of flows, nor the presentation of subprocesses.
2) **Analytical model** – prepared for the purpose of analyzing tasks that will be performed when creating and implementing the process in an organization. In this case, attention is paid to the use of data types, subprocesses, flow types, gateways and tasks.
3) **Executable model** – the most detailed business process model. It should include as much information as possible about the implemented business process.

Figure 2 illustrates an example model of the payment process in the training company. It consists of a small number of activities (represented by rounded rectangles): the actor of the process chooses the payment method, performs the payment. Then, the flow goes to the logical gateway (diamond), which indicates that two results of the decision are possible – the payment is successful or, if the actor does not have enough funds on the account, the payment is rejected. Arrows represent sequence flows – the order of actions performed in the process, i.e. the priority of activity execution, as well as time dependencies [22]. The line that connects the logical gateway to the activity *Realize Payment* has an additional cross section – this line determines the default sequence flow. The circle with a thin edge and the bold circles represent events – the start event and the end event, respectively. Figure 2, due to the very small amount of details presented and the lack of decomposition into individual actions can be considered an illustrative model.

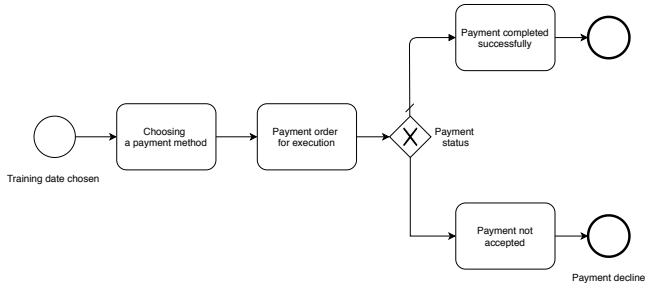


Figure 2. Example BPMN process model.

B. BPMN Choreography Diagrams

An interaction of two processes or two major participants in a process, can be represented using BPMN in the form of a collaboration diagram or a choreography. A BPMN collaboration diagram is, in fact, a combination of two or more pools with message flows between them.

Figure 3 shows an example collaboration diagram of confirming the customer's enrollment for an open training. There are two process participants in this case – a customer and a customer service employee, which is why there are two pools in the model that communicate by exchanging messages.

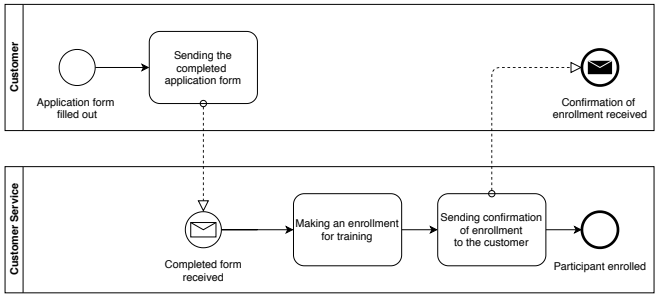


Figure 3. Example BPMN collaboration diagram of registration confirmation.

A choreography is a kind of process, but it differs significantly in purpose and behavior from a standard BPMN process model, which usually presents step-by-step activities. Choreographies focus rather on the ordered flow of information and the interactions between two participants of the process or two processes. The difference is also that in a standard process model, we can present the actions of one major actor, and choreography requires the presence of at least two. Therefore, it is impossible not to notice the very important relationship between choreography and BPMN pools. Since a pool is a graphic representation of one participant in the process, choreographies will take place only between pools. The following graphical elements are used in choreography diagrams:

- choreography activity – presented using a rectangle with rounded corners divided into three parts. The sender and recipients of the message are placed in the upper and lower part. It does not matter which actor is in which part, but the section representing message recipients is filled with a dark background, e.g. gray. The middle part contains the name of the activity being carried out;
- complex choreographies - a type of task that consists of various choreography tasks. It may also appear as several exchanges of messages between process actors;
- events;
- sequence flows;
- logical gateways.

Figure 4 shows the corresponding choreography diagram for this process. Two participants take part in the process, which is why only two actors appear in the choreography.

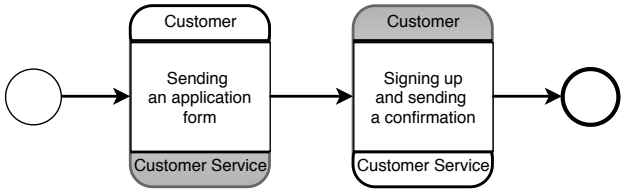


Figure 4. Example BPMN choreography diagram of registration confirmation.

IV. Event-driven Process Chain (EPC)

Event-Driven Process Chain is another notation that is used in modeling, redesigning facilitating of business processes, as well as controlling and organizing workflows. EPC was provided as part of the work on the ARIS method by August-Wilhelm Scheer from the University of Saarland in the early 1990s [23]. A model prepared using EPC is an ordered diagram of events and functions, combined flows and logical operators: OR, XOR or AND [24]. Additional passive elements, such as documents, systems, tools and data objects, can be used to refine the model [25].

The biggest advantage of the Event-Driven Process Chain language is its simplicity and intuitiveness [26]. What is more, the syntax does not include too many graphic elements, and thus easy to interpret. There were also attempts to formalize

the semantics of EPC [27], [28]. Although EPC is considered an informal notation, analysts successfully use it to prepare professional and detailed business process models [29].

The diagram presented in Figure 5 is an example of a business process model prepared using the EPC language.

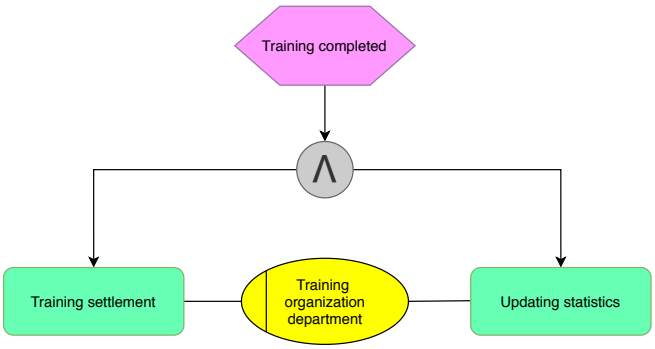


Figure 5. Example EPC diagram.

V. Comparison of notations using the process models of a training company

Event-driven Process Chain [30] was the first to be released. Work on the original version of Unified Modeling Language began four years later, in 1994 [15]. However, the Business Process Model and Notation language (currently the most popular notation of business process modeling) is the youngest – its first version was released in 2004 [20]. Over the years, managing organizations have developed and updated them – for example, UML has been modified more than 15 times, and BPMN notation has gained 4 subsequent versions since 2004 – there are 5 in total, the current version 2.0.2 was released in 2014 [21]. The EPC language has not changed since its introduction. BPMN and EPC were created and developed for convergent purposes. First of all, they are to enable graphical representation of processes that take place in organizations for the stakeholders taking part in their implementation. Unified Model Language is a notation dedicated to creating models of information systems, used in software engineering. The multitude of diagrams that it offers allows the modeler to present a complete IT system in a view of many models. However, this does not preclude using UML to create business process diagrams – this is successfully practiced.

A. Comparison of graphical elements

Unified Model Language, Business Process Model and Notation, as well as Event-driven Process Chain have a number of different graphical elements from which business process models are built. These elements, although different in appearance, name or adopted rules of use, play convergent roles in the models.

B. Comparison based on Kruchten's 4+1 view model

"4 + 1" [31] is a view model presented by Philippe Kruchten, used to compare views of system specifications and a description of software architecture. Using this tool, it is possible to analyze an IT system - from five concurrent views, each of which deals with a different set of issues. These views present the perspectives of different users of the created software (business, suppliers or end users) [16]. The views included in the 4+1 model are:

1) Logical view – describes the object model of the process, occurs at the conceptual level.
2) Process view – presents aspects of concurrency and process synchronization, also applies to the conceptual level.
3) Development view – describes static organization of software in a development environment [31].
4) Physical view – presents software mapping on hardware.
5) Use case view – presents usage scenarios of the system.

Figure 6 shows a 4+1 view model architecture.

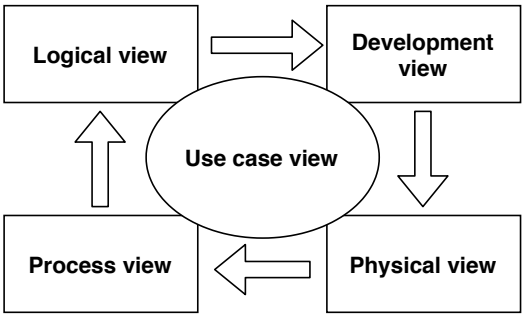


Figure 6. The 4+1 view model.

The first four views are used to to register design decisions, the fifth allows the user to illustrate them and then verify [31]. Figure 7 compares the use of UML (activity diagram), BPMN (process and choreography) and EPC notation using the "4 + 1" view model. A filled diamond means that the diagram is used in the particular view, while a partially filled diamond stands for the possibility to use a certain diagram in the particular view. As it can be seen in the figure, process modeling notations are not present in the physical view.

C. Comparison of process models used in the training company

The models used for our research represent two business processes implemented in a mid-sized training company from Krakow, Poland. This company deals with the sale and organization of open and closed training in project, portfolio, risk and change management standards etc. The first process describes the confirmation of an open training, in which customer service employees take part. Figure 8 presents this process model in the EPC notation.

The second process presents actions taken in the workflow of booking a trainer, performed by customer service or representatives of the sales department, sales director and the trainer himself. Figure 9 presents this process model in the EPC notation.

For our analysis, we used diagrams in UML (version 2.0), BPMN (version 2.0) and EPC notations.

	UML Activity Diagram	BPMN Diagram	BPMN Choreography Diagram	EPC Diagram
Logical view	◇	◇	◇	◇
Process view	◆	◆	◆	◆
Development view	◇	◈	◈	◇
Physical view	◇	◇	◇	◇
Use case view	◈	◈	◇	◈

Figure 7. Comparison of different process representation in terms of the 4+1 view model.

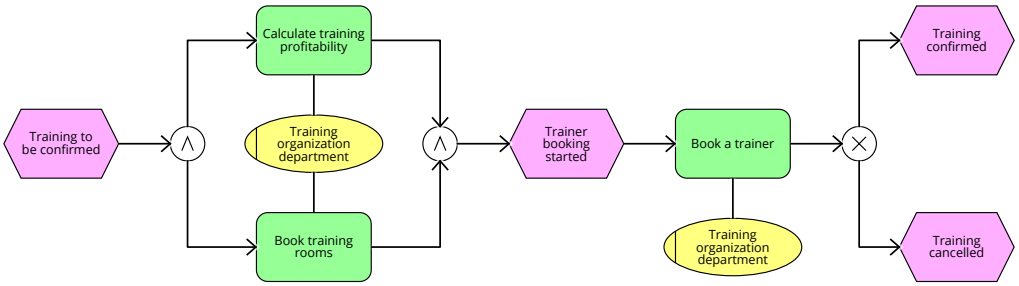


Figure 8. Open training confirmation process EPC model.

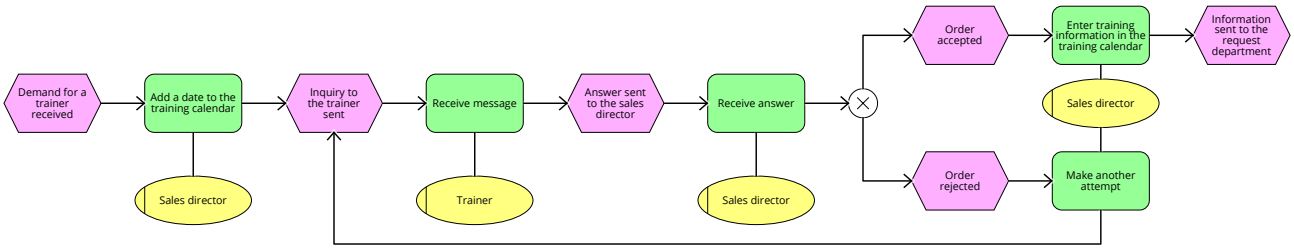


Figure 9. Booking a trainer for training process EPC model.

The process of confirming open training usually starts seven days before the planned date. This is the moment when, based on specific factors, employees decide to implement the planned training, move participants to another date or cancel the training completely. The result of the whole process is a message directed to the client – *the date you signed up for has been confirmed, the training will take place* or *the date you have chosen has not been confirmed, the training has been canceled*.

The difference that should be mentioned at the beginning of the comparison is the color schemes. UML and BPMN do not use colored elements to prepare models, which are usually color-independent. However, according to EPC, it is good practice to use colors for specific groups of elements. All examples differ at the beginning: the start of the process according to UML notation is initiated by the unlabelled point "start", and according to BPMN the initial event should be signed. Starting in a process prepared using EPC is presented using a named event. Sequence flow (in BPMN nomenclature) and control flows (according to UML and EPC) fall into operators that divide flows into two parallel paths. During their implementation, an attempt will be made to book a training room and the profitability of the training will be calculated. Income calculation and room booking are a series of separate activities, the process in question does not go into their details. BPMN notation, unlike the other two, offers a subprocess element (represented by a rectangle with a plus), which is used to present such cases [20]. In the first model, a fork divides the role of the operator dividing the flow, which is then merged. The BPMN diagram uses parallel gateways, and

the model prepared in EPC uses AND logical operators. In the first two models, flows continue to the next logical gateways. EPC notation, unlike BPMN and UML, does not allow logical operators to connect with each other, so the control flow leads into an event. At this point it is worth recalling that the EPC notation clearly defines what elements can be found at the output of logical operators. In the first and second examples, the places where the flow is split into two alternative paths, and the paths themselves are signed. The EPC did not accept labeling of logical operators or outgoing flows from them. In the BPMN scheme, one of the flows is marked with additional diagonal lines. This is the default process flow. UML and EPC notations do not have such a tool. The control flow in the UML model ends at a point called control end, the sequence flow in the BPMN model at the end event point. The EPC model began with an event and it must also end with one.

The second model represents the trainer booking process. It is an undertaking in which three organizational units of a training company are most often involved – customer service or sales department, sales director and a representative of the coaching team. These are separate departments, they work independently implementing their own processes. In this summary, it's worth looking at the ways in which notations it is possible to present roles in processes. The trainer's reservation is made in two cases - when the sales department representative finalizes the sale of the closed training and in cooperation with the client confirms the proposed dates or when the open training has implementation potential. This process always has one output – the selected lecturer accepts the order from the Commercial Director, the trainer is reserved, the selected date can be confirmed. The process begins with an e-mail request from a representative of one of the above departments. Thanks to BPMN notation, it is possible to clearly present the workflow or information between employees, individual organizational units of the enterprise or separate organizations. Aiming at a high level of detail, it was decided to apply them to this process. In the middle diagram, the responsibilities in the process are represented by pools. Each actor in the process – customer service, sales department, sales director or trainer – is presented as a separate pool with messages flowing between them. The UML activity diagram, unfortunately, does not offer such wide possibilities. In the EPC schema, determining the responsibility is possible using a graphic element organizational unit, while the notation for this element does not assume presenting message flows between units. The use of pools also enables choreography for the process. This is a separate type of process, which is representable only in BPMN notation.

VI. Analysis of the survey organized among the staff of the training company

Eight employees participated in the study (about 30% of the company's full-time team). Respondents were presented with 6 models of business processes in UML, BPMN and EPC notation: open training confirmation and trainer booking. Study participants take part in these processes on a daily basis in the performance of their duties. Each interview lasted between 9 and 15 minutes, interviews were conducted individually with each participant. The respondent answered 9 questions related to the mentioned set of diagrams.

Only one respondent did not recognize the processes shown in the models. The others correctly named the diagrams presented to them and were able to embed processes in time – to determine the moment in the organization's activity when the process would be carried out.

The models were divided into two groups: one for each process. The employees were asked to choose the most readable diagram from each group that contained different diagram types. In the group of the open training confirmation process, the EPC model was most often selected, paying attention to its colors, which were to catch the eye. It was also claimed that thanks to the colors the model is more readable and helps to find in the process: "I look and know what's going on, I know where I am". After the detailed questions, it turned out that the subjects incorrectly interpret the meaning of colors in EPC schemes – one wrongly thought that one color was the task of one actor of the process or a given color means positive (confirmation of the training) and another negative (canceling the training). Two respondents chose the UML model praising the simplicity and transparency of performance, and only one respondent indicated the BPMN scheme. In the trainer reservation process group, four respondents decided to choose the BPMN model. Respondents praised the use of swimlanes, agreeing that with this more complex process involving several company departments, a clear presentation of responsibility plays an important role. Three respondents chose the EPC diagram, again paying attention to the attractive nature of the model's color scheme. It is worth mentioning that these respondents called themselves visual learners. Nobody decided to choose the UML activity diagram.

When asked about which model is the most understandable and useful for the respondents, for the first group, they most often chose the BPMN model. This diagram was indicated three times. The use of swimlanes and the variety of graphic elements used were appreciated. Two respondents chose UML models, paying attention to the exact and legible way of describing events. Others indicated EPC schemes. In the second group, six respondents considered the BPMN model to be the most understandable. Here again, a clear and accurate division of roles in the process was praised. One respondent chose the EPC scheme.

Respondents were asked to indicate the elements that they think make models less readable, drew attention to the use of colors in EPC notation. They were subjects who did not choose schemes prepared according to EPC notation, as well as people who were presented with the way in which they should interpret the colors of the graphic elements of this language. They also exchanged transverse lines placed on the flows presented in the BPMN models (default process flow), and also pointed out too extensive event names in UML. At the same time, the method of constructing messages in BPMN diagrams was appreciated.

In many cases, the respondents said that they did not

understand the meaning and, consequently, the actions of logical operators and gates and dashed lines in BPMN models (message flows). People familiar with logic understood the logical operators of EPC notation. Study participants complained about the lack of signatures under the elements 'start' and 'end' in UML diagrams, and completely non-intuitive 'fork' and 'control merging'. Everyone reported the need to present a legend by which they could learn the functions of unknown elements. When asked what could help them in understanding the process, they pointed to the need to sign logical operators in models for UML and BPMN, and to name the points initiating and terminating UML processes. In the answers to this question, the legend appeared again as an element necessary for the correct interpretation of processes.

Participants asked if they prefer models containing more or less details differed in the answers. The first respondent admitted that due to her professional experience and good knowledge of the presented processes, her basic model is sufficient. Others pointed out that models with a high level of detail do not leave room for their own interpretation, which is not recommended for self-organizing teams. They also claimed that by accurately presenting the process, it is easier to determine my place and responsibility, "I am able to be more independent and organize myself." Two respondents mentioned that models that included a large amount of details can be useful when introducing new employees to teams.

Respondents asked about whether they prefer colored (EPC diagrams) or black and white models mostly chose the former. However, they pointed out that elements that would be of the same color should mean the same or present the same process results (positive/negative result). They again admitted that a legend would be useful that would describe the meaning of individual colors. Two respondents said that colored schemes are easier to orientate and find, colors help categorize relevant groups of elements. Two respondents chose black and white process diagrams.

Respondents noted that models that clearly define responsibility for specific process steps are definitely more useful and make the schemes clearer. Everyone agreed that BPMN notation offering 'pool and track' tools handles role presentation best. The models in which the pools were used were assessed as the most clear, the participants immediately pointed out the clear division of roles in the process. The element of 'organizational units' was less often interpreted in an appropriate manner. The respondents did not immediately understand the function they play in the models.

VII. Notation interoperability

As there are several notations for business process modeling, the possibility of converting models between them is an important research topic.

In the case of transformations between BPMN and EPC, it is important to notice that one-to-one translation pattern cannot be used here, as there are syntax differences, e.g. in EPC each function must be followed by an event [32]. There are methods based on transforming rules from EPC to BPMN models [33]. However, during transformation from EPC to BPMN, the information content may change what may result in a slight information loss due to process model transformation [34]. Thus, recently new methods have been developed and new transformation rules proposed which should minimize the information loss [35].

There are also a number of papers concerning UML to BPMN [36]–[40] and BPMN to UML [41]–[43] transformation.

The conversions between the selected three notations can be dome through some other standard such as BPEL [44] or spreadsheets [45]. Workflow patterns constitute another effective way of transformation. However, as noticed by Grigorova and Mironov [46], [47], there is still a gap between some pattern constructs available in the notations, such as cancel activity (not supported by EPC), persistent trigger and generalized AND (not supported by UML).

Khudori and Kurniawan [48] conducted a broad survey on business process transformation techniques and stated that none of the existing techniques supports a truly complete transformation between the process modeling notations.

VIII. Conclusions

In this work, we discussed and compared the three most popular business process modeling notations — Unified Modeling Language, Business Process Model and Notation and Event-Driven Process Chain. Our work included also a comparative study of the above-mentioned notations and a survey conducted among the employees of a mid-sized training company, as well as an overview of notation interoperability approaches present in the current research work.

Unified Modeling Language can be successfully used to prepare business process models. It offers two diagrams that can be used for this purpose: use cases and activities. The latter presents the process as a sequence of actions and actions that the control flow connects to decision blocks.

Business Process Model and Notation is the richest notation in terms of the number of graphic elements. Thanks to this, the person preparing models can very accurately reflect the nature of the designed activity or event, present the type of task being performed, mark the default process flow or present the required data objects. BPMN offers also a way to present roles in the process in a very clear and legible way to the user, in form of pools and lanes. It also lets users create choreography diagrams that present information flows and interactions between pools in a more user-friendly way.

Event-Driven Process Chain is characterized by a relatively small number of elements that can be used to build a diagram, as well as the practice of using colors. After all, it allows users to create models even for complex business processes, while assigning responsibilities to specific functions, using organizational units.

As future work, we plan to calculate and compare complexity metrics of the analyzed business process models, as well as to conduct a more detailed survey among a larger set of companies from the SME sector.

REFERENCES

[1] P. Sliż, "Robotization of business processes and the future of the labor market in poland-preliminary research," *Organization and Management*, no. 2 (185), pp. 67–79, 2019.

[2] R. Gabryelczyk, *ARIS w modelowaniu procesów biznesu*. Difin, 2006.

[3] OMG, "Unified Modeling Language (OMG UML) version 2.2. superstructure," Object Management Group, Tech. Rep. formal/2009-02-02, February 2009.

[4] J. Hunt, *Guide to the Unified Process featuring UML, Java and Design Patterns*. Springer, 2003.

[5] M. Fowler, *UML Distilled: A Brief Guide to the Standard Object Modeling Language*, 3rd ed. Addison-Wesley Professional, 2003.

[6] D. Pilone and N. Pitman, *UML 2.0 in a Nutshell*. O'Reilly, 2005.

[7] M. Owen and J. Raj, "BPMN and Business Process Management. Introduction to the new business process modeling standard." OMG, Tech. Rep., 2006, www.bpmn.org.

[8] N. Russell, W. M. P. van der Aalst, A. H. M. ter Hofstede, and P. Wohed, "On the suitability of UML 2.0 activity diagrams for business process modelling," in *Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling – Volume 53*, ser. APCCM '06. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2006, pp. 95–104.

[9] D. Jäger, A. Schleicher, and B. Westfechtel, "Using uml for software process modeling," in *Software Engineering – ESEC/FSE '99*, ser. Lecture Notes in Computer Science, O. Nierstrasz and M. Lemoine, Eds. Springer Berlin Heidelberg, 1999, vol. 1687, pp. 91–108.

[10] G. Engels, A. Förster, R. Heckel, and S. Thöne, "Process modeling using UML," *Process-Aware Information Systems*, pp. 85–117, 2005.

[11] M. Razavian and R. Khosravi, "Modeling variability in business process models using UML," in *Proceedings of the fifth International Conference on Information Technology: New Generations, 2008. ITNG 2008*, 2008, pp. 82–87.

[12] M. A. Kose and M. Ozkaya, "Towards extending uml's activity diagram for the architectural modeling, analysis, and implementation," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 639–648.

[13] A. Derezińska and Ł. Zaremba, "Approaches to semantic mutation of behavioral state machines in model-driven software development," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2018, pp. 863–866.

[14] F. Hunka and J. Matula, "Towards paired transactions modeling," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2016, pp. 1153–1158.

[15] T. Gzik, "Modelowanie procesów biznesowych w UML," *Business Process Management Portal*, 2018.

[16] K. Kluza, P. Wiśniewski, K. Jobczyk, A. Ligęza, and A. Suchenia (Mroczek), "Comparison of selected modeling notations for process, decision and system modeling," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. IEEE, 2017, pp. 1095–1098.

[17] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.

[18] P. Y. H. Wong and J. Gibbons, "Formalisations and applications of bpmn," *Science of Computer Programming*, vol. 76, no. 8, pp. 633–650, 2011.

[19] OMG, "Business Process Model and Notation (BPMN): Version 2.0 specification," Object Management Group, Tech. Rep. formal/2011-01-03, January 2011.

[20] S. Drejewicz, *Zrozumieć BPMN modelowanie procesów biznesowych*. Wydawnictwo Helion, 2012.

[21] *Business Process Model and Notation (BPMN)*, 2nd ed., Object Management Group, 12 2013, version 2.0.2 contains a minor change to Clause 15.

[22] E. M. Sanfilippo, S. Borgo, and C. Masolo, "Events and activities: Is there an ontology behind BPMN?" in *FOIS*, 2014, pp. 147–156.

[23] A. W. Scheer, *Aris: Business Process Modeling*, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2000.

[24] M. Rosemann and W. M. P. van der Aalst, "A configurable reference modelling language," *Information Systems*, vol. 32, no. 1, pp. 1–23, 2007.

[25] A. Amjad, F. Azam, M. W. Anwar, W. H. Butt, and M. Rashid, "Event-driven process chain for modeling and verification of business requirements–a systematic literature review," *IEEE Access*, vol. 6, pp. 9027–9048, 2018.

[26] D. M. Riehle, S. Jannaber, A. Karhof, O. Thomas, P. Delfmann, and J. Becker, "On the de-facto standard of event-driven process chains: How epc is defined in literature," *Modellierung 2016*, 2016.

[27] W. M. P. van der Aalst, "Formalization and verification of event-driven process chains," *Information and Software Technology*, vol. 41, no. 10, pp. 639–650, 1999.

[28] W. M. P. van der Aalst, J. Desel, and E. Kindler, "On the semantics of EPCs: A vicious circle," in *Proceedings of the EPK 2002: Business Process Management Using EPCs*, Bonn, Trier, Germany, November 2002, pp. 71–80.

[29] P. Pasamonik, "Modelowanie procesów biznesowych zorientowane na czynności," *Zeszyty Naukowe Wyższej Szkoły Informatyki*, vol. 9, no. 2, pp. 102–116, 2010.

[30] A.-W. Scheer, *Architecture of integrated information systems: foundations of enterprise modelling*. Springer Science & Business Media, 2012.

[31] P. B. Kruchten, "The 4+ 1 view model of architecture," *IEEE software*, vol. 12, no. 6, pp. 42–50, 1995.

[32] M. Cheung and J. Hidders, "Round-trip iterative business process modelling between bpa and bpms tools," *Business Process Management Journal*, 2011.

[33] W. Tscheschner, "Transformation from epc to bpmn," *Business Process Technology*, vol. 1, no. 3, pp. 7–21, 2006.

[34] O. Levina, "Assessing information loss in epc to bpmn business process model transformation," in *2012 IEEE 16th International Enterprise Distributed Object Computing Conference Workshops*. IEEE, 2012, pp. 51–55.

[35] A. N. Khudori and T. A. Kurniawan, "Transforming epc aris markup language into bpmn metadata," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*. IEEE, 2019, pp. 358–363.

[36] J. Chanda, A. Kanjilal, S. Sengupta, and S. Bhattacharya, "Fam2bp: Transformation framework of uml behavioral elements into bpmn design element," in *International Conference on Computer Science and Information Technology*. Springer, 2011, pp. 70–79.

[37] Y. Wautelet and S. Poelmans, "An integrated enterprise modeling framework using the rup/uml business use-case model and bpmn," in *IFIP Working Conference on The Practice of Enterprise Modeling*. Springer, 2017, pp. 299–315.

[38] A. Kalnins and V. Vitolins, "Use of UML and model transformations for workflow process definitions," *arXiv preprint cs/0607044*, 2006.

[39] M. Argañaraz, A. Funes, and A. Dasso, "An mda approach to business process model transformations," *Electronic Journal of SADIO (EJS)*, vol. 9, pp. 24–48, 2010.

[40] J. Pulgar and M. C. Bastarrica, "Transforming multi-role activities in software processes into business processes," in *International Conference on Business Process Management*. Springer, 2016, pp. 372–383.

[41] M. A. Cibran, "Translating BPMN models into UML activities," in *International Conference on Business Process Management*. Springer, 2008, pp. 236–247.

[42] N. Q. Bao, "A proposal for a method to translate BPMN model into UML activity diagram," in *13th International Conference on Business Information Systems*, 2010.

[43] L. Aversano, C. Grasso, and M. Tortorella, "Managing the alignment between business processes and software systems," *Information and Software Technology*, vol. 72, pp. 171–188, 2016.

[44] K. Grolinger, M. A. Capretz, A. Cunha, and S. Tazi, "Integration of business process modeling and web services: a survey," *Service Oriented Computing and Applications*, vol. 8, no. 2, pp. 105–128, 2014.

[45] P. Wiśniewski, K. Kluza, E. Kucharska, and A. Ligęza, "Spreadsheets as interoperability solution for business process representation," *Applied Sciences*, vol. 9, no. 2, p. 345, 2019.

[46] K. Grigorova and K. Mironov, "Bridging the gap between different interfaces for business process modeling," *International Journal of Computer and Information Engineering*, vol. 9, no. 12, pp. 2479–2482, 2015.

[47] ——, "Conversion of business process models using workflow patterns," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2018, pp. 763–766.

[48] A. N. Khudori and T. A. Kurniawan, "Business process model transformation techniques: A comprehensive survey," *Advanced Science Letters*, vol. 24, no. 11, pp. 8606–8612, 2018.

Software, System and Service Engineering

THE S3E track emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This track investigates both established traditional approaches and modern emerging approaches to large software production and evolution.

For decades, it is still an open question in software industry, how to provide fast and effective software process and software services, and how to come to the software systems, embedded systems, autonomous systems, or cyber-physical systems that will address the open issue of supporting information management process in many, particularly complex organization systems. Even more, it is a hot issue how to provide a synergy between systems in common and software services as mandatory component of each modern organization, particularly in terms of IoT, Big Data, and Industry 4.0 paradigms.

In recent years, we are the witnesses of great movements in the area of software, system and service engineering (S3E). Such movements are both of technological and methodological nature. By this, today we have a huge selection of various technologies, tools, and methods in S3E as a discipline that helps in a support of the whole information life cycle in organization systems. Despite that, one of the hot issues in practice is still how to effectively develop and maintain complex systems from various aspects, particularly when software components are crucial for addressing declared system goals, and their successful operation. It seems that nowadays we have great theoretical potentials for application of new and more effective approaches in S3E. However, it is more likely that real deployment of such approaches in industry practice is far behind their theoretical potentials.

The main goal of Track 5 is to address open questions and real potentials for various applications of modern approaches and technologies in S3E so as to develop and implement effective software services in a support of information management and system engineering. We intend to address interdisciplinary character of a set of theories, methodologies, processes, architectures, and technologies in disciplines such as: Software Engineering Methods, Techniques, and Technologies, Cyber-Physical Systems, Lean and Agile Software Development, Design of Multimedia and Interaction Systems, Model Driven Approaches in System Development, Development of Effective Software Services and Intelligent Systems, as well as applications in various problem domains. We invite researchers from all over the world who will present their contributions, interdisciplinary approaches or case studies related to modern approaches in S3E. We express an interest in gathering scientists and practitioners interested in applying these disciplines in industry sector, as well as public and government sectors, such as healthcare, education, or security services. Experts from all sectors are welcomed.

TOPICS

Submissions to S3E are expected from, but not limited to the following topics:

- Advanced methodology approaches in S3E – new research and development issues
- Advanced S3E Process Models
- Applications of S3E in various problem domains – problems and lessons learned
- Applications of S3E in Lean Production and Lean Software Development
- Total Quality Management and Standardization for S3E
- Artificial Intelligence and Machine Learning methods in advancing S3E approaches
- S3E for Information and Business Intelligence Systems
- S3E for Embedded, Agent, Intelligent, Autonomous, and Cyber-Physical Systems
- S3E for Design of Multimedia and Interaction Systems
- S3E with User Experience and Interaction Design Methods
- S3E with Big Data and Data Science methods
- S3E with Blockchain and IoT Systems
- S3E for Cloud and Service-Oriented Systems
- S3E for Smart Data, Smart Products, and Smart Services World
- S3E in Digital Transformation
- Cyber-Physical Systems (8^{th} Workshop IWCPS-8)
- Software Engineering (41^{th} IEEE Workshop SEW-41)
- Advances in Programming Languages (8^{th} Workshop WAPL'21)

TRACK CHAIRS

- **Luković, Ivan,** Unniversity of Belgrade, Serbia
- **Kardas, ,** Geylani, Ege University International Computer Institute, Turkey
- **Mazzara, Manuel,** Innopolis University, Russia

PROGRAM CHAIRS

- **Bowen, Jonathan,** Museophile Ltd., United Kingdom
- **Hinchey, Mike** (Lead Chair), Lero-the Irish Software Engineering Research Centre, Ireland
- **Szmuc, Tomasz,** AGH University of Science and Technology, Poland
- **Zalewski, Janusz,** Florida Gulf Coast University, United States

- **Seyed Hossein Haeri,** Catholic University of Louvain, Louvain-la-Neuve, Belgium and University of Bergen, Norway

PROGRAM COMMITTEE

- **Ahmad, Muhammad Ovais,** Karlstad University, Sweden
- **Challenger, Moharram,** University of Antwerp, Belgium
- **Dejanović, Igor,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Derezinska, Anna,** Warsaw University of Technology, Poland
- **Dutta, Arpita,** NIT ROURKELA, India
- **García-Mireles, Gabriel,** Universidad de Sonora, Mexico
- **Göknil, Arda,** SINTEF Digital, Norway
- **Heil, Sebastian,** Technische Universitüt Chemnitz, Germany
- **Erata, Ferhat,** Yale University, United States
- **Escalona, M.J.,** University of Seville, Spain
- **Essebaa, Imane,** Faculté des Sciences et Techniques Mohammedia, Morocco
- **Hanslo, Ridewaan,** University of Pretoria, South Africa
- **Jarzebowicz, Aleksander,** Gdansk University of Technology, Poland
- **Kaloyanova, Kalinka,** University of Sofia, Bulgaria
- **Karolyi, Masaryk University,** IBA, Czechia
- **Katic, Marija,** University of London, United Kingdom
- **Khlif, Wiem,** FSEGS, Tunisia
- **Kolukısa Tarhan, Ayça,** Hacettepe University, Turkey
- **Krdzavac, Nenad,** University of Belgrade, Serbia
- **Marcinkowski, Bartosz,** University of Gdansk, Poland
- **Milosavljevic, Gordana,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Misra, Sanjay,** Covenant University, Nigeria
- **Morales Trujillo, Miguel Ehécatl,** University of Canterbury, New Zealand
- **Ozkan, Necmettin,** Kuveyt Turk Participation Bank, Turkey
- **Ozkaya, Mert,** Yeditepe University, Turkey
- **Ristic, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Rossi, Bruno,** Masaryk University, Czech Republic
- **Sanden, Bo,** Colorado Technical University, United States
- **Shilov, Nikolay,** Innopolis University, Russia
- **Sierra Rodríguez, José Luis,** Universidad Complutense de Madrid, Spain
- **Torrecilla-Salinas, Carlos,** IWT2, Spain
- **Varanda Pereira, Maria João,** Instituto Politécnico de Bragança, Portugal

Joint 41st IEEE Software Engineering Workshop and 8th International Workshop on Cyber-Physical Systems

THE IEEE Software Engineering Workshop (SEW) is the oldest Software Engineering event in the world, dating back to 1969. The workshop was originally run as the NASA Software Engineering Workshop and focused on software engineering issues relevant to NASA and the space industry. After the 25th edition, it became the NASA/IEEE Software Engineering Workshop and expanded its remit to address many more areas of software engineering with emphasis on practical issues, industrial experience and case studies in addition to traditional technical papers. Since its 31st edition, it has been sponsored by IEEE and has continued to broaden its areas of interest.

One such extremely hot new area are Cyber-physical Systems (CPS), which encompass the investigation of approaches related to the development and use of modern software systems interfacing with real world and controlling their surroundings. CPS are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. CPS systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The joint workshop aims to bring together all those researchers with an interest in software engineering, both with CPS and broader focus. Traditionally, these workshops attract industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practices. This joint edition will also provide a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics

The workshop aims to bring together all those with an interest in software engineering. Traditionally, the workshop attracts industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practice. The workshop provides a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics of interest include, but are not limited to:

- Experiments and experience reports
- Software quality assurance and metrics
- Formal methods and formal approaches to software development
- Software engineering processes and process improvement
- Agile and lean methods
- Requirements engineering
- Software architectures
- Design methodologies
- Validation and verification
- Software maintenance, reuse, and legacy systems
- Agent-based software systems
- Self-managing systems
- New approaches to software engineering (e.g., search based software engineering)
- Software engineering issues in cyber-physical systems
- Real-time software engineering
- Safety assurance & certification
- Software security
- Embedded control systems and networks
- Software aspects of the Internet of Things
- Software engineering education, laboratories and pedagogy
- Software engineering for social media

Technical Session Chairs

- **Bowen, Jonathan,** Museophile Ltd., United Kingdom
- **Hinchey, Mike** (Lead Chair), Lero-the Irish Software Engineering Research Centre, Ireland
- **Szmuc, Tomasz,** AGH University of Science and Technology, Poland
- **Zalewski, Janusz,** Florida Gulf Coast University, United States

Program Committee

- **Ait Ameur, Yamine,** IRIT/INPT-ENSEEIHT, France
- **Cicirelli, Franco,** Dimes - Unical, France
- **Ehrenberger, Wolfgang,** University of Applied Science, Germany
- **Gomes, Luis,** Universidade NOVA de Lisboa, Portugal
- **Gracanin, Denis,** Virginia Tech, United States
- **Havelund, Klaus,** Jet Propulsion Laboratory, California Institute of Technology, United States
- **Hsiao, Michael,** Virginia Tech, United States
- **van-Katwijk, Jan,** TU Delft, The Netherlands

- **Trybus, Leszek,** Rzeszow University of Technology, Poland
- **Vardanega, Tullio,** University of Padua, Italy
- **Velev, Miroslav,** Aries Design Automation, United States

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 205–210
DOI: 10.15439/2021F125
ISSN 2300-5963 ACSIS, Vol. 26

Towards Energy-aware Cyber-Physical Systems Verification and Optimization

Reza Soltani, Eun-Young Kang, Juan Esteban Heredia Mena
University of Southern Denmark
The Mærsk Mc-Kinney Møller Institute
SDU Software Engineering, Odense, Denmark
Email: {resol, eyk, jehm}@mmmi.sdu.dk

***Abstract*—Optimizing CPS behavior in terms of energy consumption can have a significant impact on system reliability. The environment influences the system's behavior, and neglecting the environmental behavior has an indirect negative impact on optimizing the system's behavior. In this work, to increase the system's flexibility, the behavior of the environment is modeled dynamically to apply the disorderliness of its behavior. The resulting models are formally verified. By examining the past environmental behavior and predicting its future behavior, energy optimization is done more dynamically. The verification results acquired using a UPPAAL-SMC show that the optimization of system behavior by predicting the environmental behavior has been successful. Our approach is demonstrated using a case study within an I4 setting.**

I. Introduction

CYBER-PHYSICAL SYSTEMS (CPS) are continuously developing and have become an integral part of Industry 4.0 (I4). The I4 revolution relies on the interconnectivity of machines and automation of processes to improve factory productivity. Modern industry uses assets designed to do different tasks; for example, a robotic arm can assemble, palletize, and solder. Consequently, many production cells can perform the same activity, but each cell's production time differs. For instance, total automation, sustainable production, and efficient scheduling are challenges that need to be addressed to achieve the I4 objective.

According to the United Nations Goals, production sustainability is one of the goals for industry [1]. One of the key factors in CPS is energy consumption, and neglecting energy consumption limits the reliability and safety of such systems. We differentiate between two types of techniques for reducing the energy consumption of a system: whether a CPS has to reconstruct its behavior or make structural modifications such as using light materials, different types of batteries, and efficient motors. Structural modifications can reduce energy consumption but are costly; instead, behavioral changes modify the device software and are cheap.

One of the challenges in CPS analysis is predicting the environmental behavior and modeling the environment since its behavior may change depending on various factors. Although CPS is highly affected by the environment, it has no control over it. For example, in an autonomous vehicle system, the system's performance is based entirely on the analysis of environmental behavior. Knowing information such as how often we are likely to see each sign or which sign is less likely to be seen based on the current situation allows the system to make optimal decisions. For example, the speed can increase with a gentler slope if it is likely to see a stop sign. Sometimes no strong logic can be found for modeling the environment. In this regard, the challenge is run-time data collection, and the system's ability to dynamically model the environment based on these data increases the system's flexibility. One way to do this is to use learning algorithms to study the past and predict future environmental behavior.

In this paper, energy-aware timed behaviors of CPS are specified in Stochastic Hybrid Automata (SHA) [2]. By considering the cost of each behavior, the amount of energy consumption of the system for different modes is obtained. In this way, The minimum and maximum energy consumption of the system can be calculated. Due to the dependence of energy consumption on environmental behavior and the disorderliness of this behavior, the environment with which CPS is associated is dynamically modeled.

Our earlier works [3]–[6] verified the safety of CPS (including the controller and physical parts) and analyzed performance in terms of time and resource constraints: In order to model the unpredictable behaviors of environment, fixed probabilities were considered and allocated to the possible transitions. Whereas our current work utilizes dynamic probabilities instead of the predetermined probability. The main contributions of this paper are:

- Formal modeling of energy-aware timed behaviors of CPS in SHA that captures discrete and continuous behaviors of both the controller and its environment and verifying the system using a statistical model checker, UPPAAL-SMC.
- Dynamic modeling of the environment to increase the flexibility of the system against the disorderliness behavior of the environment.
- Optimize the behavior of the system in terms of energy consumption by predicting the future environmental behavior.

In this work, we use the dynamic probability for each environmental incident automatically updated after a certain number of iterations. According to the past behavior of the environment, the system can predict its future behavior and

make decisions based on it. Thus, the environmental behavior of the system is modeled more dynamically.

The rest of the paper is organized as follows. Section II presents our methodology. The industrial case study we used to demonstrate the applicability of our framework is presented in Section III. Section IV shows the modeling and verification results. Section V provides related works, and Section VI describes research challenges and intended future works.

II. Methodology

A. Energy Optimization

The behavior of CPS can be optimized by improving the behavior of its components without structural changes, such as energy modes. We conduct optimization by addressing energy consumption; We model the continuous behavior of the system through SHA, which enables us to find the best path of having a specific behavior with minimum energy consumption. By using statistical analysis, the probability of energy optimization has been investigated up to a certain confidence degree. The behavior of the environment, which directly impacts the energy consumption process, has also been studied.

B. Modeling environmental behavior

To model and predict the environmental behavior of CPS, it is essential to learn from its previous behavior. As discussed in Section I, the dynamic probability is used because the behavior of the environment can change, and modeling it with fixed probabilities will significantly simplify and limit the system model. Therefore, the current work uses variable probabilities for each environmental incident by considering its behavioral disorderliness. These probabilities are updated in each period using closed-loop control and by analyzing the output of the system. To do this, the occurrence time of each event is stored in an array of integers. After a certain number of iterations, this array is sorted using the ascending sort algorithm, and the number of occurrences of each event is calculated. Based on that, the probability of its recurrence in the future is predicted. This possibility will be updated automatically on each iteration. Thus, instead of using fixed and predetermined probabilities for each event, by considering the system's past behavior (environment) and predicting future behavior based on it, dynamic probabilities updated in each iteration are used.

III. Case Study

In this section, we present a case of study that is part of a production plant. Fig. 1 shows a reduced system architecture of a factory. In non-automatized factories, operators manufacture and package the final products. In parallel, the distribution schedule and routes of distribution are defined. In the next stage, operators sort elements for distribution via a truck.

The Industry 4.0 Laboratory at the University of Southern Denmark (SDU) [1] is working in a production line for drone assembling [7]. The current state of the project does not consider

[1]https://www.sdu.dk/en/forskning/i40lab

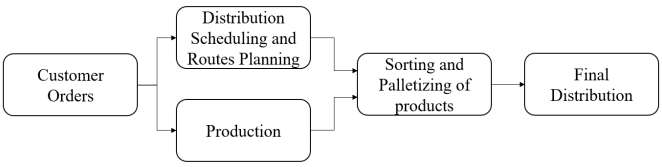


Fig. 1. Reduced System Architecture of a Factory

the distribution of the drones. We aim to add the capability of distribution to the production plant. The distribution system selects the elements from a storing warehouse, classifies and orders them. The expected final result is a pile of drone boxes ready to be picked by a truck.

In the proposed distribution system, the drones are obtained from the automated warehouse Effimat [8]. An ER mobile manipulator [9] picks the products from the warehouse and places them in the B&R's magnetic distribution conveyor. A robotic manipulator UR3e [10] sortes and palletizes the packages using distribution schedule and route plan. Fig. 2 shows the elements that compose the distribution system.

We particularly focus on the sorting capability, which is done by the manipulator UR3e. The manipulator UR3e is an anthropomorphic robotic arm made by Universal Robots [10]. The robot UR3e possesses six axes of rotation, as is shown in Fig. 3.

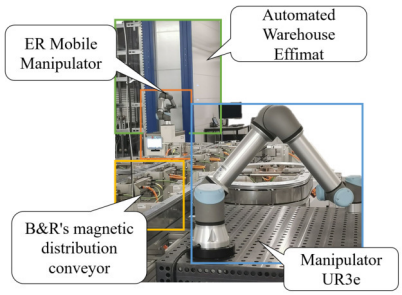


Fig. 2. Distribution system and its components

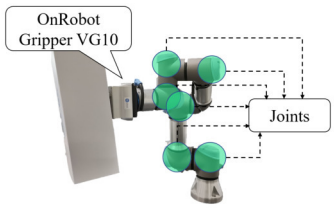


Fig. 3. Robot UR3e joints and griper VG10

In the sorting activity, the robot picks the objects from position *A* and then places the package in position *B* or *C* according to the distribution plan. The frequency of packages depends on the customer orders and the distribution plan. The system includes a position idle. When the robot is in

the idle position, it consumes less energy than in the other positions. One of the objectives of this study is to design an energy optimal robot behavior based on environmental behavior. Therefore, we measure the energy consumption of each of the robot movements and the rate of consumption of the robot in a static position, which is presented in Table I.

TABLE I
ENERGY CONSUMPTION OF EACH OF THE ROBOT MOVEMENTS

Movement / position	Energy Consumption
A to B	374.0 [J]
A to C	293.3 [J]
B to A	196.1 [J]
C to A	125.7 [J]
B to Idle	198.6 [J]
C to Idle	145.0 [J]
Idle to A	112.3 [J]
A	40.20 [J/s]
B	39.87 [J/s]
C	40.45 [J/s]
Idle	37.45 [J/s]

IV. MODELING AND VERIFICATION RESULTS

The described case study in Section III, is modeled in SHA, including the energy-aware timed transitions and the amount of energy consumed in each state. The operational semantics of both the system and its environment are formally specified in SHA, and the models are verified against given requirements through UPPAAL-SMC [11] The verification results are displayed in Table II and discussed in Section IV-B

A. Modeling energy-aware behavior and environmental behavior prediction

Fig. 4 shows the robotic arm automaton. Initially, the robotic arm is in idle mode, and with the arrival of the package, it receives a message through the synchronization channel *move?*. In this case, each of the six joints of the robotic arm has to change the angle of their position to put the arm in the ideal position for state A; then, it will be ready to grab the package. Depending on where the package goes, it will reset its joints' angles again and move the package to the desired location. In this model, the amount of energy consumption for each transmission is calculated discretely according to Table I. The robotic arm consumes energy per time unit when it is in one of the three different modes (Idle, StandbyB, StandbyC). The amount of energy consumption for each of the three modes is calculated continuously. For example, the energy consumption per time unit when the robotic arm is in idle mode is calculated using formula *energy'==Idle*.

In this SHA model, the variant *FB* is the functional behavior of the system, which can have an integer value between 0 to 3, representing moving up, moving down, going idle, and coming back from idle behaviors, respectively. The variant *GoR* is the grabber's behavior which can have an integer value of 0 or 1, representing grabbing the box or releasing it, respectively. After moving box from position *A* to *B* or *C*, it is decided

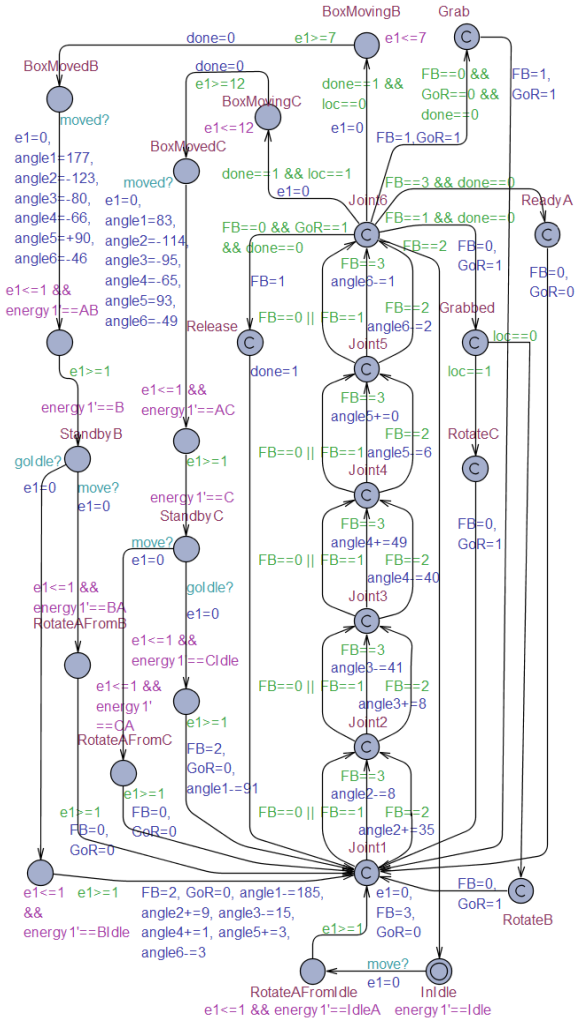


Fig. 4. Robotic arm automaton

whether to stay in the same state (StandbyB or StandbyC in Fig. 4) or to go into the idle mode to save energy.

Such a decision is made with the help of the environment model. Fig. 5 shows the SHA of the environment: 5 different periods are defined, which indicate the arrival time of the packages. The probability of occurrence of each is a variable of type integer, whose value is updated every 20 times the packages have arrived, and henceforth we call this period iteration.

The general goal of the environment SHA is: (1) to model the behavior of the environment dynamically to include its disorderliness behavior; (2) to change the behavior of the system based on the analysis of the past environmental behavior and to decide on the future behavior of the system based on a prediction. In the first case, five variables *TI1*, *TI2*, *TI3*, *TI4*, and *TI5* are used for the probability of occurrence of each time interval as shown in Fig. 5. These variables are changing in each iteration, and this increases the possibility of including critical factors such as environmental disorderliness

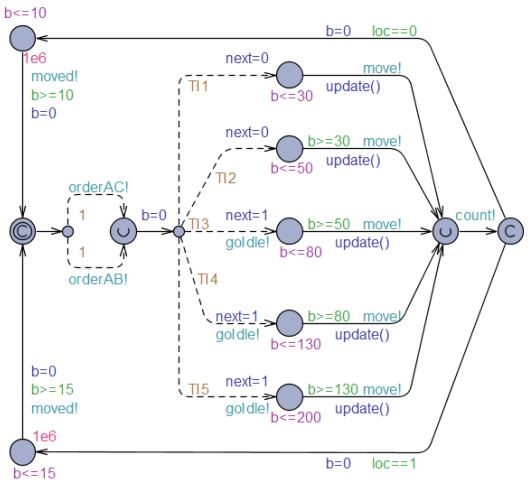


Fig. 5. Environment automaton

in modeling. In the second case, using the *Update()* function, each time the boxes arrive is saved in the array, and every 20 times, at the end of the iteration, this SHA uses this function to decide on the behavior of the system for the next iteration. Algorithm 1 shows how the *Update()* function works.

In Algorithm 1, *env* is an array of length 20 that saves the box arrival time. When the 20th box has arrived, it counts the number of times a box arrived in each time interval and updated the probabilities accordingly. Suppose the system goes into the idle mode and a box enters position *A* earlier than a specific time. In that case, the system consumes more energy because the transition cost to enter to and exit from the idle mode is more than staying in standby mode. For this reason, this function makes a decision using the calculated probabilities for the next iteration. It should be noted that in some transitions, energy consumption may increase. Still, the purpose of this prediction is to determine whether, on average, a particular transition (going idle) will decrease or increase energy consumption. Verification results show that this algorithm has been successful in reducing energy consumption.

Algorithm 1 Decide about system behavior based on a prediction of environmental behavior

```
void update()
if counter >= 19 then
    env[19]=b;
    TI1=0; TI2=0; TI3=0; TI4=0; TI5=0;
    counter=0; i= 0; j=1;
    while i < 20 do
        while j < 20 do
            if env[i] >= env[j] then
                q1 = sort[i]; sort[i] = sort[j]; sort[j] = q1;
            end
            j++;
        end
        i++;
    end
    i= 0;
    while i < 20 do
        if env[i] <= threshold1 then
            TI1++;
        end
        if env[i] <= threshold2 and env[i] > threshold1 then
            TI2++;
        end
        if env[i] <= threshold3 and env[i] > threshold2 then
            TI3++;
        end
        if env[i] <= threshold4 and env[i] > threshold3 then
            TI4++;
        end
        if env[i] > threshold4 then
            TI5++;
        end
        i++;
    end
      q2= TI1+TI2;
    q3= TI3+TI4+TI5;
    if q2 >= q3 then
        goIdle=0;
    else
        goIdle=1;
    end
else
    env[counter]=b;
    counter++;
end
```

B. Verification results

In this section, we discuss the verification results that are given in Table II: *energy* indicates the system's energy consumption in the presence of idle mode and *energy2* indicates the system's energy consumption in the absence of idle mode. The first five queries are related to energy consumption. As presented in the verification results, by predicting the behavior of the environment and changing the system's behavior accordingly, the amount of energy consumption is reduced by a probability of close to 100 (query 3). For example, in the table, in 3 hours (10800 seconds) and with an average of 20 runs, 60 kJ of energy consumption has decreased. In some transitions, there is a possibility that energy consumption in the case of having idle mode increases (queries 4 and 5).

As illustrated in Table I, the cost of going idle from position *C*, and coming back to position *A* is more than the cost of staying in standby mode for some time and then going to position *A*. Therefore, it is beneficial to move to the idle mode when the duration of staying in that mode is longer than a certain amount of time. The proposed method examines whether such behavior is beneficial or detrimental to the system. Since the proposed algorithm makes decisions for each iteration (for every 20 boxes), and the decision is based on an average reduction in energy consumption, in some cases, the variable *energy* may be more than *energy2* (especially at the beginning due to small differences between energy and energy2), because by predicting the behavior of the environment, decisions are made for the entire next iteration. This is why query 5 (which indicates that energy consumption is always lower by having idle mode) is invalid. However, energy consumption generally decreases over time.

The modeling and verification results (first five queries) show that, in the case of being aware of the environmental behavior, the system can optimize the energy consumption by improving the behavior of its components. The second group of verification results (queries 6 to 16) is related to the safety properties of the robotic arm. For example, in each of the different states, such as *Idle*, *A*, *B*, and *C*, the joints of the arm must be at a certain angle in order to achieve the ideal

TABLE II
VERIFICATION RESULTS

	Query	Result
1	**E[<=10800; 20] (max:energy)**	2899.54
2	**E[<=10800; 20] (max:energy2)**	2959.38
3	**Pr[t<=10800](<>energy<=energy2)**	[0.901855,1]
4	**E<> Behavior.StandbyC and Behavior2.StandbyC and energy >= energy2**	Valid
5	**A[] Behavior.StandbyC and Behavior2.StandbyC and energy <= energy2**	Invalid
6	**A[](Behavior.InIdle imply Behavior.angle1==-8 and Behavior.angle2==-79 and Behavior.angle3==-87 and Behavior.angle4==-105 and Behavior.angle5==87 and Behavior.angle6==-51)**	Valid
7	**A[] (Behavior.ReadyA imply Behavior.angle1==-8 and Behavior.angle2==-87 and Behavior.angle3==-128 and Behavior.angle4==-56 and Behavior.angle5==87 and Behavior.angle6==-52)**	Valid
8	**A[] (Behavior.StandbyC imply Behavior.angle1==83 and Behavior.angle2==-114 and Behavior.angle3==-95 and Behavior.angle4==-65 and Behavior.angle5==93 and Behavior.angle6==-49)**	Valid
9	**A[] (Behavior.StandbyB imply Behavior.angle1==177 and Behavior.angle2==-123 and Behavior.angle3==-80 and Behavior.angle4==-66 and Behavior.angle5==90 and Behavior.angle6==-46)**	Valid
10	**A[] (Behavior.BoxMovedC imply Behavior.e1<=15)**	Valid
11	**A[] (Behavior.BoxMovedB imply Behavior.e1<=10)**	Valid
12	**Behavior.BoxMovedC –> Behavior.Grab**	Valid
13	**Behavior.BoxMovedB –> Behavior.Grab**	Valid
14	**Behavior.InIdle –> Behavior.ReadyA**	Valid
15	**Behavior.InIdle –> Behavior.Grab**	Valid
16	**A[] not deadlock**	Valid

position.

V. RELATED WORKS

For optimizing the energy consumption, authors in [12]–[15] have focused on behavioral techniques, such as time scaling, task scheduling, and break-release scheduling. In the time scaling technique, which refers to modifying energy consumption by scaling the task execution time, The optimization technique is not straightforward because the relation between energy consumption and task execution time is not linear.

Authors in [16], presented a scheme for energy transfer between nodes. The authors used a solar panel with a certain efficiency and energy loss to charge the nodes. The solar panel was considered to be in clear-sky daily irradiation, which is almost ideal. Even issues such as energy loss can be increased or decreased under the influence of environmental conditions.

Authors in [4], [6], predict the energy consumption of the system and examine whether the energy consumption is in a certain range or not. In these works, the performance of the system is based on the analysis of environmental behavior. Both of these works could be improved by considering the behavior of the environment and its impact on energy consumption. Also, in [17], which examines the impact of security on system safety, the probability of a successful message is permanently fixed, something that may not always be true in reality.

VI. DISCUSSION AND FUTURE WORK

Among the important challenges that CPS face is environmental behavior. The behavior of CPS is highly dependent on how the environment proceeds. In our case study, reducing energy consumption by changing the system's behavior depends on whether the boxes arrive earlier or later than a specific time. As mentioned in Section IV, the amount of energy consumption for the case study is reduced when the idle mode is not used in the case that the boxes' arrival time to the position *A* is earlier than a specific time. Since the box arrival time is unpredictable, such environment behaviors cannot be decided by the system. This presents a challenge to the improvement of energy consumption. To solve this challenge, we dynamically modeled the environment's behavioral disorderliness and decided on the next iteration continuously.

In this work, we analyze the reduction of energy consumption by improving the behavior of the components without structural changes by using cost-optimal reachability analysis. Due to the dependence of system behavior on the environment, we examine the behavior of the environment from two aspects. First, the behavior of the environment may be quite irregular, so using fixed, predefined probabilities for incidents makes the existing system much simplified. For this purpose, to increase the system's flexibility against the disorderliness behavior of the environment, it is modeled dynamically. Second, the system can optimize its behavior by predicting the behavior that the environment may have in the future. So, by examining the past environmental behavior, the system can dynamically adjust its behavior to optimize its behavior in terms of energy consumption. We demonstrate the applicability of our framework on an industrial case study within the I4 Lab at SDU, which can also be extended to other assets in the CPS and IoT domains.

To increase system reliability, the system must be able to operate despite the behavioral disorder of the environment. This study demonstrates that it is possible to improve system performance by predicting its future behavior. Therefore, as future work, we intend to use machine learning (ML) in the decision-making process [18] to improve the way we predict environmental behavior and check the mathematical expectation of each incident. This helps the system make a decision, based on whether it will gain a profit or a loss

by choosing a specific route. Moreover, by including the Entropy criterion [19], we will measure the severity of this disorderliness as well as the efficiency of our decision-making algorithm in different conditions.

Since such systems communicate with the main server through IoT communication protocols, we will also include security checks. As a continuation of our previous work [17], which showed how a breach in cybersecurity could endanger safety, we will improve both the safety and security aspects of CPS with IoT features. As ongoing work, we will: (1) Use ML algorithms in the decision-making process to increase the efficiency of the decision; (2) Measure the severity level of the disorderliness of the environment by including the Entropy criterion; (3) Improve and integrate safety and security in CPS & IoT verification; (4) Demonstrate the applicability with other verification/optimization tools supporting these techniques.

Acknowledgement: This research has been performed under the EF-CPS: Trustworthy CPS & IoT within I4 project, and "Energy-efficient Programming of Collaborative Robots" project funded by ELFORSK.

REFERENCES

[1] U. Nations, "The 17 goals for sustainable development," 2015. [Online]. Available: https://sdgs.un.org/goals

[2] J. Lygeros and M. Prandini, "Stochastic hybrid systems: A powerful framework for complex, large scale applications," *European Journal of Control*, vol. 16, p. 583–594, 11 2010.

[3] E.-Y. Kang, D. Mu, and L. Huang, "Probabilistic verification of timing constraints in automotive systems using uppaal-smc," in *Integrated Formal Methods*, ser. EuroSys '10. Cham: Springer International Publishing, 2018, pp. 236–254.

[4] E.-Y. Kang, D. Mu, L. Huang, and Q. Lan, "Model-based analysis of timing and energy constraints in an autonomous vehicle system," in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2017, pp. 525–532.

[5] ——, "Verification and validation of a cyber-physical system in the automotive domain," in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2017, pp. 326–333.

[6] E.-Y. Kang, L. Huang, and D. Mu, "Formal verification of energy and timed requirements for a cooperative automotive system," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1492–1499. [Online]. Available: https://doi.org/10.1145/3167132.3167291

[7] S. C. Jepsen, T. I. Mørk, J. Hviid, and T. Worm, "A pilot study of industry 4.0 asset interoperability challenges in an industry 4.0 laboratory," in *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2020, pp. 571–575.

[8] "Effimat automated warehouses." [Online]. Available: https://effimat.com/

[9] "Enabled robotics." [Online]. Available: https://www.enabled-robotics.com/

[10] "Collaborative robots universal robots." [Online]. Available: https://www.universal-robots.com/

[11] A. David, K. Larsen, A. Legay, M. Mikučionis, and D. Poulsen, "Uppaal smc tutorial," *International Journal on Software Tools for Technology Transfer*, vol. 17, 01 2015.

[12] D. Meike, M. Pellicciari, G. Berselli, A. Vergnano, and L. Ribickis, "Increasing the energy efficiency of multi-robot production lines in the automotive industry," *IEEE International Conference on Automation Science and Engineering*, pp. 700–705, 2012.

[13] A. Vergnano, C. Thorstensson, B. Lennartson, P. Falkman, M. Pellicciari, F. Leali, and S. Biller, "Modeling and optimization of energy consumption in cooperative multi-robot systems," *IEEE Transactions on Automation Science and Engineering*, vol. 9, no. 2, pp. 423–428, 2012.

[14] D. Meike, M. Pellicciari, and G. Berselli, "Energy efficient use of multirobot production lines in the automotive industry: Detailed system modeling and optimization," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 798–809, 2014.

[15] M. Pellicciari, G. Berselli, F. Leali, and A. Vergnano, "A method for reducing the energy consumption of pick-and-place industrial robots," *Mechatronics*, vol. 23, no. 3, pp. 326–334, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.mechatronics.2013.01.013

[16] A. Gamatié, G. Sassatelli, and M. Mikučionis, *Modeling and Analysis for Energy-Driven Computing using Statistical Model-Checking; Design, Automation and Test in Europe Conference, Virtual, France.*, 02 2021.

[17] L. Huang and E.-Y. Kang, "Formal verification of safety & security related timing constraints for a cooperative automotive system," in *Fundamental Approaches to Software Engineering*, R. Hähnle and W. van der Aalst, Eds. Cham: Springer International Publishing, 2019, pp. 210–227.

[18] A. Kuhnle and G. Lanza, "Application of reinforcement learning in production planning and control of cyber physical production systems," in *Machine Learning for Cyber Physical Systems*, J. Beyerer, C. Kühnert, and O. Niggemann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 123–132.

[19] H. Li, "Entropy reduction via communications in cyber physical systems: How to feed maxwell's demon?" in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 2206–2210.

Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems pp. 211–218
DOI: 10.15439/2021F90
ISSN 2300-5963 ACSIS, Vol. 26

Decentralized Controller for Software Interconnected System Subject to Malicious Attacks

Pushkar Kishore
Dept. of C.S.E.
NIT Rourkela
Odisha, India
518CS1002@nitrkl.ac.in

Swadhin Kumar Barisal
1. Dept. of C.S.E., NIT Rourkela.
2. S'O'A Deemed to be University,
Odisha, India
swadhinbarisal@gmail.com

Durga Prasad Mohapatra
Dept. of C.S.E.
NIT Rourkela
Odisha, India
durga@nitrkl.ac.in

***Abstract*—This paper examines the decentralized controller for an interconnected software system subject to malicious attacks. The security of software interconnected systems (SIS) subject to malicious attacks is discussed using Event-Triggered Mechanism (ETM). We design a novel ETM with decentralized feedback for managing resources and keeping the system stable during attacks. We use Numenta-Hierarchical Temporal Memory (N-HTM) for monitoring the ETM values. In addition, numerical simulation of the service provider system is considered for illustrating our model's effectiveness. Experiments reveal that our model stabilizes the system after an average of 2s from the launch of the last attack. As a result, the average consumption of the resources is reduced by 70%.**

***Index Terms*—Decentralized control, Software interconnected systems, Event-triggered control, Numenta-Hierarchical Temporal Memory**

I. Introduction

Interconnected systems consist of a set of coupled subsystems that are physically distributed. We use decentralized control due to its better flexibility, scalability, and reliability compared to centralized control. In these years, decentralized control schemes have been used for dealing with complex interconnected systems [1].

Rapid development in computers and communication directed the rise of Software Control Systems (SCs). Most works on distributed system communication assumed that the quality of service of the communication would ensure stable communication. The objective of our model is to maintain prefixed controller performance during attacks. Different kinds of attacks are examined in security control domains such as denial-of-service (DoS) attacks [2], replay attacks, and deception attacks [3]. Liu et al. [2] concentrated on the stabilization problem for communication systems enduring intermittent DoS jamming attacks. During replay attacks, the data from the operator to the actuator is maliciously repeated. An et al. [4] examined a secure state estimation model based on an adaptive switched mechanism during deception attacks. Wang et al. [5] modeled deception attacks utilizing norm-bounded functions conditioned on the state of the system and developed a resilient control for neural control systems. Ding et al. [6] examined distributed recursive filter against deception attacks utilizing a gradient method. Unlike a simplistic communication system with just one independent controller, Software Interconnected System (SIS) has various subsystems and controllers, making it tough to analyze its performance during malicious attacks. In point-to-point communication, performance is hardly maintained by the controller during non-ideal data transmission. In these years, researchers are trying to improve the Quality of Service (QoS) of the software for a better Quality of Control (QoC). During the last decade, the event-triggered mechanism helped in balancing between QoS and QoC for control systems and sampled-data control systems [7]. Our model is different from the time-triggered mechanism (TTM) in terms of execution frequency of the Event-Triggered Mechanism (ETM). When the frequency of execution of ETM is reduced, resource consumption is controlled.

The existing state-of-the-art works highlight that ETC mainly relies on absolute error, relative error, and some additional measuring parameters. If the error is beyond a predefined threshold, then data-releasing is done. Fei et al. [8] investigated cloud-aided active suspension control where the ETM threshold depends on the bandwidth use. Tian et al. [9] designed a hybrid-triggered scheme which was based on random switching within TTM and ETM and achieved a commending tradeoff among QoC and QoS.

We measure the performance of ETM using a parameter termed as Data Releasing Rate (DRR). Data Release Sample Ratio (DRSR) is the ratio of the number of data releases to the number of data-sampling in a defined period. After reviewing the earlier state-of-the-art works, it was concluded that ETM could effectively reduce the DRR. Whenever an attack occurs on the software system, the controller needs more frequent data to stabilize the system again. We apply a technique, namely Numenta Hierarchical Temporal Memory (N-HTM) [10] which can find and spot anomalous patterns for data where simplistic techniques such as thresholds generate substantial false positives and false negatives. It helps set thresholds; otherwise, the delayed transmission will reduce the system life due to rapid temperature fluctuations. Until the ETM responds with a change in data, the service provider will get massive requests due to malicious attacks and heat the system. As the controller gets feedback from ETM, it reduces the load rapidly, leading to quick cooling. The cycle of rapid temperature changes deteriorates the system's life. A simple ETM will not be sufficient, and designing a resilient ETM for

SCS subject to malicious attacks is challenging. The above issues motivate us to design an efficient ETM.
This article introduces a decentralized controller for interconnected software systems subject to malicious attacks. The main contributions of this article are as follows:

1) Malicious attacks on the software are considered.
2) A novel ETM is proposed where the control unit receives the least amount of feedback defined using N-HTM and guarantees desired control performance during malicious attacks.
3) DRR during the run-time is maintained at a moderate level.
4) A decentralized security output feedback control technique is proposed for stabilizing the system during malicious attacks.

The remaining part of this paper is organized as follows: Section II discusses the integrated proposed model of the SIS, ETM, and malicious attacks. Section III manifests the design considerations of decentralized, resilient control for SIS subject to malicious attacks. Section IV signifies results, advantages and effectiveness of our proposed model using service provider model, Section V considers related state-of-the-art works, section VI highlights threats to validity of our model and Section VII concludes the paper.

II. Integrated proposed model of SIS, ETM, and malicious attacks

Before going in-depth, we summarize the notations that will be used in the paper. $\mathbb{R}^n$ is used for n-dimensional Euclidean space, $\mathbb{R}^{nxm}$ for a set of n x m matrices, $\|.\|$ represents Euclidean norm, P^T is the transpose of a matrix P, $E\{\beta\}$ evaluates the expectation of the stochastic variable β, $diag_N\{X_i\} = diag\{X_1, X_2, ..., X_N\}$, $col_n\{x_i\} = \left[x_1^T, ..., x_N^T\right]^T$ and $*$ represents the symmetric term in a matrix.

A. System Description

Consider an interconnected system S depicted in Figure 1:

$$\dot{x}(t) = Ax(t) + Bu(t) + f(t, x(t)) \tag{1}$$

$$y(t) = Cx(t) \tag{2}$$

where A, B and C are matrices, $f(t, x(t))$ represents the coupling between interconnected systems, state of the system: $x(t) = col_N\{x_i(t)\}$, input to the system: $u(t) = col_N\{u_i(t)\}$, output of the system: $y(t) = col_N\{y_i(t)\}$, $x_i(t)\epsilon\mathbb{R}^{n_i}(\sum_{i=1}^{N} n_i = n)$, $y_i(t)\epsilon\mathbb{R}^{p_i}(\sum_{i=1}^{N} p_i = p)$, $u_i(t)\epsilon\mathbb{R}^{m_i}(\sum_{i=1}^{N} m_i = m)$, $p < n$ and C is row full rank. Here it is assumed that coupling function satisfies Equation 3:

$$||f_i(t, x(t))|| \leq \delta_i^2 \|F_i x(t)\| \tag{3}$$

where δ_i is scalar and F_i is a known matrix.

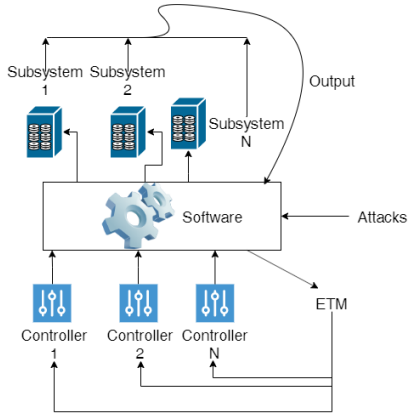


Fig. 1: The framework of software decentralized control system

B. Novel ETM

Figure 1 shows the output of the subsystem provided to the software and monitored by the ETM. The ETM monitors the output and accordingly decides whether to inform the controller or not. We use N-HTM for calculating the raw anomaly score of outputs of subsystems, as anomaly score indicates the deviation between actual and predicted output. First, an anomaly score is computed from the intersection between predicted and actual sparse vectors. Then, we compute the anomaly likelihood value from the window of the last W raw anomaly scores. N-HTM models this distribution as a rolling normal.

$$W = \sum S_t \tag{4}$$

where, t = 1, 2, 3, 4, ... The sample mean and variance are continuously calculated and updated using Equations 5 and 6.

$$\mu_t = \frac{\sum_{i=0}^{i=w-1} s_{t-i}}{k} \tag{5}$$

$$\delta_t^2 = \frac{\sum_{i=0}^{i=w-1} (s_{t-i} - \mu_t)^2}{k-1} \tag{6}$$

where, w is the last window length and k is the number of instances. Then an average of recent anomaly scores is evaluated using Equation 7, and the anomaly is confirmed by applying a threshold to the Gaussian tail probability.

$$L_t = 1 - Q(\frac{\mu_t{}' - \mu_t}{\delta_t}) \tag{7}$$

where $\mu_t' = \frac{\sum_{i=0}^{i=w-1} s_{t-i}}{j}$ and $j < k$. Anomalous behavior will be reported if $L_t \geq 1$.

Remark 1: We use different output values, i.e., y from other ETC [11]. Anomalous data is searched within a sequence of outputs, and an alert is sent to the controller for altered output. Anomalous finder like N-HTM tackles fluctuations caused by noises and disturbances.

Remark 2: Compared to the conventional ETM, our model sends packets sporadically to the controller when the system is under attack or sensing external disturbance. The above

process will ensure better QoC, and DRR average value will be lower through runtime.

C. Malicious Attack Model

Decentralized control will break the controller problems by utilizing multiple controllers. In our work, the decentralized scheme of the subsystems is defined using Equation 8.

$$u_i(t) = K_i y_i(t) \tag{8}$$

where, K_i is the local controller gain of the subsystems.
Remark 3: The centralized output feedback control is defined using Equation 9.

$$u_i^c(t) = \sum_{j=1}^{N} k_{ij} y_j(t) \tag{9}$$

The control information transmitted via software is vulnerable to attack. The control input to the subsystem during the malicious attack is defined using Equation 10.

$$\breve{u}_i(t) = u_i(t_k) \pm a_i(t) u_i(t_k) \tag{10}$$

where, $t\varepsilon\,[t_k + \tau_k, t_{k+1} + \tau_{k+1}]$, $a_i(t_k)$ is the attack that tampers the control input at time t_k, τ_k is the software processing delay having value between η and expectation of τ. Gu et al. [11] had proved that malicious attack will remain undetected when $\|a_i(t)\| \leq \varsigma^3/N$, where ς is a scalar value.
Remark 4: The malware which attacks intermittently is effective due to these reasons.

1) The probability of the continuous attack being detected is more prominent than random intermittent attacks. E.g., Trojan is tougher to detect due to its stealth and intermittent attacking behavior, while worms are easily detected due to continuous attack [1].
2) The attack is obstructed due to the underlying operating system on which the software is running.
3) The objective of the attack is to destroy the control system; thus, the malicious instances need to alter the input so that the dropout of the attack signal looks like a standard transmission.

Remark 5: $a_i(t)$ varies between 0 to 1 depending on whether the data transmitted through the system is benign or malicious. The intensity of attacks happening on each subsystem is defined using the number of malicious samples active at any point.
From the above discussions, the control input during malicious attacks is finalized using Equation 11.

$$\breve{u}(t) = K y(t_k) \pm a(t) K y(t_k) \tag{11}$$

where, $K = diag_N\,\{K_i\}$.

[1]https://www.websecurity.digicert.com/security-topics/difference-between-virus-worm-and-trojan-horse

D. Overall Model

Combining Equations 1, 2 and 8, we define SIS using Equation 12.

$$\dot{x}(t) = Ax(t) + B(I + a(t))Ky(t_k) + f(t) + \sum_{i=1}^{N} (\bar{a}_i - a_i(t))BL_iKy(t_k) \tag{12}$$

where, $L_i = diag\,\{0..0I0..0\}$; $I's$ location depends on the value of i. We define $e(t_k, l) = x(t_k) - x(t_k + l)$ and $\eta_t = t - (t_k + l)$. Then, Equation 12 is transformed to Equation 13.

$$\dot{x}(t) = Ax(t) + B(I + a(t))KC\,[e(t_k, l) + x(t - \eta(t))] + f(t) + \sum_{i=1}^{N} (\bar{a}_i - a_i(t))BL_iKC[e(t_k, l) + x(t - \eta_t))] \tag{13}$$

A SIS is termed stable whenever $x^T(t)Px(t) \leq \varsigma^2$, $\forall t \geq t_0 + T$ for $\|a_i(t)\| \leq \varsigma^3/N$. Here, $P > 0$ and $T > 0$. The primary purpose of our work is to build a controller and ETM such that SIS is stable in the presence of malicious attacks.

III. Controller Design

In this section, we develop the controller together with ETM for SIS when malicious attacks are happening. Conditions will be formed and represented in terms of a set of linear matrix inequalities. At first, we define $\zeta = \begin{bmatrix} x^T(t) & x^T(t-\eta_1) & x^T(t-\eta(t)) & x^T(t-\eta_2) & a^T(t) \end{bmatrix}$. $\Im_i$ represents a compatible row-matrix with the i^{th} block as identity matrix and other as zero matrices, e.g. $\Im_3 = \begin{bmatrix} 0 & 0 & I & 0 & 0 \end{bmatrix}$. Later on, we discuss some lemmas which are applied in our designs.
Lemma 1 [11]: Let $\eta(t)\varepsilon\,[\eta_1, \eta_2]$, $x(t)\varepsilon\mathbb{R}^n$ be some positive matrices like $R_1\varepsilon\mathbb{R}^{nxn}$, $R_2\varepsilon\mathbb{R}^{nxn}$ and matrix $U\varepsilon\mathbb{R}^{nxn}$. Then the inequalities are given in Equation 14:

$$\begin{aligned} -\eta_1 \int_{t-\eta_1}^{t} \dot{x}^T(s) R_1 \dot{x}(s) ds &\leq \zeta^T(t) \Re_1 \zeta(t) \\ -(\eta_2 - \eta_1) \int_{t-\eta_2}^{t-\eta_1} \dot{x}^T(s) R_2 \dot{x}(s) ds &\leq \zeta^T(t) \Re_2 \zeta(t) \end{aligned} \tag{14}$$

Where, $\Re_1 = -(\Im_1 - \Im_2)^T R_1 (\Im_1 - \Im_2)$ and $\Re_2$ is solved using Equation 15.

$$\Re_2 = -\begin{bmatrix} \Im_2 - \Im_3 \\ \Im_3 - \Im_4 \end{bmatrix}^T \begin{bmatrix} R_2 & * \\ U & R_2 \end{bmatrix} \begin{bmatrix} \Im_2 - \Im_3 \\ \Im_3 - \Im_4 \end{bmatrix} \tag{15}$$

Lemma 2: For some given constants, η_1, η_2, ϱ, ς, σ, k, with ETM and SIS is stable whenever there exists matrices like $P > 0$, $\psi > 0$, $Q_1 > 0$, $Q_2 > 0$, $R_1 > 0$, $R_2 > 0$, a matrix U, a positive scalar ϵ fulfilling:

$$\begin{bmatrix} \Gamma_1 & * & * \\ ZA_0 & -Z & * \\ \Gamma_2 & 0 & -\Gamma_3 \end{bmatrix} < 0 \tag{16}$$

Where, Γ_1 is defined as:

$$\Gamma_1 = \begin{bmatrix} \Gamma_{11} & * & * & * & * \\ R_1 & \Gamma_{22} & * & * & * \\ \Gamma_{31} & \Gamma_{32} & \Gamma_{33} & * & * \\ 0 & -U & \Gamma_{43} & \Gamma_{44} & * \\ \Gamma_{51} & 0 & \Gamma_{53} & 0 & \Gamma_{55} \end{bmatrix} \tag{17}$$

$\Gamma_{11} = PA + A^T P + Q_1 + Q_2 - R_1 + \epsilon\delta^2 F^T F + \varsigma P$
$\Gamma_{22} = -Q_1 - R_1 - R2$
$\Gamma_{31} = C^T K^T B^T P(I + a)$
$\Gamma_{32} = R_2 + U$
$\Gamma_{33} = -2R_2 - U - U^T + 4\sigma\psi$
$\Gamma_{43} = \Gamma_{32}$
$\Gamma_{44} = -Q_2 - R_2$
$\Gamma_{51} = \Gamma_{31}$
$\Gamma_{53} = 2\sigma(1 + \varrho)\psi C$
$\Gamma_{55} = -(1 - \sigma - 2\varrho\sigma)\psi$
$A_0 = \begin{bmatrix} A & 0 & B(I + a)KC & 0 & Ba \end{bmatrix}$
$A_{2i} = \begin{bmatrix} 0 & 0 & BL_iKC & 0 & -2BL_i \end{bmatrix} \forall i\epsilon\{1, n\}$

$$\Gamma_2 = \begin{bmatrix} ZA_{21} \\ . \\ . \\ . \\ ZA_{2N} \end{bmatrix}$$

$Z = \eta_1^2 R_1 + (\eta_2 - \eta_1)^2 R_2$
$\Gamma_3 = diag\{Z, ..., Z\}$ N times
$\psi = C^T \tilde{\psi} C$

Now, we evaluate the values of gain and parameters of ETM.

Theorem 1: For some given constants, η_1, η_2, ϱ, ς, σ and ϵ, we have a stable SIS with ETM during malicious attacks when $\tilde{\psi} > 0$, $\tilde{Q}_j > 0$, $\tilde{R}_1 > 0$, $\tilde{R}_2 > 0$, Y, $\tilde{U}$ and V.

$$\Gamma = \begin{bmatrix} \tilde{\Gamma}_1 & * & * & * \\ \tilde{A}_1 & -\tilde{Z}_0 & * & * \\ \tilde{\Gamma}_2 & 0 & \tilde{\Gamma}_3 & * \\ \tilde{\Gamma}_4 & 0 & 0 & -\epsilon I \end{bmatrix} < 0 \tag{18}$$

$$CX = VC \tag{19}$$

Where,

$$\tilde{\Gamma}_1 = \begin{bmatrix} \tilde{\Gamma}_{11} & * & * & * & * \\ \tilde{R}_1 & \tilde{\Gamma}_{22} & * & * & * \\ \tilde{\Gamma}_{31} & \tilde{\Gamma}_{32} & \tilde{\Gamma}_{33} & * & * \\ 0 & -\tilde{U} & \tilde{\Gamma}_{43} & \tilde{\Gamma}_{44} & * \\ \tilde{\Gamma}_{51} & 0 & \tilde{\Gamma}_{53} & 0 & \tilde{\Gamma}_{55} \end{bmatrix} \tag{20}$$

$\tilde{\Gamma}_{11} = AX + A^T X + \tilde{Q}_1 + \tilde{Q}_2 - \tilde{R}_1 + \varsigma X$
$\tilde{\Gamma}_{22} = -\tilde{Q}_1 - \tilde{R}_1 - \tilde{R}2$
$\tilde{\Gamma}_{31} = C^T Y^T B^T (I + a)$
$\tilde{\Gamma}_{32} = \tilde{R}_2 + \tilde{U}$
$\tilde{\Gamma}_{33} = -2\tilde{R}_2 - \tilde{U} - \tilde{U}^T + 4\sigma\tilde{\psi}$
$\tilde{\Gamma}_{43} = \tilde{\Gamma}_{32}$
$\tilde{\Gamma}_{44} = -\tilde{Q}_2 - \tilde{R}_2$
$\tilde{\Gamma}_{51} = \tilde{\Gamma}_{31}$
$\tilde{\Gamma}_{53} = 2\sigma(1 + \rho)\tilde{\psi} C$
$\tilde{\Gamma}_{55} = -(1 - \sigma - 2\rho\sigma)\tilde{\psi}$
$\tilde{\Gamma}_4 = \begin{bmatrix} \epsilon\delta FX & 0 & 0 & 0 & 0 \end{bmatrix}$
$\tilde{A}_0 = \begin{bmatrix} AX & 0 & B(I + a)YC & 0 & Ba \end{bmatrix}$
$\tilde{A}_{2i} = \begin{bmatrix} 0 & 0 & BL_iYC & 0 & -WBL_i \end{bmatrix} \forall i\epsilon\{1, n\}$

$$\tilde{\Gamma}_2 = \begin{bmatrix} \tilde{A}_{21} \\ . \\ . \\ . \\ \tilde{A}_{2N} \end{bmatrix}$$

$\tilde{\Gamma}_3 = diag\left\{\tilde{Z}_1, ..., \tilde{Z}_N\right\}$
$\tilde{Z}_i = -2\alpha_i X + \alpha_i^2 \tilde{Z}$

The gain is defined using Equation 21 and weight matrix using 22.

$$K = YV^{-1} \tag{21}$$

$$\bar{\psi} = (CC^T)^{-1} CX^{-1} \tilde{\psi} X^{-1} C^T (CC^T)^{-1} \tag{22}$$

Theorem 2: For some given constants, η_1, η_2, ϱ, ς, σ, ϵ and φ , we have a stable SIS with ETM during malicious attacks when $\tilde{\psi} > 0$, $\tilde{Q}_j > 0$, $\tilde{R}_1 > 0$, $\tilde{R}_2 > 0$, Y, $\tilde{U}$ and V. The linear inequalities holding at this stage is:

$$\Gamma < 0 \tag{23}$$

$$\begin{bmatrix} -\varphi I & * \\ CX - VC & -I \end{bmatrix} \tag{24}$$

IV. Results and Effectiveness

In this section, we discuss the service provider model, which manifests the advantages and effectiveness of our model. Figure 2 represents the architecture of the service provider system. The above model comprises four subsystems with a

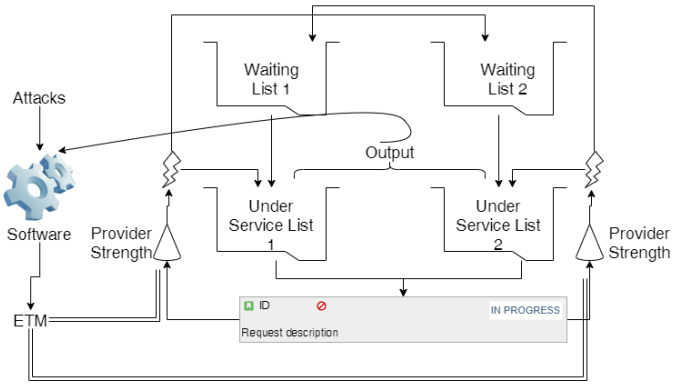


Fig. 2: Architecture of Service Provider System

waiting list containing requests to be served later and **under service list** processing the requests. Requests are stored in a pool and sent to the waiting list or **under service list**. Requests passing through right provider strength are passed to **under service list 2** and **waiting list 1**. Similarly, requests passing through left provider strength will be sent to the **under service length 1** and **waiting list 2**. Requests are transferred to the **under service** list from their respective waiting list at a fixed rate. Our objective is to keep the number of requests at 50% of the capacity of the list size such that the server keeps running smoothly. The limit of 50% can be varied according to our server processing power. Provider Strength is varied across

run-time to keep **under service lists'** performance at their best. The outputs from the **under service** list are sent to the software prone to malicious attacks. After getting the feedback from the software, ETM decides the following vector of parameters for provider strengths to ensure fulfilling our objective.
The system can be modelled using following equations:

$$\begin{aligned}\frac{dr_{usr1}}{dt} &= (c_{usr1}/C_{usr1})(r_{usr1})^2 + (c_{wl1}/C_{usr1})(r_{wl1})^2 + \frac{\eta_1}{C_{usr1}}v_1 \\ \frac{dr_{usr2}}{dt} &= (c_{usr2}/C_{usr2})(r_{usr2})^2 + (c_{wl2}/C_{usr2})(r_{wl2})^2 + \frac{\eta_2}{C_{usr2}}v_2 \\ \frac{dr_{wl1}}{dt} &= (c_{wl1}/C_{wl1})(r_{wl1})^2 + \frac{(1-\eta_1)}{C_{wl1}}v_1 \\ \frac{dr_{wl2}}{dt} &= (c_{wl2}/C_{wl2})(r_{wl2})^2 + \frac{(1-\eta_2)}{C_{wl2}}v_2 \\ y_{usr1} &= k_c r_{usr1}, y_{usr2} = k_c r_{usr2} \end{aligned} \quad (25)$$

where, $C_{usr1} = C_{wl1} = C_{usr2} = C_{wl2} = 1000$ representing number of requests going in the sub-system, $c_{usr1} = c_{wl1} = c_{wl2} = c_{usr2} = 50$ representing numbers of requests going out of the sub-system, $k_c = 0.5$, r_* shows the maximum holding capacity of the subsystems in terms of number of requests and η_* along with v_* represents the provider strength. We keep the operating range of the system as:

$$0 \leq r_{usr1} \leq 800, 0 \leq r_{usr2} \leq 800, 0 \leq r_{wl1} \leq 500, 0 \leq r_{wl2} \leq 500 \quad (26)$$

For the minimal case, we consider following parameters: r_{usr1} = 600, r_{usr2} = 600, r_{wl1} = 200, r_{wl2} = 200, v_1 = 2, v_2 = 2, η_1 = 0.7 and η_2 = 0.3. Then, we obtain the following system equations:
$x_{usr1} = r_{usr1}$ - 600, $x_{usr2} = r_{usr2}$ - 600, $x_{wl1} = r_{wl1}$ - 200, $x_{wl2} = r_{wl2}$ - 200, $u_1 = v_1$ - 2 and $u_2 = v_2$ - 2.
Then Equation 25 is modified as follows:

$$\begin{aligned}\dot{x}_{usr1} &= 0.05(x_{wl1}+200)^2 - 0.05(x_{usr1}+600)^2 + 0.0007(u_1+2) \\ \dot{x}_{usr2} &= 0.05(x_{wl2}+200)^2 - 0.05(x_{usr2}+600)^2 + 0.0003(u_2+2) \\ \dot{x}_{wl1} &= -0.05(x_{wl1}+200)^2 + 0.0003(u_1+2) \\ \dot{x}_{wl2} &= -0.05(x_{wl2}+200)^2 + 0.0007(u_2+2) \end{aligned} \quad (27)$$

From 26, we define the range of the variables as:

$$-600 \leq x_{usr1} \leq 200, -600 \leq x_{usr2} \leq 200, -200 \leq x_{wl1} \leq 300, -200 \leq x_{wl2} \leq 300 \quad (28)$$

From Equation 27 and 28, we redefine the system model as:

$$\begin{aligned}\dot{x}_{usr1} &= 0.05x_{wl1}^2 + 20x_{wl1} - 0.05x_{usr1}^2 - 60x_{usr1} + 0.0007u_1 \\ \dot{x}_{usr2} &= 0.05x_{wl2}^2 + 20x_{wl2} - 0.05x_{usr2}^2 - 60x_{usr2} + 0.0003u_2 \\ \dot{x}_{wl1} &= 0.05x_{wl1}^2 - 20x_{wl1} + 0.0003u_1 \\ \dot{x}_{wl2} &= 0.05x_{wl2}^2 - 20x_{wl2} + 0.0007u_2 \\ y_{usr1} &= 0.5x_{usr1}, y_{usr2} = 0.5x_{usr2} \end{aligned} \quad (29)$$

From Equations 28 and 29, we restate Equation 1 as:

$$A = \begin{bmatrix} -60 & 20 & 0 & 0 \\ 0 & -20 & 0 & 0 \\ 0 & 0 & -60 & 20 \\ 0 & 0 & 0 & -20 \end{bmatrix}$$
$$B = \begin{bmatrix} 0.0007 & 0 \\ 0 & 0 \\ 0 & 0.0003 \\ 0 & 0 \end{bmatrix} F = \begin{bmatrix} 39 & 0 & 0 & 0 \\ 0 & 0.05 & 0 & 0 \\ 0 & 0 & 60 & 0 \\ 0 & 0 & 0 & 0.06 \end{bmatrix} \quad (30)$$

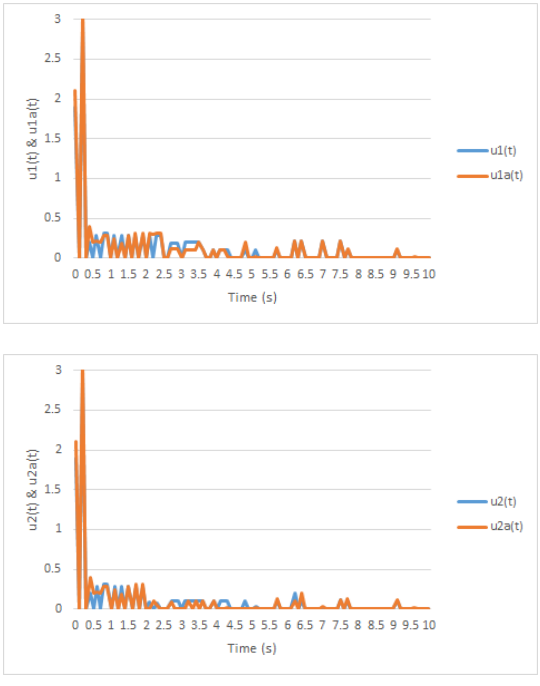
We define $x_i = [C_{usri}, C_{wli}]^T$. The subsystem S can be described using Equation 1 and satisfies

$$||f_i(t, x(t))|| \leq 0.01 \, ||F_i x(t)|| \quad (31)$$

The measured output of the subsystem S_i is:

$$y_i(t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} x_i(t) \forall i = 1, 2 \quad (32)$$

Here, the system is a nonself-regulating due to the presence



(a) Sequence of attacks on subsystem wl2 and usr2

Fig. 3: Sequence of attacks done on the subsystem

of two positive provider units in the system. In Figure 2,

TABLE I: Malware dataset

Sl. No.	Malware family	Number of samples
1	Backdoor	1352
2	Worm	559
3	Trojan	2394
4	Virus	809
Total number of samples		5114

control signal is passed through a software. We choose software processing delay η = 1 ms and expectation of τ = 10 ms. We select the parameters of ETM as σ = 0.1 and ϱ = 0.2.

Suppose our software is attacked by malicious instances described in Table I. These samples are provided by VirusTotal [2]. In case of $\|a_i(t)\| \leq \varsigma^3/N$, ς is 0.1. Figure 3 depicts the sequences of attacks done on the in/out puts of subsystems.

From Equation 21 and 22, we can obtain the gain and weight matrix as described below:

$$k = \begin{bmatrix} -0.8 & 1.2 \\ 1.2 & -0.8 \end{bmatrix} \tag{33}$$

$$\bar{\psi} = \begin{bmatrix} 0.000704 & 0.000384 \\ 0.000064 & -0.000256 \end{bmatrix} \tag{34}$$

As per the attacks depicted in Figure 3, and with initial values determined from r_{usr1}, r_{wl1}, r_{usr2} and r_{wl2} = [100, 150, 150, 100], we can get the state response of each subsystem using the parameters described above and depicted in Figure 4. It is observed that subsystems achieve stability after 6s even under attack conditions. The data release sequence is depicted in Figure 5. It is observed that the average DRR is around 30% of the total time. Thus, it is clear that our proposed model effectively provides service even when the software is under attack. The reduction of DRR achieved using ETM assist in saving computational resources.

Now, we will illustrate the beneficial aspect of our proposed model using:

1) Provider performance.
2) DRR.

We alter the execution of the subsystems by attacking software from 6s to 10s. Now, we will study the response of the subsystem under two cases:

1) *Case 1*: The model proposed by Gu et al. [11], where σ = 0.1 and ϱ = 0.2 and dependent on ETM.
2) *Case 2*: The model proposed by us with similar parameters and dependent on N-HTM based ETM.

Our proposed model's ETM is a bit less sensitive to variable external disturbances compared to other ETM (as shown in case 1). The sensitivity is measured by the number of transmissions sent to the providers. The average DRR is low, but instantaneous higher DRR is required sometimes to defy and counter the effects of the attack. From Figure 6 and 7, we

[2]https://www.virustotal.com/

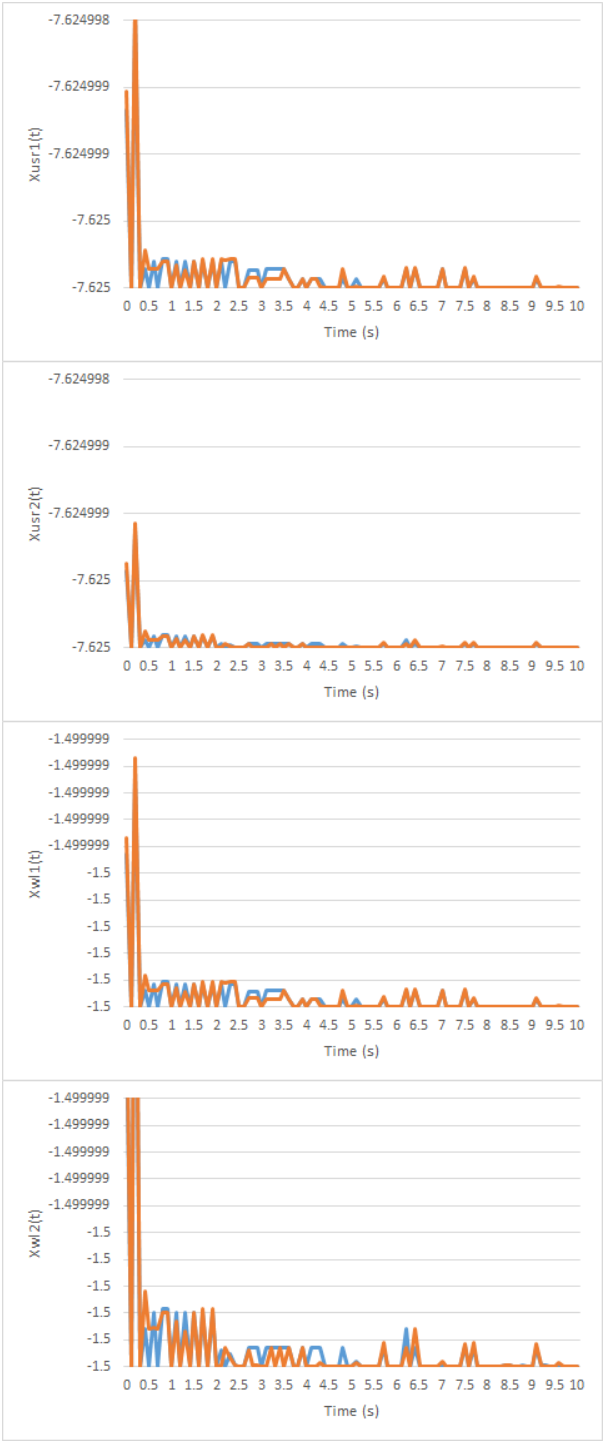


Fig. 4: State response of each subsystem

conclude that the average DRR is higher in Case 1 compared to Case 1 during the extreme attack duration of 6-10s. It is also concluded that the system providers with proposed N-HTM based ETM receive less information in case of the attack on software than other ETM. However, the anomaly score returned by the ETM covers up the lower DRR and saves

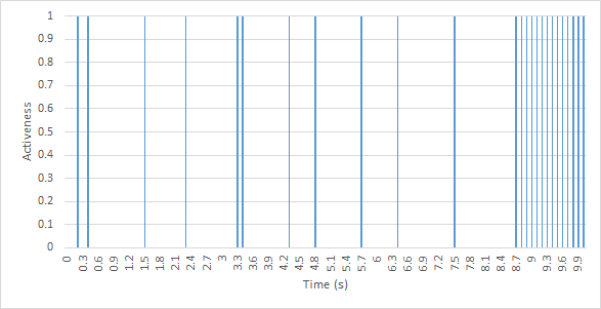


Fig. 5: Release by ETM

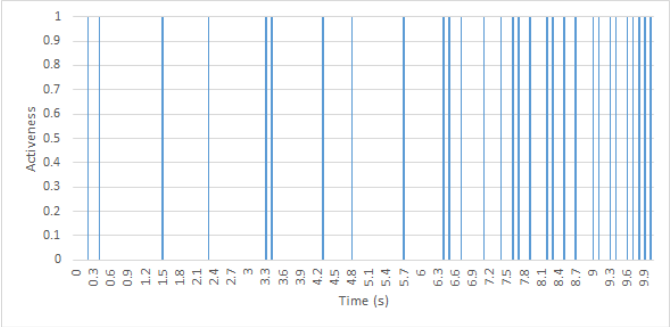


Fig. 6: Release by ETM (Case 2)

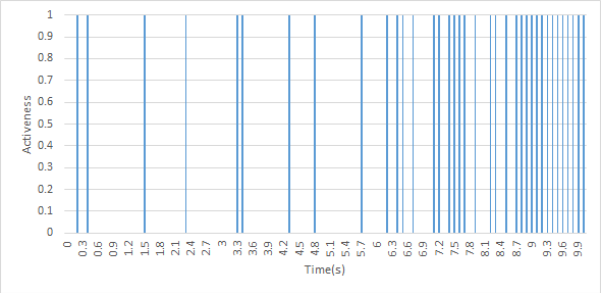


Fig. 7: Release by ETM (Case 1)

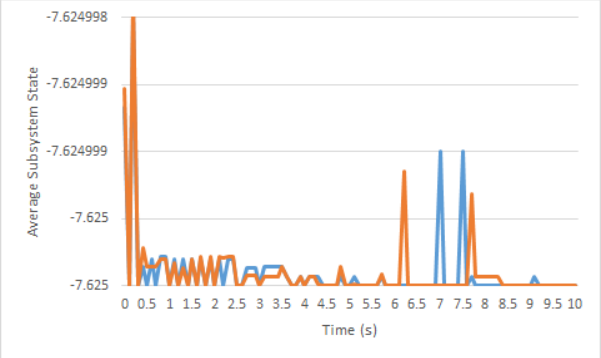


Fig. 8: Average state response (Case 2)

resources. An anomaly score provides exact rectification for the inputs such that the system reaches stability or remains stable. However, the ETM in case 1 needs to send many data for higher-order attacks and needs much time for stabilization. Therefore, the above practice will deplete a lot of resources and inefficient for longer running time. Thus, we can achieve better performance using our proposed model. From Figure 8, it is concluded that on average, the subsystems will stabilize after 8^{th} second, i,e., 2 seconds after the extensive attack had occurred. However, for Case 1, the system stabilizes at the last moment, i.e., 9^{th} second. However, the software may not always fail from an attack as it depends on penetration level. For the above case, N-HTM will not create any threat alert.

V. Related State-of-the-art Works

Liang et al. [12] studied the quantized cooperative control problem for multiagent systems with unknown gains. They designed a speed function and observed that the tracking errors converge to a prescribed compact set in a given finite time. Niu et al. [1] proposed an adaptive neural-network-based dynamic surface control (DSC) method for stochastic interconnected nonlinear non-strict-feedback systems. The proposed controllers guarantee that the closed-loop stochastic interconnected system is probably semi-globally bounded stable.

Liu et al. [2] estimated the security of distributed state for nonlinear networked systems against denial-of-service attacks. An event-triggered scheme and a quantization mechanism were employed to reduce network burden. Lyapunov stability theory was used for ensuring the exponential stability of the estimation error systems. A numerical example was considered for testing the feasibility of their proposed method.

Tang et al. [13] investigated the tracking control of mobile robots under the presence of malicious denial-of-service attacks. Some explicit characterizations were considered for the duration and frequency property of malicious DOS attacks. They developed a set of event-triggering conditions for ensuring tracking convergence. A practical experiment is conducted by tracking the control of an amigobot mobile robot under the presence of malicious attacks. Ye et al. [3] investigated the detection problem of false data injection attacks in cyber-physical systems (CPSs) with white noise. For ensuring the stability of CPSs during false data-injection attacks, a summation (SUM) detector was proposed. The SUM detector utilized the current compromised as well as historical information for identifying the threat. An improved false data-injection attack with a time-variable increment coefficient was also developed. Some simulations were conducted for ascertaining the effectivity of the SUM detector.

Gu et al. [11] studied the security of NIS in the presence of cyber-attacks based on a new ETM. They designed a novel ETM and a decentralized output feedback control (DOFC) scheme to keep NIS stable in the presence of cyber-attacks. They reduced the data-release rate, consequently reducing network bandwidth, battery supply, and computation. Numerical simulations were done to illustrate the effectiveness of their technique.

For the generation of anomaly scores in sequential data,

Ahmad et al. [10] developed a technique termed N-HTM. It is predominantly used for real-time applications dedicated to finding out the anomalies in data streams of their respective domains. They had demonstrated that their system was efficient, produced accurate results in the presence of noisy data, detected subtle temporal anomalies and minimized false positives, and adaptable to statistical change in the data. Kishore et al. [14] proposed an incremental malware detection model for meta-feature API and system call sequence. They used the N-HTM for generating the anomaly score of each element in the sequence of system and API calls. The detection accuracy of 95.2% achieved using N-HTM was also the motivating factor for using N-HTM.

VI. Threats to Validity

In this section, we identify some possible threats to the validity of our approach. First, the probability of attacks on each of the sub-systems is kept constant. It is done to ensure that N-HTM learns the patterns properly and provides anomaly scores accurately. If the attack probability is different, then the N-HTM will be inaccurate due to improperly learning multiple different distribution patterns at once [10]. However, we can make the model work by considering other attention mechanism based models. Malicious samples that can attack only network channels will be ineffective. While studying the response of sub-systems, we define stability at the negative values. It seems unusual, but the state of all sub-systems is maintained at a negative point from the beginning. Due to these reasons, values at base of the state response graph represent stable state of the sub-systems. For the values provided in the ETM release graph, 1 represents active, and 0 represents no release. These results depend on the response of the software if an attack occurs on any of the sub-systems.

At last, the centralized controller will be least prone to attacks and maintain stability during the duration of attack [11]. However, the continuous attack will block the service providers from responding when the decision-maker is under continuous attack. This problem can be covered using a decentralized controller with the least time required to attain stability and reduce resource consumption with ETM guided by N-HTM.

VII. Conclusions and Future work

This paper proposes a model that stabilizes the Software InterConnected Model (SIS), having decentralized controllers during malicious attacks. A new Event-Triggered Mechanism (ETM) is designed, having Numenta-Hierarchical Temporal Memory at its back-end. N-HTM seems effective in reducing the average Data Release Rate (70% reduced) and even stabilizes the system by providing anomalous scores for guiding the following input to the providers. Due to the reduction in average DRR, we reduce the power usage of battery supply, disk utilization, and processor computation time. Controller gain and ETM parameters assist in stabilizing the model based on the response of ETM obtained during the presence of attacks. Controller gain and ETM parameters are obtained using stochastic analysis and Lyapunov stability theory. Lastly, we choose a service provider system to check the effectiveness of our proposed model and observe that subsystems stabilize after 2s from the launch time of the last attack. N-HTM based ETM helped in reducing DRR and resource consumption by 70%.

As future work, we will try with other anomaly detectors like Long Short term memory, Bidirectional Long Short-term memory as the base of ETM. The above configuration of the system resembles an edge-computing computation where a network-based communication channel is not required.

References

[1] B. Niu, H. Li, Z. Zhang, J. Li, T. Hayat, and F. E. Alsaadi, "Adaptive neural-network-based dynamic surface control for stochastic interconnected nonlinear nonstrict-feedback systems with dead zone," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1386–1398, 2018. doi: 10.1109/TSMC.2018.2866519

[2] J. Liu, W. Suo, L. Zha, E. Tian, and X. Xie, "Security distributed state estimation for nonlinear networked systems against dos attacks," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 3, pp. 1156–1180, 2020. doi: 10.1002/rnc.4815

[3] D. Ye and T.-Y. Zhang, "Summation detector for false data-injection attack in cyber-physical systems," *IEEE transactions on cybernetics*, vol. 50, no. 6, pp. 2338–2345, 2019. doi: 10.1109/tcyb.2019.2915124

[4] L. An and G.-H. Yang, "Secure state estimation against sparse sensor attacks with adaptive switching mechanism," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2596–2603, 2017. doi: 10.1109/tac.2017.2766759

[5] K. Wang, E. Tian, J. Liu, L. Wei, and D. Yue, "Resilient control of networked control systems under deception attacks: a memory-event-triggered communication scheme," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 4, pp. 1534–1548, 2020. doi: 10.1002/rnc.4837

[6] D. Ding, Z. Wang, D. W. Ho, and G. Wei, "Distributed recursive filtering for stochastic systems under uniform quantizations and deception attacks through sensor networks," *Automatica*, vol. 78, pp. 231–240, 2017. doi: 10.1201/9780429465901-16

[7] J. Xu, Y. Tang, W. Yang, F. Li, and L. Shi, "Event-triggered minimax state estimation with a relative entropy constraint," *Automatica*, vol. 110, p. 108592, 2019. doi: 10.1016/j.automatica.2019.108592

[8] Z. Fei, X. Wang, M. Liu, and J. Yu, "Reliable control for vehicle active suspension systems under event-triggered scheme with frequency range limitation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019. doi: 10.1109/tsmc.2019.2899942

[9] E. Tian, Z. Wang, L. Zou, and D. Yue, "Probabilistic-constrained filtering for a class of nonlinear systems with improved static event-triggered communication," *International Journal of Robust and Nonlinear Control*, vol. 29, no. 5, pp. 1484–1498, 2019. doi: 10.1002/rnc.4447

[10] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.04.070

[11] Z. Gu, J. H. Park, D. Yue, Z.-G. Wu, and X. Xie, "Event-triggered security output feedback control for networked interconnected systems subject to cyber-attacks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020. doi: 10.1109/tsmc.2019.2960115

[12] H. Liang, Y. Zhang, T. Huang, and H. Ma, "Prescribed performance cooperative control for multiagent systems with input quantization," *IEEE Transactions on cybernetics*, vol. 50, no. 5, pp. 1810–1819, 2019. doi: 10.1109/tcyb.2019.2893645

[13] Y. Tang, D. Zhang, D. W. Ho, W. Yang, and B. Wang, "Event-based tracking control of mobile robot with denial-of-service attacks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 9, pp. 3300–3310, 2018. doi: 10.1109/tsmc.2018.2875793

[14] P. Kishore, S. K. Barisal, and D. P. Mohapatra, "An incremental malware detection model for meta-feature api and system call sequence," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020. doi: 10.15439/2020f73 pp. 629–638.

Young Researchers Workshop on Artificial Intelligence and Cybersecurity

WORKSHOP is intended for young researchers only—Engineering or MSc students—who intend to design computer systems, in the domains of, broadly understood, Artificial Intelligence or Cybersecurity.

The presentations are aimed to be practical-application-oriented—and should consist of a short theoretical introduction, accompanied by a demonstration of a system prototype. The acceptance of the participants will be based on the short description of ideas that are going to be presented/discussed, submitted by July 16th 2021 (UTC+12). The length of the description should be between 500 and 800 words long (Extended Abstract of about 2 pages). It should be formatted according to the instructions for the authors, found at the For Authors page within the FedCSIS conference portal. Presentations will be divided into topical sessions. Awards will be given to the most valuable presentation in each session. Selected presentations will be shown (in a shortened version) to the general audience of the FedCSIS conference.

TOPICS

Invited are proposals covering industrial applications and academic research that include, but are not limited to the following topics:

- Artificial Intelligence:
  - Natural Language Processing
  - Language Models
  - Machine Translation
  - Computer Vision
  - Data Mining and Knowledge Discovery
  - Neural Networks and Deep Learning
  - Reinforcement Learning
  - Web Mining and Social Networks
  - Evolutionary Algorithms and Evolutionary Computation
- Cybersecurity:
  - Cryptography and cryptanalysis
  - Digital right management and data protection
  - Threats and countermeasures for cybercrimes
  - Cyber and physical security infrastructures
  - Steganography and watermarking
  - Digital forensics and crime science
  - Misuse and intrusion detection
  - Cloud and big data security
  - Computer network security

TECHNICAL SESSION CHAIR

- **Jassem, Krzysztof,** Adam Mickiewicz University, Poznań, Poland

DEPUTY CHAIR

- **Piłka, Tomasz,** Adam Mickiewicz University, Poznań, Poland

Preproceedings of the Federated Conference on Computer Science and Information Systems pp. 221–222
DOI: 10.15439/2021F146
ISSN 2300-5963 ACSIS, Vol. 26

Speech sound detection employing deep learning

Cezary Polak, Jakub Mańkowski, Wiktor Uciński,
Patryk Schramka, Mikołaj Mysiakowski
Gdańsk University of Technology
Faculty of Electronics Telecommunication and Informatics
Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: {s165516, s172466, s160299,
s168827, s165771}@student.pg.edu.pl

Adam Kurowski
Gdańsk University of Technology
Faculty of Electronics Telecommunication and Informatics
Multimedia Systems Department
Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: adakurow@pg.edu.pl

***Abstract*—The primary way of communication between people is speech, both in the form of everyday conversation and speech signal transmitted and recorded in numerous ways. The latter example is especially important in the modern days of the global SARS-CoV-2 pandemic when it is often not possible to meet with people and talk with them in person. Streaming, VoIP calls, live podcasts are just some of the many applications that have seen a significant increase in usage due to the necessity of social distancing. In our paper, we provide a method to design, develop, and test the deep learning-based algorithm capable of performing voice activity detection in a manner better than other benchmark solutions like the WebRTC VAD algorithm, which is an industry standard based mainly on a classic approach to speech signal processing.**

I. Introduction

VOICE transmission-based techniques are constantly being improved to enable the broadcast of the human voice with the highest quality possible while reducing the demand for transmission bandwidth. In mobile networks, despite multimedia content being a more and more prominent part of transmitted data, voice transmission is still a crucial and basic functionality. To reduce the extensive occupation of the transmission bandwidth, speech detection methods have been employed in the above-mentioned applications to detect and transmit only the speech-containing part of the conversation, which leads to a decrease in bandwidth use. This class of algorithms is called voice activity detectors (VADs). Despite the significant development of techniques related to deep learning, the task is often performed by relatively simple heuristic algorithms. In literature, it is possible to find examples of VAD applications for which authors directly stress, that the choice of the VAD used for carrying out e.g. the speech quality enhancement task directly determines how good the output of such enhancement algorithms is [1].

II. Data acquisition

The first step to design and develop a voice activity detection algorithm employing any machine learning technique is to gather the database containing speech recordings and interfering signals. In our case, a signal obtained from an Internet podcast was used. It was an excerpt of an interview with president Andrzej Duda carried out by Karol Paciorek [2]. The length of the interview is 1 hour 16 minutes and 48 seconds. The recording was downsampled to the sampling rate of 16 kHz to reduce the required memory and processing power of the VAD algorithm. Additionally, procedures for introducing additive white Gaussian noise (AWGN) were designed. They were designed in such a way, that an arbitrary value of signal-to-noise ratio (SNR) can be obtained. Also, an additional recording containing cocktail party type of noise was also obtained from the Freesound online sound archive [3]. Such choice of interfering signals makes it possible to represent a wide range of real-world cases that may present difficulties for the algorithm under test in performing the voice activity detection task.

Next, recordings were annotated. In the process, all parts of the recording that contain speech fragments were marked with the appropriate label by the authors. Annotation was performed for signal frames of 200 ms as a human can easily hear if such frame contains speech. For the VAD algorithm, each of 200 ms frames was split into 20 ms frames, as this is one of the typical lengths for which VAD algorithms operate. Each of 10 short 20 ms frames derived from the single 200 ms long frame had the same label as the long, 200 ms one. Such a database is then used to generate test recordings containing combinations of speech signals and AWGN with varying SNR levels and ones contaminated by the cocktail party-type noise.

The data collected in the aforementioned process were intended to be used as the input to the convolutional neural network (CNN), and therefore they had to be parameterized. For each frame, an MFCC-gram was calculated. It was calculated with the FFT frame length of 32 points, and overlap factor equal to 0.75, a number of MFCC coefficients was set to 10. Such processing resulted in obtaining an MFCC-gram matrix having shape of 37 parameters x 10 parameters associated with every 20 ms long frame. For calculation, a librosa Python library was used. [4]. The interview used as the data source for our study was recorded with a studio-grade microphone, therefore we considered it to be not contaminated by any significant amount of noise.

III. The experiment

Both the clean and noise-contaminated audio frames were processed by two algorithms. Namely, the reference algorithm which in our case was the WebRTC VAD algorithm [5], and

TABLE I
ACCURACIES OF TWO TESTED VAD ALGORITHMS FOR THE CASE OF INPUT SIGNAL CONTAINING NO NOISE, AND FOR THE INPUT SIGNAL CONTAINING THE AWGN OR THE COCKTAIL-PARTY NOISE.

VAD algorithm	noise type	SNR [dB]						
		15	10	5	0	-5	-10	-15
WebRTC	**no noise**	0.937						
	AWGN	0.936	0.938	0.940	0.943	0.943	0.940	0.925
	cocktail	0.936	0.937	0.937	0.936	0.928	0.919	0.882
CNN-based	**no noise**	0.964						
	AWGN	0.946	0.919	0.842	0.696	0.606	0.565	0.548
	cocktail	0.925	0.850	0.706	0.606	0.568	0.555	0.549

the algorithm designed by the authors which was based on convolutional neural networks (CNNs). Processing employing neural networks was implemented with the use of TensorFlow Python library [6]. The structure of a CNN used as a VAD algorithm was as follows:

1) a convolutional layer containing 32 channels with (2,2) filters and ReLu activation function, followed by a (2,2) max pooling operation with a stride parameter set to (2,2), and a batch normalization layer,
2) a group of layers identical to 1),
3) a group of layers identical to 1),
4) a flattening layer,
5) a dense layer with 64 neurons with ReLu activation function,
6) a dropout layer with dropout coefficient of 0.3,
7) an output dense layer containing 2 neurons (as output is encoded in a one-hot manner), having softmax activation function.

IV. RESULTS

Input dataset consisted of 200190 audio frames, 53.88% (107860 frames) of the examples present in the dataset were associated with the speech signal presence, 46.12% (92330 frames) were associated with a so-called silence by which we mean frames not containing any speech signal. The input dataset was split into training (60% of examples), validation (20% of examples), and test (20% of examples) subsets. Data were divided in a stratified manner, so proportions of speech and silence in each of them were similar to ones in the original dataset. The neural network was trained for 100 epochs. The final accuracy achieved by the algorithm for the training set was 0.983, for the validation a dataset accuracy of 0.964 was achieved.

Performances of both the CNN-based, and the WebRTC VAD were evaluated on speech signals which were either noise-free recordings obtained from the podcast, or the audio fragments contaminated by additional noise. The CNN-based VAD was evaluated only on examples from the test dataset, as the test dataset was the only piece of data not used in the process of training. WebRTC VAD was evaluated on the whole dataset, as no training was necessary in its case. Two types of noise were used to contaminate the input signal, namely the AWGN, and the cocktail-party noise. The signal-to-noise ratio (SNR) of the noise signal added to the inputs of both tested algorithms was varied from 15 dB up to -15 dB with a decrement of 5 dB. Results of all tests carried out in our experiment are shown in Tab. I

V. CONCLUSION

Results obtained with the use of a CNN-based VAD algorithm are promising for conditions of nonexistent or low-amplitude noise (SNR = 15 dB, AWGN noise) if compared to the WebRTC VAD algorithm. Therefore, the algorithm proposed in our paper may be used for automatic labeling of signals for research purposes, as it is capable of obtaining better accuracies than the WebRTC VAD used as a baseline for our study. On the other hand, even for the SNR of 10 dB the performance of CNN-based algorithms drops below the baseline results and degrades in a significantly more pronounced manner for SNRs lower than 10 dB. For the cocktail-party type of noise, this degradation is even worse than for the AWGN. Therefore, the use of simple CNN-based VADs is not encouraged in the case of noisy environments. Possible countermeasures are, e.g. use of noise-contaminated audio frames for the training of the algorithm which may be an interesting future work to be carried out in the case of our research. We also plan to test our approach on signals other than on-line interviews, such as conversation of three or more people with fragments of simultaneous speech coming from two or more speakers.

REFERENCES

[1] H. Haneche, B. Boudraa, and A. Ouahabi, "A new way to enhance speech signal based on compressed sensing," *Measurement*, vol. 151, p. 107117, 2020. doi: https://doi.org/10.1016/j.measurement.2019.107117. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224119309832

[2] K. Paciorek. Andrzej Duda o: LGBT, TVP, koronawirusie, głosach po Bosaku i o szansach w starciu z Trzaskowskim (in Polish). Youtube (Imponderabilia channel). [Online]. Available: https://www.youtube.com/watch?v=Izxj72bg4A4

[3] Freesound. Party Sounds recording from the online royalty free recordings archive. [Online]. Available: https://freesound.org/people/FreqMan/sounds/23153/

[4] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[5] GitHub. Python interface to the WebRTC voice activity detector. [Online]. Available: https://github.com/wiseman/py-webrtcvad

[6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

Preproceedings of the Federated Conference on Computer Science and Information Systems pp. 223–226
DOI: 10.15439/2021F145
ISSN 2300-5963 ACSIS, Vol. 26

Endoscopy Image Retrieval by Mixer Multi-Layer Perceptron

Quoc-Huy Trinh
University of Science, VNU-HCM, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
Email: 20120013@student.hcmus.edu.vn

Minh-Van Nguyen
University of Science, VNU-HCM, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
Email: 20127094@student.hcmus.edu.vn

***Abstract*—In Computer Vision, the Image Retrieval task is one of the interests of researchers, particularly medical image retrieval and endoscopy images. With the development of the Convolution Neural Network and Vision Transformer Technique, there are many proposals for using these techniques to make Image Retrieval Task and achieve a competitive result. In this paper, we propose a method that using Mixer Multi-Layer Perceptron architecture (Mixer-MLP) to build an Image Retrieval System with Medical images, particularly Endoscopic Images. This System base on the Classification process of Mixer-MLP architecture to generate vector representation for similarity calculation. The research result achieves competitively with efficient training time.**

I. Introduction

IMAGE RETRIEVAL TASK is the topic that using images to Retrieve the Image in the database[1]. In the Medical system, with the widespread in using digital imaging and storing, it causes difficult to query these large databases, this is the reason why there are high necessaries to use a content-based Image Retrieval system.[7]

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images[17]. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words[18]. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation.

In recent years, the number of people that have been affected by colorectal cancer (CLC) is increasing. It is also on a third of the world for many years[12]. However, can we diagnose and prevent CLC is a crucial issue for the health organization[15]. Some studies illustrate that almost 95% of CLC is from the adenomatous polyp. The resection of Colorectal adenomatous polyps can reduce the CLC. On the other hand, the best way to deal with CLC is early diagnosis and have straight treatment. Nowadays, with the growth of the number of people that have CLC, digital imaging technique is applied to store the endoscopic images[13]. However, the doctor finds difficulty in querying the database because of the number of images in the database.[3]

Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City

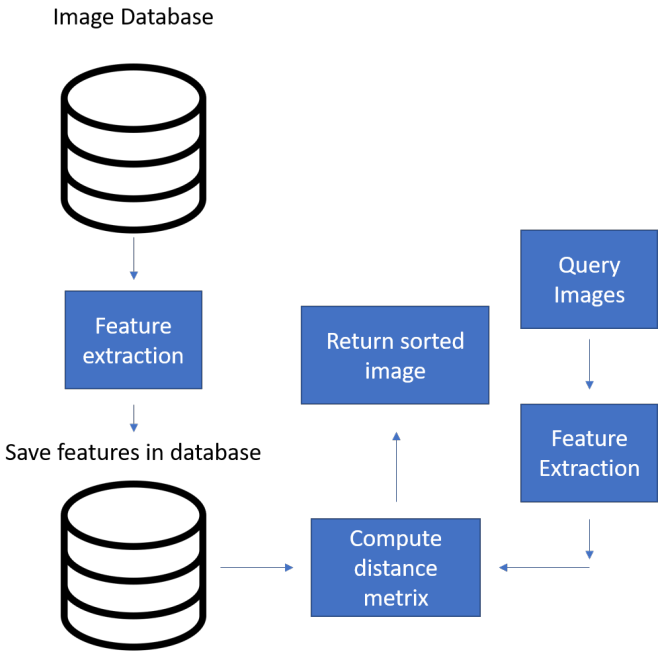


Fig. 1. Image Retrieval system

The figure above explain the image retrieval system. The method that we propose depend on this model.

Due to the development of Convolution Neural Network (CNN), there are many deep architectures applied in the feature-vector generating process such as ResNet, DenseNet and EfficientNet, etc. [5]. In early 2021, there is a method of using Mixer Multi-Layer Perceptron (Mixer-MLP) is proposed to classify images. [2]

In this paper, we build an Image Retrieval system on Endoscopic Images with a training process on Mixer-MLP architecture and approach a method to generate image vector representation from the model trained before.

Different from the previous Mixer-MLP architecture that is proposed in May 2021. We propose an architecture that base on the previous architecture, by cutting the classify layers, we add an embedding layer to generate the feature vector to represent the images feature. This architecture has the merit that it is based exclusively on multi-layer perceptrons.

II. Related Work

In our methods, we use some previous work to build our system. To help our research briefly, we review some past research to our best knowledge.

A. Mixer-MLP

In half of 2021, Googe AI has published an architecture that uses a variety of Multi-Layer Perceptrons (MLPs). There are two types of layers in this architecture: MLPs apply independently to image patches, MLPs apply across patches. This model achieves the competitive score on the image classification benchmark. This result is quite positive when compared with state-of-the-art models.[2]

B. Content-based Image Retrieval

Content-based Image Retrieval is a well-studied problem in computer vision, with retrieval problems generally divided into two groups: category-level retrieval and instance-level retrieval[8]. Given a query image of the Sydney Harbour Bridge, for instance, category-level retrieval aims to find any bridge in a given dataset of images, whilst instance-level retrieval must find the Sydney Harbour bridge to be considered a match.[6]

C. Similarity Search

To have accurate retrieval, they propose diversity similarity methods such as cosine similarity, Euclid distance, Manhattan distance. In this paper, we use the traditional Similarity search method to retrieve the images in the database collections.[9]

D. Image classification by Deep Learning models

In recent years, there is a variety of Deep Convolution Neural Network architecture to classify images, almost the result of that architecture achieve competitive. In later years, methods use state-of-the-art architecture while they combine Convolution Neural Network with Transformer to get the better result[14]. However, we can have a new approach by using Multi-layer Perceptrons to each patch of the images instead of using Convolution layers, lead to the method of Mixer-MLP. [2]

III. Dataset

To evaluate our system, we experiment with Kvasir Dataset. The Kvasir Dataset is collected using endoscopic equipment at Vestre Viken Health Trust (VV) in Norway. The VV consists of 4 hospitals and provides health care to 470.000 people. One of these hospitals (the Bærum Hospital) has a large gastroenterology department from where training data have been collected and will be provided, making the dataset larger in the future. Furthermore, the images are carefully annotated by one or more medical experts from VV and the Cancer Registry of Norway (CRN).[3]
The dataset consists of 80000 images in 10 folds for cross-validation in the training and evaluating process. 80000 images are split into eight classes: dyed-lifted-polyps, dyed-resection-margins, esophagitis, normal-cecum, normal-pylorus, normal-z-line, polyps and ulcerative-colitis.[3]

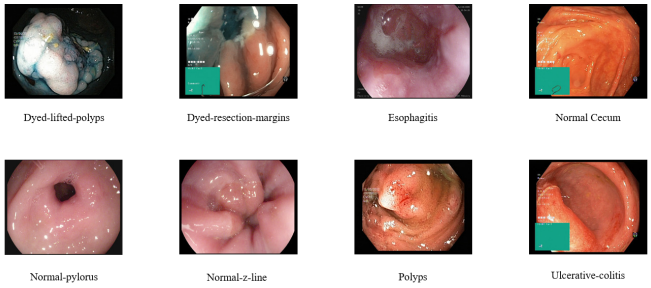


Fig. 2. The sample of Kvasir Dataset.

In our method, we propose a system with two parts: Collection generation and Retrieval. Meanwhile, we also approach to training classification model for the Kvasir dataset by using Mixer-MLP architecture.

IV. The Proposed Approach

A. Training Model with Mixers-MLP

1) Data Preparation: After loading data, we resize all the images to the size (150,150), then we split the dataset into the training set and validation set in the ratio of 0.75:0.25. After resizing and splitting the validation set, we rescale the data pixel down to be in the range [-1,1] by divide by 127.5.

2) Environment Setup: We prepare the training process on GPU Nvidia P100. The data is load with the batch size is 32, and preprocessing after loading.
With model initialization, we set the parameter for initializing the model in the following table:

B. Generating vectors for data representation

To generating the vector from the database and the vector for the query, we propose to replace the classification layers of Mixer-MLP with a Dense layer that enables to generate a vector for representation, then we will flatten the vector output to one dimension to facilitate similarity calculation.

TABLE I
Mixer-MLP initialization.

Parameter	Value
Number of MLPs block	8
Number of Patches	5
Input shape	(150,150,3)
Channel Mixing Unit	256
Token Mixing Unit	2048
Projection Unit	512
Patch size	5
Learning Rate	0.001
Optimizer	Adam

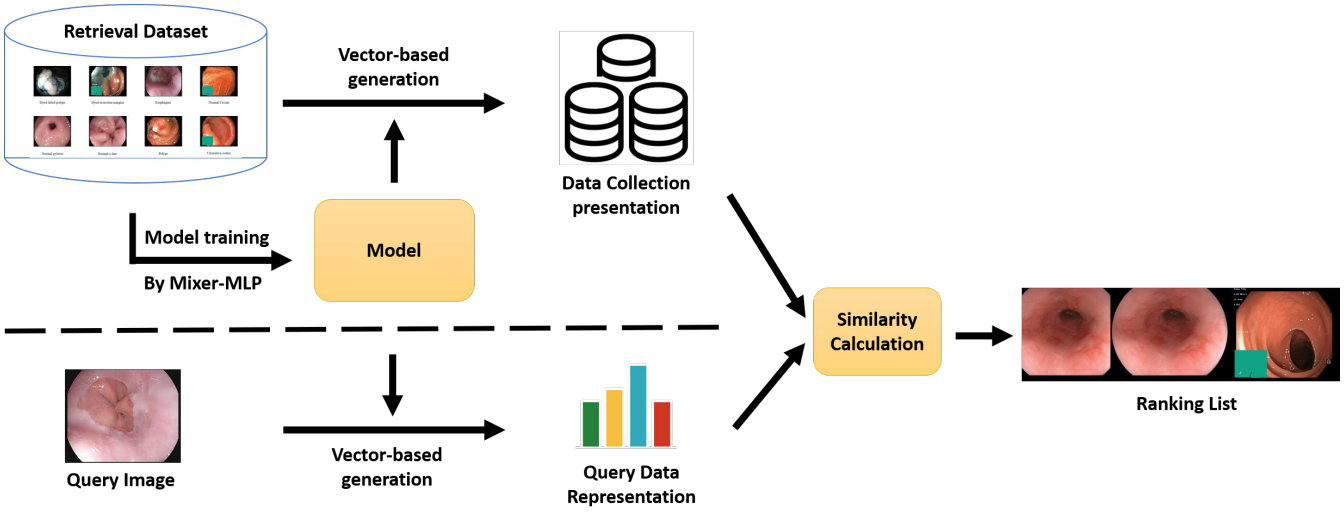


Fig. 3. Retrieval system by using Mixer MLP classification

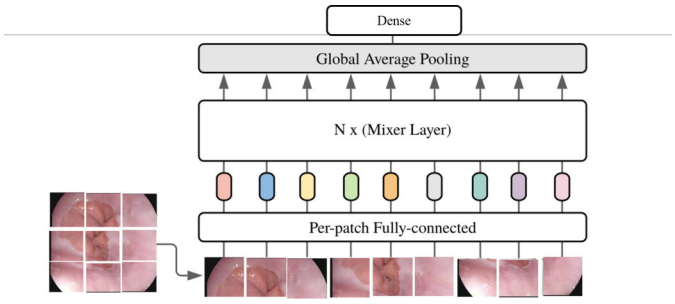


Fig. 4. Architecture of vector extraction for data representation

C. Similarity Calculation

To calculate between the query vector and database collection vector, we use cosine similarity with the following formula:[8]

$$Similarity(A,B) = cos(\phi) = \frac{\vec{A}\times\vec{B}}{|\vec{A}|\times|\vec{B}|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where:
A: is the vector A that has n elements
B: is the vector B that has n elements
The value of similarity is in the range from 0 to 1, depend on the value of cosine value. This can help cost efficient in computing.
After calculate the distance between the query vector and database collection vectors, the result is sorted to give back the rank of the result.

D. Data Augmentation

Data Augmentation is vital in data preparation process. Data Augmentation improve the number of data by adding sligtly modified copy of already exist data or newly created synthetic data from existing data to decrease the probability of the Overfitting problem, we use augmentation to generate the data randomly by random flip images and random rotation with an index of 0.2..[16]

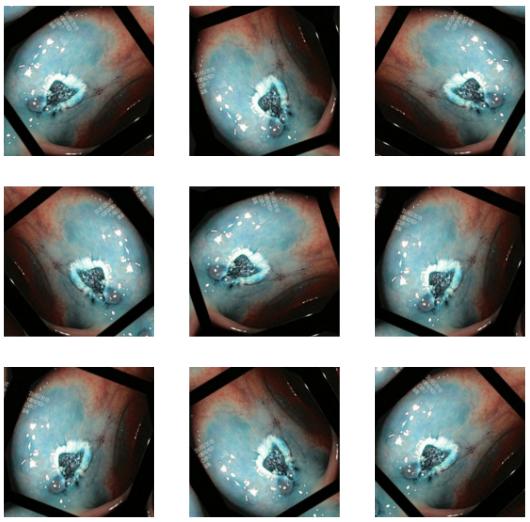


Fig. 5. Data after augmentation

V. Evaluation and Discussion

A. Model training

After 10 epochs, we got the following loss. The performance of the model on the Kvasir dataset is quite competitive, this can lead to the vector generation step can get a competitive score.

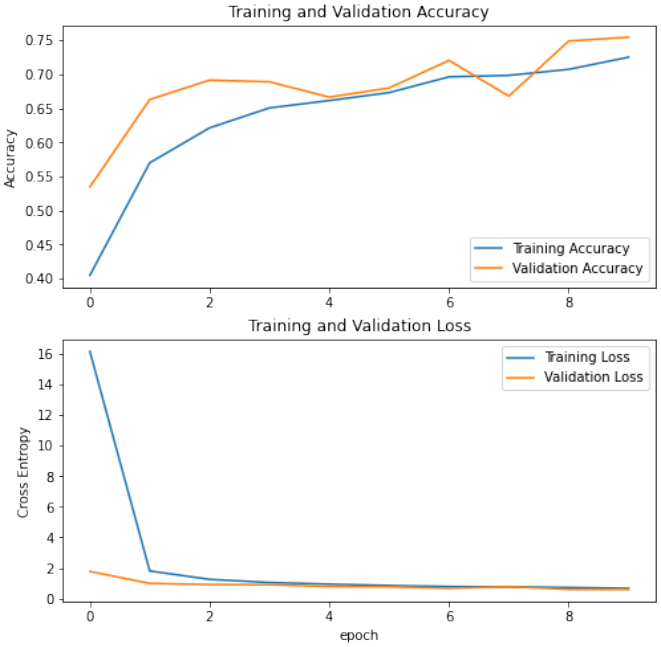


Fig. 6. Retrieval system by using Mixer MLP classification

The result on the test set achieves 0.8251 on the test set with 4000 images. This result can ensure the vector generation process of the system.

B. Discussion

Although our method gets a competitive score, there are some drawbacks in our methods: the training time gets long with 458ms/step, we can custom layers in the architecture to accelerate the computing cost. We can get more layers or can ensemble more backbones to achieve higher results.
The model can open a new approach to the Image Retrieval Task. However, there we have to deal with drawbacks by using this proposal:
- We have to improve the accuracy of the Mixer-MLP model to improve the quality of the vector generation process.
- We can improve the time for generating the vector representation for collection and query.

VI. Conclusion

In general, our research proposes the method that using Mixer Multilayers Perceptron to extract the feature of endoscopic images for image retrieval system. Our method achieves a competitive result on content-based retrieval. By using the classification model of Mixer-MLP and cutting the classification layer, we can generate a vector to represent the data feature. That will enable a new approach in the Content-based Image Retrieval task. Moreover, our method inspires the new approach to extract the feature for retrieval task instead of using the previous methods such as Deep Convolution Neural Network architecture or Vision Transformer.

VII. Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

[1] Fan Yang, Ryota Hinami, Yusuke Matsui, Steven Ly, Shin'ichi Satoh, Efficient Image Retrieval via Decoupling Diffusion into Online and Offline Processing.
[2] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy, Mixer-MLP: An all-MLP architecture for Vision.
[3] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, Pål Halvorsen, Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection, In MMSys'17 Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS), Pages 164-169 Taipei, Taiwan, June 20-23, 2017.
[4] Filip Radenovic Giorgos Tolias Ond ´ ˇrej Chum, Fine-tuning CNN Image Retrieval with No Human Annotation
[5] Jinyun Lu , Image Retrieval Based on ResNet and KSH,Advances in Intelligent Systems Research, volume 147.
[6] Huiyi Hu, Wenfang Zheng, Xu Zhang, Xinsen Zhang, Jiquan Liu, Weiling Hu, Huilong Duan, Jianmin Si, Content-based gastric image retrieval using convolutional neural networks, Imaging Systems and Technology, pages 439-449.
[7] Sun Q, Yang Y, Sun J, Yang Z, Zhang J, eds. Using deep learning for content-based medical image retrieval. Paper presented at: Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications; 2017: International Society for Optics and Photonics.
[8] JC Felipe, AJ Traina, C Traina, eds. Retrieval by content of medical images using texture for tissue identification. Paper presented at: 16th IEEE Symposium Computer-Based Medical Systems, 2003 Proceedings; 2003: IEEE.
[9] A Rashno, S Sadri, eds. Content-based image retrieval with color and texture features in neutrosophic domain. Paper presented at: 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA); 2017 19–20 2017.
[10] Cai Y, Li Y, Qiu C, Ma J, Gao X. Medical image retrieval based on convolutional neural network and supervised hashing. IEEE Access. 2019; 7: 51877- 51885.
[11] Hasan MM, Islam N, Rahman MM. Gastrointestinal polyp detection through a fusion of contourlet transform and neural features. J King Saud Univ-Comput Info Sci. 2020.
[12] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in Proceedings of the 24th International Conference on World Wide Web. ACM, 2015, pp. 1067–1077.
[13] F Sommen, S Zinger, EJ Schoon, eds. Computer-Aided Detection of Early Cancer in the Esophagus Using HD Endoscopy Images. Medical Imaging 2013: Computer-Aided Diagnosis. Vol. 8670. Florida: International Society for Optics and Photonics; 2013.
[14] Yu D, Seltzer ML, Li J, Huang J-T, Seide F. Feature learning in deep neural networks-studies on speech recognition tasks. arXiv. 2013;13013605.
[15] Nini Rao, Hongxiu Jiang, Chengsi Luo: Review on the Applications of Deep Learning in the Analysis of Gastrointestinal Endoscopy Images., Article in IEEE Access - September 2019
[16] Chung Y-A, Weng W-H. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. arXiv. 2017;171108490.
[17] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, Andrew Zisserman, Thinking Fast and Slow: Efficient Text-to-Visual Retrieval With Transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9826-9836
[18] JohnA.HForrestN.D.CFinlaysonD.J.CShearman, ENDOSCOPY IN GASTROINTESTINAL BLEEDING, the Lancet, Volume 304, Issue 7877, 17 August 1974, Pages 394-397.

Preproceedings of the Federated Conference on Computer Science and Information Systems pp. 227–228
DOI: 10.15439/2021F142
ISSN 2300-5963 ACSIS, Vol. 26

Applying Machine Translation Methods in the Problem of Automatic Text Correction

Wojciech Jarmosz
Adam Mickiewicz University
in Poznań
ul. Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
Email: wojjar3@st.amu.edu.pl

***Abstract*—This document describes the problem of automatic text corrections. The author presents a classification of errors, a process of correcting texts and a proof of concept as a containerized version of machine translation system - MSuedin.**

I. Introduction

A. Groups of errors

TEXT CORRECTION is a broad topic, with many approaches leading to slightly different results. According to the classification introduced in Naber[1], we can divide errors into four main categories:

- **spelling errors** - words that don't belong to the language. Analysis doesn't require knowledge about the context.
- **grammatical errors** - known also as real-word errors, which break the grammatical rules of the language. Very often, correction of this type of error requires analysis of context.
- **stylistic errors** - they don't break grammatical or morphological rules of the language but include repetitions, colloquialisms, and overcomplicated structures.
- **semantic errors** - information not related to facts and knowledge about the world. Such errors are very difficult to detect and correct. This type of error demands building some sort of knowledge model.

B. Text correction process

Building error correction systems requires a process. According to Kukich[2], it can include:

- error detection,
- generating candidates to correction,
- rating candidates,
- choosing candidate with the highest score.

C. Approaches to automatic text correction

Over the years, approaches to creating text correction systems have evolved. Rule-based architecture, used for example in MS Word, is expensive, complex, and requires linguists involved in defining such rules. On the other hand, we have statistical approaches and classification methods, used more commonly in contemporary tools. Each of these approaches is good in some specific areas, such as selecting the best candidate from the confusion set, correcting prepositions or articles. Nowadays, the most effective and widely used method is based on machine-translation techniques, including the usage of the transformers models. The author of this article focuses on the implementation of MarianNMT[3] transformer-based system - MSuedin[4], and prepares a docker image for it, allowing the user to easily run such a program with different parameters, on multiple GPUs without worrying about system dependencies and framework integrations.

II. Proof of concept

Author prepared a docker image for the MSuedin system using the docker-nvidia, CUDA toolkit, virtualenv and python3. The container was built and run on Linux Ubuntu 20.04 with installed CUDA Toolkit 11.3 and 19GB GPU. Containerization makes the program independent from the host operating system and other dependencies. The technique is called "write once, run anywhere". It helps our grammar correction system to scale better and start up faster. Instead of working on system configuration and program installation, machine learning engineers can now focus on providing high-quality training data.

III. Potential usage

The project can be used to support the translator's job by reducing time spent on correcting errors manually (it can takes many hours) and form a good basis to create automatic text correction systems with any language corpus due to virtualization and containerization.

IV. Research status

MSuedin[5] - one of the winning GEC systems in the BEA 2019 shared task is an example of a transformer-based machine translation system that corrects English grammatical errors. On the other hand, there are no such systems for niche languages like Polish, German or Finnish, which will provide satisfactory results in correcting text errors.

V. Research goals

The main goal of the research is to prepare a containerized version of the machine translation system for text error correction, which can be run on multiple graphic cards. Another goal is to train models on the language corpora less popular than English.

VI. Future experiments

The author will build a corpus, containing pairs of sentences with and without the error. To increase the number of samples, there may be also a need to generating synthetic data using some sort of thesaurus or confusion sets. After that, the author will divide the dataset into a train and test set in a proportion of 8:2. Using available tools in the MarianNMT framework, the model will be trained (there will be a need for multiple GPU usage). The model will be loaded in a containerized version of MSuedin and evaluated using F-score and accuracy metrics.

VII. Conclusion

A successful experiment includes preparing a model and running it in the containerized version of MSuedin. Error correction of the text is a very interesting topic, which has broad and measurable usage in real-world scenarios. I believe that time-consuming tasks such as grammar error correction can be successfully automated.

References

[1] D. Naber, A rule-based style and grammar checker, GRIN Verlag 2003.
[2] K. Kukich, Techniques for automatically correcting words in text, ACM Computing Surveys 1992
[3] https://marian-nmt.github.io
[4] https://github.com/grammatical/pretraining-bea2019
[5] R. Grundkiewicz, M. Junczys-Dowmunt, K. Heafield, Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data, BEA 2019

Preproceedings of the Federated Conference on Computer Science and Information Systems pp. 229–230
DOI: 10.15439/2021F143
ISSN 2300-5963 ACSIS, Vol. 26

Voice Controlled Games – The approach and challenges of implementing speech recognition and voice control in games

Dominik Strzałko
Adam Mickiewicz University in Poznań
Email: dominikstrzalko@gmail.com

***Abstract*—The subject of voice controlled games is quite underrated and exciting. Even though there are not that many papers focusing on that particular subject, we can find many papers describing both, concept of voice recognition, and controlling the game using natural language. The most illustrative example of usage of that kind of control are games in VR/AR/MR and games for people with disabilities. Moreover, almost every game can benefit from voice commands e.g. for controlling the user interface or units in strategy game. Therefore, voice controlled games are an interesting and innovative concept for game designers and developers.**

I. Introduction

THE SUBJECT of voice controlled games is quite underrated and exciting. Even though there are not that many papers focusing on that particular subject, we can find many papers describing both, concept of voice recognition, and controlling the game using natural language (e.g. a popular game "Façade" [1]). One might argue that speech control in games are even somehow part of the future of gaming industry. The most illustrative example of usage of that kind of control are games in Virtual Reality (VR)/Augmented Reality (AR)/Mixed Reality (MR) (that are getting more and more popular [2]) and games for people with disabilities. Moreover, almost every game can benefit from voice commands e.g. for controlling the user interface or units in strategy game. Therefore, voice controlled games are an interesting and innovative concept for game designers and developers . This paper will focus on different aspects of this large and broad topic.

II. The Paper

The whole paper will be built from 4 distinct chapters.

A. Beginning

The first part to the paper will outline the short history of voice commands in the game industry and show how this feature was used across time. From the first games that used this technology in some way (like "Command: Aces of the Deep" [3]), to the most recent productions on the market (like games built with "Just AI" tools [4]).

B. Components

The next part is going to focus on describing the main components of the speech recognition: Natural Language Understanding (NLU) and Speech-to-Text (STT). It will also present various tools and libraries that are commonly used to implement those modules into applications. As for NLU, the description will concentrate on LUIS by Microsoft Azure, Dialogflow by Google and author's own module created with the help of spaCy and Keras (this solution is still in the conceptual phase). As For the Speech-to-Text module, the paper will focus only on already created services like STT by Microsoft Azure, IBM and Google. It is important to mention that all selected technologies are compatible and working well with different commercial game engines. Regarding game engines, the paper will also shortly describe those engines which will later be used to implement the modules.

C. Implementation

The plans for the implementation process will take up the succeeding and largest part of this paper. A step by step explanation will be given to the process of implementing the different tools/technologies into those game engines and also comparing the accuracy and other metrics of the Speech-to-Text and Natural Language Understanding tools. It will be very important to determine the most accurate and easy to implement tools, also for the sake of future game developers that will be interested in implementing those tools into their games. The most ambitious version of the implementation plans involve the development of a plugin that will be available under an open-source license and will be easily integrated to the games in the engine of author's choice (Where unity engine is the most probable option).

D. Demo Games

The final part will feature author's own demo games that will be created especially for the project. The process of building them will be present in the previous chapter but the last part of the paper will sum up the technologies used for each game and describe how the voice control influences them. The planned demo games include with high probability: a VR Game, an AR Game, a game dedicated for people with disabilities, a normal game with both standard control system and voice control system, and potentially, a game for visually impaired people (idea inspired by the game "Ptolem's Singing Catacombs" by Narayana Walters, aka Mizizizi [5]), were player uses only his voice and hearing to navigate and progress through the game. The genres of the games will vary, to

show that developer can create not enchant a simple puzzle game with voice commands, but also a more complicated and interesting products for different demographics.

III. Conclusion

To sum up, this paper is aimed at promoting the voice control among other game developers and designers, and helping them create more immersive games for everyone. This involves the idea of helping people with disabilities enjoy one of the largest source of entertainment, which are: video games.

References

[1] Michael Mateas and Andrew Stern, "Façade", https://www.playablstudios.com/facade
[2] Virtual Reality (VR) - statistics & facts, https://www.statista.com/topics/2532/virtual-reality-vr
[3] Teemu Kiiski, "Voice Games: The History of Voice Interaction in Digital Games"
[4] Just-AI, https://just-ai.com/en/voice-games
[5] Narayana Walters, "Ptolem's Singing Catacombs", https://nartier.itch.io/ptolems-singing-catacombs

Preproceedings of the Federated Conference on Computer Science and Information Systems p. 259
DOI: 10.15439/2021F144
ISSN 2300-5963 ACSIS, Vol. 26

Implementation of the game model of the Polish Ekstraklasa team using machine learning techniques

Jakub Pogodziński
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Uniwersytetu Poznańskiego 4 Street, 61-614 Poznań, Poland
Email: jakpog1@st.amu.edu.pl

***Abstract*—The aim of the thesis is to create a model defining the style of play of a team playing in the Polish Ekstraklasa**

I. Introduction

THE AIM of the thesis is to create a model defining the style of play of a team playing in the Polish Ekstraklasa. The limitation to the highest Polish league class is dictated by the differences in the style of play depending on the league. The model is to be created on the basis of data about the team's game (e.g. XGoals, xAssists, distance run). The data will mainly come from the StatsBomb portal. To build the model, supervised and unsupervised learning techniques will be used and compared to find the relationship between the team's statistics and the determination of its playing style. Supervised learning requires expert knowledge to determine the relationship that is being sought. Unsupervised learning can allow to find new features that are not yet named and the relationship between them has not yet been noticed.

II. Potential applications

Building a model of the team's game will allow you to identify the strengths and weaknesses of the team, to prepare tactically for a particular opponent and to find players who match the style of play of the team.

III. State of research

In the article "Introducing Similar Team Search" [1]. the possibility to search for similar clubs based on statistics is shown. There is only the possibility to compare clubs. There are no other correlations between them. In the article Balla M. Which teams are the most similar to Bayern Munich? [2]. (https://totalfootballanalysis.com/article/teams-similar-bayern-munich) in the magazine Total Football Analysis there is a comparison of clubs that are the most similar to Bayern Munich. The author writes that 25 metrics from WyScout were taken for analysis, without specifying which metrics, and then using the K-means algorithm the data were grouped. The distance between the sets was then counted. The article by Marton Ball throws an interesting perspective on comparing ensembles. Unfortunately, it is not given in sufficient detail how the conclusions presented were reached.

IV. Demonstration of the need for own research

The articles presented above are very poor and only present a concrete way of comparing the teams' playing styles. What is missing is a more universal approach, a model that will describe the differences and similarities between teams in a more global way.

V. Discuss the planned experiment - e.g. what the training data will be, what the test data will be, what methods / algorithms will be used

The data will come from the StatsBomb portal. These will be individual player metrics and aggregate metrics for the whole team in each league game over the past few seasons. They will be split into a training and testing set. The data will be grouped and a comparison of clustering algorithms (Ward, K-means) and principal component analysis algorithms will take place. The data will be compared with supervised learning algorithms, where labels will be assigned to the data based on the expert knowledge gathered.

VI. Summary - what we expect to gain from the project and why our experiment is worth doing

In conclusion, football is increasingly based on statistics and analysis. New tools allow a better understanding of the game. Better understanding translates into more effective work on tactics and better sporting, financial and social (fan satisfaction) results. This project aims to create a model to determine a team's playing pattern, highlighting its strengths and weaknesses. Football teams will be able to adapt better to matches, thus increasing their chances of achieving victory both from a one-match perspective and by winning more matches, they will increase their chances of achieving the best possible final result.

References

[1] Statsbomb, "Introducing Similar Team Search" *https://statsbomb.com/2020/12/introducing-similar-team-search*
[2] Balla M., "Which teams are the most similar to Bayern Munich?" *https://totalfootballanalysis.com/article/teams-similar-bayern-munich,*

Preproceedings of the Federated Conference on Computer Science and Information Systems pp. 233–234
DOI: 10.15439/2021F147
ISSN 2300-5963 ACSIS, Vol. 26

Training of neural machine translation model to apply terminology constraints for language with robust inflection.

Jakub Konieczny
Adam Mickiewicz University
ul. Wieniawskiego 1, 61-712 Poznań, Poland
Email: jakkon6@st.amu.edu.pl

***Abstract*—The goal of this study is to explore the transformer's capability of domain translation into a morphologically rich language. Satisfactory translation into Polish requires inflection by tense, number, and person, taking into account six declination cases. The ideal outcome of this study would be to prove that the method proposed by Dinu is capable of training the transformer to translate English to Polish in domain-specific scenarios. Achieving metrics similar to Nowakowski would result in a "zero-shot" translator with a considerably higher translation speed**

I. Introduction

THE goal of this study is to explore the transformer's capability of domain translation into a morphologically rich language. Satisfactory translation into Polish requires inflection by tense, number, and person, taking into account six declination cases. The ideal outcome of this study would be to prove that the method proposed by Dinu [1] is capable of training the transformer to translate English to Polish in domain-specific scenarios. Achieving metrics similar to Nowakowski [2] would result in a "zero-shot" translator with a considerably higher translation speed.

II. Potential applications

There is no shortage of uses for machine translation in general. It is reasonable to believe that low latency domain-specific translation will be in demand. Due to the translation speed equal to unconstrained translation, "train-by" models Dinu [1] can be adopted in simultaneous translation, broadening their usability. Institutions like EU Parliament could heavily benefit from such a system.

III. Related work

Among Slavic languages, Polish is not as well researched as Russian when it comes to machine translation. Nowakowski [2] investigated the application of Constrained Beam Search proposed by Anderson [3] and Hokamp and Liu [4] for English to Polish translation in domain-specific applications. This work examines "train-by" models described by Dinu [1] in the same context. The study, hopefully, will compliment Nowakowski's [2] findings.

IV. Motivation

The "train-by" approach does not give the premise of achieving translation and inflection quality superior to the Constrained Beam Search method. What it ensures, however, is translation speed equal to that of unconstrained translation. Bearing that in mind, and the fact, that both solutions allow translations of phrases that were not part of training vocabulary, it seems that if capable of ensuring similar quality, the "train-by" approach is worth consideration in most real-life scenarios.

V. Experiment

The main idea behind the experiment is to determine the approach that, applying lexical constraints, allows to achieve the best correctness of term inflection. To evaluate the solution, the following metrics will be used: BLEU [5], Term Rate [6], as well as Placement Rate, Duplication Rate and Inflection Rate [2]. The training corpus consists of Europarl v8, EUBookshop v2, JRC-Acquis v3.0, TildeMODEL v2018, Wikipedia v1.0, and most of DGT v2019 corpora, filtered with Bicleaner 2 and Bifixer 3 [7]. The corpus icludes 3103819 segments. Test sets are made of 1000 and 1104 segment pairs, validation sets consist of 2000 sentences from DGT corpus. The lexical constraints will be extracted from the Compendium of Accounting in Polish & English [8] and will consist of 1197 term pairs. The architecture of choice will be a standard NMT transformer. It is plausible that to reduce the computational power needed for training, parameter sharing will be taken advantage of, as described by Takase and Kiyono [9].

VI. Summary

The goal of the study is to acquire a better understanding of what might be a state-of-the-art approach to translating into Polish in domain-specific scenarios. The experiment will also bring new knowledge about the expectable quality of translation into a language with robust inflection. The low latency of the method that will be used gives a premise, that the model will be feasible to apply in production environments.

REFERENCES

[1] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Jassem Nowakowski. Neural machine translation with inflected lexicon. 2021.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *CoRR*, abs/1612.00576, 2016.

[4] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. *ArXiv*, abs/1704.07138, 2017.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[6] Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[7] Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[8] Robert Patterson. Compendium of accounting in polish & english. 2015.

[9] Sho Takase and Shun Kiyono. Lessons on parameter sharing across layers in transformers. *ArXiv*, abs/2104.06022, 2021.

Author Index